UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Daline Zat

ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO PARA MINERAÇÃO DE DADOS EDUCACIONAIS

Caxias dos Sul

2012

Daline Zat

ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO PARA MINERAÇÃO DE DADOS EDUCACIONAIS

Trabalho de Conclusão de Curso para obtenção do Grau de Bacharel em Ciência da Computação da Universidade de Caxias do Sul.

Carine Geltrudes Webber
Orientadora

Caxias do Sul

2012

AGRADECIMENTOS

Para chegar até aqui contei com a ajuda de diversas pessoas que me auxiliaram nessa caminhada, me apoiando nos momentos mais difíceis em que tudo parecia dar errado. Sem elas não teria conseguido. Queria agradecer os meus pais que sempre incentivaram os meus estudos desde a minha infância. A minha mãe Marlei que foi a primeira pessoa que me apoiou na decisão de sair de casa para estudar. Ao meu pai Norberto que sempre esteve presente, sendo um alicerce nas horas de aperto. Ao meu noivo Janderson Ferreira, um agradecimento especial, pois ele é o maior responsável por eu não ter largado tudo na metade, sempre acreditou em mim, mesmo quando eu não acreditava. Se hoje sou uma profissional da área de informática a culpa é dele, ele sempre esteve ao meu lado nos momentos de desespero quando eu achava que ia reprovar nas disciplinas do curso, mas no fim, eu me descabelava por nada, sempre me ajudando em todos os momentos sendo um companheiro, um amigo, em fim uma pessoa especial. As minhas antigas colegas de casa Adriana e Jucilaine que tiveram um papel importante no início da minha caminhada, pois se tornaram a minha família nos primeiros anos fora de casa. Aos meus colegas de faculdade Raquel, Luís e Wagner por todas as horas de estudo dedicadas aos trabalhos de faculdade nos finais de semana e a amizade construída no decorrer desses anos, que ficará para a vida toda. Ao meu colega de trabalho e amigo Diego por emprestar livros sobre mineração de dados para a monografia. A minha amiga Fabiana por me ajudar na revisão da monografia com dicas e palavras de apoio. Ao professor Ricardo Dorneles, que nesse período de monografia, sempre teve uma palavra de conforto quando eu começava a desanimar usando a seguinte frase: “No fim vai dar tudo certo, se não deu ainda é porque não é o fim”. A minha orientadora Carine que teve muita paciência comigo no começo da monografia, quando eu estava perdida, sempre me conduzindo para fazer um bom trabalho.

“A Grande Conquista é o resultado de
pequenas vitorias que passam despercebidas.”

Paulo Coelho

SUMÁRIO

RESUMO

A mineração de dados educacionais é um campo de pesquisa que vem adquirindo destaque dentro da área de mineração de dados. Ela é uma disciplina que busca obter novas informações através de dados educacionais com o intuito de desenvolver e fortalecer as teorias cognitivas de ensino-aprendizagem. O grande volume dos dados educacionais disponíveis dificulta a análise manual dos mesmos, por isso são necessárias técnicas automáticas para fazer essa análise. Dentre estas técnicas destaca-se: a predição, o agrupamento, a mineração relacional, a descoberta com modelos e a destilação de dados para o julgamento humano. Sendo uma das mais importantes, a técnica de agrupamento consiste em formar grupos de dados com grande similaridade entre si e uma grande dissimilaridade entre elementos de grupos diferentes. Este trabalho apresenta uma revisão bibliográfica sobre mineração de dados educacionais e o uso de técnicas de agrupamento de dados, apresentando um estudo comparativo dos algoritmos de agrupamento, tais como: k-média, maximização da expectativa, modelo imunológico e métodos hierárquicos. O algoritmo k-média é o mais conhecido dentre os algoritmos de agrupamento. Ele forma os grupos visando minimizar a distância entre os elementos do grupo em relação ao centro. A maximização da expectativa é um algoritmo de estimativa e possui o objetivo de encontrar o melhor ajuste de um modelo para um conjunto de dados através da estimativa da máxima verossimilhança. O modelo imunológico procura formar grupos, levando em consideração uma maior homogeneidade entre os elementos do mesmo grupo e uma maior heterogeneidade entre os elementos de grupos diferentes, utilizando dois conceitos da área de Sistemas Imunológicos Artificiais. Os algoritmos hierárquicos agrupam os elementos em uma estrutura de árvore, organizando os grupos em formato hierárquico, resultando assim uma sequência aninhada de partições. Neste estudo, foram utilizadas as ferramentas WEKA (Mark et al., 2009) e R (Chambers, 2008). Além disso, três conjuntos de dados públicos: Geometry, Chinese Tone Study e Álgebra I 2006 foram analisados. Os resultados da execução dos algoritmos foram tabulados e analisados através dos critérios de homogeneidade e separação. A análise dos critérios identificou que os algoritmos não estão suficientemente preparados para trabalhar com os dados educacionais. Dentre todos os testes o algoritmo imunológico foi o que apresentou melhores resultados em relação aos critérios de homogeneidade e separação.

Palavras-chave: agrupamento de dados, mineração de dados educacionais, mineração de dados, k-média, maximização da expectativa, hierárquico, sistema imunológico artificial.

ABSTRACT

Educational data mining is an important research field that have been growing in the data mining area. It is a subject that aims to obtain new information through educational data in order to develop and strengthen the cognitive theories of teaching and learning. The large amount of educational data available makes it difficult to manually analyze it, so automated techniques are necessary to do this analysis. Among these techniques are: prediction, clustering, relational mining, the discovery with models and the data distillation for human judgment. Being one of the most important, the clustering technique consists in forming groups of data with high similarity between themselves and a great dissimilarity between elements of different groups. This work presents a review on educational data mining techniques and the use of data clustering, presenting a comparative study of clustering algorithms such as k-means, expectation maximization, immunological model and hierarchical methods. The k-means is the most known among the clustering algorithms. It forms groups in order to minimize the distance between the elements of the group in relation to the center. The expectation maximization algorithm is an estimation algorithm and has the goal of finding the best fit of a model to a data set by estimating the maximum likelihood. The immunological model aims forming groups, taking into consideration a greater homogeneity among elements of the same group and greater heterogeneity between elements of different groups, using two concepts from the Artificial Immune Systems area. The hierarchical algorithms cluster elements in a tree structure, organizing groups in hierarchical format, thus resulting in a nested sequence of partitions. In this study, we used the WEKA (Mark et al., 2009) and R (Chambers, 2008) tools. In addition, three public datasets: Geometry, Chinese Tone Study and Algebra I 2006 were analyzed. The results of the execution of the algorithms were tabulated and analyzed through the criterion of homogeneity and separation. The criterion analysis showed that the algorithms are not sufficiently prepared to work with the educational data. Among all tests the immunological algorithm showed the best results in relation to the criteria of homogeneity and separation.

Keywords: data clustering, educational data mining, data mining, k-means, expectative maximization, hierarchical, artificial immune systems.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

Siglas	Significado em Português	Significado em Inglês
API	Interface de Programação de Aplicativos	*Application Programming Interface*
CCTI	Centro de Computação e Tecnologia da Informação	
EDM	Conferência Internacional de Mineração de dados Educacional	*International Conference on Educational Data Mining*
EM	Maximização da Expectativa	*Expectation Maximization*
GNU	Licença Pública Geral	*General Public License*
IA	Inteligência Artificial	*Artificial Intelligence*
KC	Componentes de Conhecimento	*Knowledge Componentes*
MD	Mineração de Dados	*Data Mining*
MDE	Mineração de Dados Educacional	*Educational Data Mining*
NCES	Centro Nacional para Estatística da Educação	*National Center for Education Statistic*
SIA	Sistemas Imunológicos Artificias	*Artificial Immune Systems*
SBIE	Simpósio Brasileiro de Informática na Educação	
SIGKDD	Grupo de Interesse Especial sobre Descoberta de Conhecimento e Mineração de Dados	*Special Interest Group on Knowledge Discovery and Data Mining*
TCC	Trabalho de Conclusão de Curso	
WEKA	Waikato Ambiente para a Análise do Conhecimento	Waikato Environment for Knowledge Analysis

1 INTRODUÇÃO

A Mineração de Dados (MD) é uma etapa da descoberta do conhecimento em banco de dados que busca novas e potenciais informações a partir de uma grande quantidade de dados (Fayyad, 1995). Dentro dessa área, há um campo em evolução que desperta interesse científico identificado como Mineração de Dados na Educação (MDE). Esse campo busca explorar dados de ambientes educacionais com o intuito de compreender os estudantes, as suas formas de aprendizado, entre outras investigações. Com o aumento das informações educacionais, analisá-las sem a utilização de métodos automáticos de MD seria inviável durante um processo de ensino-aprendizagem. Os dados educacionais possuem uma complexidade elevada em comparação a outros dados como, por exemplo, dados de comércio eletrônico. Muitas vezes nestes dados está expressa a forma de aprendizagem de um conjunto de estudantes ou o desempenho individual de cada estudante.

As informações podem ser extraídas de dados educacionais através de métodos como: previsão, agrupamento (*clustering*), mineração relacional, descoberta com modelos e destilação de dados para julgamento humano (Baker, 2010).

No contexto deste trabalho foi estudada a técnica agrupamento de dados que busca agrupar objetos existentes em conjuntos de dados (*datasets*) formando grupos menores com maior similaridade entre si. A escolha dos algoritmos de agrupamento ocorre através da possibilidade de trabalhar com dados complexos sem um conhecimento prévio do conjunto de dados (aprendizagem não supervisionada).

Existem diversos algoritmos de agrupamento tais como: K-média, Maximização da Expectativa (EM), Imunológico, entre outros. Todos estes vem sendo aplicados em diversas áreas, com diversas publicações inclusive em eventos específicos de MDE. Porém, não está claro como cada algoritmo de agrupamento pode contribuir com esta área. Quais algoritmos produzem os melhores resultados? Que resultados são esses e como avaliá-los?

Nesta direção, este trabalho propõe um estudo comparativo entre os principais algoritmos de agrupamento e ferramentas como o WEKA (Mark et al., 2009) e o R (Chambers, 2008), que implementam esses algoritmos, aplicando-os em conjuntos de dados educacionais públicos. Tais conjuntos de dados foram retirados do PSLC DataShop (Koedinger et al., 2010) com o intuito fornecer material para teste com relação a outros trabalhos científicos na área e também identificar qual ou quais algoritmos são mais apropriados para a MDE em geral.

1.1 OBJETIVO DO TRABALHO

O objetivo desse trabalho é fornecer material para teste com relação a outros trabalhos científicos através de um estudo comparativo de algoritmos de agrupamento, identificando os mais apropriados para a mineração de dados educacionais. Visando atingir o objetivo geral apresentado, o trabalho será dividido em cinco objetivos específicos:

- Estudo da mineração de dados educacionais;
- Estudo das técnicas de agrupamento;
- Planejamento do estudo comparativo dos algoritmos de agrupamento;
- Execução do experimento;
- Avalição os resultados do experimento.

1.2 ORGANIZAÇÃO DO DOCUMENTO

O trabalho está organizado em seis capítulos. O primeiro introduz os assuntos abordados nesse trabalho. O segundo fala sobre mineração de dados na educação onde explica o que é MDE, os seus processos, as suas aplicações, as áreas chave, as técnicas e ferramentas utilizadas na MDE. O terceiro capítulo apresenta os algoritmos de agrupamento explicando as etapas do processo de agrupamento juntamente com os métodos particionais e os seus algoritmos, os métodos hierárquicos e o algoritmo imunológico. O quarto capítulo apresenta a proposta do experimento onde mostra os conjuntos de dados, as ferramentas, os algoritmos escolhidos e os critérios de avaliação que serão utilizados para avaliar os resultados apresentados na execução dos algoritmos. O quinto capítulo apresenta os passos para executar o experimento explicando os passos efetuados para a preparação dos dados, adaptação dos algoritmos, desenvolvimento dos critérios de avaliação e os estudos de caso para cada conjunto de dados. O sexto capítulo mostra uma síntese do que foi visto neste trabalho e estudos futuros.

2 MINERAÇÃO DE DADOS NA EDUCAÇÃO

A mineração de dados busca novas e potenciais informações a partir de uma grande quantidade de dados armazenados em repositórios, bancos de dados e *data warehouses*. Esse processo utiliza métodos tradicionais de análise de dados com algoritmos sofisticados (Tan et al, 2009), por isso é considerada uma mistura de estatística, IA (Inteligência Artificial) e base de pesquisa (Pregibon, 1997).

A mineração de dados no campo da Educação, segundo a *International Educational Data Mining Society'* (2012), é uma disciplina emergente focada no desenvolvimento de métodos que exploram dados vindos de ambientes educacionais. Esses dados tem uma complexidade elevada para se trabalhar, pois constituem um domínio pouco formalizado, podendo ter vários níveis de hierarquia juntamente com outros fatores como: tempo, sequência e contexto.

2.1 PROCESSO DA MINERAÇÃO DE DADOS

Segundo Tan et al (2009) o processo de descoberta do conhecimento em banco de dados consiste em converter dados brutos em informações úteis, através do fluxo de passos identificado como: pré-processamento, mineração de dados e pós-processamento.

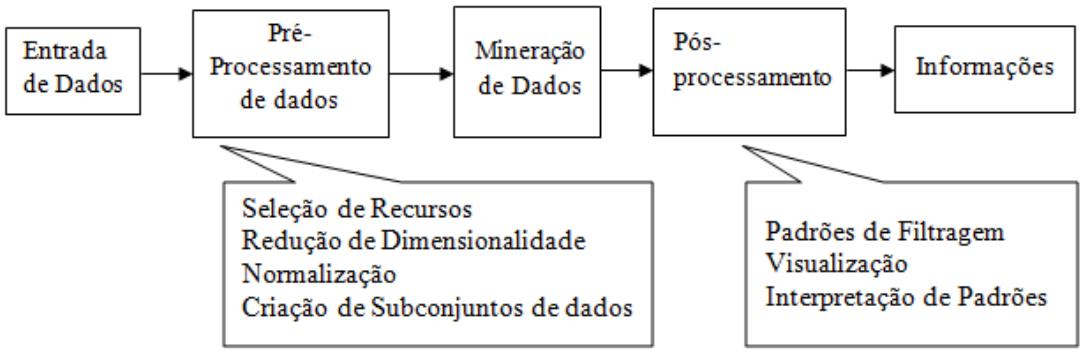

Figura 1 - O processo de descoberta do conhecimento em banco de dados (Tan et al, 2009).

O pré-processamento transforma os dados brutos em dados apropriados para a análise. Essa transformação ocorre através da limpeza dos dados que consiste em retirar informações duplicadas, na seleção de registros, na fusão de dados, na identificação de características com relevância e exclusão de dados que não se aplicam à pesquisa em questão. Esse processo é considerado o passo mais trabalhoso dentro da descoberta do conhecimento.

Após a filtragem dos dados no pré-processamento, o processo de mineração de dados é efetuado e os resultados são integrados aos sistemas de apoio à decisão para o pós-processamento, onde as informações de maior relevância são identificadas.

As análises sobre os resultados da MD revelam padrões de dados importantes que contribuem em diversas áreas tais como: Negócio, Medicina, Ciência, Engenharia, Comércio Eletrônico e Educação (Tan et al, 2009)**.**

2.2 DIFERENÇAS ENTRE MINERAÇÃO DE DADOS E MINERAÇÃO DE DADOS EDUCACIONAIS

A MD já demonstrou sucesso na área de comércio eletrônico e começou a ser aplicada na área educacional com grandes resultados (Srivastava et al., 2000). As formas de descoberta de informações são parecidas entre as duas áreas, no entanto Romero e Ventura (2007) identificaram algumas diferenças relevantes entre elas:

- Domínio: As aplicações em comércio eletrônico focam em orientar as compras enquanto uma aplicação educacional foca na aprendizagem dos estudantes.

- Dados: Normalmente os dados são mais simples nas aplicações de comércio eletrônico vindos de servidores web através de *logs* de acesso. Já nas aplicações educacionais, as informações podem ter relação com outros dados, serem focadas na interação dos estudantes com um sistema educacional, possuindo níveis de hierarquia significante, por exemplo, dados retirados de um curso *online* onde este curso possui vários capítulos organizados em sessões que possuem vários conceitos. Os dados gerados podem ser de modelos de usuários diferentes levando em consideração aspectos pedagógicos de cada estudante, elevando assim a complexidade dos dados em questão (Romero e Ventura, 2010).

- Objetivo: A MD voltada para o comércio eletrônico busca retornar um lucro crescente, medido através de dinheiro, na MDE busca-se melhorar as formas de aprendizado.

- Técnicas: Nas aplicações educacionais, dados diferentes são encontrados, por isso, a MDE precisa de um tratamento diferenciado juntamente com técnicas especiais na busca do processo de aprendizagem, possibilitando também a aplicação de técnicas tradicionais.

2.3 MINERAÇÃO DE DADOS NA EDUCAÇÃO

A MDE é um campo novo de pesquisa dentro da MD. Ela analisa dados de ambientes educacionais, através de sistemas de aprendizagem interativos, sistemas de tutores inteligentes, educação à distância e dados administrativos referentes a escolas e universidades entre outras aplicações, com o objetivo de descobrir padrões ou evidências científicas sobre estudantes e as formas de aprendizagem (IEEE, 2012). Nessa área encontram-se vários autores. Os que mais apresentam destaque são: Cristobal Romero, Sebastian Ventura e Ryan Baker com diversas publicações e citações.

Desde 2008 os pesquisadores da área de MDE se reúnem anualmente em conferências internacionais. Em 2009 o *Journal of Educational Data Mining* teve a sua primeira publicação (Scheuer e Mclaren, 2011). Já no ano de 2011 o grupo *International Working Group on Educational Data Mining* fundou a *International Educational Data Mining Society'* com o intuito de apoiar o desenvolvimento científico dessa área.

A MDE trabalha com diferentes grupos analisando as informações educacionais de vários ângulos e visões buscando a descoberta do conhecimento para auxiliar os professores a conduzirem melhor suas turmas, compreenderem mais o processo de aprendizagem dos estudantes e a melhoria dos métodos de ensino. Além disso, fornecerem *feedback* aos estudantes através de reflexões sobre a aprendizagem (Romero e Ventura, 2010).

A descoberta do conhecimento forma um ciclo onde os educadores e estudantes são responsáveis por manter a concepção, planejamento, construção e manutenção dos sistemas educacionais, cada um com o seu papel. Observa-se que outros grupos também acabam envolvidos nesse ciclo. Romero e Ventura (2007 e 2010) definiram esses grupos como:

- Orientado aos Alunos/Estudantes: Possui o objetivo de indicar atividades que aprimoram cada vez mais a aprendizagem dos estudantes juntamente com recomendações de livros, cursos e atividades semelhantes já executadas pelo próprio estudante ou por seus colegas.

- Orientado aos Professores/Educadores: Foca-se em obter um *feedback* sobre o ensino através de análise referente ao comportamento dos estudantes verificando quais são as maiores dificuldades, os erros comuns, o desempenho particular de cada estudante. Isso tudo para construir um padrão de aprendizagem regular ou irregular montando grupos, com o objetivo de identificar as atividades com melhor eficácia conforme a necessidade de cada grupo.

- Orientada aos Pesquisadores Educacionais e Desenvolvedores de Cursos: Avalia o curso com o intuito de melhorar o conteúdo, a aprendizagem do estudante e a eficácia do processo de aprendizagem. Busca também construir modelos de estudantes e tutores com o objetivo de recomendar as atividades que melhor se encaixam a cada estudante, além de desenvolver ferramentas específicas para analisar os dados educacionais.

- Orientada a Organizações/Universidades/Companhias Privadas de Treinamento: Foca-se em melhorar a tomada de decisão na busca de objetivos específicos como, por exemplo, selecionar os candidatos mais qualificados que irão ingressar na universidade, indicação de curso para determinadas áreas visando à importância para a mesma, entre outros.

- Orientada a Administradores de Distrito Escolar/Rede/Sistemas: Procura a melhor maneira de organizar a instituição, cursos, recursos (humanos e materiais), além de se preocupar com programas educacionais avaliando os professores e o currículo dos cursos. Busca também melhorar os *websites* deixando-os mais próximo dos usuários.

Dentre muitos estudos na área de MDE destacam-se quatro áreas chave de aplicação identificada por Baker (2010): melhoria de modelos de estudantes, descoberta ou melhoria de modelos da estrutura do conhecimento do domínio, apoio pedagógico oferecido pelo software de aprendizagem e descoberta científica da aprendizagem e do estudante.

A melhoria de modelos de estudantes procura identificar informações diferenciadas de cada estudante referente às características ou estado envolvendo conhecimento, motivação, atitude, entre outros. A identificação dessas diferenças é útil para as pesquisas que envolvem softwares educacionais, onde é preciso essa individualidade. Com o avanço das pesquisas na MDE, é possível conseguir modelos de alto nível referente ao comportamento dos estudantes, em relação à aprendizagem como, por exemplo, identificar que um estudante cometeu um erro apesar de possuir a habilidade necessária para a execução da tarefa. Esses modelos de estudantes são considerados mais completos, pois, aumentam a capacidade de prever o conhecimento e o desempenho futuro dos estudantes permitindo, aos pesquisadores estudarem os fatores que levam os estudantes a determinadas escolhas no processo de aprendizagem.

A descoberta ou melhoria de modelos da estrutura do conhecimento do domínio implica na descoberta rápida e precisa de modelos, combinando o algoritmo de espaço com

modelos de estruturas psicométricas. Geralmente é aplicado em problemas de predição com o propósito de descoberta de modelos, um exemplo de aplicação, seria prever se as ações individuais são corretas ou incorretas utilizando modelos de domínio diferentes.

O apoio pedagógico oferecido pelo software de aprendizagem busca descobrir qual apoio pedagógico é mais eficaz na aprendizagem. A decomposição da aprendizagem é um tipo de mineração relacional, aplicada em curvas exponenciais de desempenho de aprendizagem relacionando o sucesso com a quantidade oferecida de apoio pedagógico a cada estudante.

A descoberta científica da aprendizagem e do estudante trabalha em conjunto com as questões identificadas nas três primeiras áreas chave. Além disso, encontram-se estudos voltados para a descoberta científica envolvendo a descoberta com modelos através dos dados educacionais. Pode-se aplicar essa pesquisa em métodos de decomposição de aprendizagem, ou também em pesquisas que estudam o quanto os fatores de estado ou características de um estudante foram melhores preditores em um sistema de jogo, por exemplo.

Todas essas aplicações se preocupam em oferecer melhores recursos para a aprendizagem dos estudantes.

2.4 TÉCNICAS DA MDE

As pesquisas na área de MDE podem utilizar técnicas específicas de MDE e ou técnicas comuns de MD para chegar à resolução das quatro áreas de aplicação citadas na seção anterior. Essas técnicas foram classificadas e explicadas por Baker (2010) nas seguintes categorias: predição, agrupamento, mineração relacional, descoberta com modelos e destilação de dados para o julgamento humano. Cada categoria é descrita a seguir.

2.4.1 Predição

Na predição o objetivo é desenvolver um modelo onde se pode inferir sobre um único aspecto de dados (variável prevista) para uma combinação de outros aspectos de dados (variáveis preditoras). É necessário ter um rótulo para a variável de saída de um determinado conjunto de dados sendo usada de duas formas: na primeira forma o rótulo tem um valor específico para um conjunto de dados e, na segunda forma, busca-se considerar o grau de importância que realmente esse rótulo possui visando à proximidade.

Há duas aplicações de predição com maior relevância dentro da MDE. A primeira aplica-se em pesquisas que tentam prever os resultados de aprendizagem dos estudantes

usando informações subjacentes dos rótulos em construtores sem previsão de intermediários ou fatores mediadores. A segunda aplicação busca prever se o valor de saída não está no contexto desejável, um exemplo, seria não encontrar mais os dados rotulados em repositórios anteriores.

2.4.2 Agrupamento

O agrupamento procura encontrar conjuntos de dados que se agrupam naturalmente por alguma similaridade gerando diversos grupos menores, sendo muito útil pelo motivo de não conhecer com antecedências as categorias existentes nos conjuntos de dados que serão analisados.

Os grupos possuem diversas formas de granularidade dentro da MDE, por exemplo, diversas escolas podem ser agrupadas para investigar as suas semelhanças e diferenças, estudantes de diversos cursos, escolas ou universidades podem ser agrupados para investigar as suas semelhanças, ou as ações dos estudantes podem ser agrupadas para investigar padrões de comportamento.

Os algoritmos de agrupamento seguem algumas linhas, porém existem duas linhas utilizadas com mais frequência na MDE. Uma linha começa sem hipóteses anteriores sobre os grupos usando o algoritmo K-média e a outra linha inicia a partir de uma hipótese específica através de estudos anteriores usando o EM.

A técnica de agrupamento será aprofundada no próximo capítulo com o intuito de verificar os resultados de sua utilização em relação a dados educacionais.

2.4.3 Mineração Relacional

A mineração relacional procura descobrir a relação entre variáveis existentes em um conjunto de dados com um número elevado de variáveis, buscando identificar em determinado momento quais as variáveis que são fortemente associadas a uma única variável ou identificar qual a relação mais forte entre duas variáveis.

Identificam-se quatro categorias de mineração relacional:

- Mineração de regra de associação: busca encontrar regras que possuem uma condição gerando um resultado (*if-then*), por exemplo se o aluno está feliz então ele tem um desempenho elevado.
- Mineração de correlação: procura encontrar correlação linear entre variáveis, por exemplo, positivo e negativo.

- Mineração de padrões sequenciais: tem como objetivo de encontrar associação temporal entre eventos, por exemplo, determinar qual o comportamento de um estudante conduz um caminho de aprendizagem, identificando o evento de interesse relacionado.
- Mineração de dados casual: possui o intuito de descobrir se um evento foi a causa de outro evento através da análise de covariância entre dois eventos.

A técnica de mineração de dados relacional precisa satisfazer dois critérios identificados como: significância estatística e interesse. A significância estatística utiliza padrões estatísticos de teste como *F-Test* para avaliar o grande número de relacionamentos encontrados. O critério de interesse procura reduzir o conjunto de regras, correlações e relações casuais na MD com o intuito de determinar quais os resultados são mais importantes em determinados momentos.

2.4.4 Descoberta com Modelos

Na descoberta com modelos, o modelo de um fenômeno é desenvolvido utilizando técnicas como: predição, agrupamento e engenharia do conhecimento. Esse modelo é usado como um componente em outras análises tais como: predição ou mineração relacional.

Na situação da predição, as predições do modelo criado são usadas como variáveis dentro do preditor prevendo uma nova variável. Já na situação da mineração relacional, são estudados os relacionamentos entre as predições do modelo criado e as variáveis adicionais. Isto permite ao pesquisador estudar o relacionamento entre uma construção complexa e várias construções perceptíveis.

É comum na descoberta com modelos aproveitar a generalização validada de um modelo de previsão em contextos, por exemplo, Baker (2007) utilizou previsões de um sistema de jogo trabalhando com dados de um ano inteiro vindos de um software educacional para estudar se os fatores de estado ou traço (característica) eram melhores preditores de como um estudante se sairia no jogo do sistema.

2.4.5 Destilação de Dados para o Julgamento Humano

A destilação de dados para julgamento humano é a técnica em que o ser humano faz inferências sobre os dados apresentados, pois, isso foge ao alcance dos métodos inteiramente automatizados da MD. Nessa área trabalha-se com a visualização das informações sendo que essas informações são diferenciadas em relação aos outros campos da MD.

Os dados que passam pelo julgamento humano possuem dois propósitos principais: identificação e classificação. O propósito da identificação se encaixa quando o ser humano identifica facilmente padrões, no entanto esses padrões são difíceis de expressar formalmente como, por exemplo, a curva de aprendizagem, que exibe o número de oportunidades para uma habilidade versus o desempenho formando uma curva de progressão.

Além das áreas chave e das técnicas apresentadas, as pesquisas na MDE precisam de dados "considerados ecologicamente válidos", isto significa que eles são adquiridos em ambientes educacionais fornecidos por colégios e universidades. Esses dados podem ser encontrados em repositórios públicos como o PSLC *DataShop* e do Centro Nacional para Estatística da Educação (NCES - *National Center for Education Statistic*) que garantem essa característica (Baker, 2010).

2.5 TRABALHOS RELACIONADOS

Dentro da área de MDE encontra-se várias pesquisas tanto internacionais como nacionais. Internacionalmente podem-se visualizar algumas pesquisas apresentadas anualmente na Conferência Internacional de Mineração de Dados Educacional (*International Conference on Educational Data Mining* - EDM) e nacionalmente encontra-se o Simpósio Brasileiro de Informática na Educação (SBIE) e revistas como a Renote Novas Tecnologias na Educação.

A Conferência Internacional de Mineração de Dados Educacional acontece todos os anos desde 2008 expondo os estudos em andamento na área de MDE. Os dez artigos mais relevantes da conferência de 2011 no contexto deste trabalho foram tabulados e são apresentados na tabela 1 a seguir (Pechenizkiy et al., 2011).

Tabela 1 – Trabalhos Relacionados na Mineração de dados Educacional

Autores	Título	Síntese	
Nguyen Thai-Nghe, Tomás Horváth e Lars Schmidt-Thieme	**Factorization Models for Forecasting Student Performance**	Busca prever o desempenho dos estudantes através do desenvolvimento de um novo modelo de previsão com fatoração tensor levando em consideração 3 modos estudante/tarefa/tempo .	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	Compara o algoritmo EM com os algoritmos TFMAF (Tensor Factorization - Moving Average Forecasting) e TFF (Tensor Forecasting of Factorization) sendo que o modelo TFF foi criado para essa comparação.	PSLC DataShop usando os conjuntos de dados Algebra 2008-2009 e Bridge to Algebra 2008-2009.	Alemanha
Autores	**Título**	**Síntese**	
Nan Li, Wiliam Cohen, Kenneth R. Koedinger e Noboru Matsuda	**A Machine Learning Approach for Automatic Student Model Discovery**	Propõem um método que descobre automaticamente modelos de estudante não sendo humano-gerado. O sistema usa um agente de aprendizagem de máquina do estado-da-arte chamado de SimStudent, para adquirir o conhecimento da habilidade e cada habilidade corresponde a um componente do conhecimento (KC) que os estudantes precisam aprender.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	Compara o modelo gerado pelo SimStudent com o Modelo humano-gerado e um estudante real; Utiliza a medida do erro quadrático RMSE e AIC para ajuste dos dados dos estudantes para avaliação transversal.	PSLC DataShop usando os conjuntos de dados de Álgebra	Estados Unidos
Autores	**Título**	**Síntese**	
Michel Desmarais	**Conditions for effectively deriving a Q-Matrix from data with Non-negative Matrix Factorization**	Propõe validar a eficiência de gerar uma Q-matriz a partir da matriz não-negativa de fatorização (NMF)	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	-Utiliza algoritmos simples de agrupamento com o algoritmo NMF; -Utiliza o R para apresentar os grupos.	Os dados são retirados de dois sites: http://alumni.cs.ucr.edu/~titus/ e http://www.professeurs.polymtl.ca/michel.desmarais/Papers/EDM2011/scripts.html.	Canadá

Autores	Título	Síntese	
Dave Barker-Plummer, Richard Cox e Robert Dale	**Student Translations of Natural Language into Logic: The Grade Grinder Translation Corpus Release 1.0**	Apresenta traduções de linguagem natural para lógica de primeira ordem.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	Utiliza um livro LPL que é um sistema tutor com vários exercícios para traduzir frase que estão em linguagem natural para linguagem lógica de primeira ordem.	O sistema armazena os dados no repositório Grade Grinder.	Estados Unidos, Austrália e Reino Unido
Autores	**Titulo**	**Síntese**	
Yue Gong e Joseph Beck	**Items, skills, and transfer models: which really matters for student modeling?**	Analisa qual é a melhor forma para modelar o conhecimento dos alunos, usando a dificuldade do item e dificuldade da habilidade.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	- Usa predição; - Cria um modelo com base no modelo PFA-item. - Cria um modelo PFA-skill para avaliar a habilidade. Obs. PFA - Performance Factors Analysis	PSLC DataShop usando o conjunto de dados Álgebra - 2005-2006 e o conjunto de dados do sistema de tutoria ASSISTments	Estados Unidos
Autores	**Título**	**Síntese**	
Roberto Martinez Maldonado, Kalina Yacef, Judy Kay, Ahmed Kharrufa e Ammar Al-Qaraghuli	**Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop**	Analisa a interações homem-máquina através dos logs criados por grupos usando o recurso tabletop (mesa inteligente) multiusuário e a exploração de duas aproximações diferentes para considerar as ações físicas do toque.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	Técnica de agrupamento usando o método hierárquico aglomerativo.	Dados coletados em uma escola primária, sendo que no total de 18 estudantes participaram formando 6 grupos de 3 elementos cada.	Austrália, Reino Unido e Malásia

Autores	Título	Síntese	
Samad Kardan e Cristina Conati	**A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces**	Criar um modelo automático para descobrir e reconhecer testes padrões relevantes durante a interação do estudante com software educacional.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	- Usa agrupamento não supervisionado através do algoritmo K-média sendo que a inicialização dos centróides ocorre através de um algoritmo genético; - Usa classes de regras de associação através do algoritmo Hotspot que está disponível no WEKA; - Implementa uma ferramenta em Python com chamadas para WEKA para executar o algoritmo Hotspot.	Utiliza um conjunto com 65 estudantes.	Canadá
Autores	**Título**	**Síntese**	
Sujith M. Gowda, Jonathan P. Rowe, Ryan S.J.D. Baker, Min Chi e Kenneth R. Koedinger	**Improving Models of Slipping, Guessing, And Moment-By-Moment Learning with Estimates of Skill Difficulty**	Analisa modelos supervisionados de aprendizagem de máquina treinados a partir de novas características referentes à habilidade-dificuldade, em comparação aos modelos da suposição do contexto original que possuem as características originais do preditor. Investiga se os novos modelos produzem previsões melhores em relação a suposições dos estudantes, os deslizamentos (falhas), entre outros.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	- Regressão linear para consistência dos dados com a análise original; - Classificação dos dados através de análise Bayesiana.	Dados de dois tutores inteligentes o Middle School Mathematics Cognitive Tutor com 232 estudantes e o Genetics Cognitive Tutor com 71 estudantes.	Estados Unidos

Autores	Título	Síntese	
Shubhendu Trivedi, Zachary A. Pardos, Gábor N. Sárközy e Neil T. Heffernan	**Spectral Clustering in Educational Data Mining**	Apresenta a aplicação do algoritmo de agrupamento Espectral na MDE.	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	Algoritmo de agrupamento espectral.	Os dados são do Tutor ASSISTments referente ao ano escolar 2004-05 e recolhidos em duas escolas de Massachusetts, contendo 628 estudantes.	Estados Unidos
Autores	**Título**	**Síntese**	
Bahador Nooraei B., Zachary A. Pardos, Neil T. Heffernan e Ryan S.J.D. Baker	**Less Is More: Improving the Speed and Prediction Power of Knowledge Tracing by Using Less Data**	Busca responder a questão "quão sensível é o modelo KT à quantidade dos dados utilizados na sua formação?".	
	Algoritmos/Técnicas	**Origem dos Dados**	**País**
	-Usa o modelo Knowledge Tracing (KT) aplicando o algoritmo Expectation Maximization (EM) e a ferramenta Bayes Net Toolbox for Matlab;	Os dados são da Carnegie Learning Bridge to Álgebra Tutor referente o ano de 2007-2008 e do Cognitive Tutor for Genetics com 76 estudantes;	Estados Unidos

Para a seleção desses artigos utilizou-se como critério de relevância: a origem dos dados, ferramentas, algoritmos de agrupamento, autores renomados, universidades.

No Brasil, Prado Lima et al (2011) propõem um estudo sobre o desenvolvimento cognitivo individual e em grupo através de análise automática com o objetivo de promover a integração das teorias de ensino-aprendizagem com ambientes virtuais. Esse estudo foi aplicado em uma disciplina de Introdução a Programação (84 horas) através de um experimento, onde o processo de ensino-aprendizagem depende do raciocíno desenvolvido pelo estudante.

O experimento foi dividido em três etapas:

- A primeira etapa os estudantes estavam com 36 horas de curso e precisaram resolver um problema individualmente.
- A segunda etapa foi aplicada com 40 horas de curso onde os estudantes formaram duplas e precisaram analisar a solução proposta de um problema identificando erros existentes nesta solução.

– A terceira etapa ocorreu com 44 horas de curso onde cada estudante precisava analisar novamente uma solução proposta de um problema, porém individualmente.

O intervalo reduzido de tempo teve o intuito de avaliar a evolução dos estudantes num curto período de tempo com atividades variadas. Para cada etapa do experimento foi aplicado o algoritmo de agrupamento EM gerando grupos onde se observou a evolução dos estudantes em cada etapa do experimento.

Os resultados apresentados pelo EM identificaram grupos de alunos com o mesmo estilo de aprendizagem e padrões de comportamento similares, formando assim perfis de comportamento através da análise automática para auxiliar os professores a direcionarem o ensino de forma ágil.

2.6 FERRAMENTAS

Na busca pelo conhecimento precisa-se de ferramentas adequadas para a extração de informações visando qualidade e facilidade no processo de pesquisa. Encontram-se ferramentas como WEKA e o R, conhecidas mundialmente por possuírem diversos algoritmos para agrupamento de dados, qualidade na apresentação dos resultados, facilidade de uso e portabilidade em diversos sistemas operacionais. Além dessas ferramentas será apresentada uma nova ferramenta que trabalha com agrupamento de dados chamado de algoritmo Imunológico.

2.6.1 WEKA

O WEKA é um software livre com código aberto (*open source*) desenvolvido por pesquisadores da universidade de Waikato localizada na Nova Zelândia, dentro das especificações *General Public License* (GNU), utilizado na mineração de dados onde no ano de 2005 recebeu o SIGKDD *Data Mining and Knowledge Discovery Service Award* (Mark et al., 2009).

O desenvolvimento do WEKA surgiu a partir de um projeto financiado pelo governo da Nova Zelândia apresentado no final do ano de 1992, com o intuito de desenvolver técnicas de aprendizagem de máquina e investigar as suas aplicações nas principais áreas da economia da Nova Zelândia. A ideia inicial de utilização do software seria na indústria agrícola (Mark et al., 2009).

A primeira versão do WEKA foi desenvolvida nas linguagens de programação C, Prolog com interface TCL/TK. O seu primeiro lançamento interno ocorreu em 1994. Em 1996 foi o primeiro lançamento público na versão 2.1 a partir desse lançamento ocorreram mais dois lançamentos antes da decisão de reescrever o software para JAVA. No ano de 1999 foi lançada a versão 3.0 não gráfica escrita em JAVA, em 2003 foi lançada a versão 3.4 gráfica e atualmente encontra-se na versão 3.6 (Mark et al., 2009).

Os dados analisados pelo WEKA podem vir de várias fontes sendo, por exemplo: arquivos, URLS e banco de dados. O formato padrão de dados do WEKA é o ARFF, mas também aceita CSV, o formato LibSVM e C4.5.

Encontra-se um conjunto abrangente de algoritmos de aprendizagem de máquina e pré-processamento de dados que permite os usuários rapidamente testar e comparar os resultados obtidos nas suas pesquisas. Além dos algoritmos já implementados para classificação, agrupamento, regras de associação e seleção de atributos, pode-se incluir novos algoritmos através da API utilizando *plugins* existentes e a facilidade de integração apresentada pelo software.

2.6.1.1 Algoritmos Implementados no WEKA

O WEKA possui vários algoritmos divididos em quatro grupos: classificação, agrupamento, regras de associação e seleção de atributos apresentados por Witten e Frank (2000):

- Classificação: Possui a tarefa de organizar vários objetos em categorias pré-definidas utilizando os seus algoritmos subdivididos em: Bayesianos, Árvores, Regras, Funções, Classificadores Preguiçosos e a última categoria de chamada de Classificadores Variados (*Miscellaneous*).

- Agrupamento: Identifica grupos semelhantes de um conjunto de dados. O WEKA possui implementado os seguintes algoritmos:
  - EM: Os grupos usam a maximização da expectativa;
  - Cobweb: Implementa o cobweb e o algoritmo classit;
  - Farthest First: Implementa o algoritmo de passagem mais distante do primeiro sendo rápido e simples, com um modelo aproximado do K-média;
  - Make Density Based Clusterer: Retorna uma probabilidade de distribuição e densidade;

- Simple K-means: Usa o método K-média, porém o número de grupos precisa ser especificado por parâmetros;
- CLOPE: É um sistema de agrupamento rápido (Mark, 2009);
- Sequential Information Bottleneck: Desenvolvido para documentos de agrupamento (Mark, 2009);

- Regras de Associação: Procura identificar as associações e as relações de correlação existente em cada item de um banco de dados (Han e Kamber, 2001). Possuindo três algoritmos: *Apriori*, *Predictive Apriori* e *Tertius.*

- Seleção de Atributos: A maioria dos algoritmos de máquina sabe quais são os atributos adequados para a tomada de decisão. Existem casos onde a máquina não consegue chegar a uma definição, nesse momento entra a seleção de atributos onde se seleciona o atributo avaliador e o método de pesquisa para a tomada de decisão. Um exemplo da aplicação da seleção de atributos é quando a árvore de decisão chega a um nível de profundidade elevado e, nesse momento, há poucos dados disponíveis para a tomada de decisão. Nesse caso, pode-se ter a escolha do atributo errado, por isso usa-se a escolha manual de atributos.

2.6.2 R

O R surgiu a partir da linguagem S, originalmente desenvolvida no laboratório Bell por Rick Becker, John Chambers e Allan Wilks. É considerado um conjunto de software para manipulação de dados, cálculos, apresentações gráficas, de código aberto sendo um projeto *Free Software Foundation's* e GNU (Venables et al., 2011).

O ambiente R é caracterizado como um sistema planejado e coerente para análise de dados. Apresenta um manuseio eficiente dos dados, pois possui ferramentas para a análise que se integram com linguagens de programação como C, C++ e Fortan, usando-as em tempo de execução ou podendo manipulá-las com objetos do R. Além de ter um conjunto de operadores para cálculos com vetores, matrizes e instalações gráficas (Chambers, 2008).

Muitas vezes, o R é considerado um software estatístico devido à ampla variedade de técnicas estatísticas tanto clássicas quanto modernas, tais como: modelagem linear ou não linear, testes clássicos de estatística, análise de *time-series*, classificação, agrupamento e técnicas de gráficos (Chambers, 2008). Esse software está disponível nas plataformas Unix,

Windows e MacOs podendo ser executado e compilado. Atualmente está na versão 2.15.0 lançada em 2012 (R, 2012).

O R trabalha com diversos tipos de entrada de dados, desde arquivos de texto (ASCII), seguindo por formatos (Excel, SAS, SPSS, XML) e até mesmo acessar bancos de dados como MySQL, Oracle, SqlServer, entre outros através das bibliotecas pertencentes a cada banco de dados. Para cada tipo de arquivo, o R tem uma função específica onde consegue acessar o arquivo e manipulá-lo conforme a necessidade (R Development Core Team, 2012).

2.6.2.1 Pacotes de Agrupamento no R

O R possui diversos pacotes com métodos de agrupamento como, por exemplo, os pacotes *cluster, flexclust, trimcluster e mclust* que implementam algoritmos particionais e os pacotes *hclust, flashclust, agner* e *diana* que implementam algoritmos hierárquicos (Girke, 2012; R, 2012).

Os algoritmos k-média e EM são os mais conhecidos dentro do método particional e estão nos pacotes descritos a seguir. O algoritmo k-média está implementado em 3 pacotes: *stats, flexclust* e *trimcluster*, já o algoritmo EM está no pacote *mclust.* Cada pacote possui uma chamada de função diferente para os algoritmos, o k-média no pacote *stats* é chamado pela função kmeans(), para o pacote *flexclust* é chamado por kcca() e para o pacote *trimcluster* é chamado por trimkmeans(). O algoritmo EM no seu pacote é chamado pela função em(), tendo variações conforme o passo. A instalação do R possui somente alguns pacotes padrões como o *stast*, geralmente os pacotes precisam ser instalados antes da utilização (Girke, 2012; R, 2012).

Para o método hierárquico, os pacotes *hclust* e *agnes* realizam agrupamento aglomerativo e o pacote *diana* realiza agrupamento divisivo. O pacote *flashClust* realiza agrupamento aglomerativo no entanto a sua velocidade foi melhorada de 50 a 100 vezes em comparação com o pacote *hclust* (Girke, 2012). Esses são alguns dos pacotes disponíveis para execução no R. Cada um deles possui as suas peculiaridades e uma forma de execução.

2.6.3 Imunológico

O algoritmo Imunológico para agrupamento de dados foi desenvolvido por Raquel Machado em conjunto com o núcleo de pesquisa do Centro de Computação e Tecnologia da Informação (CCTI) da Universidade de Caxias do Sul, baseado em Sistemas Imunológicos

Artificias (SIA). Esse algoritmo foi desenvolvido na linguagem JAVA para desktop, possuindo o seu código proprietário, estruturado em três camadas (Machado, 2011).

A interface de execução possui parâmetros de entrada para o processamento do algoritmo. A primeira informação a ser preenchida é o local onde se encontra o conjunto de dados a ser analisado, a segunda opção é a seleção entre os algoritmos imunológico ou variante de k-média, se for escolhido o algoritmo imunológico será habilitado 3 parâmetros específicos para preenchimento sendo eles: *Mutation Factor* responsável pelo percentual de mutação, *Number of Clonel* define o número de clones a serem feitos e *Initial Iteracions* é o número inicial de iterações para ser realizado com o algoritmo variante do k-média (Machado, 2011).

Após o preenchimento desses parâmetros, encontram-se opções comuns aos dois algoritmos como: a normalização nos dados caso seja de interesse na análise em questão, *Number of Clusters* que indica quantos grupos serão formados, *Metric* responsável pelo cálculo da medida de distância para a formação dos grupos e, por fim, a *Stop Conditions*, ou seja, a condição de parada que possui duas opções: a primeira opção *Runs* é definida por iteração, ou seja, recebe o número de iterações que o algoritmo deve executar e a segunda opção *Threshold* é definida por limiar onde se deve informar o índice de homogeneidade e separação para que o algoritmo processe até atingir os índices indicados (Machado, 2011).

Além dos parâmetros, a interface possui quatro botões. O botão *Start* que inicia a execução do algoritmo, o botão *Close* finaliza a execução e fecha o programa, o botão *About* abre uma nova janela com informações referentes ao programa e o botão *Clean* que limpa os parâmetros preenchidos retornando ao estado inicial (Machado, 2011).

A figura 2 apresenta a interface do algoritmo imunológico com todos os parâmetros preenchidos.

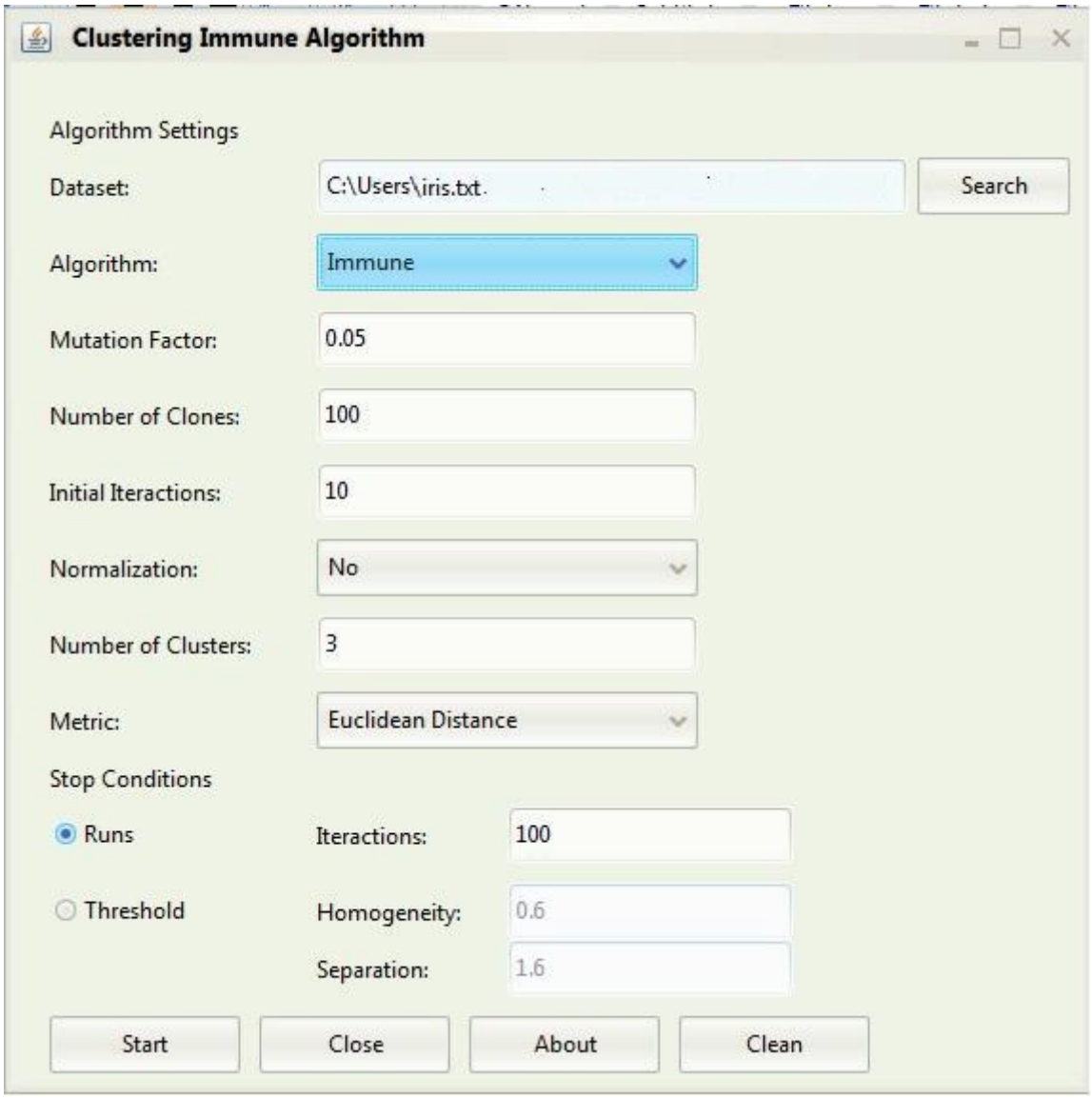

Figura 2 - Interface inicial do algoritmo imunológico (Machado, 2011).

3 AGRUPAMENTO DE DADOS

Agrupar objetos faz parte da vida dos seres humanos desde a infância, onde se consegue dividir objetos através das semelhanças ou outras características existentes entre si. Uma criança consegue separar fotografias de casas, prédios, carros, animais e plantas formando assim grupos divididos em categorias (Tan et al., 2009).

Agrupamento de dados possui diversas definições encontradas na literatura, a seguir algumas delas são apresentadas:

- *Agrupamento de dados é a organização de uma coleção de padrões (geralmente representado como um vetor de medições, ou um ponto em um espaço multidimensional) em grupos com base na similaridade* (Jain et al., 1999).

- *Um grupo é um conjunto de entidades semelhantes e entidades pertencentes a grupos diferentes não semelhantes* (Everitt, 1993; Mertz, 2006).

- *Um grupo é um agrupamento de pontos no espaço de teste de tal maneira que a distância entre quaisquer dois pontos em um mesmo grupo é menos que a distância entre qualquer ponto desse grupo e um outro ponto qualquer não pertencente a ele* (Everitt, 1993; Mertz, 2006).

- *Grupos podem ser descritos como regiões conectadas de um espaço multidimensional contendo uma alta densidade relativa de pontos, separados de outras regiões por uma região contendo uma baixa densidade relativa de pontos* (Everitt, 1993; Mertz, 2006).

Segundo Duarte (2011), a partir das definições existentes na literatura é formado um conceito partilhado entre elas, onde o processo de agrupamento consiste em organizar elementos de um conjunto de dados em classes ou grupos homogêneos, de forma que exista uma alta similaridade intragrupo e uma baixa similaridade intergrupo, ou seja, exista uma grande similaridade entre elementos do mesmo grupo e uma grande dissimilaridade entre elementos de grupos diferentes. Assim, um grupo será um subconjunto de elemento que são similares entre si e que diferem dos elementos pertencentes a outros grupos.

A figura 3 mostra um conjunto de pessoas representando dados de um conjunto que serão divididos em grupos.

Figura 3 - Conjunto de dados para aplicação de algoritmos de agrupamento (Keogh, 2003).

A partir da aplicação dos algoritmos de agrupamento formaram-se dois grupos menores. O primeiro grupo mostra as pessoas divididas em relação ao convívio formando duas subcategorias classificadas em família e escola, já o segundo agrupou as pessoas entre homens e mulheres.

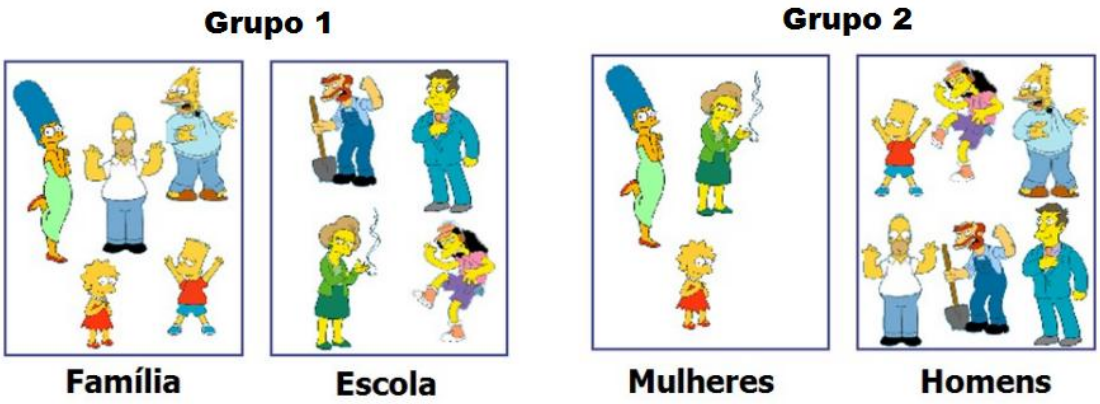

Figura 4 - Grupos formados após a aplicação de algoritmos de agrupamento (Keogh, 2003).

A utilização de algoritmos de agrupamento vem crescendo nos últimos anos, pois a cada dia geram-se mais informações tornando a análise manual desses dados uma atividade complexa. O intuito dessa análise é descobrir e formar novos padrões utilizados em diversas áreas tais como: biologia, estatística, clima, negócios, psicologia e medicina (Tan et al., 2009).

3.1 APRENDIZAGEM DE MÁQUINA

Os algoritmos de agrupamento são considerados algoritmos de aprendizagem de máquina, pois conseguem adquirir conhecimento através de formas automáticas. O processo de aprendizagem de máquina possui cinco estratégias: hábito, instrução, dedução, analogia e indução (Metz, 2006).

A estratégia de indução é dividida em três modos de aprendizagem: supervisionada, semi-supervisionada e não supervisionada. A aprendizagem supervisionada possui os atributos de classe que rotulam um conjunto de dados com o intuito de minimizar o risco funcional, a não supervisionada não possui esses rótulos e a semi-supervisionada possui somente alguns rótulos (Xu et al., 2005; Metz, 2006).

Os algoritmos de agrupamentos adquirem conhecimento através da descoberta de padrões identificando características similares sem conhecer previamente o conjunto de dados. Isso faz com que eles se enquadrem no modo de aprendizagem não supervisionado que busca o conhecimento a partir de observação, descoberta ou análise exploratória de dados (Metz, 2006).

3.2 ETAPAS DO PROCESSO DE AGRUPAMENTO

O processo de agrupamento é dividido nas etapas: pré-processamento de dados e seleção de variáveis, seleção da medida de similaridade/dissimilaridade, execução/desenvolvimento do algoritmo de agrupamento, avaliação dos resultados e interpretação dos resultados. Essas definições foram geradas a partir das seguintes literaturas: Jain et al. (1999), Junior (2006), Metz (2006), Oliveira (2008) e Naldi (2010).

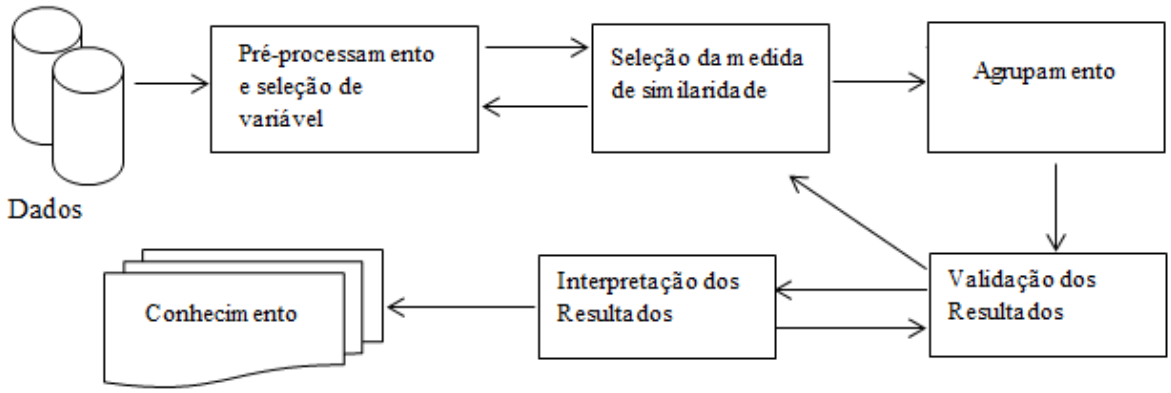

Figura 5 - Etapas do processo de agrupamento.

- Pré-Processamento de dados e seleção de variáveis: Seleciona atributos distintos ou variáveis em um conjunto de dados, onde serão aplicados os algoritmos de agrupamento. Aplicando se necessária alguma transformação nos dados como: conversão, normalização, remoção de duplicidade, entre outras.

- Seleção da Medida de Similaridade/Dissimilaridade: É utilizada no processo de criação dos grupos gerando influência nos agrupamentos. Uma forma de cálculo desta medida é a partir de uma função específica aplicada entre pares de elementos. A maneira mais comum de calcular a medida de similaridade/dissimilaridade é através da distância euclidiana.

- Execução dos Algoritmos de Agrupamento: Nessa etapa, escolhe-se qual o tipo de algoritmo de agrupamento será utilizado na geração dos grupos e aplica-se no conjunto de dados. Dentre os diversos algoritmos de agrupamentos destacam-se os algoritmos hierárquicos e particionados.

- Avaliação dos Resultados: Avalia a qualidade dos resultados obtidos através de índices estatísticos ou comparação com outros algoritmos sempre considerando critérios de avaliação. Busca validar a qualidade dos grupos ou estimar o grau que a estrutura resultante possui sobre o conjunto de dados.

- Interpretação dos Resultados: Examina-se os grupos resultantes na busca de identificar um significado prático para cada grupo na descoberta do conhecimento, sendo necessária a participação de um especialista nesta etapa, pois trata-se de uma tarefa complexa. Esta etapa pode ser desconsiderada conforme o objetivo do grupo.

3.3 MÉTODOS DE AGRUPAMENTO

Há diversos métodos de agrupamento encontrados na literatura, onde a escolha de cada método deve levar em consideração os tipos de dados disponíveis e a aplicação desejada, não existindo um método que seja ideal para todos os tipos de dados e aplicações (Jain et al., 1999; Junior, 2006).

Dentre os métodos de agrupamento existentes há duas categorias que possuem maior destaque sendo classificadas como métodos particionais e métodos hierárquicos. A escolha de

cada método ocorre de acordo com a estrutura dos dados envolvidos (Xu et al., 2005; Santos, 2009; Naldi, 2010).

3.3.1 Particionais

Informalmente os métodos particionais consistem em particionar um conjunto de dados em um número determinado de grupos uma única vez (Oliveira, 2008). Estes métodos possuem algumas condições que devem ser satisfeitas: cada grupo deve conter pelo menos um elemento; e cada elemento deve pertencer a exatamente um único grupo (Mello, 2008).

Formalmente os métodos particionais são definidos da seguinte forma: dado um conjunto n de dados, determine as k partições iniciais, onde cada uma dessas partições representa um grupo de objetos com $k \leq n$. Nessas k partições, cada grupo deve possuir ao menos um elemento e cada elemento deve pertencer a somente um grupo. O número de partições k é o parâmetro de entrada sendo escolhido conforme o problema abordado podendo interferir na eficiência do algoritmo utilizado (Oliveira, 2008; Mello, 2008).

A partir do parâmetro inicial, são construídas k partições iniciais e através de uma técnica de realocação iterativa, o algoritmo tenta melhorar o esquema de partições movimentando os elementos de um grupo para o outro da seguinte forma: dado o número de partições k e um conjunto de n elementos, inicialmente seleciona-se arbitrariamente k pontos que definem os k primeiros grupos. Os n-k elementos restantes são associados aos k grupos. A partir dos grupos formados, são calculados novos centros para representar os grupos. Um critério ou medida de erro é calculado para o agrupamento formado. Esse processo de encontrar novos centros e realocação dos elementos é repetido até que a medida de erro não possa mais ser reduzida (Mello, 2008).

O critério ou medida de erro usado a cada movimentação dos elementos avalia a qualidade do particionamento procurando formar grupos mais homogêneos, orientando o algoritmo a alocar os elementos similares no mesmo grupo e os elementos dissimilares em grupos diferentes, diminuindo a soma dos desvios dos elementos aos centros de cada grupo (Oliveira, 2008; Mello, 2008; Machado, 2011).

O cálculo do critério procura encontrar um valor ótimo para a qualidade do agrupamento, mas para isso seria necessário avaliar todas as possíveis combinações de partições o que pode ser inviável computacionalmente. Por isso, muitos métodos trabalham com heurísticas para a escolha das partições (Junior, 2006; Mello, 2008).

A utilização de métodos particionais apresentam vantagens e desvantagens. As duas principais vantagens na utilização desses métodos são: a possibilidade de um elemento trocar de grupo para outro com a evolução do algoritmo e a possibilidade de utilizar conjuntos de dados maiores visto que não é preciso guardar a matriz de similaridade na memória durante o processamento como, por exemplo, nos métodos hierárquicos. A principal desvantagem é que eles não trabalham muito bem em conjuntos de dados onde os grupos resultantes são de formas complexas e diferentes tamanhos (Pereira, 2007; Maximiliano e Cordeiro, 2008).

Segundo Pereira (2007), os dois algoritmos mais populares são o k-média e o EM.

3.3.1.1 K-média ou K-means

O algoritmo K-média foi desenvolvido por Hantingan (1975), sendo um dos mais conhecidos dentre os métodos particionais. Ele divide um conjunto de dados em *k* partições de modo que a similaridade intragrupo seja alta e a similaridade intergrupo seja baixa (Macqueen, 1967), com o objetivo de encontrar a melhor divisão de *p* dados em *k* grupos, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada. Para isso, ele agrupa dados em um número de grupos predefinido e os grupos são representados por um vetor centróide, que indica o ponto central daquele grupo (Mello, 2008; Naldi, 2010; Machado, 2011). O k-média pode ser dividido em quatro passos principais:

1. Recebimento do parâmetro de entrada com o número de partições;
2. Definição do centróide através da medida da distância no espaço multidimensional;
3. Definição dos grupos de acordo com o ponto médio;
4. Alocação dos elementos nos grupos.

A execução do k-média começa após o recebimento do parâmetro de entrada que define a quantidade de partições *k*. A partir disso, ocorre a escolha aleatória de *k* elementos para representar os centros dos grupos, após essa definição dos centros, os elementos restantes são associados aos centróides mais próximos, essa associação é realizada através da distância entre um elemento e o centróide. A cada novo passo, o algoritmo calcula novamente a média e define o novo centróide, realocando novamente os elementos nos grupos de maior similaridade. Esse processo de recalcular as médias dos grupos e realocar os elementos é repetido até que o critério de parada seja satisfeito.

Um critério de parada pode ser encontrado na função-objetivo do k-média sendo descrita pela Equação 1, em que c_i é o centróide do grupo C_i e $d(x_j, c_i)$ é a distância euclidiana ao quadrado entre o elemento x_j e o centróide c_i. A meta do algoritmo é minimizar a distância euclidiana ao quadrado entre cada ponto e o centróide do grupo ao qual ele pertence, ou seja, minimizar o valor da função-objetivo *obj* dada pela Equação 1 para *k* grupos. Essa função-objetivo tende a formar grupos de formato híper esférico do mesmo tamanho e bem separados (Naldi, 2010).

$$\min_{c_1,\dots c_k} obj = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, c_i)$$

Equação 1 – Função-objetivo do k-média.

Um pseudocódigo do k-média contextualizado por Naldi (2010) pode ser visto no algoritmo 1. Esse algoritmo é uma típica versão do k-média utilizando como critério de parada a função-objetivo exposta na equação 1, onde os valores resultantes decrescem fazendo com que não haja diferença significativa nos valores de c_i entre duas iterações consecutivas do algoritmo.

Algoritmo 1: Algoritmo particional *k*-média

Seja *k* o número de grupos e *t* o número máximo de iterações

Requer $k \geq 2$ e $t \geq 1$

1. inicializar os centroides dos grupos $c_1, c_2, \dots, c_k$;
2. repita
3. para todos $x_j \in X$ faça
4. para todos c_i faça
5. calcula a dissimilaridade $d(x_j, c_i)$;
6. fim para
7. adiciona o objeto x_j ao grupo C_i para o qual $i = arg\ \min_{l=1}^{k} d(x_j, c_l)$;
8. fim para
9. para todos c_i faça
10. recalcula c_i como centroide do grupo C_i;
11. fim para
12. até convergência ser alcançada ou o número de iterações exceder um dado limite *t*

A figura 6 ilustra um exemplo de execução do *k*-média no plano para *k=3*.

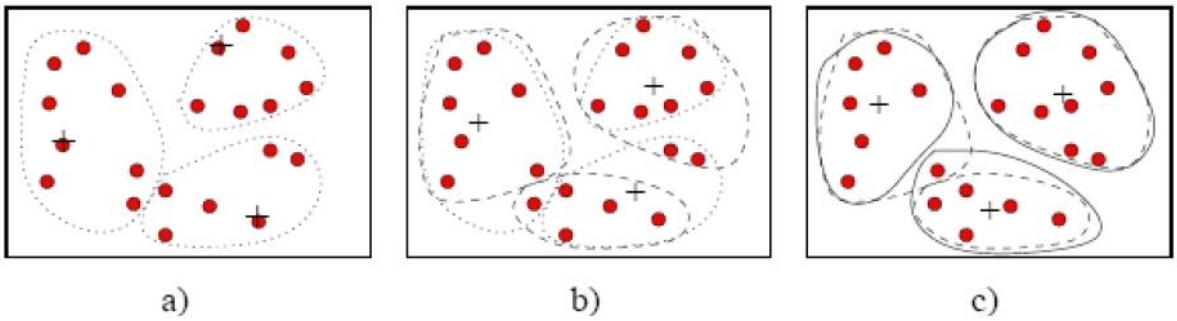

Figura 6 - Exemplo da execução do algoritmo de agrupamento k-média (Junior, 2006).

Os *k* centróides são representados através do sinal de "+". Na Figura 6(a), os *k* elementos são escolhidos aleatoriamente como centróides. Em seguida, o algoritmo associa os elementos restantes aos centróides. A linha pontilhada indica os grupos formados. Na Figura 6(b), são calculadas as médias dos grupos formados e definidos novos centróides. Novamente, de acordo com esses centróides, os elementos são realocados nos grupos. O algoritmo segue até que, conforme apresentado na Figura 6(c), os elementos não mudam mais de grupo, isto é, o valor do erro quadrático não pode mais ser reduzido.

O algoritmo K-média é conhecido pela sua simplicidade de implementação sendo assim um dos mais encontrados na literatura com diversas variações como, por exemplo, o K-medianas (Kauffman et al., 1989) que é uma escolha natural quando o cálculo das médias não é possível, pois o k-média limita-se a dados descritos por atributos quantitativos (Junior, 2006; Machado, 2011).

Além da sua simplicidade, outras vantagens do k-média são a sua rápida convergência e a sua escalabilidade apresentada em relação aos dados. No entanto, apresenta desvantagens como: a necessidade de estimar o número de grupos na inicialização da execução do algoritmo e não possui imunidade a ruídos (Lenz, 2009; Machado 2011).

3.3.1.2 EM

O algoritmo EM surgiu a partir da unificação de diversos trabalhos sendo apresentado por Demspter et al (1977). Ele procura encontrar o melhor ajuste de um modelo para um conjunto de dados através da estimativa da máxima verossimilhança dos parâmetros, ou seja, estima parâmetros que sejam os mais consistentes em relação aos dados da amostra no sentido de maximizar a função de verossimilhança. Cada grupo é representado matematicamente por uma distribuição de probabilidade de forma que não haja fronteiras entre os grupos (Luna, 2004; Duarte, 2008; Lenz, 2009; Machado, 2011).

De uma forma geral o algoritmo EM avalia se alguma variável foi observada em um determinado momento e outra variável não. É possível utilizar os casos observados para aprender e predizer os valores não observados. Além dessa tarefa, ele pode ser utilizado para variáveis cujos valores nunca foram observados, sempre e quando seja conhecida a forma geral da distribuição de probabilidade das variáveis (Arruda, 2011).

Esse algoritmo utiliza um processo de iteração para que possam ser aproximadas soluções de equações não lineares, a partir de um valor inicial predefinido, até que haja convergência a um valor final. O valor ideal para a convergência seria um valor máximo global, porém esse valor depende dos parâmetros iniciais sendo assim pode-se obter um valor máximo local que não é desejado (Lenz, 2009).

O cálculo que busca a convergência ocorre a partir da função de verossimilhança e do estimador da máxima verossimilhança apresentados por Luna (2004). A função de verossimilhança com n variáveis aleatórias $X_1, X_2, \dots, X_n$ é definida como a densidade conjunta de *n* variáveis aleatórios, ditas $f x_1, \dots, x_n(x_1, \dots, x_n; \theta)$, considerada como função de θ. Em particular, se $X_1, \dots, X_n$ é uma amostra aleatória da densidade $f(x;\ \theta)$, então a função de verossimilhança é $f(x_1;\ \theta) f(x_2;\ \theta)$ ... $f(x_n;\ \theta)$. Sendo que a notação usada na função de verossimilhança para demostrar que é uma função θ deve seguir os seguintes exemplos $L(\theta;\ x_1, \dots, x_n)\ ou\ L\ (.; x_1, \dots, x_n)$. O estimador da máxima verossimilhança é definido da seguinte forma: seja a função de verossimilhança $L(\theta) = L(\theta;\ x_1, \dots, x_n)$ para as variáveis aleatórios $X_1, X_2, \dots, X_n$. Se $\widehat{\theta}$ [onde $\hat{\theta} = \hat{\vartheta}(x_1, x_2, \dots, x_n)$ é uma função das observações $x_1, \dots, x_n$] é o valor de θ em $\overline{\Theta}$ (universo dos θs) que maximiza $L(\theta)$, então $\widehat{\Theta} = \vartheta(X_1, X_2, \dots, X_n)$ é o estimador de máxima verossimilhança de θ, e $\hat{\theta} = \vartheta(x_1, \dots, x_n)$ é a estimativa de máxima verossimilhança de θ para as amostra de $x_1, \dots, x_n$.

O algoritmo EM pode ser dividido em dois passos principais sendo eles:

1. Passo E: Encontram-se os valores esperados das estatísticas suficientemente para os dados completos Y, dado os dados incompletos Z e as estimativas atuais dos parâmetros.
2. Passo M: Utilizam-se estas estatísticas para fazer uma estimativa de máxima verossimilhança.

Pode-se descrever formalmente o EM da seguinte forma: Considere que $X = \{x_1, \dots, x_m\}$ são os dados observados de um conjunto de m instâncias ocorridas independentes, $Z = \{z_1, \dots, z_m\}$ são os dados não observados nesta instância e seja $Y = X \cup Z$ o total de dados. Pode-se tratar Z como uma variável aleatória cuja distribuição de

probabilidades depende do conjunto de parâmetros desconhecidos θ e dos dados observados X. Y é uma variável aleatória, pois, é definida em função da variável aleatória Z. Para descrever a forma geral do algoritmo EM, denota-se a hipótese dos parâmetros atuais, θ por h e a hipótese revisada, que é estimada a cada iteração do algoritmo, por h' (Luna, 2004; Arruda, 2011).

O algoritmo EM busca a hipótese h' de máxima verossimilhança, isto é, que maximize $E[\ln P(Y|h')]$. Está função é dividida em três passos (Luna, 2004):

1. $P(Y|h')$ é a verossimilhança de todos os dados Y, dada à hipótese h'.
2. Maximizar o logaritmo desta quantidade de $\ln P(Y|h')]$ também maximiza $P(Y|h')$.
3. Introduz o valor esperado de $E[\ln P(Y|h')]$ pois o total de dados Y é, ele próprio, uma variável aleatória.

Este valor esperado é calculado sobre a distribuição de probabilidade de Y, que é determinada pelos parâmetros desconhecidos θ. Sendo que os dados Y são uma combinação dos dados observados X e não observados Z, obtêm-se o valor de $E[\ln P(Y|h')]$ sobre a distribuição de probabilidade de Y, que é determinada pelos valores conhecidos X mais a distribuição de probabilidade de Z. Em geral a distribuição de probabilidades de Y não é conhecida, pois ela é determinada pelos parâmetros θ que se deseja estimar. Entretanto o algoritmo EM usa sua hipótese atual h no lugar do parâmetro θ atual para determinar a distribuição de probabilidade de Y. Assim, considere uma função $Q(h'|h)$ que dá $E[\ln P(Y|h')]$ como função de h', pela suposição que $\theta = h$ e dada à porção dos dados observados X dos dados Y (Luna, 2004, Arruda, 2011).

Escreve-se esta função Q na forma $Q(h'|h)$ para indicar que ela é definida em parte pela suposição que a hipótese atual h é igual a θ (Luna, 2004; Arruda, 2011). A função Q pode ser vista na equação 2:

$$Q(h'|h) = E[\ln(P(Y|h')|h, \mathrm{X})]$$

Equação 2 – Função Q para estimar a distribuição de probabilidade.

Assim, o algoritmo EM repete os dois passos seguintes até a convergência:

Passo E: Calcula $Q(h'|h)$ utilizando a hipótese atual h e os dados observados X para estimar a distribuição de probabilidade de Y utilizando a equação 2.

Passo M: Troca a hipótese *h* pela *h'* que maximize a função Q:

$$h = \arg \max_{h'} Q(h'|h)$$

Equação 3 – Maximização da função Q.

Quando a função Q é contínua, o algoritmo EM converge para um ponto estacionário da função de verossimilhança $P(Y|h')$. Quando esta função tem um único máximo, o algoritmo convergirá para esta estimativa de máxima verossimilhança global para *h'*, caso contrário o algoritmo poderá convergir para o máximo local (Luna, 2004; Arruda, 2011).

O pseudocódigo do EM contextualizado por Duarte (2008) pode ser observado no algoritmo 2, onde a linha 3 utiliza a fórmula de máxima verossimilhança para identificar a distribuição de probabilidade dos elementos.

Algoritmo 2: Algoritmo particional EM

Entrada: X- conjunto de dados com *n* elementos;
K- número de grupos;
Saída: P- Agrupamento de dados final;
Procedimento

1. escolher arbitrariamente K centros/distribuídos como solução inicial
2. repetir
3. recalcular a distribuição de probabilidade dos elementos de acordo com a solução atual
4. atualizar as distribuições dos grupos de acordo com a nova distribuição de probabilidade dos elementos.
5. até ser atingido critério de convergência

A figura 7 representa a execução do EM seguindo os passos principais: E e M. Observa-se que na figura (a) inicia-se com dois centros escolhidos aleatórios, a figura (b) representa o passo E, a figura (c) representa o passo M e, a partir disso, o algoritmo executa os passos E e M até a convergência se tornar insignificante terminado a execução do algoritmo.

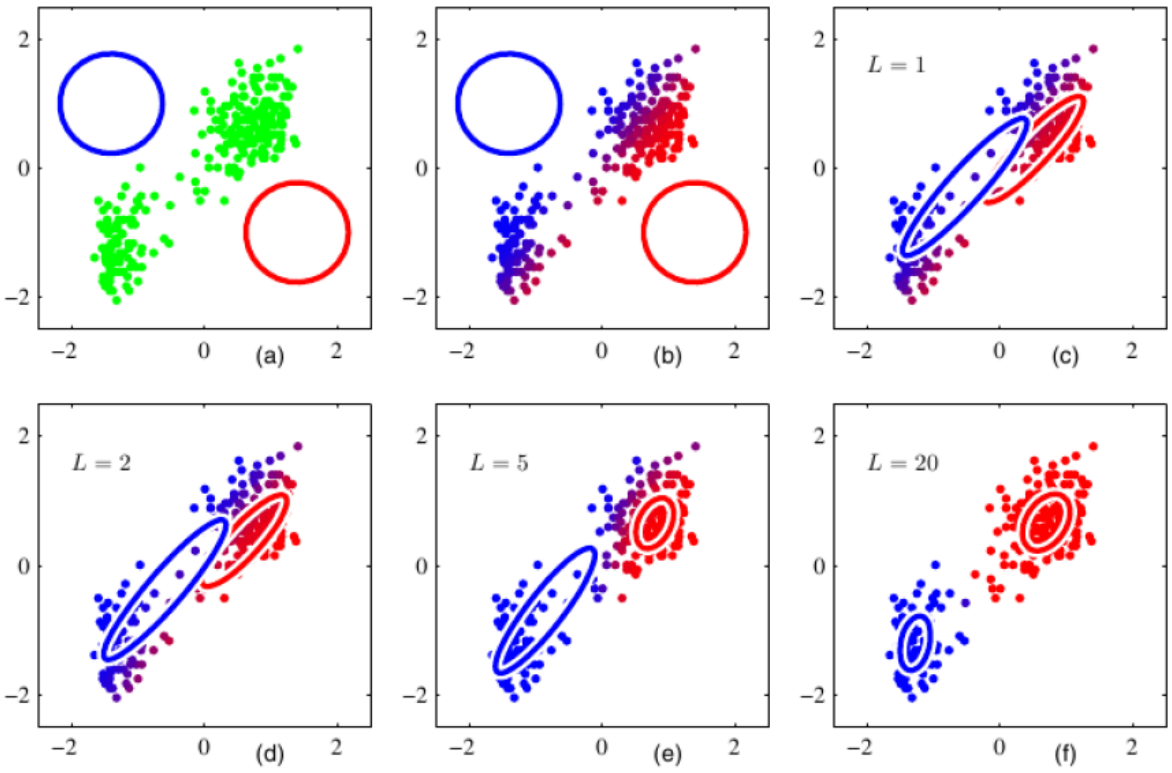

Figura 7 - Exemplo da execução do algoritmo de agrupamento EM (Arruda, 2011).

Uma vantagem no algoritmo EM é à possibilidade de não informar o número de grupos iniciais como parâmetro inicial. A desvantagem apresentada por ele é a geração de um máximo local quando os parâmetros iniciais são informados.

O algoritmo EM atualmente começou a ganhar maior destaque, pois possui um forte embasamento estatístico.

3.3.2 Hierárquicos

Os métodos de agrupamento hierárquicos trabalham agrupando os elementos em uma estrutura de árvore de grupos organizando-os em formato hierárquico. Eles resultam em uma sequência aninhada de partições criando uma hierarquia de classes de dados, dependendo da maneira como esta é construída, pode-se classificar os métodos hierárquicos em aglomerativos ou divisivos (Mello, 2008; Naldi, 2010).

Os resultados dos algoritmos hierárquicos geralmente são apresentados como uma estrutura de árvore binária ou um dendograma, apesentando os elementos analisados e como são agrupados a cada execução do algoritmo. A raiz do dendograma representa o conjunto de dados inteiro, os nós folha representam os indivíduos, os nós intermediários representam a importância da proximidade entre os indivíduos. A altura do dendograma expressa à distância

entre um par de indivíduos ou entre um par de grupos ou ainda entre um indivíduo e um grupo. O resultado do agrupamento pode ser obtido cortando-se o dendograma em diferentes níveis. Esta forma de representação fornece a visualização da estrutura de grupos e descrições informativas, especialmente quando há realmente relações hierárquicas nos dados como, por exemplo, dados de pesquisa sobre evolução de espécies (Junior, 2006; Mello, 2008).

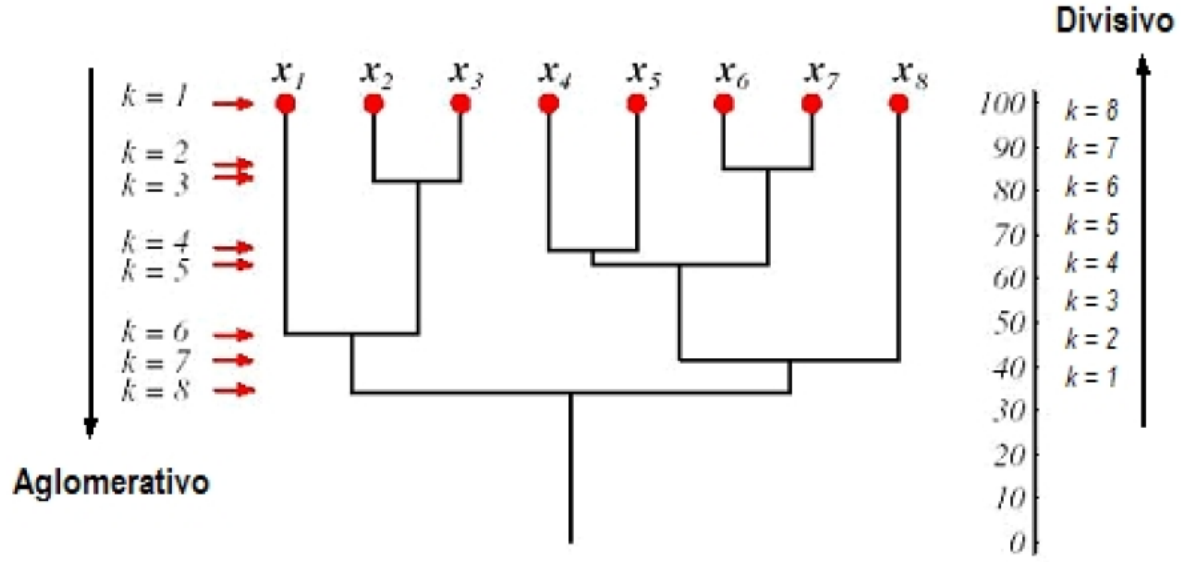

Figura 8 - Exemplo dos Algoritmos Hierárquicos Aglomerativo e Divisivo na Representação de um Dendograma (Junior, 2006).

Os métodos aglomerativos (*bottom-up)* constroem a hierarquia de grupos iniciando com *n* grupos e cada grupo com um único elemento, ou seja, inicialmente temos o mesmo número de grupos que o número de elementos. A cada passo o algoritmo une dois grupos de acordo com um critério de similaridade entre eles, isto é, os grupos mais similares são aglomerados, formando novos grupos. O algoritmo segue essa iteração até que algum limite de similaridade intergrupos (entre grupos) seja alcançado ou até que seja formado um único grupo com todos os elementos (Pereira, 2007; Mello, 2008).

Os grupos são unidos baseados na similaridade, distância entre eles, que geralmente é calculada através da distância entre os centros dos grupos, ou através da distância do par de dados mais próximo, ou ainda através da média da distância entre todos os pares de dados dos dois grupos. A escolha da união dos grupos dos algoritmos aglomerativos geralmente depende do conjunto de dados e dos objetivos da aplicação (Pereira, 2007; Rezende et al., 2011).

Existem vários algoritmos de agrupamentos hierárquicos aglomerativos, os mais conhecidos são: método de ligação simples (*Single Link*) que utiliza o critério do vizinho mais próximo, método de ligação completa (*Complete Link*) que utiliza o critério do vizinho mais

distante e o método da medida das distâncias (*Average Link*) que utiliza a média da distância entre dois grupos. A principal diferença entre eles está no critério de seleção do par de grupos mais próximo. O pseudocódigo do algoritmo 3 foi contextualizado por Rezende et al (2011), sendo um típico modelo de algoritmo hierárquico aglomerativo onde na linha 4 aplica-se esses critérios (Rezende et al., 2011).

Algoritmo 3: Agrupamento hierárquico aglomerativo

Entrada: X = $\{x_1, x_2, \dots, x_n\}$**:** conjunto de dados;
Saída: S = $\{P_1, \dots, P_n\}$: lista de agrupamentos formados;

1. fazer com que cada elemento x ∈ X seja um grupo;
2. computar a dissimilaridade entre todos os pares distintos de grupos;
3. repita
4. selecionar o par de grupos mais próximo;
5. unir os dois grupos para formar um grupo G;
6. computar a dissimilaridade entre G e os outros grupos;
7. até (obter um único grupo com todos os elementos);

Os métodos divisivos (*top-down*) seguem a filosofia oposta dos métodos aglomerativos. Estes partem do nível mais alto da hierarquia, pois iniciam com um único grupo contendo *n* elementos, após esse grupo é dividido em grupos menores de acordo com a similaridade entre eles. Essas divisões são realizadas iterativamente até que algum limiar de similaridade seja alcançado ou até que todos os elementos pertençam a *n* grupos de um único elemento (Pereira, 2007; Mello, 2008).

A complexidade apresentada na estratégia dos métodos divisivos faz com eles não sejam muito utilizados na literatura, pois eles crescem exponencialmente em relação ao tamanho do conjunto de dados isso faz com que a sua utilização se torne inviável em conjuntos de dados grandes (Rezende et al., 2011).

Uma característica dos métodos hierárquicos é não reavaliar as junções e separações dos grupos então uma vez que uma junção ou separação de grupos é realizada, esta nunca mais pode ser desfeita. Essa característica pode ser uma desvantagem, caso o critério para unir ou dividir os grupos não seja apropriado (Mello, 2008).

Encontram-se outras desvantagens como: redução dos impactos dos pontos fora do padrão (*outliers*), desempenho, complexidade, entre outras. Algumas vantagens também são observadas como: disponibilizar diversas medidas de similaridade, simplicidade na visualização dos resultados, entre outras (Mello, 2008; Machado, 2011).

3.3.3 Imunológico

O algoritmo Imunológico foi desenvolvido por Machado (2011) com o objetivo de formar grupos de dados bem definidos constituídos por instâncias similares através de conjuntos e dados escolhidos. Esse algoritmo leva em consideração uma maior homogeneidade entre os elementos do mesmo grupo e uma maior heterogeneidade entre os elementos de grupos diferentes. Além disso, utiliza dois conceitos principais com origem na área de SIA: a teoria da seleção clonal e o princípio da seleção negativa (Machado, 2011).

A teoria da seleção clonal refere-se à resposta imune adaptativa, onde os linfócitos se ligam a antígenos que irão se proliferar através de "clones". Esses clones podem ser divididos em células de plasma com vários anticorpos de curta duração de vida ou em células de memória para auxiliar no reconhecimento de um antígeno que já passou pelo organismo. Segundo essa teoria, o sistema é capaz de se modificar, conforme um repertório inicial de células imunes e experiências que o mesmo teve com o ambiente. Isso é realizado através de um processo de aprendizagem que envolve uma população de adaptação de unidades de informação onde cada uma representa uma solução do problema (Brownlee, 2011; Machado, 2011).

A teoria da seleção negativa procura fornecer tolerância às células próprias evitando assim que as células do corpo se destruam. Esse problema é conhecido como discriminação do próprio/não-próprio. O paradigma imunológico do próprio/não-próprio não é intuitivo, pois ele propõe uma modelagem de um domínio desconhecido a partir do que é conhecido, sendo realizada a partir de modelos de mudanças e anomalias de dados desconhecidos. Os modelos criados pelo paradigma próprio/não-próprio são usados para monitorar a existência de dados normais ou fluxos de dados buscando novos padrões para os dados desconhecidos (Forrest et al., 1994; Brownlee, 2011; Machado, 2011).

O algoritmo imunológico pode ser definido da seguinte forma: Dado um conjunto de dados, a representação de entrada dos dados é feita para cada linha *i* do conjunto de dados. Cada linha representa um antígeno definido por Ag_i. O conjunto de antígenos pode ser descrito como $Ag_i = \{Ag_1, Ag_2, \dots, Ag_i\}$ que representa os elementos a serem agrupados. O processamento dos dados ocorre a partir da criação dos linfócitos para a verificação da afinidade entre os antígenos, onde está afinidade diz se o antígeno pertence ou não a determinado grupo. A saída dos dados é representada pelos grupos de anticorpos $Ab_j = \{Ab_1, Ab_2, \dots, Ab_j\}$ que possuem um ou mais antígenos, sendo que cada anticorpo é definido por Ab_j e j é o número do grupo.

O pseudocódigo apresentado no algoritmo 4 contextualizado por Machado (2011) mostra o algoritmo imunológico para agrupamento de dados. A execução do algoritmo inicia com *n* antígenos correspondentes as *n* linhas do conjunto de dados a serem agrupadas e cada linha será representada por um elemento. Aplicando a normalização nos dados caso seja selecionada a opção na tela de seleção. Em seguida é criado um conjunto de linfócitos onde serão selecionados alguns antígenos aleatoriamente que servirão como base inicial para os grupos. Após isso os linfócitos, serão confrontados com os antígenos para formar os grupos verificando a afinidade entre os linfócitos e o antígeno em questão, através do cálculo de similaridade.

A partir das iterações iniciais os linfócitos serão clonados e mutados, recalculando a afinidade entre estes e os antígenos, assim poderá ser verificado se os grupos melhoraram ou não conforme alguns índices de validação. Repetindo essa iteração até que se atinja determinada condição de parada definida pelo usuário. O resultado do algoritmo será um grupo de anticorpos que são os elementos agrupados.

Algoritmo 4: Imunológico

Entrada: Antígenos
Saída: Grupo de Anticorpos

1. ler arquivo de entrada (antígenos);
2. normalizar dados (quando solicitado);
3. selecionar linfócitos iniciais através do conjunto de dados original;
4. formar agrupamento inicial;
5. calcular homogeneidade de cada grupo;
6. calcular homogeneidade global;
7. calcular separação;
8. repetir
   - 8.1. mover linfócitos para centro do grupo;
   - 8.2. formar novo agrupamento;
   - 8.3. calcular homogeneidade de cada grupo;
   - 8.4. calcular homogeneidade global;
   - 8.5. calcular separação;
   - 8.6. selecionar o melhor agrupamento através dos índices calculados (homogeneidade e separação);

   até atingir critério de parada
9. repetir
   - 9.1. clonar linfócitos;
   - 9.2. gerar mutação dos novos linfócitos;
   - 9.3. formar novos grupos;
   - 9.4. calcular homogeneidade de cada grupo;
   - 9.5. calcular homogeneidade global;
   - 9.6. calcular separação
   - 9.7. selecionar o melhor agrupamento através dos índices calculados (homogeneidade e separação)

até atingir o critério de parada
10. retorna melhor agrupamento (anticorpos)

A opção de escolha de parada do algoritmo imunológico é uma das vantagens em relação aos outros algoritmos apresentados, pois podem ocorrer convergências diferentes conforme o tipo de parada, sendo necessário executar diversos testes. Outras vantagens são: os índices de homogeneidade e separação disponíveis para a execução do algoritmo. Uma desvantagem encontrada é a necessidade de informar mais 3 parâmetros para execução do algoritmo que podem interferir nos resultados dos agrupamentos.

4 PLANEJAMENTO DO EXPERIMENTO

O experimento a ser desenvolvido utilizará três conjuntos de dados públicos retirados do PSCL DataShop, onde a integridade dos dados é garantida possuindo diversas publicações usando dados desse repositório como base. Além dos conjuntos de dados, será utilizado o algoritmo imunológico com a sua ferramenta própria e os algoritmos k-média, EM e hierárquicos disponíveis nas ferramentas WEKA e R. Após, os algoritmos serão aplicados nos dados com o auxílio das ferramentas e os resultados serão avaliados conforme os critérios de homogeneidade e separação.

4.1 PREPARAÇÃO DO EXPERIMENTO

A partir do estudo efetuado no decorrer desse trabalho, apresenta-se a proposta do experimento comparativo de algoritmos de agrupamento, que inicialmente foi dividida em seis etapas:

- Na primeira etapa selecionou-se os algoritmos de maior relevância seguindo alguns critérios exemplificados a seguir:
  - K-média: por ser um dos algoritmos mais conhecidos dentro da técnica de agrupamento;
  - EM: por ser um algoritmo estatístico e não precisar da informação do número de grupos.
  - Imunológico: por ser um algoritmo novo que utiliza conceitos da área de SIA;
  - Hierárquicos: por trabalhar bem com dados em formato hierárquico, que pode ser ou não o caso dos dados educacionais;

- Na segunda etapa escolheu-se as ferramentas WEKA e R pela facilidade de trabalhar com algoritmos de agrupamento e a possibilidade de trabalhar diretamente com a API de cada uma delas.

- Na terceira etapa identificou-se os conjuntos de dados públicos: *Geometry, Chinese Tone Study* e Álgebra I 2006. A escolha desses conjuntos ocorreu, pois os dados estão divididos em vários formatos deixando a análise ainda mais complexa, seguindo assim a ideia inicial do experimento que é verificar o desempenho dos algoritmos.

- Na quarta etapa escolheu-se dois critérios de avaliação: homogeneidade e separação.
- A quinta etapa efetuou a execução do experimento;
- Na sexta etapa apresenta o resultado do experimento. Por se tratar de um estudo quantitativo, as conclusões serão apresentadas em tabelas mostrando o desempenho de cada algoritmo e concluindo quais algoritmos obtiveram melhores resultados.

A quinta e sexta etapas (execução e análise do experimento) foram dividida em 8 atividades principais, apresentadas no workflow da figura 9.

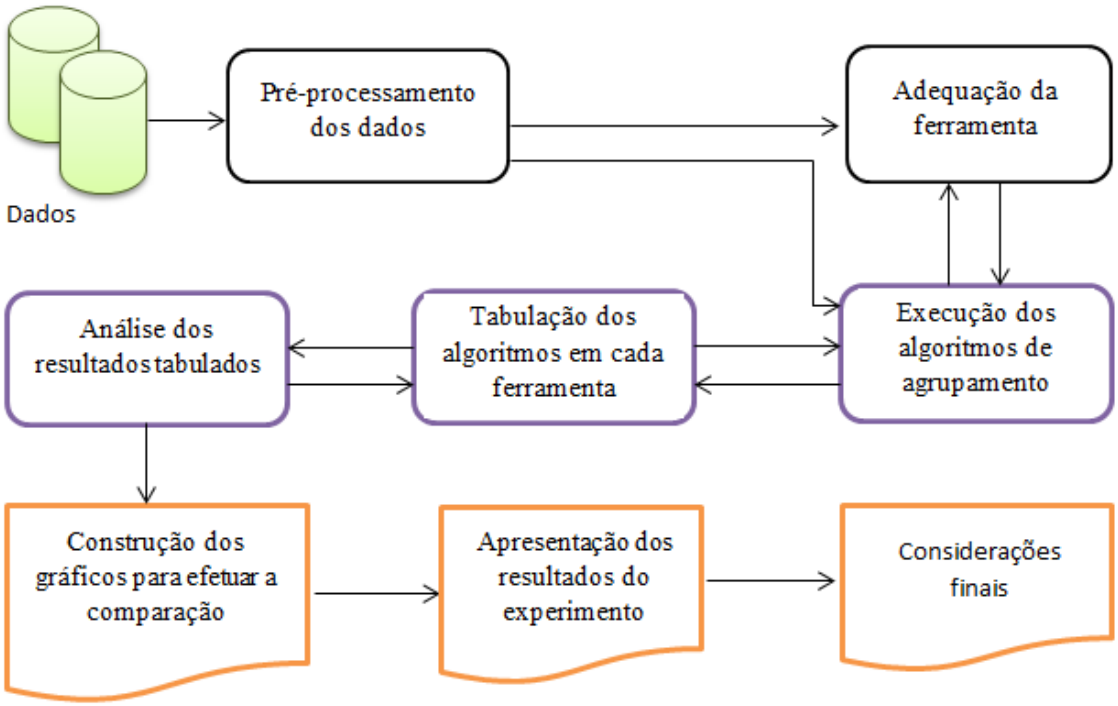

Figura 9 - Workflow de atividades para execução da proposta.

As atividades do workflow são brevemente descritas a seguir:

1. Os dados passarão por um pré-processamento para remover alguma sujeira ou duplicidade de registros existentes, aplicar a normalização (caso seja necessário), entre outros tratamentos;
2. A adequação das ferramentas consiste em ajustar os critérios de avaliação nos algoritmos, entre outros eventos que venham a ocorrer;

3. Execução dos algoritmos modificados;
4. A cada algoritmo executado os dados serão tabulados;
5. A análise dos resultados e as etapas 3 e 4 ocorrerão diversas vezes para obter-se um resultado confiável;
6. A construção das tabelas utilizará os resultados das etapas 4 e 5 para demonstrar o desempenho dos algoritmos;
7. A apresentação dos resultados comparará as tabelas construídas na etapa 6 e identificará os algoritmos mais relevantes para a área de MDE;
8. As considerações finais apresentarão se o experimento obteve sucesso ou não. Identificando os pontos chave da contribuição do experimento para outros trabalhos científicos na área.

4.2 CONJUNTO DE DADOS GEOMETRIA

O primeiro conjunto de dados escolhido é o *Geometry*. Ele é composto por dados retirados de um ano de curso de Geometria ocorrido no período de 1996-1997 tendo como principal investigador Ken Koedinger. Esse conjunto foi formado a partir de arquivos que simulam o passo a passo dos estudantes, sendo que os tempos apresentados no conjunto são simulados. No entanto como o estudo desse trabalho não foca na predição dos tempos de aprendizagem, a utilização dos mesmos não ocasionará problemas (Koedinger et al., 2010).

Para a exportação do arquivo foram selecionados todos os modelos de conhecimento e todas as datas disponíveis. O arquivo é constituído por dados de 59 estudantes, 6.778 instâncias, dividido em 29 modelos de conhecimento e cada modelo possui as suas áreas de conhecimento subdivididas em níveis. Alguns modelos de conhecimento podem ser observados nas colunas das tabelas 2 e 3 apresentadas a seguir, identificados com KC (knowledge componentes, componentes de conhecimento) no início da descrição (Koedinger et al., 2010).

As tabelas 2 e 3 exemplificam o conjunto de dados *Geometry* mostrando algumas linhas e colunas. A tabela 2 mostra diversas linhas sobre um estudante e os exercícios resolvidos por ele, já a tabela 3 mostra diversos alunos agrupados por uma área de conhecimento específica.

A tabela 2 apresenta alguns dados selecionados para exemplificar o conjunto com as seguintes colunas: o número da linha do conjunto de dados, o código fictício do estudante, o nome da etapa em que se encontra o exercício, o resultado classificado em correto ou incorreto, a coluna área representa se o exercício possui fórmula de área ou não e a última

coluna mostra qual parte geométrica o exercício está avaliando como, por exemplo: retângulo, paralelogramo, quadrado, etc.

Tabela 2 – Demonstração dos dados agrupados por um estudante.

Linha	Código Estudante	Nome Etapa	Resultado	KC (Área)	KC (Livro de Texto)
13	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(AREA QUESTION1)	CORRECT	Non-area formula	parallelogram-area
14	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(HEIGHT QUESTION2)	CORRECT	Non-area formula	parallelogram-area
15	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(BASE QUESTION3)	CORRECT	Non-area formula	parallelogram-area
16	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(AREA QUESTION1)	CORRECT	Non-area formula	square-area
17	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(HEIGHT QUESTION2)	CORRECT	Non-area formula	square-area
18	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(BASE QUESTION3)	CORRECT	Non-area formula	square-area
19	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(DOOR-AREA QUESTION1)	INCORRECT	Non-area formula	rectangle-area
20	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(DOOR-AREA QUESTION1)	CORRECT	Non-area formula	rectangle-area
21	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(WALL-AREA QUESTION1)	INCORRECT	Non-area formula	rectangle-area
22	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(WALL-AREA QUESTION1)	CORRECT	Non-area formula	rectangle-area
23	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(PAINTED-AREA QUESTION1)	INCORRECT	Area formula	compose-by-addition
24	Stu_0a8e3638e3c0deb4e5e49c72286a5b83	(PAINTED-AREA QUESTION1)	CORRECT	Area formula	compose-by-addition

Fonte: Conjunto de dados retirados de <http://pslcdatashop.org> (Koedinger et al., 2010).

A tabela 3 é estruturada a partir da seleção do passo *Diameter Question1* onde é subdividida em alguns níveis de conhecimento, sendo possível observar dois deles apresentados nas colunas *KC (Original)* e *KC (DecompArithDiam)*. Além dessas colunas, pode-se observar que a terceira coluna apresenta os passos de resolução do problema.

Tabela 3 – Demonstração dos dados agrupados por área de conhecimento.

Linha	Código Estudante	Passos	Resultado	KC (Original)	KC (DecompArithDiam)
5454	Stu_c43d4a17398b2667daacdc70c76cf8ef	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5473	Stu_c43d4a17398b2667daacdc70c76cf8ef	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5717	Stu_ca875f87b1e56812e275dee7791967ec	1	INCORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5718	Stu_ca875f87b1e56812e275dee7791967ec	2	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5908	Stu_cda7a650c5856cf2f6738072447d7825	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5967	Stu_cda7a650c5856cf2f6738072447d7825	1	INCORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
5968	Stu_cda7a650c5856cf2f6738072447d7825	2	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6087	Stu_d4eaf56d62b41278e02f509a10c03610	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-given
6415	Stu_d70bdfbefdabf0ea991cda211f667069	1	INCORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6416	Stu_d70bdfbefdabf0ea991cda211f667069	2	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6435	Stu_d70bdfbefdabf0ea991cda211f667069	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6628	Stu_e7fef5e04a919abde875b9c1a186e77f	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6702	Stu_e7fef5e04a919abde875b9c1a186e77f	1	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6743	Stu_ee503bf91e442ebb22015cbebffc5e88	1	INCORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal
6744	Stu_ee503bf91e442ebb22015cbebffc5e88	2	CORRECT	ALT:CIRCLE-DIAMETER	circle-diam-from-subgoal

Fonte: Conjunto de dados retirados de <http://pslcdatashop.org> (Koedinger et al., 2010).

Inicialmente os autores do conjunto de dados avaliaram quinze habilidades encontradas na coluna KC (*Original)*. Essas habilidades, denominadas regra de produção, são descritas a seguir (Cen et al, 2006):

- *Circle-area*: A partir do raio deve-se encontrar a área do círculo;
- *Circle-circumference*: A partir do diâmetro deve-se encontrar a circunferência de um círculo;
- *Circle-diameter*: A partir do raio ou circunferência deve-se encontrar o diâmetro de um circulo;
- *Circle-radius:* Encontre o raio através da área, circunferência ou diâmetro;
- *Compose-by-addition*: Em a+b=c, a partir de dois valores a, b ou c, encontrar o terceiro;
- *Compose-by-multiplication*: Em a*b=c, a partir de dois valores a, b ou c, encontrar o terceiro;

- *Parallelogram-area*: A partir da base e altura deve-se encontrar a área do paralelogramo;
- *Parallelogram-side*: A partir da área e altura (ou base) deve-se encontrar a base (ou altura);
- *Pentagon-area*: A partir de um lado e do apótema deve-se encontrar a área de um pentágono;
- *Pentagon-side*: A partir da área e do apótema deve-se encontrar o lado (ou o apótema);
- *Trapezoid-area*: A partir da altura e duas bases deve-se encontrar a área de um trapézio;
- *Trapezoid-base*: A partir da área e altura deve-se encontrar a base de um trapézio;
- *Trapezoid-height*: A partir da área e da base deve-se encontrar a altura de um trapézio;
- Triangle-area: A partir da base e da altura deve-se encontrar a área de um triângulo;
- Triangle-side: A partir da base e do lado deve-se encontrar a altura de um triângulo.

Após a definição das regras de produção foram definidos 4 fatores de dificuldade para avaliar os problemas. Esses fatores são apresentados na tabela 4 e descritos a seguir por Cen et al (2006):

- *Embed:* Indica se uma forma é incorporada em outro formato, isto é, dois problemas exigem a mesma regra de produção *Circle-area* em algum dos seus passos. Em um dos problemas o círculo é incorporado a um quadrado, já no outro problema o círculo é apresentado isoladamente. Isso significa que os estudantes podem encontrar maior dificuldade ao calcular a área do círculo quando ela está inserida em outra figura;
- *Backward*: Refere-se à apresentação da fórmula, isto é, a ordem em que ela aparece. A forma no sentido para frente do *Compose-by-addition* é S = S1 + S2, já a sua forma no sentido inverso é S1 = S - S2;
- *Repeat:* Indica se a rega de produção foi utiliza mais de uma vez na solução do mesmo problema;
- *FigurePart*: Indica qual parte da figura geométrica será calculada.

Tabela 4 – Fatores de dificuldade.

Nome do Fator	Valor do Fator
Embed	alone, embed
Backward	forward, backward
Repeat	initial, repeat
FigurePart	area, area-difference, area-combination, diameter, circumference, radius, side, segment, base, height, apothem

Fonte: Tabela retirada do artigo (Cen et al, 2006).

As regras de produção e os fatores de dificuldade podem ser considerados os dados com maior relevância encontrados no conjunto *Geometry,* pois eles representam a base para a construção dos grupos e as análises que serão executadas. Os dados apresentados acima são apenas uma parte das informações que serão analisadas pelos algoritmos de agrupamento.

4.3 CONJUNTO DE DADOS LÍNGUA CHINESA

O segundo conjunto de dados escolhido é o *Chinese Tone Study*. Esse conjunto possui informações sobre os estudantes que estão aprendendo a sua segunda língua. Nesse caso a língua estudada é a Chinesa. O aprendizado dessa língua trabalha com o recurso de tons que representa um desafio maior para quem deseja aprendê-la (Koedinger et al., 2010).

Os dados pertencem ao projeto *Chinese Tone Study* que busca testar hipóteses de aprendizagem baseadas no pressuposto de que se as características críticas do contorno tonal forem atendidas, a aprendizagem torna-se mais fácil. Para isso, o teste projetado foi dividido em três condições onde os estudantes pudessem perceber os diferentes tons. A cada condição os estudantes receberam *feedback* referente ao seu desempenho. O principal investigador deste projeto é Ying Liu, o projeto foi realizado em estudantes da escola CMU no período de 06 de setembro de 2005 a 12 de abril de 2006 (Koedinger et al., 2010).

Para a exportação do arquivo, foram selecionados todos os modelos de conhecimento e todas as datas disponíveis. O arquivo é constituído por dados de 97 estudantes, sendo 48.443 instâncias, dividido em três modelos de conhecimento e cada modelo possui as suas áreas de conhecimento subdivididas em níveis. Os dados foram retirados no dia 08 de junho de 2012 e o projeto ainda estava em andamento, sendo assim modificações nas informações podem ocorrer nesse conjunto de dados, porém não constarão neste trabalho (Koedinger et al., 2010).

A tabela 5 apresenta algumas linhas de um estudante, onde essas linhas foram agrupadas pela lição 1. A coluna Proble refere-se a 2 problemas resolvidos pelo estudante em questão, a coluna P representa quantos passos o estudante percorreu até acertar a resposta do problema, a coluna Resposta está classificada em três níveis: correto, incorreto e ajuda (*hint)*

onde demonstra que o estudante procurou ajuda para encontrar a resposta e a última coluna apresenta o *feedback* para o estudante a cada passo.

Tabela 5 - Demonstração dos dados agrupados pela lição do curso.

Linha	Código do Estudante	Proble	P	Resposta	Feedback
5949	Stu_3a1e655328599b6a1208115f2271c942	l01w01	1	INCORRECT	No, this is not correct. Please click on "hint" in the upper right corner for help.
5950	Stu_3a1e655328599b6a1208115f2271c942	l01w01	2	CORRECT	Correct! Now click on Done to go to the next one.
5951	Stu_3a1e655328599b6a1208115f2271c942	l01w01	1	HINT	Focus on the pitch change and listen to the sound again.
5952	Stu_3a1e655328599b6a1208115f2271c942	l01w01	2	CORRECT	Correct! Now click on Done to go to the next one.
5953	Stu_3a1e655328599b6a1208115f2271c942	l01w01	1	CORRECT	No HInt or Other Message Text
5954	Stu_3a1e655328599b6a1208115f2271c942	l01w02	1	INCORRECT	No, this is not correct. Please click on "hint" in the top right corner for help.
5955	Stu_3a1e655328599b6a1208115f2271c942	l01w02	2	INCORRECT	No, this is not correct. Please click on "hint" in the top right corner for help.
5956	Stu_3a1e655328599b6a1208115f2271c942	l01w02	3	CORRECT	Correct! Now click on Done to go to the next one.
5957	Stu_3a1e655328599b6a1208115f2271c942	l01w02	1	CORRECT	No HInt or Other Message Text

Fonte: Conjunto de dados retirados de <http://pslcdatashop.org> (Koedinger et al., 2010).

Esse conjunto de dados possui a sua divisão principal em lições e cada lição possui os seus problemas com um determinado peso. Para cada problema o estudante pode efetuar mais de um passo e cada passo possui uma ajuda com dicas para a resolução dos problemas. Um exemplo dessa ajuda pode ser vista na tabela 4 na linha 5954. O estudante respondeu o problema 101w02 e foi comunicado que a resposta estava incorreta e poderia solicitar ajuda. Porém, mesmo com a opção de ajuda, ele preferiu assinalar novamente uma resposta sendo comunicado que a sua escolha estava incorreta. O estudante acertou a questão na terceira tentativa, sem recorrer à opção de ajuda disponível.

4.4 CONJUNTO DE DADOS ÁLGEBRA

O terceiro conjunto de dados escolhido é a Álgebra I 2006. Ele foi selecionado, pois possui uma quantidade relativa de números e equações matemáticas, aumentado ainda mais a complexidade dos dados. Esse conjunto é usado no projeto *A Multimodal Interface for Solving Equations*, tendo como sua principal investigadora Lisa Anthony e aplicação na escola *Central Westmoreland Carreira e Centro de Tecnologia* (CWCTC) no período de 12 de outubro a 20 de dezembro de 2006. O projeto compara padrões cognitivos de um tutor de Álgebra que trabalha com exemplos, tendo como ideia principal de incentivar o estudante a praticar um tipo de problema levando-o a observar o seu aprendizado e desempenho com o passar do tempo.

Para a exportação do arquivo, foram selecionados todos os modelos de conhecimento e todas as datas disponíveis. O arquivo é composto por dados de 31 estudantes, sendo 12.568 instâncias, dividido em quatro modelos de conhecimento, porém, só dois desses modelos possuem dados com subdivisões.

A tabela 6 demonstra algumas linhas e colunas do conjunto de dados selecionado. A coluna L é o identificador da linha no conjunto, a coluna P é o passo em que o estudante está na resolução do problema, a coluna R apresenta o resultado da resolução do problema, as colunas Seleção, Ação e Ent fazem parte da resolução do problema em questão. Pode-se observar que os passos da resolução a partir da linha 15 começam na coluna Seleção onde a estratégia inicia com o problema em si e, nas próximas linhas pode-se observar que a cada iteração a coluna Seleção vai agregando as colunas Ação e Ent até obter a solução completa.

Tabela 6 – Demonstração dos dados agrupados pelo problema.

L	Código Estudante	Problema	Nome Etapa	P	R	Seleção	Ação	Ent.
15	Student 09tSMme Session 1	Eg40 -6x+8 = 8	-6x+8 = 8	1	OK	-6x+8 = 8 strategic	subtract	8
16	Student 09tSMme Session 1	Eg40 -6x+8 = 8	-6x+8-8 = 8-8	1	OK	-6x+8-8 = 8-8 strategic	clt	-6x+8-8
17	Student 09tSMme Session 1	Eg40 -6x+8 = 8	-6x = 8-8	1	OK	-6x = 8-8 strategic	clt	8-8
18	Student 09tSMme Session 1	Eg40 -6x+8 = 8	-6x = 0	1	OK	-6x = 0 strategic	divide	-6
19	Student 09tSMme Session 1	Eg40 -6x+8 = 8	-6x/-6 = 0/-6	1	OK	-6x/-6 = 0/-6 strategic	rf	-6x/-6
20	Student 09tSMme Session 1	Eg40 -6x+8 = 8	x = 0/-6	1	OK	x = 0/-6 strategic	rf	0/-6

Fonte: Conjunto de dados retirados de <http://pslcdatashop.org> (Koedinger et al., 2010).

A escolha desse conjunto de dados foi proposital para observar como os algoritmos de agrupamento se comportam quando o conjunto possui muito números e equações matemáticas elevando o grau de dificuldade para agrupar os dados.

4.5 CRITÉRIOS DE AVALIAÇÃO

Os grupos formados pelos algoritmos selecionados serão avaliados pelos critérios de homogeneidade e separação, entre outros. A homogeneidade consiste em ter os dados mais similares possíveis em cada grupo, por isso quanto menor o resultado da equação da homogeneidade mais similar será o grupo. A separação consiste em ter um grupo mais distante do outro, assim, quanto maior o resultado da equação da separação maior será a dissimilaridade entre os grupos.

A homogeneidade é calculada como a distância média entre cada perfil de expressão gênica e do centro do grupo que pertence, onde g_i é o elemento *i*-ésimo de cada grupo, $C(g_i)$

é o centro do grupo que g_i pertence, N_{gene}é o número total de instâncias e D é a função da distância a ser aplicada. A seguir equação 4 apresenta a fórmula da homogeneidade (Chen et al., 2002; Machado, 2011):

$$H_{ave} = \frac{1}{H_{gene}} \sum D(g_i, C(g_i))$$

Equação 4 – Fórmula da homogeneidade

A separação é calculada como a distância média ponderada entre os centros dos grupos, onde C_i e C_j são os centros dos grupos do *i*-ésimo e *j*-ésimo, e Nc_i e Nc_j são números de instâncias nos grupos do *i*-ésimo e *j*-ésimo e D é a função da distância a ser aplicada. A seguir equação 5 apresenta a fórmula da separação (Chen et al., 2002; Machado, 2011):

$$S_{ave} = \frac{1}{\sum_{i \neq j} Nc_i Nc_j} \sum_{i \neq j} Nc_i Nc_j \, D(C_i, C_j)$$

Equação 5 – Fórmula da separação

A partir dessas equações consegue-se demonstrar a solidez dos grupos observando que H_{ave} e S_{ave} são fortemente ligados, enquanto S_{ave} representa a distância entre os grupos de forma crescente, já H_{ave} representa união dos dados do grupo de forma decrescente.

A função de distância D utilizará como medida a distância euclidiana (Chen et al., 2002), que será apresentada a seguir pela equação 6, onde:

- D_{ij} é a distância euclidiana do elemento *i* para o elemento *j;*
- *l* é o índice do atributo do vetor de *d* atributos dos elementos; e
- x_{il} é o l-ésimo atributo do elemento *i* (Mello, 2008).

$$D_{ij} = \sqrt{\sum_{l=1}^{d} (x_{il} - y_{jl})^2}$$

Equação 6 – Distância euclidiana.

A distância euclidiana é uma das mais utilizadas em algoritmos de agrupamento. Ela visa expressar a distância geométrica Euclidiana entre os exemplos em um espaço multidimensional, tentando encontrar grupos de elementos de formato esférico (Mertz, 2006; Mello, 2008; Machado, 2011).

5 EXPERIMENTO

O experimento iniciou com a escolha dos atributos e os ajustes necessários para as instâncias de cada conjunto de dados. Após foram desenvolvidas as fórmulas de homogeneidade e separação juntamente com a adequação dos algoritmos pertencentes à ferramenta WEKA. Por fim, desenvolveu-se um software que lê um arquivo com os agrupamentos efetuados pela ferramenta R e gera os resultados dos cálculos da homogeneidade e separação.

Com os dados e as ferramentas prontas passou-se aos estudos de casos para cada conjunto de dado, executando os algoritmos.

5.1 PREPARAÇÃO DOS CONJUNTOS DE DADOS

A preparação dos conjuntos de dados começou com a análise e identificação dos atributos de maior relevância, além da remoção de caracteres especiais como: “(“, “)”, “:”, entre outros. Isto foi feito sempre observando a necessidade da remoção pois encontram-se muitas equações no conjunto de dados Álgebra, impossibilitando a remoção do carácter “(”, por exemplo. Após essa preparação inicial, começou-se a adaptação dos conjuntos de dados para cada ferramenta.

A tabela 7 apresenta os dados no seu formato original sem manipulação.

Tabela 7 – Dados antes da conversão

Problem Name	Step Name	Back	Figure-part	Repeat	KC(Original)
RECTANGLE_ABCD	(AREA QUESTION1)	0	Área	initial	ALT:PARALLELOGRAM-AREA
RECTANGLE_ABCD	(HEIGHT QUESTION2)	1	Height	initial	ALT:PARALLELOGRAM-SIDE
RECTANGLE_ABCD	(BASE QUESTION3)	1	Base	repeat	ALT:PARALLELOGRAM-SIDE
BUILDING_A_SIDEWALK	(LARGE-RECTANGLE-AREA QUESTION1)	0	Área	repeat	ALT:PARALLELOGRAM-AREA
BUILDING_A_SIDEWALK	(POOL-AREA QUESTION1)	0	Área	repeat	ALT:PARALLELOGRAM-AREA
BUILDING_A_SIDEWALK	(SIDEWALK-AREA QUESTION1)	1	area-difference	repeat	ALT:COMPOSE-BY-ADDITION
BUILDING_A_SIDEWALK	(POOL-AREA QUESTION2)	0	Área	repeat	ALT:PARALLELOGRAM-AREA

A tabela 8 exibe a estrutura do arquivo de dados para o algoritmo imunológico. O arquivo de dados possui a seguinte formatação:

- Não possui cabeçalho;
- Atributos separados por vírgula;
- Ao final de cada linha deve possuir o carácter ponto e vírgula;

– Os valores dos atributos devem ser numéricos. Exceto os valores pertencentes ao último atributo.

Tabela 8 – Arquivo de dados para o algoritmo Imunológico.

0,0,0,0,0,0,0,0,0,0,0,ALT:PARALLELOGRAM-AREA;
0,0,1,0,0,0,1,0,1,0,0,ALT:PARALLELOGRAM-SIDE;
0,0,2,0,0,0,1,0,2,0,1,ALT:PARALLELOGRAM-SIDE;
1,0,3,0,0,0,0,0,0,0,1,ALT:PARALLELOGRAM-AREA;
1,0,4,0,0,0,0,0,0,0,1,ALT:PARALLELOGRAM-AREA;
1,0,5,0,0,1,1,0,3,1,1,ALT:COMPOSE-BY-ADDITION;
1,0,7,0,0,0,0,0,0,0,1,ALT:PARALLELOGRAM-AREA;

A tabela 9 apresenta a estrutura do arquivo de dados para a ferramenta R. O arquivo de dados possui a seguinte formatação:

– Cabeçalho opcional;

– Atributos separados por vírgula;

– Valores dos atributos devem ser numéricos.

Tabela 9 – Dados convertido para a ferramenta R.

0,0,0,0,0,0,0,0,0,0,0
0,0,1,0,0,1,0,1,0,1,0,0
0,0,2,0,0,1,0,1,0,2,0,1
1,0,3,0,0,0,0,0,0,0,1
1,0,4,0,0,0,0,0,0,0,1
1,0,5,0,0,2,1,1,0,3,1,1
1,0,6,0,0,0,0,0,0,0,1

A tabela 10 apresenta a estrutura do arquivo de dados para a ferramenta WEKA. O arquivo de dados é dividido em duas partes: cabeçalho e dados.

1) Cabeçalho é composto pelos seguintes itens:
   – Nome que identifica o arquivo;
   – Para atributos do tipo categórico, é preciso identificar os valores distintos e cita-los no cabeçalho atribuindo um nome para o atributo;
   – Marco inicial dos dados no arquivo.
2) Dados:
   – Atributos separados por vírgula.

Tabela 10 – Arquivos de dados para a ferramenta WEKA.

@relation geometria @attribute problemname {rectangle_abcd, building_a_sidewalk} @attribute problemview {1} @attribute stepname {area_question1,height_question2,base_question3, large_rectangle_area_question1,pool_area_question1,sidewalk_area_question1, pool_area_question2} @attribute attemptatstep {1} @attribute outcome {correct, incorrect} @attribute kcoriginal {alt_parallelogram_area, alt_parallelogram_side, alt_compose_by_addition} @attribute kcarea{area_formula, non_area_formula} @attribute cffactorbackward {0,1} @attribute cffactorembeddedness {alone, embedded} @attribute cffactorfigurepart {area, height, base, area_difference} @attribute cffactorfiguretype {0, rectangle} @attribute cffactorrepeat {initial, repeat} @data
rectangle_abcd,1,area_question1,1,correct,alt_parallelogram_area,non_area_formula,0,alone, area,rectangle,initial
rectangle_abcd,1,height_question2,1,correct,alt_parallelogram_side,non_area_formula,1,alon e,height,rectangle,initial
rectangle_abcd,1,base_question3,1,correct,alt_parallelogram_side,non_area_formula,1,alone, base,rectangle,repeat
building_a_sidewalk,1,large_rectangle_area_question1,1,correct,alt_parallelogram_area,non_ area_formula,0,alone,area,rectangle,repeat
building_a_sidewalk,1,pool_area_question1,1,correct,alt_parallelogram_area,non_area_form ula,0,alone,area,rectangle,repeat
building_a_sidewalk,1,sidewalk_area_question1,1,correct,alt_compose_by_addition,area_for mula,1,alone,area_difference,0,repeat
building_a_sidewalk,1,pool_area_question2,1,correct,alt_parallelogram_area,non_area_form ula,0,alone,area,rectangle,repeat

5.2 PREPARAÇÃO DOS SOFTWARES DE CONVERSÃO DOS CONJUNTOS DE DADOS

Para facilitar a adaptação dos dados, desenvolveu-se dois softwares de conversão. Os softwares lêem dados de um arquivo XLS e geram um arquivo TXT, respeitando as estruturas das ferramentas WEKA, R, e Imunológico.

O primeiro software, como pode ser visto na figura 10 converte os dados e gera o arquivo respeitando a formatação exigida pela ferramenta R e o algoritmo Imunológico. É preciso identificar cada valor diferente de cada atributo e substituir o mesmo por um número inteiro sequencial, separando cada atributo por vírgula. Além disso, o software possui duas opções especiais para o algoritmo Imunológico: a primeira permite incluir o carácter especial

ponto e vírgula no fim de cada linha e a segunda possibilita a escolha do atributo classificatório.

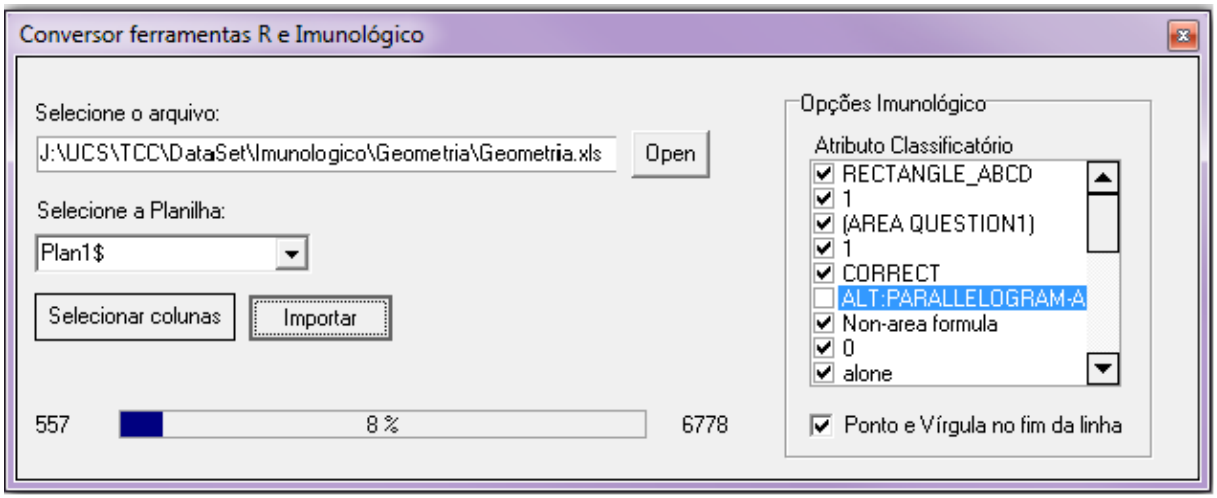

Figura 10 - Software de conversão para a ferramenta R e o algoritmo Imunológico.

O segundo software, como pode ser visto na figura 11, gera o arquivo respeitando a formatação exigida pela ferramenta WEKA.

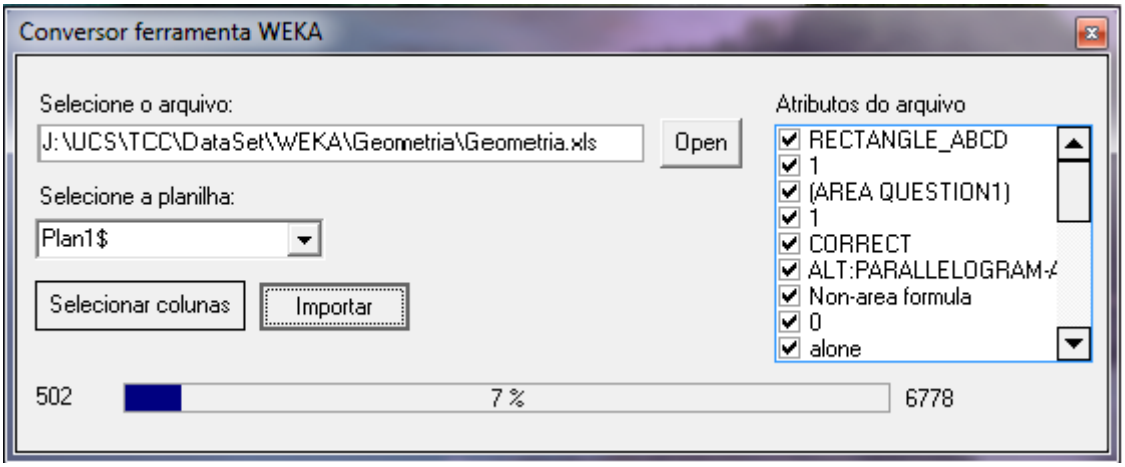

Figura 11 - Software de conversão para a ferramenta WEKA.

5.3 PREPARAÇÃO DOS ALGORITMOS E DESENVOLVIMENTOS

A preparação dos algoritmos e os desenvolvimentos foram constituídos pelas seguintes atividades:

- Desenvolvimento das fórmulas de homogeneidade e separação;
- Implementação das modificações necessárias nos algoritmos K-média, EM para utilizarem as fórmulas desenvolvidas;

– Desenvolvimento de um software para calcular as fórmulas de homogeneidade e separação para os agrupamentos efetuados na ferramenta R.

5.3.1 Fórmulas de Homogeneidade e Separação

A fórmula da homogeneidade possui o cálculo dividido em dois passos:

1) Calcula-se a homogeneidade de cada grupo através dos seguintes passos:
   – Calcula-se a distância entre cada instância e o centróide através da fórmula da distância euclidiana disponível no WEKA;
   – Somam-se as distâncias calculadas;
   – Após aplica-se a seguinte fórmula (1/total de instâncias do grupo) * (o somatório das distâncias);
2) Em seguida, calcula-se a homogeneidade global com os seguintes passos:
   – Para cada grupo formado, multiplica-se o valor da sua homogeneidade pelo número de instâncias do grupo;
   – Somam-se as multiplicações;
   – Após aplica-se a seguinte fórmula (1/total de instâncias do conjunto) * (somatório das homogeneidades).

A fórmula da separação é calculada da seguinte forma:

Repetir esses passos para todos os grupos que ainda não foram selecionados:

   – Selecionar dois grupos;
   – Armazenar o total de instâncias do primeiro grupo;
   – Armazenar o total de instâncias do segundo grupo;
   – Calcular a distância entre os centróides dos dois grupos;
   – Multiplicar o total de instâncias dos dois primeiros grupos com a distância entre os centróides;
   – Somar os valores gerados;
   – Calcular um denominador, onde se soma a multiplicação ocorrida entre o número total de instâncias dos dois primeiros grupos;

Por fim, aplicar a fórmula (1/denominador) * (somatório das distâncias).

As fórmulas foram desenvolvidas na linguagem de programação JAVA utilizando a IDE Eclipse versão 4.2.0.

Os métodos desenvolvidos para os cálculos de homogeneidade e separação podem ser vistos nos anexo: A (método do cálculo da homogeneidade do grupo), B (método do cálculo da homogeneidade global) e C (método do cálculo da separação).

5.3.2 Ferramenta WEKA

A ferramenta WEKA é um software de código aberto que possibilita os desenvolvedores e pesquisadores utilizarem os seus algoritmos para pesquisas. Dentre todos os algoritmos encontrados no WEKA escolheu-se o K-média e o EM para serem testados no experimento proposto.

Para executar os algoritmos K-média e EM, foi necessário criar uma classe que inicializa cada um dos algoritmos como se estivessem rodando pela ferramenta WEKA, após modificou-se cada um deles para calcularem a homogeneidade e separação.

O algoritmo K-média foi modificado ao final do método *bluiderCluster* que é o responsável por fazer os agrupamentos. A modificação ocorreu antes do retorno dos grupos formados. Neste momento foi incluída a chamada dos cálculos das fórmulas de homogeneidade e separação.

O algoritmo EM possuiu duas modificações no final do método *iterate*:

1) Criação dos centróides probabilísticos utilizados nas fórmulas de homogeneidade e separação;
2) Chamada do cálculo das fórmulas de homogeneidade e separação.

O centróide probabilístico é formado pelos valores mais significativos de cada atributo do grupo. Isto é, analisam-se os valores de cada atributo de todas as instâncias pertencentes ao grupo e identifica-se o valor que se repetiu mais vezes formando assim o centróide.

Os códigos utilizandos no experimento foram retirados do pacote weka-src.jar versão 3.6.7 e as classes de inicialização foram desenvolvidos na linguagem de programação JAVA utilizando a IDE Eclipse versão 4.2.0.

A figura 12 exibe o processo executado para chegar ao resultado dos cálculos da homogeneidade e separação envolvendo os algoritmos EM e K-média.

O processo é dividido em 4 etapas principais:

1) Leitura do arquivo com as instâncias para formar os grupos;
2) Execução dos algoritmos;
3) Execução dos cálculos de homogeneidade e separação;
4) Geração do arquivo com os resultados.

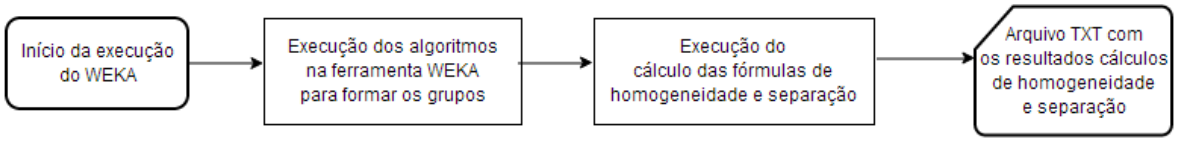

Figura 12 - Workflow referente ao processo executado para chegar ao resultado dos cálculos da homogeneidade e separação no WEKA.

5.3.3 Ferramenta R

Desenvolveu-se um software que lê um arquivo TXT com os agrupamentos efetuados pelos algoritmos da ferramenta R juntamente com o arquivo que contém o conjunto de dados original e calcula as fórmulas de homogeneidade e separação. O arquivo TXT possui a identificação das instâncias por um número referente ao grupo que elas pertencem. Um exemplo deste arquivo pode ser visto no anexo D.

O software desenvolvido possui duas funcionalidades principais:

1) Criar os centróides probabilísticos utilizados nas fórmulas de homogeneidade e separação;
2) Chamar o cálculo das fórmulas de homogeneidade e separação.

O centróide probabilístico é formado pelos valores mais significativos de cada atributo do grupo. Isto é, analisam-se os valores de cada atributo de todas as instâncias pertencentes ao grupo e identifica-se o valor que se repetiu mais vezes formando assim o centróide.

A figura 13 exibe o processo executado para chegar ao resultado dos cálculos da homogeneidade e separação envolvendo o uso da ferramenta R e software desenvolvido.

O processo é dividido por 2 etapas principais:

1) Utilização da ferramenta R:
   - Leitura do arquivo TXT com as instâncias para formar os grupos;
   - Execução dos algoritmos no R;
   - Formação do arquivo com os agrupamentos.
2) Execução do software de cálculo da homogeneidade e separação:
   - Leitura do arquivo com os agrupamentos formados pelo R;
   - Leitura do arquivo que contém o conjunto de dados original;
   - Execução do software de cálculo;
   - Geração do arquivo com os resultados.

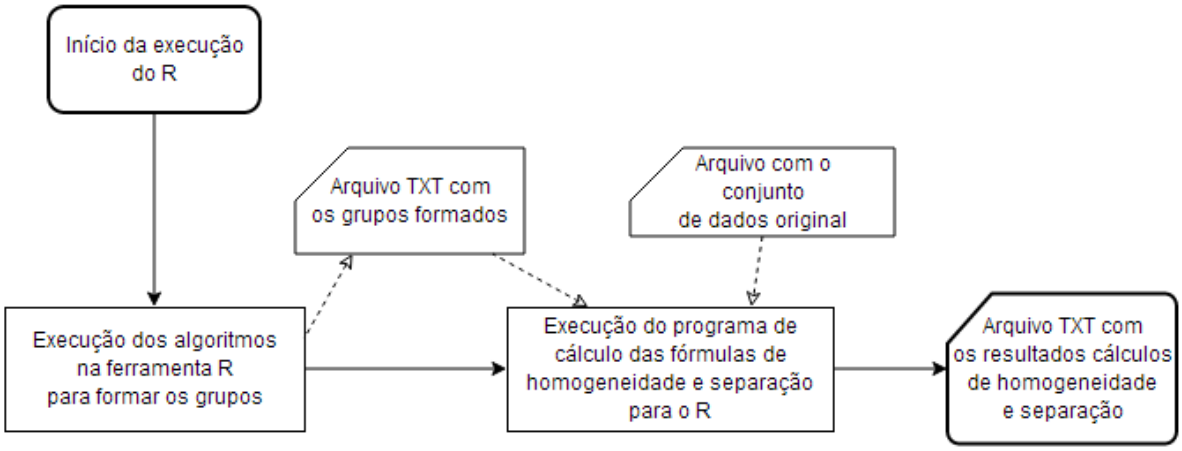

Figura 13 - Workflow referente o processo executado para chegar ao resultado dos cálculos de homogeneidade e separação.

5.3.4 Algoritmo Imunológico

O algoritmo imunológico dentre os algoritmos escolhidos é o mais completo e o menos alterado, pois possui os critérios de homogeneidade e separação utilizados como condição de parada durante a execução.

Após os primeiros testes, o algoritmo imunológico apresentou um consumo elevado de memória. O consumo de memória foi causado pelo acumulo de objetos temporários. Esses objetos são criados a cada iteração para identificar a melhora nos índices de homogeneidade e separação dos grupos formados. Foi necessário modificar a implementação original eliminando esses objetos quando a formação dos grupos não apresentava melhora nos índices.

5.4 ESTUDOS DE CASOS

O estudo de caso foi dividido inicialmente por conjunto de dados e, em sequência, por ferramenta. Para cada conjunto de dados foram aplicados todos os algoritmos pertencentes ao experimento e efetuadas as análises sobre os resultados.

Foram aplicados testes exaustivos para cada conjunto de dados até identificar o número de grupos a ser analisado no experimento. O número de grupos variou conforme a melhor solução identificada pelo valor do *log-likelihood* (valor do log da máxima verossimilhança) estimado pelo EM e pelos resultados apresentados K-média.

5.4.1 Conjunto de Dados Geometria

O estudo de caso para o conjunto de dados Geometria iniciou com a seleção dos atributos considerados mais relevantes, sendo escolhidos os seguintes:

1. Identificação do tipo do problema;
2. Número de resoluções do problema;
3. Identificação da questão do problema;
4. Número de repetições do problema;
5. Resultado da questão (correto ou incorreto);
6. Área de conhecimento avaliada;
7. Identificação de existência de fórmula não na questão;
8. Ordem de apresentação da fórmula na questão;
9. Existência de um ou mais formatos geométricos sobrepostos na questão;
10. Identificação da parte geométrica calculada na questão;
11. Identificação do tipo da figura geométrica;
12. Identificação da repetição da regra de produção.

Com os atributos selecionados e convertidos para cada tipo de ferramenta, inicia-se a execução dos algoritmos. A lista completa dos atributos pode ser encontrada no anexo E.

5.4.1.1 Algoritmo EM

Para a execução do algoritmo EM modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 8 grupos considerando os resultados dos testes preliminares efetuados para o conjunto de dados. Os demais parâmetros permaneceram com as opções *default*.

A tabela 11 apresenta os centróides probabilísticos gerados pelo algoritmo EM, bem como os atributos de maior relevância para cada grupo.

Tabela 11 – Centróides probabilísticos gerados pelo algoritmo EM.

Grupos	Centróides
0	trapezoid_abcd,1,area_question1,1,correct,alt_circle_area,non_area_formula,0, alone, area, circle, initial
1	pogs,1,shaded_area_question1,1,correct,alt_compose_by_addition,area_formula,1, embedded,area_difference,0,initial
2	painting_the_wall,1,wall_area_question1,1,correct,alt_parallelogram_area, non_area_formula, 0,embedded,area,rectangle,initial
3	circle_o,1,circumference_question1,1,correct,alt_circle_circumference, area_formula, 0,alone,circumference,circle,initial
4	pentagon_abcde,1,apothem_question2,1,correct,alt_pentagon_side, non_area_formula, 1,alone,base,pentagon,initial
5	designing_a_quilt,1,white_area_question1,1,correct, alt_compose_by_multiplication, area_formula, 0,alone,area_combination,0,initial
6	trogs,1,area_question1,1,correct,alt_triangle_area,non_area_formula,0,alone,area, triangle, initial
7	circle_o,1,radius_question1,1,correct,alt_circle_radius,area_formula,1,alone, radius, circle, initial

Os grupos formados pelo EM ficaram prioritariamente agrupados pelo tipo de problema e subsequentemente pela questão abordada no problema (respectivamente o primeiro e o terceiro atributos). Um exemplo disto ocorre nos grupos 3 e 7, onde o primeiro atributo de valor *circle_o* repete, porém identificou-se em cada grupo que o terceiro atributo possuía dois valores significativos agrupando assim pelos valores: *circumference_question1* e *radius_question1*.

Analisando os centróides de cada grupo, identificou-se que os atributos 1 (tipo do problema), 3 (questão abordada no problema), 5 (resultados das questões) e 6 (área de conhecimento) foram os mais relevantes na formação dos grupos. Nessa primeira análise, levou-se em consideração o grande número de valores que cada atributo pode receber.

A tabela 12 apresenta a distribuição em percentual dos valores do atributo 1 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica o tipo de cada problema.

Tabela 12 - Distribuição dos valores do atributo 1 do conjunto de dados Geometria nos grupos formados pelo EM.

Atributo\Grupos	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
RECTANGLE_ABCD	0%	0%	0%	0%	7%	0%	4%	0%
BUILDING_A_SIDEWALK	0%	14%	36%	0%	0%	0%	0%	0%
SQUARE_ABCD	0%	0%	0%	0%	5%	0%	3%	0%
SQUARE_AREA	0%	0%	0%	0%	0%	0%	1%	0%
TRAPEZOID_ABCD	**21%**	0%	0%	0%	13%	0%	0%	0%
TRAPEZOID_AREA	6%	0%	0%	0%	0%	0%	0%	0%
TRAPEZOID_HEIGHT	7%	0%	0%	0%	0%	0%	0%	0%
TRAPEZOID_BASE	0%	0%	0%	0%	10%	0%	0%	0%
TRIANGLE_ABC	0%	0%	0%	0%	8%	0%	5%	0%
TRIANGLE_TRIANGLE	0%	0%	0%	0%	0%	0%	7%	0%
TRIANGLE_RECTANGLE	0%	0%	0%	0%	0%	8%	6%	0%
TRIANGLE_AREA	0%	0%	0%	0%	0%	0%	4%	0%
TRIANGLE_HEIGHT	0%	0%	0%	0%	2%	0%	0%	0%
TRIANGLE_BASE	0%	0%	0%	0%	2%	0%	0%	0%
PENTAGON	8%	0%	0%	0%	21%	0%	0%	0%
PENTAGON_ABCDE	9%	0%	0%	0%	**23%**	0%	0%	0%
LAWN_SPRINKLER	3%	0%	0%	0%	0%	0%	0%	0%
LAWN_SPRINKLER_2	3%	5%	6%	0%	0%	0%	0%	0%
DESIGNING_A_QUILT	0%	0%	0%	0%	0%	**70%**	19%	0%
COVERING_POOL	0%	0%	0%	9%	0%	0%	0%	15%
DOG_ON_A_ROPE	3%	0%	0%	6%	0%	0%	0%	0%
TROGS	0%	12%	0%	17%	0%	0%	**22%**	0%
WATERING_VEGGIES	3%	4%	0%	0%	0%	0%	4%	0%
POGS	6%	**24%**	0%	0%	0%	0%	10%	0%
PAINTING_THE_WALL	0%	19%	**47%**	0%	0%	0%	0%	0%
CIRCLE_O	13%	0%	0%	**38%**	0%	0%	0%	**44%**
CIRCLE_DIAMETER	4%	0%	0%	7%	0%	0%	0%	10%
CIRCLE_AREA	0%	0%	0%	13%	0%	0%	0%	10%
CIRCLE_CIRCUMFERENCE	3%	0%	0%	0%	0%	0%	0%	15%
CIRCLE_RADIUS	1%	0%	0%	5%	0%	0%	0%	0%
ONE_CIRCLE_IN_CIRCLE	2%	0%	5%	0%	0%	0%	0%	0%
ONE_CIRCLE_IN_SQUARE	1%	5%	0%	0%	0%	0%	2%	0%
TWO_CIRCLES_IN_CIRCLE	5%	6%	0%	0%	0%	8%	0%	0%
TWO_CIRCLES_IN_SQUARE	1%	6%	0%	0%	0%	9%	3%	0%
RECTANGLE_AREA	0%	0%	0%	0%	0%	0%	1%	0%
RECTANGLE_HEIGHT	0%	0%	0%	0%	1%	0%	0%	0%
RECTANGLE_BASE	0%	0%	0%	0%	0%	0%	0%	0%
PARALLELOGRAM_ABDE	0%	0%	0%	0%	6%	0%	4%	0%
PARALLELOGRAM_AREA	0%	0%	0%	0%	0%	0%	1%	0%
PARALLELOGRAM_HEIGHT	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 13 apresenta a distribuição em percentual dos valores do atributo 3 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica cada questão dos problemas.

Tabela 13 - Distribuição dos valores do atributo 3 do conjunto de dados Geometria nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7
AREA_QUESTION1	**48%**	0%	0%	0%	0%	0%	**40%**	0%
HEIGHT_QUESTION2	0%	0%	0%	0%	16%	0%	0%	0%
BASE_QUESTION3	0%	0%	0%	0%	14%	0%	0%	0%
LARGE_RECTANGLE_AREA_QUESTION1	0%	0%	6%	0%	0%	0%	0%	0%
POOL_AREA_QUESTION1	0%	0%	6%	0%	0%	0%	0%	0%
SIDEWALK_AREA_QUESTION1	0%	5%	0%	0%	0%	0%	0%	0%
POOL_AREA_QUESTION2	0%	0%	6%	0%	0%	0%	0%	0%
LARGE_RECTANGLE_AREA_QUESTION2	0%	0%	5%	0%	0%	0%	0%	0%
SIDEWALK_AREA_QUESTION2	0%	4%	0%	0%	0%	0%	0%	0%
POOL_AREA_QUESTION3	0%	0%	6%	0%	0%	0%	0%	0%
LARGE_RECTANGLE_AREA_QUESTION3	0%	0%	6%	0%	0%	0%	0%	0%
SIDEWALK_AREA_QUESTION3	0%	4%	0%	0%	0%	0%	0%	0%
DOOR_AREA_QUESTION1	0%	0%	7%	0%	0%	0%	0%	0%
WALL_AREA_QUESTION1	0%	0%	**9%**	0%	0%	0%	0%	0%
PAINTED_AREA_QUESTION1	0%	7%	0%	0%	0%	0%	0%	0%
DOOR_AREA_QUESTION2	0%	0%	7%	0%	0%	0%	0%	0%
WALL_AREA_QUESTION2	0%	0%	7%	0%	0%	0%	0%	0%
PAINTED_AREA_QUESTION2	0%	6%	0%	0%	0%	0%	0%	0%
DOOR_AREA_QUESTION3	0%	0%	8%	0%	0%	0%	0%	0%
WALL_AREA_QUESTION3	0%	0%	7%	0%	0%	0%	0%	0%
PAINTED_AREA_QUESTION3	0%	6%	0%	0%	0%	0%	0%	0%
LONGER_BASE_QUESTION2	0%	0%	0%	0%	22%	0%	0%	0%
HEIGHT_QUESTION3	16%	0%	0%	0%	0%	0%	0%	0%
APOTHEM_QUESTION2	0%	0%	0%	0%	**23%**	0%	0%	0%
SIDE_QUESTION3	0%	0%	0%	0%	19%	0%	0%	0%
BASE_QUESTION1	0%	0%	0%	0%	0%	8%	0%	0%
WHITE_AREA_QUESTION1	0%	0%	0%	0%	0%	**9%**	0%	0%
AREA_QUESTION2	6%	0%	0%	0%	0%	0%	6%	0%
GREEN_AREA_QUESTION1	0%	0%	0%	0%	0%	8%	0%	0%
PURPLE_AREA_QUESTION1	0%	0%	0%	0%	0%	7%	0%	0%
WHITE_AREA_QUESTION2	0%	0%	0%	0%	0%	8%	0%	0%
PURPLE_AREA_QUESTION2	0%	0%	0%	0%	0%	7%	0%	0%
GREEN_AREA_QUESTION2	0%	0%	0%	0%	0%	8%	0%	0%
AREA_QUESTION3	0%	0%	0%	0%	0%	0%	6%	0%
WHITE_AREA_QUESTION3	0%	0%	0%	0%	0%	7%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7
PURPLE_AREA_QUESTION3	0%	0%	0%	0%	0%	7%	0%	0%
GREEN_AREA_QUESTION3	0%	0%	0%	0%	0%	7%	0%	0%
RADIUS_QUESTION1	0%	0%	0%	0%	0%	0%	0%	**35%**
CIRCUMFERENCE_QUESTION1	0%	0%	0%	**21%**	0%	0%	0%	0%
DIAMETER_QUESTION2	0%	0%	0%	10%	0%	0%	0%	0%
CIRCUMFERENCE_QUESTION2	0%	0%	0%	10%	0%	0%	0%	0%
RADIUS_QUESTION3	0%	0%	0%	0%	0%	0%	0%	20%
DIAMETER_QUESTION3	0%	0%	0%	13%	0%	0%	0%	0%
CIRCUMFERENCE_QUESTION3	0%	0%	0%	12%	0%	0%	0%	0%
RADIUS_QUESTION4	0%	0%	0%	0%	0%	0%	0%	17%
DIAMETER_QUESTION4	0%	0%	0%	0%	0%	0%	0%	15%
AREA_QUESTION4	6%	0%	0%	0%	0%	0%	0%	0%
DIAMETER_QUESTION1	0%	0%	0%	9%	0%	0%	0%	0%
WATERED_LAWN_AREA_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%
TOTAL_LAWN_AREA_QUESTION1	0%	0%	6%	0%	0%	0%	0%	0%
UNWATERED_LAWN_AREA_QUESTION1	0%	5%	0%	0%	0%	0%	0%	0%
SQUARE_AREA_QUESTION1	0%	0%	0%	0%	0%	0%	11%	0%
TROG_HEIGHT_QUESTION1	0%	0%	0%	6%	0%	0%	0%	0%
TROG_AREA_QUESTION1	0%	0%	0%	0%	0%	0%	4%	0%
SCRAP_METAL_AREA_QUESTION1	0%	13%	0%	0%	0%	0%	0%	0%
SQUARE_AREA_QUESTION2	0%	0%	0%	0%	0%	0%	7%	0%
TROG_HEIGHT_QUESTION2	0%	0%	0%	5%	0%	0%	0%	0%
TROG_AREA_QUESTION2	0%	0%	0%	0%	0%	0%	4%	0%
SCRAP_METAL_AREA_QUESTION2	0%	11%	0%	0%	0%	0%	0%	0%
SQUARE_AREA_QUESTION3	0%	0%	0%	0%	0%	0%	7%	0%
TROG_HEIGHT_QUESTION3	0%	0%	0%	4%	0%	0%	0%	0%
TROG_AREA_QUESTION3	0%	0%	0%	0%	0%	0%	4%	0%
SCRAP_METAL_AREA_QUESTION3	0%	11%	0%	0%	0%	0%	0%	0%
WATERED_AREA_QUESTION1	3%	0%	0%	0%	0%	0%	0%	0%
TOTAL_GARDEN_QUESTION1	0%	0%	0%	0%	0%	0%	4%	0%
UNWATERED_AREA_QUESTION1	0%	4%	0%	0%	0%	0%	0%	0%
POG_AREA_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%
POG_AREA_QUESTION2	2%	0%	0%	0%	0%	0%	0%	0%
POG_AREA_QUESTION3	2%	0%	0%	0%	0%	0%	0%	0%
CIRCLE_AREA_C_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%
CIRCLE_AREA_A_QUESTION1	2%	0%	5%	0%	0%	0%	0%	0%
CIRCLE_AREA_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%
SHADED_AREA_QUESTION1	0%	**16%**	0%	0%	0%	0%	0%	0%
CIRCLE_AREA_O_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%
CIRCLE_AREA_S_QUESTION1	2%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7
CIRCLE_RADIUS_AB_QUESTION1	0%	0%	0%	0%	0%	6%	0%	0%
SQUARE_BASE_QUESTION1	0%	0%	0%	0%	0%	7%	0%	0%
UNSHADED_AREA_QUESTION1	0%	0%	0%	0%	0%	4%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 14 apresenta a distribuição em percentual dos valores do atributo 5 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica as respostas de cada questão (correta ou incorreta). O percentual encontrado nos valores do atributo não se mostram muito decisivos na formação dos grupos.

Tabela 14 - Distribuição dos valores do atributo 5 do conjunto de dados Geometria nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7
CORRECT	76%	81%	90%	77%	77%	82%	87%	76%
INCORRECT	24%	19%	10%	23%	23%	18%	13%	24%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 15 apresenta a distribuição em percentual dos valores do atributo 6 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo representa a área de conhecimento dos problemas resolvidos.

Tabela 15 - Distribuição dos valores do atributo 6 do conjunto de dados Geometria nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7
ALT_PARALLELOGRAM_AREA	0%	0%	**92%**	0%	0%	0%	44%	0%
ALT_PARALLELOGRAM_SIDE	0%	0%	0%	0%	18%	0%	0%	0%
ALT_COMPOSE_BY_ADDITION	0%	**98%**	5%	0%	0%	8%	0%	0%
ALT_TRAPEZOID_AREA	18%	0%	0%	0%	0%	0%	0%	0%
ALT_TRAPEZOID_BASE	0%	0%	0%	0%	24%	0%	0%	0%
ALT_TRAPEZOID_HEIGHT	17%	0%	0%	0%	0%	0%	0%	0%
ALT_TRIANGLE_AREA	0%	0%	0%	0%	0%	0%	**54%**	0%
ALT_TRIANGLE_SIDE	0%	0%	0%	17%	13%	0%	0%	0%
ALT_PENTAGON_AREA	18%	0%	0%	0%	0%	0%	0%	0%
ALT_PENTAGON_SIDE	0%	0%	0%	0%	**44%**	0%	0%	0%
ALT_COMPOSE_BY_MULTIPLICATION	0%	0%	0%	0%	0%	**90%**	0%	0%
ALT_CIRCLE_RADIUS	0%	0%	0%	0%	0%	0%	0%	**81%**
ALT_CIRCLE_AREA	**47%**	0%	0%	0%	0%	0%	0%	0%
ALT_CIRCLE_CIRCUMFERENCE	0%	0%	0%	**47%**	0%	0%	0%	0%
ALT_CIRCLE_DIAMETER	0%	0%	0%	35%	0%	0%	0%	17%
Total	100%	100%	100%	100%	100%	100%	100%	100%

Durante as análises identificou-se que existem atributos fortemente vinculados ao primeiro atributo. Um exemplo disto é o atributo 10 que identifica a parte geométrica calculada na questão, isto é, se o valor do atributo 1 for *circle_o*, o atributo 10 possuirá valores referentes às fórmulas de cálculo de um circulo, tais como: *radius*, *area*, *circumference* e *diameter*.

A última etapa da análise sobre os resultados do algoritmo EM envolveu os cálculos de homogeneidade e separação. Quanto mais próximo de zero for o valor do resultado melhor será a homogeneidade. Já na separação é o inverso, isto é, quanto mais distante de zero for o valor melhor será o resultado. A homogeneidade global apresentada pelo EM foi de 1,8306 e a separação foi de 2,4558. Os valores da homogeneidade de cada grupo podem ser observados na tabela 16.

Tabela 16 - Homogeneidades dos grupos formados pelo EM.

Grupo	Nº Instâncias	%	Homogeneidade
0	1512	22%	1,9444
1	767	11%	1,6910
2	587	9%	1,5670
3	782	12%	1,8336
4	1126	17%	2,0919
5	652	10%	1,7012
6	837	12%	1,7875
7	515	8%	1,6621
Total	6778	100%	

5.4.1.2 Algoritmo K-média

Para a execução do algoritmo k-média modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 8 grupos considerando os resultados dos testes preliminares efetuados para o conjunto de dados. Os demais parâmetros permaneceram com as opções *default*.

A tabela 17 apresenta os centróides gerados pelo algoritmo k-média. Observa-se algumas diferenças entre os centróides gerados pelo algoritmo k-média e o algoritmo EM, um exemplo disto é a formação do grupo 6 a partir do valor *pentagon*. Este valor não apareceu em nem um centróide gerado pelo algoritmo EM.

Tabela 17 - Centróides gerados pelo algoritmo K-média.

Grupos	Centróides
0	trapezoid_abcd,1,longer_base_question2,1,correct,alt_parallelogram_area, non_area_formula,1,alone,area,circle,initial
1	circle_o,1,circumference_question1,1,correct,alt_circle_circumference, area_formula,0,alone,circumference,circle,initial
2	designing_a_quilt,2,green_area_question1,1,correct, alt_compose_by_multiplication, area_formula,0,embedded,area_combination, 0, initial
3	designing_a_quilt,1,white_area_question1,1,correct, alt_compose_by_multiplication, area_formula,0,alone,area_combination,0,repeat
4	trapezoid_abcd,1,area_question1,1,correct,alt_parallelogram_area, non_area_formula,0,alone,area,circle,repeat
5	pogs,1,shaded_area_question1,1,correct,alt_compose_by_addition,area_formula,1 ,embedded, area_difference,0,initial
6	pentagon,1,apothem_question2,1,correct,alt_pentagon_side,non_area_formula,1, alone, apothem, pentagon, initial
7	pentagon_abcde,2,side_question3,1,correct,alt_pentagon_side,non_area_formula, 1, alone, side, pentagon, repeat

Os grupos formados pelo k-média ficaram agrupados prioritariamente pelo atributo 1 e, em sequencia, pelo atributo 3. Através da análise efetuada nos centróides, identifica-se dois casos onde o atributo 1 repete-se, isso pode ser visualizado nos grupos 0 e 4 com o valor *trapezoid_abcd* e nos grupos 2 e 3 com o valor *designing_a_quilt.*

A tabela 18 exibe a distribuição em percentual dos valores do atributo 1 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica o tipo de cada problema.

Tabela 18 - Distribuição dos valores do atributo 1 do conjunto de dados Geometria nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7
RECTANGLE_ABCD	4%	0%	0%	0%	2%	0%	0%	0%
BUILDING_A_SIDEWALK	6%	0%	0%	0%	6%	15%	0%	0%
SQUARE_ABCD	3%	0%	0%	0%	2%	0%	0%	0%
SQUARE_AREA	0%	0%	0%	0%	0%	0%	0%	0%
TRAPEZOID_ABCD	**13%**	0%	0%	0%	**11%**	0%	0%	8%
TRAPEZOID_AREA	0%	0%	0%	0%	6%	0%	0%	0%
TRAPEZOID_HEIGHT	4%	0%	0%	0%	2%	0%	0%	0%
TRAPEZOID_BASE	5%	0%	0%	0%	0%	0%	0%	2%
TRIANGLE_ABC	4%	0%	0%	0%	3%	0%	0%	2%
TRIANGLE_TRIANGLE	0%	0%	0%	0%	4%	0%	0%	0%
TRIANGLE_RECTANGLE	1%	0%	3%	4%	3%	2%	0%	0%
TRIANGLE_AREA	0%	0%	0%	0%	2%	0%	0%	0%
TRIANGLE_HEIGHT	1%	0%	0%	0%	0%	0%	0%	0%
TRIANGLE_BASE	1%	0%	0%	0%	0%	0%	0%	0%
PENTAGON	0%	0%	0%	0%	8%	0%	**60%**	21%
PENTAGON_ABCDE	0%	0%	0%	0%	8%	0%	40%	**57%**
LAWN_SPRINKLER	0%	0%	0%	0%	3%	0%	0%	0%
LAWN_SPRINKLER_2	2%	0%	0%	0%	2%	5%	0%	0%
DESIGNING_A_QUILT	3%	0%	**81%**	**75%**	5%	0%	0%	0%
COVERING_POOL	3%	8%	0%	0%	0%	0%	0%	4%
DOG_ON_A_ROPE	0%	5%	0%	0%	3%	0%	0%	0%
TROGS	6%	10%	1%	11%	4%	12%	0%	0%
WATERING_VEGGIES	3%	0%	0%	0%	1%	4%	0%	0%
POGS	6%	0%	0%	0%	3%	**24%**	0%	0%
PAINTING_THE_WALL	9%	0%	0%	0%	6%	20%	0%	0%
CIRCLE_O	7%	**51%**	0%	0%	7%	0%	0%	3%
CIRCLE_DIAMETER	2%	6%	0%	0%	3%	0%	0%	1%
CIRCLE_AREA	2%	11%	0%	0%	0%	0%	0%	1%
CIRCLE_CIRCUMFERENCE	5%	0%	0%	0%	0%	0%	0%	2%
CIRCLE_RADIUS	1%	5%	0%	0%	0%	0%	0%	0%
ONE_CIRCLE_IN_CIRCLE	1%	2%	2%	0%	1%	0%	0%	0%
ONE_CIRCLE_IN_SQUARE	1%	0%	0%	0%	1%	5%	0%	0%
TWO_CIRCLES_IN_CIRCLE	2%	2%	5%	4%	1%	7%	0%	0%
TWO_CIRCLES_IN_SQUARE	2%	1%	9%	6%	1%	6%	0%	0%
RECTANGLE_AREA	0%	0%	0%	0%	1%	0%	0%	0%
RECTANGLE_HEIGHT	0%	0%	0%	0%	0%	0%	0%	0%
RECTANGLE_BASE	0%	0%	0%	0%	0%	0%	0%	0%
PARALLELOGRAM_ABDE	3%	0%	0%	0%	2%	0%	0%	0%
PARALLELOGRAM_AREA	0%	0%	0%	0%	0%	0%	0%	0%
PARALLELOGRAM_HEIGHT	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 19 apresenta a distribuição em percentual dos valores do atributo 3 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica cada questão dos problemas.

Tabela 19 - Distribuição dos valores do atributo 3 do conjunto de dados Geometria nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7
AREA_QUESTION1	0%	0%	3%	0%	**67%**	0%	0%	0%
HEIGHT_QUESTION2	9%	0%	0%	0%	0%	0%	0%	0%
BASE_QUESTION3	8%	0%	0%	0%	0%	0%	0%	1%
LARGE_RECTANGLE_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
POOL_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
SIDEWALK_AREA_QUESTION1	0%	0%	0%	0%	0%	5%	0%	0%
POOL_AREA_QUESTION2	1%	0%	0%	0%	1%	0%	0%	0%
LARGE_RECTANGLE_AREA_QUESTION2	1%	0%	0%	0%	1%	0%	0%	0%
SIDEWALK_AREA_QUESTION2	0%	0%	0%	0%	0%	5%	0%	0%
POOL_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
LARGE_RECTANGLE_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
SIDEWALK_AREA_QUESTION3	0%	0%	0%	0%	0%	5%	0%	0%
DOOR_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
WALL_AREA_QUESTION1	2%	0%	0%	0%	1%	0%	0%	0%
PAINTED_AREA_QUESTION1	0%	0%	0%	0%	0%	8%	0%	0%
DOOR_AREA_QUESTION2	2%	0%	0%	0%	1%	0%	0%	0%
WALL_AREA_QUESTION2	1%	0%	0%	0%	1%	0%	0%	0%
PAINTED_AREA_QUESTION2	0%	0%	0%	0%	0%	6%	0%	0%
DOOR_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
WALL_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
PAINTED_AREA_QUESTION3	0%	0%	0%	0%	0%	6%	0%	0%
LONGER_BASE_QUESTION2	**12%**	0%	0%	0%	0%	0%	0%	2%
HEIGHT_QUESTION3	9%	0%	0%	0%	2%	0%	0%	8%
APOTHEM_QUESTION2	0%	0%	0%	0%	0%	0%	**81%**	5%
SIDE_QUESTION3	0%	0%	0%	0%	0%	0%	19%	**72%**
BASE_QUESTION1	1%	0%	2%	4%	0%	2%	0%	0%
WHITE_AREA_QUESTION1	0%	0%	10%	**10%**	0%	0%	0%	0%
AREA_QUESTION2	4%	0%	3%	0%	3%	0%	0%	0%
GREEN_AREA_QUESTION1	0%	0%	**14%**	3%	0%	0%	0%	0%
PURPLE_AREA_QUESTION1	0%	0%	9%	7%	0%	0%	0%	0%
WHITE_AREA_QUESTION2	0%	0%	8%	9%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7
PURPLE_AREA_QUESTION2	0%	0%	6%	9%	0%	0%	0%	0%
GREEN_AREA_QUESTION2	0%	0%	8%	9%	0%	0%	0%	0%
AREA_QUESTION3	1%	0%	2%	0%	1%	0%	0%	0%
WHITE_AREA_QUESTION3	0%	0%	7%	9%	0%	0%	0%	0%
PURPLE_AREA_QUESTION3	0%	0%	6%	8%	0%	0%	0%	0%
GREEN_AREA_QUESTION3	0%	0%	7%	9%	0%	0%	0%	0%
RADIUS_QUESTION1	6%	7%	0%	0%	0%	0%	0%	5%
CIRCUMFERENCE_QUESTION1	0%	**19%**	0%	0%	0%	0%	0%	0%
DIAMETER_QUESTION2	0%	9%	0%	0%	0%	0%	0%	0%
CIRCUMFERENCE_QUESTION2	0%	9%	0%	0%	0%	0%	0%	0%
RADIUS_QUESTION3	5%	0%	0%	0%	0%	0%	0%	4%
DIAMETER_QUESTION3	0%	11%	0%	0%	0%	0%	0%	0%
CIRCUMFERENCE_QUESTION3	0%	11%	0%	0%	0%	0%	0%	0%
RADIUS_QUESTION4	2%	6%	0%	0%	0%	0%	0%	2%
DIAMETER_QUESTION4	2%	5%	0%	0%	0%	0%	0%	0%
AREA_QUESTION4	4%	0%	0%	0%	1%	0%	0%	0%
DIAMETER_QUESTION1	0%	8%	0%	0%	0%	0%	0%	0%
WATERED_LAWN_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
TOTAL_LAWN_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
UNWATERED_LAWN_AREA_QUESTION1	0%	0%	0%	0%	0%	5%	0%	0%
SQUARE_AREA_QUESTION1	3%	0%	0%	0%	2%	0%	0%	0%
TROG_HEIGHT_QUESTION1	0%	4%	0%	5%	0%	0%	0%	0%
TROG_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
SCRAP_METAL_AREA_QUESTION1	0%	0%	0%	0%	0%	13%	0%	0%
SQUARE_AREA_QUESTION2	2%	0%	0%	0%	1%	0%	0%	0%
TROG_HEIGHT_QUESTION2	0%	3%	0%	4%	0%	0%	0%	0%
TROG_AREA_QUESTION2	1%	0%	0%	0%	1%	0%	0%	0%
SCRAP_METAL_AREA_QUESTION2	0%	0%	0%	0%	0%	12%	0%	0%
SQUARE_AREA_QUESTION3	2%	0%	0%	0%	1%	0%	0%	0%
TROG_HEIGHT_QUESTION3	0%	3%	0%	2%	0%	0%	0%	0%
TROG_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
SCRAP_METAL_AREA_QUESTION3	0%	0%	0%	0%	0%	11%	0%	0%
WATERED_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
TOTAL_GARDEN_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
UNWATERED_AREA_QUESTION1	0%	0%	0%	0%	0%	4%	0%	0%
POG_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
POG_AREA_QUESTION2	1%	0%	0%	0%	0%	0%	0%	0%
POG_AREA_QUESTION3	1%	0%	0%	0%	1%	0%	0%	0%
CIRCLE_AREA_C_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
CIRCLE_AREA_A_QUESTION1	1%	2%	2%	0%	1%	0%	0%	0%
CIRCLE_AREA_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
SHADED_AREA_QUESTION1	0%	0%	0%	0%	0%	**17%**	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7
CIRCLE_AREA_O_QUESTION1	1%	0%	0%	0%	0%	0%	0%	0%
CIRCLE_AREA_S_QUESTION1	1%	0%	0%	0%	1%	0%	0%	0%
CIRCLE_RADIUS_AB_QUESTION1	0%	2%	2%	4%	0%	1%	0%	0%
SQUARE_BASE_QUESTION1	0%	1%	6%	4%	0%	0%	0%	0%
UNSHADED_AREA_QUESTION1	0%	0%	5%	3%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 20 apresenta a distribuição em percentual dos valores do atributo 5 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo identifica as respostas de cada questão (correta ou incorreta). Observando os percentuais identifica-se que o atributo não é decisivo na formação dos grupos como ocorre no algoritmo EM.

Tabela 20 - Distribuição dos valores do atributo 5 do conjunto de dados Geometria nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7
CORRECT	81%	78%	84%	82%	81%	81%	71%	79%
INCORRECT	19%	22%	16%	18%	19%	19%	29%	21%
Total	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 21 apresenta a distribuição em percentual dos valores do atributo 6 do conjunto de dados Geometria considerando os 8 agrupamentos formados. Este atributo representa a área de conhecimento dos problemas resolvidos.

Tabela 21 - Distribuição dos valores do atributo 6 do conjunto de dados Geometria nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7
ALT_PARALLELOGRAM_AREA	**24%**	0%	0%	0%	**24%**	0%	0%	0%
ALT_PARALLELOGRAM_SIDE	10%	0%	0%	0%	0%	0%	0%	0%
ALT_COMPOSE_BY_ADDITION	0%	4%	6%	4%	1%	**98%**	0%	0%
ALT_TRAPEZOID_AREA	0%	0%	0%	0%	16%	0%	0%	0%
ALT_TRAPEZOID_BASE	12%	0%	0%	0%	0%	0%	0%	2%
ALT_TRAPEZOID_HEIGHT	9%	0%	0%	0%	2%	0%	0%	8%
ALT_TRIANGLE_AREA	6%	0%	8%	0%	18%	0%	0%	0%
ALT_TRIANGLE_SIDE	7%	10%	1%	11%	0%	0%	0%	2%
ALT_PENTAGON_AREA	0%	0%	0%	0%	16%	0%	0%	0%
ALT_PENTAGON_SIDE	0%	0%	0%	0%	0%	0%	**100%**	**77%**
ALT_COMPOSE_BY_MULTIPLICATION	1%	1%	**85%**	**84%**	0%	2%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7
ALT_CIRCLE_RADIUS	13%	13%	0%	0%	0%	0%	0%	11%
ALT_CIRCLE_AREA	16%	0%	0%	0%	22%	0%	0%	0%
ALT_CIRCLE_CIRCUMFERENCE	0%	**38%**	0%	0%	0%	0%	0%	0%
ALT_CIRCLE_DIAMETER	2%	34%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%

Após as tabulações dos atributos, foram analisados os resultados dos cálculos de homogeneidade e separação obtidos pelo algoritmo k-média. A homogeneidade global gerada foi de 1,9756 e a separação foi de 2,4644. A homogeneidade de cada grupo pode ser observada na tabela 22.

Tabela 22 - Homogeneidades dos grupos formados pelo K-média.

Grupo	Nº Instâncias	%	Homogeneidade
0	2141	32%	2,2888
1	972	14%	1,8996
2	328	5%	1,6221
3	320	5%	1,6187
4	1675	25%	2,0472
5	784	12%	1,7066
6	319	5%	1,4117
7	239	4%	1,5772
Total	6778	100%	

5.4.1.3 Software R

A execução dos algoritmos na ferramenta R iniciou pelo algoritmo conhecido como vizinho mais próximo (*single*), em sequencia pelo algoritmo vizinho mais distante (*complete*) e por último o algoritmo que faz a média das distâncias (*average*). Foram gerados 8 grupos levando em consideração os testes preliminares efetuados para cada conjunto de dados.

Para executar os algoritmos na ferramenta R, utilizam-se as seguintes linhas de comando:

1. Carrega os dados: dados <- read.table("C:\\Users \\testeR.txt", head=T, sep=",", dec=".");
2. Formação dos grupos:

   c<- hclust(dist(dados[,-8]),method="single");

   c<- hclust(dist(dados[,-8]),method="complete");

   c<- hclust(dist(dados[,-8]),method="average").
3. Impressão do dendograma: plot(c);

4. Identifica qual grupo cada instância pertence: cutree(c,8).

Para cada algoritmo executado, as informações dos agrupamentos foram armazenadas em arquivo. Ao término dos agrupamentos, executou-se o software de cálculo da homogeneidade e a separação.

5.4.1.3.1 Vizinho mais próximo

Analisando os centroides probabilísticos gerados pelo algoritmo vizinho mais próximo apresentados na tabela 23, identifica-se que o algoritmo manteve o mesmo comportamento que os algoritmos EM e o K-média. Agrupando prioritariamente pelo atributo 1 e, em sequência, pelo atributo 3. No entanto, formou 4 grupos diferentes em comparação aos algoritmos da ferramenta WEKA, sendo os grupos: 4, 5, 6 e 7.

Tabela 23 - Centróides probabilísticos gerados pelo algoritmo vizinho mais próximo.

Grupos	Centróides
0	designing_a_quilt,1,area_question1,1,correct,alt_parallelogram_area, non_area_formula,0,alone, area, trapezoid, initial
1	circle_o,1,radius_question1,1,correct,alt_circle_radius,area_formula,0,alone,radius , circle,initial
2	circle_o,1,area_question4,1,correct,alt_circle_area,non_area_formula,0,alone,area, circle,initial
3	trogs,1,shaded_area_question1,1,correct,alt_compose_by_addition,area_formula,0, embedded,area,0,initial
4	circle_radius,1,area_question2,1,correct,alt_circle_area,non_area_formula,0,alone, area, circle, initial
5	one_circle_in_circle,1,circle_area_a_question1,1,correct,alt_compose_by_addition , area_formula,0,embedded,area,circle, initial
6	rectangle_area,1,area_question1,1,correct,alt_parallelogram_area,non_area_formul a,0, alone, area, rectangle, initial
7	triangle_height,1,height_question2,1,correct,alt_triangle_side,non_area_formula,1, alone, height, triangle, initial

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 2,2571 e o cálculo da separação resultou o valor 2,4147. A homogeneidade de cada grupo pode ser observada na tabela 24.

Tabela 24 - Homogeneidades dos grupos formados pelo vizinho mais próximo.

Grupo	Nº Instâncias	%	Homogeneidade
0	3946	58%	2,3489
1	1162	17%	2,0532
2	132	2%	1,2392
3	1409	21%	2,3568
4	21	0%	0,4959
5	32	0%	1,2087
6	28	0%	1,5607
7	48	1%	1,3945
Total	6778	100%	

5.4.1.3.2 Vizinho mais distante

Analisando os centróides probabilísticos apresentados na tabela 25, observa-se que o algoritmo vizinho mais distante formou grupos bem variados, isto ocorre devido à forma de cálculo utilizada neste algoritmo para a formação dos grupos.

Tabela 25 - Centróides probabilísticos gerados pelo algoritmo vizinho mais distante.

Grupos	Centróides
0	building_a_sidewalk,1,area_question1,1,correct,alt_parallelogram_area, non_area_formula,0,alone,area,rectangle,initial
1	painting_the_wall,1,longer_base_question2,1,correct,alt_parallelogram_area, non_area_formula,1,alone,area,trapezoid,initial
2	triangle_abc,1,area_question1,1,correct,alt_pentagon_area,non_area_formula,0, alone, area, triangle, initial
3	designing_a_quilt,1,apothem_question2,1,correct,alt_compose_by_multiplication, non_area_formula,0,alone,area_combination,pentagon,initial
4	circle_o,1,radius_question1,1,correct,alt_circle_radius,area_formula,0,alone, radius, circle, initial
5	trogs,1,scrap_metal_area_question1,1,correct,alt_compose_by_addition, area_formula,0,alone,area,0,initial
6	two_circles_in_circle,1,shaded_area_question1,1,correct,alt_circle_area, area_formula,0,embedded,area,circle,initial
7	triangle_height,1,height_question2,1,correct,alt_triangle_side,non_area_formula,1, alone, height, triangle, initial

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 2,1547 e o cálculo da separação resultou o valor 2,3512. A homogeneidade de cada grupo pode ser observada na tabela 26.

Tabela 26 - Homogeneidades dos grupos formados pelo vizinho mais distante.

Grupo	Nº Instâncias	%	Homogeneidade
0	911	13%	2,0284
1	976	14%	2,2969
2	844	12%	1,8949
3	1236	18%	2,2881
4	1294	19%	2,0833
5	894	13%	2,2868
6	547	8%	2,2189
7	76	1%	1,7670
Total	6778	100%	

5.4.1.3.3 Média das distâncias

A tabela 27 apresenta os centróides probabilísticos, gerados pelo algoritmo média das distâncias. Observando os centróides identifica-se que não houve repetição dos valores do atributo 1 como ocorreu nos outros algoritmos.

Tabela 27 - Centróides probabilísticos gerados pelo algoritmo média das distâncias.

Grupos	Centróides
0	building_a_sidewalk,1,area_question1,1,correct,alt_trapezoid_area, non_area_formula,0,alone,area,triangle,initial
1	painting_the_wall,1,longer_base_question2,1,correct,alt_parallelogram_area, non_area_formula,1,alone,area,trapezoid,initial
2	designing_a_quilt,1,apothem_question2,1,correct,alt_compose_by_multiplication, non_area_formula,0,alone,area_combination,pentagon,initial
3	circle_o,1,radius_question1,1,correct,alt_circle_radius,area_formula,0,alone, radius, circle,initial
4	trogs,1,scrap_metal_area_question1,1,correct,alt_compose_by_addition, area_formula,0,alone,area,0,initial
5	circle_radius,1,area_question2,1,correct,alt_circle_area,non_area_formula,0,alone, area,circle,initial
6	two_circles_in_circle,1,shaded_area_question1,1,correct,alt_circle_area, area_formula,0,embedded,area,circle,initial
7	triangle_height,1,height_question2,1,correct,alt_triangle_side,non_area_formula,1, alone,height,triangle,initial

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 2,1850 e o cálculo da separação resultou o valor 2,3907. A homogeneidade de cada grupo pode ser observada na tabela 28.

Tabela 28 - Homogeneidades dos grupos formados pelo algoritmo média das distâncias.

Grupo	Nº Instâncias	%	Homogeneidade
0	1607	24%	2,0679
1	1124	17%	2,3222
2	1215	18%	2,2870
3	1294	19%	2,0833
4	894	13%	2,2868
5	21	0%	0,4959
6	547	8%	2,2189
7	76	1%	1,7670
Total	6778		

5.4.1.4 Sistema Imunológico

Para a execução do algoritmo imunológico modificou-se os parâmetros referentes ao número total de grupos a ser gerado, formando 8 grupos considerando os resultados dos testes preliminares e selecionando a opção de normalização dos dados. Os demais parâmetros permaneceram com as opções *default*.

Os testes foram efetuados utilizando as duas condições de parada disponíveis para o algoritmo. A primeira condição de parada escolhida foi a por número de iterações atribuindo os valores: 100, 200, 500 e 1000. A segunda condição de parada utiliza índices atribuídos para a homogeneidade e separação, isto é, o algoritmo finaliza a sua execução somente quando atingir os índices atribuídos no início da execução. Os valores selecionados foram: 0,98 para a homogeneidade e 1,6 para a separação.

Ao término da execução, o algoritmo apresenta 3 formas de resultados:

1) A primeira é um arquivo chamado Cluster.txt que contem os centróides gerados e as instâncias identificadas com o seu grupo;
2) A segunda é um arquivo chamado Indices.txt, onde se encontra os índices de homogeneidade e separação gerados durante a execução juntamente com o número de iterações ocorridos durante a execução;
3) A terceira é uma tela que apresenta a matriz de confusão gerada após o término da execução. A matriz de confusão apresenta os resultados dos agrupamentos.

O algoritmo imunológico recalcula os centróides a cada nova iteração com o intuito de reavaliar os grupos formados buscando melhorar os valores da homogeneidade e separação.

As tabelas 29, 30, 31 e 32 apresentam os centróides gerados pela condição de parada referente ao número de iterações.

Tabela 29 - Centróide gerado após 100 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.6411551719797743, 0.0, 0.674992273305864, 0.12568087045616422, 0.0, 1.01490633261208, 0.0, 0.0, 0.6066063085793982, 0.8591197298331714, 0.0, alt:compose-by-multiplication;
1	0.41326137361141907, 0.0, 0.5559389829394245, 0.0, 0.0, 1.0149368837697974, 0.0, 0.0, 0.9204533834599545, 1.0109479203972707, 0.0, alt:circle-circumference;
2	0.17971340557039164, 0.1980119109071833, 0.2827431385742337, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0999085638037647, 0.5057560657559885, 0.0, alt:circle-area;
3	0.10293666694914838, 0.19839618830977163, 0.20778281976143384, 0.0, 0.9955765209884051, 0.0, 0.0, 1.0040663575385727, 0.0, 0.0, 0.0, alt:parallelogram-area;
4	0.7652326505610774, 0.0, 0.9207820251807383, 0.12511793548758157, 0.0, 1.0108668821916433, 1.0144827971452028, 0.0, 0.29954647726830763, 0.12467622661830739, 0.0, alt:compose-by-addition;
5	0.1533243620399181, 0.39822038612902233, 0.0, 0.0, 0.9960960855121269, 0.0, 0.0, 0.0, 0.0, 0.5013410287363789, 1.000224350561241, alt:trapezoid-area;
6	0.3585028856015261, 0.0, 0.3126085741920167, 0.25014055651908257, 0.9886096252268551, 0.0, 1.0210511039330927, 0.0, 0.49932773357114724, 0.7505328359042348, 0.9999143263868013, alt:pentagon-side;
7	0.6605594228394065, 0.0, 0.8259671316091256, 0.0, 0.0, 0.0, 0.0, 1.023124727226142, 0.0, 0.3729153636498991, 0.0, alt:parallelogram-area;

Tabela 30 - Centróide gerado após 200 iterações do algoritmo imunológico.

Grupos	Centróides
0	0.5933829179150854, 0.0, 0.6047759305306973, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0083195164298757, 0.0, alt:circle-area;
1	0.05140392584103154, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25045192572028446, 1.0107403866549667, alt:parallelogram-area;
2	0.02581622047093624, 0.0, 0.06502406295740806, 0.0, 0.0, 1.0192153450523307, 1.017703098932092, 0.0, 0.30145680328927466, 0.12503637454627822, 0.0, alt:compose-by-addition;
3	0.1293707318633151, 0.0, 0.28461785200584533, 0.0, 0.0, 0.0, 1.0546067433232376, 0.0, 0.10085997538717249, 0.5042918480307044, 0.0, alt:pentagon-side;
4	0.6440375499919863, 0.0, 0.738708085867225, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.6031314163978202, 0.0, alt:parallelogram-area;
5	0.02566895632369128, 0.0, 0.10334633282109905, 0.0, 0.9883833860357141, 0.9992291507533285, 1.0130960735900845, 0.0, 0.302556906803727, 0.12539162514768568, 0.998982202835113, alt:compose-by-addition;
6	0.42102845334697436, 0.0, 0.5080239052945429, 0.0, 0.0, 1.0427719859232476, 0.0, 0.0, 0.9249345460245163, 0.969728149975227, 0.0, alt:compose-by-multiplication;
7	0.40664698406861444, 0.19959065419004168, 0.340915394584919, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.9467115807318741, 0.0, alt:triangle-area;

Tabela 31 - Centróide gerado após 500 iterações do algoritmo imunológico.

Grupos	Centróides
0	0.12748710187341522, 0.20107875552641946, 0.27174400051952535, 0.2500812368317013, 1.004782575084693, 0.0, 0.9925779382420703, 0.0, 0.20123243846557556, 0.49896274216502956, 0.9992866555794188, alt:trapezoid-height;
1	0.3343614961023042, 0.19926428227684706, 0.3678643698693842, 0.0, 0.0, 1.0484119078213807, 0.0, 0.0, 0.7160796531474025, 0.12404241138341651, 0.965359305791421, alt:compose-by-multiplication;
2	0.35107810746448687, 0.0, 0.0, 0.12371854685402836, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7410271526361301, 0.9507233659191817, alt:circle-area;
3	0.5306658767604397, 0.0, 0.6697522338894308, 0.0, 0.0, 1.0691149304343337, 1.0476338064787265, 0.0, 0.3006999788921559, 0.12480594256772755, 0.0, alt:compose-by-addition;
4	0.4173669504733603, 0.0, 0.5805002616711725, 0.0, 0.0, 1.072406680523887, 0.9559784479669919, 0.0, 0.8576235607095436, 1.030767663388686, 0.0, alt:circle-diameter;
5	0.3414308964342188, 0.19513241407980064, 0.4363053685162484, 0.0, 0.0, 0.0, 0.0, 0.9476486221026469, 0.0, 0.6010790138602753, 0.0, alt:parallelogram-area;
6	0.4164380401792599, 0.0, 0.49933468947734616, 0.24721170831898312, 0.0, 1.0385310650334163, 0.0, 0.0, 0.9352318654369353, 1.0134110894288375, 0.955410371251802, alt:circle-circumference;
7	0.20407460171469163, 0.19953890878790168, 0.2682053213607225, 0.0, 0.0, 0.0, 1.0745455462110045, 0.0, 0.2019540129820942, 0.5079636177794158, 0.0, alt:pentagon-side;

Tabela 32 - Centróide gerado após 1000 iterações do algoritmo imunológico.

Grupos	Centróides
0	0.30961531057917263, 0.0, 0.3318442419534912, 0.0, 0.9725434930885186, 1.081088915210432, 0.9383271565987072, 0.9625953986334087, 0.6143381922293448, 0.8269911998511552, 0.0, alt:compose-by-addition;
1	0.341144207560701, 0.0, 0.3692516558143662, 0.0, 0.0, 0.0, 0.0, 0.9960877072258396, 0.0, 0.598339819726541, 1.0201718518033, alt:parallelogram-area;
2	0.025474752005374274, 0.0, 0.10597103553460922, 0.0, 0.0, 1.0055892631294014, 0.9470629836211023, 0.0, 0.3061552429225887, 0.12721770360048387, 1.0893671906185531, alt:compose-by-addition;
3	0.7922366009814232, 0.0, 0.9096865416682934, 0.0, 0.9209374892568889, 1.0433438033634197, 0.998602378315767, 0.0, 0.2975590562075426, 0.12582077376983214, 0.0, alt:compose-by-addition;
4	0.10330201443609492, 0.0, 0.24285007345552748, 0.0, 0.0, 0.0, 0.0, 1.0414805424936853, 0.0, 0.0, 0.0, alt:parallelogram-area;
5	0.3394328782365675, 0.0, 0.0, 0.12182525135046073, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7365471122574779, 0.0, alt:circle-area;
6	0.6606586044090 63, 0.19491154758543908, 0.7235250318064829, 0.12363603072670037, 0.0, 0.0, 0.0, 0.0, 0.0, 0.36334292170376536, 0.0, alt:circle-area;
7	0.34467361622117176, 0.19926982424761774, 0.31248832670009663, 0.0, 0.0, 0.0, 1.0343088519107626, 0.0, 0.5158990630788816, 0.7662199643769432, 0.0, alt:pentagon-side;

A tabela 33 exibe um comparativo dos valores da homogeneidade global e separação gerados a partir da condição de parada que utiliza o número de iterações. Observando os dados apresentados pelos valores iniciais da homogeneidade e separação, identifica-se que eles são decisivos para o resultado final apresentado após as iterações.

Tabela 33 - Comparativo da homogeneidade e separação versus o número de iterações.

Iterações	Homogeneidade Inicial	Homogeneidade Final	Separação Inicial	Separação Final
100	0,95002	0,94733	1,62413	1,63661
200	0,91719	0,91238	1,53461	1,56195
500	0,93376	0,92458	1,62694	1,66492
1000	0,94810	0,94261	1,60625	1,63138

A tabela 34 apresenta os centróides gerados pela condição de parada referente aos índices esperados para a homogeneidade e separação.

Tabela 34 - Centróide gerado após o algoritmo imunológico atingir a homogeneidade global de 0,98 e a separação de 1,6.

Grupos	Centróides
0	0.12849626402668035, 0.7694475636398369, 0.2864496418492838, 0.12479102810694452, 0.0, 0.0, 1.032375110367513, 0.0, 0.1007971783002508, 0.5042110363148152, 1.0253823454952065, alt:pentagon-side;
1	0.5056188673085338, 0.0, 0.62796354534133, 0.12471958230332646, 0.0, 0.0, 0.0, 1.0864576304823566, 0.0, 0.0, 0.9842257537690006, alt:parallelogram-area;
2	0.5399879548630371, 0.0, 0.48806932221681637, 0.0, 0.0, 1.0463234314103491, 1.0399949399853117, 0.0, 0.805892820217575, 0.9777116569354392, 0.0, alt:compose-by-addition;
3	0.5979254307641932, 0.0, 0.5886988565217995, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0472530424970974, 0.9359142460017279, alt:circle-area;
4	0.33732375258651576, 0.0, 0.40779965565936005, 0.0, 0.0, 1.0608294430464777, 0.0, 0.0, 0.7377112455868867, 0.12755656887727976, 0.0, alt:circle-circumference;
5	0.6381563143414759, 0.0, 0.6764133487761493, 0.12422946401863151, 0.0, 1.0280237985614622, 0.0, 1.0439979210512607, 0.5938619017673226, 0.8567084282932308, 0.0, alt:compose-by-addition;
6	0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.9344611634379875, alt:parallelogram-area;
7	0.3507471370152108, 0.3883103514322489, 0.3064845777600165, 0.0, 0.0, 0.0, 1.0180898366640154, 0.0, 0.48184341496086946, 0.7423172789444016, 0.0, alt:pentagon-side;

O algoritmo imunológico executou 354 iterações até atingir a condição de parada selecionada chegando aos valores de: 0,97998 para a homogeneidade e 1,72737 para a separação.

Nos resultados apresentados, observa-se que o valor inicial da homogeneidade e separação interferem significativamente no valor resultante para as duas condições de parada, por exemplo, o critério de parada envolvendo o número de iterações apresenta valores piores para a execução com 1000 iterações do que com a de 500 iterações. Já o critério referente ao índice para a homogeneidade e separação possui uma grande variação, pois o valor de inicialização é incerto isto dificulta a escolha dos índices para a execução do algoritmo. Pode-se ter execuções que durem dias até alcançar o índice selecionado ou execuções que durem somente uma iteração quando o índice selecionado for maior que o índice inicializado. Um exemplo disso pode ser visto na figura 14.

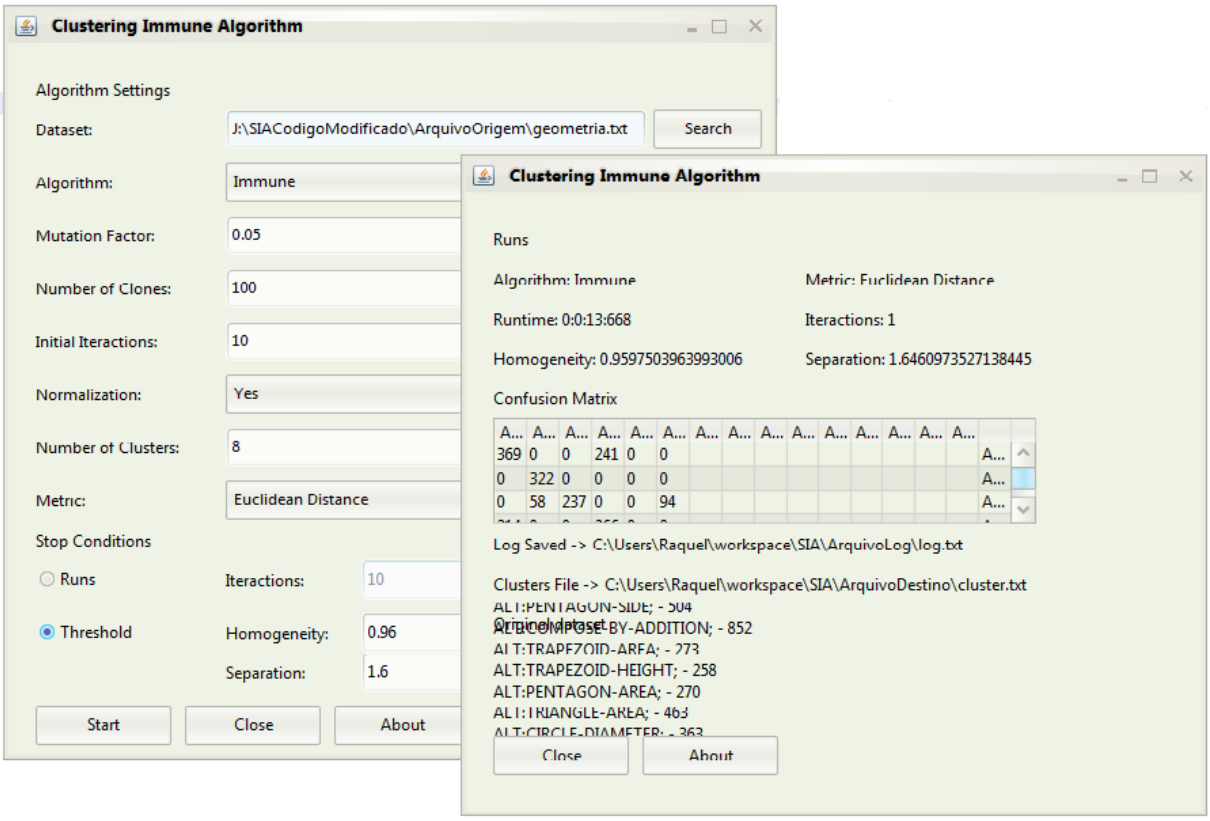

Figura 14 - Demonstração da execução do algoritmo imunológico com valor inicializado menor que o índice de homogeneidade.

5.4.1.5 Resultados

Os resultados da homogeneidade global e separação obtidos nos testes são apresentados na tabela 35. Cada coluna representa os algoritmos executados no estudo de caso. Para a homogeneidade quando mais próximo de zero for o valor melhor é o resultado, já para a separação é o inverso quando mais distante de zero melhor é o resultado. Observa-se

que o algoritmo imunológico foi o que apresentou melhores resultados para o conjunto de dados Geometria.

Tabela 35 - Resultados da homogeneidade global e separação para o conjunto de dados Geometria.

Critério/Algoritmos	EM	K-média	VP	VD	MD	Imunológico
Homogeneidade	1,8306	1,9756	2,2571	2,1547	2,1850	0,9123
Separação	2,4558	2,4644	2,4147	2,3512	2,3907	1,5679

5.4.2 Conjunto de Dados Língua Chinesa

O estudo de caso para o conjunto de dados Língua Chinesa iniciou com a seleção dos atributos considerados mais relevantes, sendo escolhidos os seguintes:

1. Identificação da lição do curso;
2. Nível da seção;
3. Identificação do problema;
4. Número de resoluções do problema;
5. Componente que originou a entrada dos dados.
6. Identificação do número de repetições por etapa do problema.
7. Resultado da questão (correto, incorreto ou hint);
8. Mensagem de feedback apresentada para o estudante.

Com os atributos selecionados e convertidos para cada tipo de ferramenta, inicia-se a execução dos algoritmos. A lista completa dos atributos pode ser encontrada no anexo F.

5.4.2.1 Algoritmo EM

Para a execução do algoritmo EM modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 10 grupos considerando os resultados dos testes preliminares efetuados. Os demais parâmetros permaneceram com as opções *default*.

A tabela 36 apresenta os centróides probabilísticos, gerados pelo algoritmo EM. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 1 e, em sequência, pelo atributo 3.

Tabela 36 - Centróides probabilísticos gerados pelo algoritmo EM.

Grupos	Centróides
0	lesson14,s2,l13_w01,1,radioGroup_UpdateRadioButton,1,INCORRECT, No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_hand_corner_for_help
1	lesson14,s2,l16_w01,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
2	lesson16,s2,l16_w02,1,radioGroup2_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
3	lesson14,s3,l14_w02,1,radioGroup2_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
4	lesson14,s3,l14_w03,1,done_ButtonPressed,1,CORRECT, No_HInt_or_Other_Message_Text
5	lesson18,s2,l16_w02,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_go_on_to_syllable_2
6	lesson18,s3,l16_w11,1,radioGroup_UpdateRadioButton,1,INCORRECT, No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_corner_for_help
7	lesson4, s2, l02_w01_swf, 1, done_ButtonPressed, 1 , CORRECT, No_HInt_or_Other_Message_Text
8	lesson16,s2,l16_w23,1,done_ButtonPressed,1,CORRECT, No_HInt_or_Other_Message_Text
9	lesson16,s3,l16_w02,1,radioGroup_UpdateRadioButton,1,INCORRECT, No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_hand_corner_for_help

Analisando os centróides de cada grupo, identificou-se que os atributos 1 (identificação da lição do curso), 4 (número de resoluções do problema), 6 (identifica o número de repetições por etapa do problema), 7 (resultado das questões) e 8 (mensagem de feedback apresenta para ao estudante) foram os mais relevantes na formação dos grupos. Nessa primeira análise, levou-se em consideração o grande número de valores que cada atributo pode receber.

A tabela 37 apresenta a distribuição em percentual dos valores do atributo 1 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. O atributo 1 forma um agrupamento macro repetindo o valor em mais de um grupo. Isto pode ser observado nos grupos 0, 1, 3 e 4 através do valor *lesson14*.

Tabela 37 - Distribuição dos valores do atributo 1 do conjunto de dados Língua Chinesa nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
lesson5	10%	4%	6%	0%	9%	4%	0%	0%	0%	0%
lesson6	0%	5%	6%	0%	11%	4%	4%	0%	0%	6%
lesson7	0%	6%	0%	8%	11%	4%	2%	0%	0%	9%
lesson8	0%	2%	3%	0%	0%	2%	1%	0%	6%	2%
lesson9	0%	2%	3%	0%	0%	2%	1%	0%	6%	6%
lesson18	0%	6%	0%	24%	20%	**12%**	**17%**	0%	0%	11%
lesson16	0%	10%	**18%**	0%	0%	12%	8%	0%	**33%**	**21%**

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
lesson13	24%	10%	17%	0%	0%	11%	6%	0%	31%	0%
lesson12	0%	1%	0%	2%	2%	1%	1%	0%	0%	4%
lesson14	**31%**	**17%**	0%	**26%**	**21%**	0%	8%	0%	0%	0%
lesson10	0%	4%	7%	0%	10%	5%	7%	0%	0%	4%
lesson15	0%	4%	16%	0%	16%	10%	2%	0%	0%	9%
lesson17	0%	7%	11%	0%	0%	7%	4%	0%	21%	10%
lesson4	0%	6%	0%	17%	0%	8%	14%	**31%**	0%	12%
lesson1	0%	4%	1%	0%	0%	0%	10%	11%	0%	0%
lesson2	19%	5%	9%	0%	0%	6%	3%	26%	0%	0%
lesson3	15%	4%	0%	22%	0%	11%	7%	31%	0%	0%
lesson11	0%	1%	1%	0%	0%	1%	2%	0%	3%	5%
unit1	0%	0%	0%	0%	0%	0%	3%	1%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 38 apresenta a distribuição em percentual dos valores do atributo 4 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica o número de resoluções do problema. Observa-se que as etapas foram resolvidas geralmente na primeira execução.

Tabela 38 - Distribuição dos valores do atributo 4 do conjunto de dados Língua Chinesa nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
1	**71%**	**81%**	**85%**	**70%**	**94%**	**77%**	**75%**	**29%**	**92%**	**86%**
2	9%	8%	7%	10%	5%	9%	8%	17%	7%	8%
3	5%	3%	2%	5%	0%	4%	4%	13%	1%	2%
4	2%	2%	1%	4%	0%	2%	3%	8%	0%	1%
5	2%	1%	1%	3%	0%	2%	3%	7%	0%	0%
6	3%	1%	1%	2%	0%	2%	3%	6%	0%	1%
7	2%	1%	1%	2%	0%	1%	2%	6%	0%	1%
8	1%	1%	1%	2%	0%	1%	1%	4%	0%	0%
9	1%	1%	1%	1%	0%	1%	1%	3%	0%	0%
10	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
11	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
12	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
13	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
14	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
15	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
18	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%
19	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 39 apresenta a distribuição em percentual dos valores do atributo 6 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica o número de repetições por etapa do problema. Observa-se que não houve muitas repetições para as etapas durante a resolução dos problemas.

Tabela 39 - Distribuição dos valores do atributo 6 do conjunto de dados Língua Chinesa nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
1	54%	80%	79%	76%	99%	74%	51%	98%	97%	59%
2	22%	11%	11%	13%	0%	15%	26%	0%	0%	22%
3	11%	4%	4%	5%	0%	6%	12%	0%	0%	11%
4	5%	2%	2%	3%	0%	3%	5%	0%	0%	4%
5	2%	1%	2%	2%	0%	1%	2%	0%	0%	1%
6	1%	0%	0%	0%	0%	0%	1%	0%	0%	1%
7	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
11	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
12	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
13	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
14	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
15	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
18	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
19	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
21	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
22	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
23	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
24	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
25	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
26	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
27	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
28	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
29	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
30	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
31	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
32	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
34	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
35	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
37	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
38	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
39	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
40	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
41	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
42	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
43	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
44	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
45	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
46	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
47	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
48	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
49	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
51	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
52	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
53	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
54	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
55	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
56	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
57	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
58	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
59	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
60	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
61	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
62	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
63	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
64	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
65	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
66	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
67	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 40 apresenta a distribuição em percentual dos valores do atributo 7 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica as respostas de cada problema (correta ou incorreta) ou a solicitação de ajuda (*hint*).

Tabela 40 - Distribuição dos valores do atributo 7 do conjunto de dados Língua Chinesa nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
CORRECT	0%	**100%**	**100%**	**100%**	**100%**	**100%**	0%	**100%**	**98%**	0%
INCORRECT	**92%**	0%	0%	0%	0%	0%	**90%**	0%	2%	**100%**
HINT	8%	0%	0%	0%	0%	0%	10%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 41 apresenta a distribuição em percentual dos valores do atributo 8 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo contem as mensagens de *feedback* apresentadas a cada etapa para os estudantes.

Tabela 41 - Distribuição dos valores do atributo 8 do conjunto de dados Língua Chinesa nos grupos formados pelo EM.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
Correct_Now_click_on_Done_to_do_the_next_one	0%	**85%**	**99%**	**99%**	0%	0%	0%	0%	0%	0%
No_HInt_or_Other_Message_Text	3%	0%	0%	0%	**99%**	0%	0%	**99%**	**99%**	0%
Correct_Now_go_on_to_syllable_2	0%	11%	0%	0%	0%	**100%**	0%	0%	0%	0%
No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_hand_corner_for_help	**81%**	0%	0%	0%	0%	0%	0%	0%	0%	**95%**
No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_corner_for_help	0%	0%	0%	0%	0%	0%	80%	0%	0%	0%
Focus_on_the_pitch_change_and_listen_to_the_sound_again	5%	0%	0%	0%	0%	0%	5%	0%	0%	0%
The_sound_pitch_goes_down_at_the_beginning_then_up_at_the_middle_ie_ it_is_falling-rising_Listen_again	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
This_is_the_third_tone_Its_pitch_goes_down_then_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
You_must_complete_the_steps_in_the_correct_orderThis_is_the_third_tone_Its_pitch_goes_down_then_up_Please_click_the_correct_answer Remember_that_you_can_ask_for_a_hint_if_you_need_help	5%	0%	0%	0%	0%	0%	2%	0%	0%	3%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
The_sound_pitch_is_going_down_ie_it_is_high-falling_Listen_again	1%	0%	0%	0%	0%	0%	1%	0%	0%	0%
Im_sorry_but_you_are_not_done_yet_Please_continue_working	2%	0%	0%	0%	0%	0%	1%	0%	0%	1%
Correct_Now_click_on_Done_to_go_to_the_next_one	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
No_this_is_not_correct_Please_click_on_hint_in_the_top_right_corner_for_help	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%
Correct_Now_go_on_to_syllable_2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_pitch_is_high-rising_Listen_again	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Focus_on_the_frequency_change_and_listen_to_the_sound_again	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
The_sound_frequency_is_high-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_pitch_is_nearly_high-flat_Listen_again	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_first_tone_Its_pitch_stays_stable_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Focus_on_the_picth_change_and_listen_to_the_sound_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_forth_tone_Its_pitch_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_is_short_and_its_pitch_is_low-flat_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Please_click_on_the_highlighted_button	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_frequency_is_going_down_ie_it_is_high-falling_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_forth_tone_Its_frequency_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Focus_on_the_pitch_change_and_listen_to_the_sound_againThe_sound_pitch_is_high-rising_Listen_againThis_is_the_second_tone_Its_pitch_goes_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_neutral_tone_It_is_short_with_little_pitch_change_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
This_is_the_fourth_tone_Its_pitch_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_frequency_goes_down_at_the_beginning_then_up_at_the_middle_ie_it_is_falling-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_is_short_and_its_frequency_is_low-flat_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_neutral_tone_Its_short_with_no_frequecy_change_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_second_tone_Its_frequency_goes_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_second_tone_Its_pitch_goes_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Após as tabulações dos atributos, foram analisados os resultados dos cálculos de homogeneidade e separação obtidos pelo algoritmo EM. O cálculo da homogeneidade global resultou o valor 1,6805 e o cálculo da separação resultou o valor 1,9995. A homogeneidade de cada grupo pode ser observada na tabela 42.

Tabela 42 - Homogeneidades dos grupos formados pelo EM.

Grupo	Nº Instâncias	%	Homogeneidade
0	3211	7%	1,8873
1	8649	18%	1,7063
2	4584	9%	1,6387
3	3345	7%	1,6616
4	6871	14%	1,5494
5	7005	14%	1,7035
6	2289	5%	1,8996
7	3436	7%	1,7107
8	5425	11%	1,5178
9	3628	7%	1,7870
Total	48443	100%	

5.4.2.2 Algoritmo K-média

Para a execução do algoritmo k-média modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 10 grupos considerando os resultados dos testes preliminares efetuados. Os demais parâmetros permaneceram com as opções *default*.

A tabela 43 apresenta os centróides gerados pelo algoritmo k-média. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 1 e, em sequência, pelo atributo 3.

Tabela 43 - Centróides gerados pelo algoritmo K-média.

Grupos	Centróides
0	lesson15,s1,l15_w12,1,radioGroup2_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
1	lesson4,s2,l04_w18_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_go_on_to_syllable_2
2	lesson13,s1,l13_w01,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
3	lesson14,s3,l14_w04,1,radioGroup_UpdateRadioButton,1,INCORRECT, No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_hand_corner_for_help
4	lesson13,s2,l13_w28,1,radioGroup2_UpdateRadioButton,2,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
5	lesson16,s2,l16_w23,1,done_ButtonPressed,1,CORRECT, No_HInt_or_Other_Message_Text
6	lesson16,s3,l16_w25,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_go_on_to_syllable_2
7	lesson3,s2,l03_w10_swf,1,radioGroup_UpdateRadioButton,3,CORRECT, Correct_Now_go_on_to_syllable_2
8	lesson18,s3,l18_w08,1,radioGroup2_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
9	lesson2,s3,l02_w10_swf,4,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_go_on_to_syllable_2

A tabela 44 apresenta a distribuição em percentual dos valores do atributo 1 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica a lição do curso.

Tabela 44 - Distribuição dos valores do atributo 1 do conjunto de dados Língua Chinesa nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
lesson5	3%	4%	4%	4%	4%	4%	4%	2%	4%	0%
lesson6	3%	5%	5%	4%	3%	5%	4%	2%	3%	0%
lesson7	3%	4%	6%	6%	2%	5%	4%	2%	4%	0%
lesson8	3%	3%	1%	1%	1%	2%	0%	1%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
lesson9	2%	2%	2%	3%	2%	2%	1%	2%	1%	0%
lesson18	5%	9%	4%	7%	6%	9%	11%	7%	**34%**	0%
lesson16	9%	8%	6%	11%	9%	**12%**	**28%**	6%	8%	0%
lesson13	4%	4%	**37%**	7%	**29%**	8%	8%	4%	7%	0%
lesson12	1%	1%	1%	2%	1%	1%	1%	1%	1%	0%
lesson14	8%	9%	7%	**19%**	9%	10%	11%	8%	10%	0%
lesson10	4%	6%	4%	3%	5%	5%	3%	4%	3%	0%
lesson15	**28%**	3%	0%	4%	2%	5%	9%	3%	0%	0%
lesson17	6%	8%	7%	5%	4%	8%	5%	4%	5%	0%
lesson4	6%	**17%**	2%	6%	7%	7%	0%	0%	7%	5%
lesson1	1%	2%	4%	1%	2%	2%	2%	2%	0%	1%
lesson2	5%	5%	4%	7%	7%	6%	1%	4%	4%	**87%**
lesson3	8%	6%	4%	8%	6%	7%	7%	**45%**	8%	7%
lesson11	1%	1%	2%	2%	1%	1%	1%	2%	0%	0%
unit1	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 45 apresenta a distribuição em percentual dos valores do atributo 4 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica o número de resoluções do problema. Observando o grupo 9 identifica-se um percentual de resoluções maior em comparação aos outros grupos.

Tabela 45 - Distribuição dos valores do atributo 4 do conjunto de dados Língua Chinesa nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
1	**81%**	**72%**	**87%**	**82%**	**78%**	**79%**	**88%**	**62%**	**83%**	1%
2	6%	10%	6%	7%	8%	9%	9%	7%	8%	21%
3	3%	4%	2%	3%	3%	3%	2%	8%	2%	11%
4	2%	2%	1%	2%	1%	2%	0%	3%	1%	**28%**
5	2%	2%	1%	1%	2%	2%	0%	6%	1%	6%
6	1%	2%	1%	2%	2%	1%	0%	6%	1%	6%
7	1%	2%	1%	1%	2%	1%	0%	2%	1%	5%
8	1%	1%	0%	1%	1%	1%	0%	1%	1%	6%
9	1%	1%	0%	0%	1%	1%	0%	2%	0%	5%
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%
11	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
12	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
13	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
14	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
15	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
18	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
19	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

A tabela 46 apresenta a distribuição em percentual dos valores do atributo 6 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica o número de repetições por etapa do problema.

Tabela 46 - Distribuição dos valores do atributo 6 do conjunto de dados Língua Chinesa nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
1	**86%**	**83%**	**79%**	**61%**	12%	**99%**	**75%**	4%	**87%**	**71%**
2	8%	11%	12%	21%	**69%**	0%	15%	19%	4%	16%
3	3%	1%	5%	11%	10%	0%	5%	**71%**	5%	7%
4	2%	3%	3%	4%	4%	0%	3%	4%	3%	3%
5	1%	1%	1%	1%	3%	0%	1%	2%	2%	2%
6	0%	0%	0%	1%	1%	0%	0%	1%	0%	0%
7	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
11	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
12	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
13	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
14	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
15	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
18	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
19	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
21	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
22	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
23	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
24	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
25	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
26	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
27	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
28	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
29	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
30	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
31	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
32	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
34	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
35	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
37	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
38	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
39	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
40	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
41	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
42	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
43	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
44	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
45	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
46	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
47	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
48	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
49	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
51	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
52	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
53	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
54	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
55	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
56	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
57	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
58	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
59	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
60	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
61	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
62	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
63	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
64	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
65	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
66	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
67	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 47 apresenta a distribuição em percentual dos valores do atributo 7 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo identifica as respostas de cada problema (correta ou incorreta) ou a solicitação de ajuda (*hint*).

Tabela 47 - Distribuição dos valores do atributo 7 do conjunto de dados Língua Chinesa nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
CORRECT	**86%**	**90%**	**87%**	1%	**80%**	**99%**	**100%**	**77%**	**95%**	**89%**
INCORRECT	13%	9%	10%	**98%**	17%	1%	0%	20%	4%	1%
HINT	1%	1%	3%	1%	2%	0%	0%	3%	0%	10%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

A tabela 48 apresenta a distribuição em percentual dos valores do atributo 8 do conjunto de dados Língua Chinesa considerando os 10 agrupamentos formados. Este atributo contem as mensagens de *feedback* apresentadas a cada etapa para os estudantes.

Tabela 48- Distribuição dos valores do atributo 8 do conjunto de dados Língua Chinesa nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
Correct_Now_click_on_Done_to_do_the_next_one	**78%**	31%	**69%**	0%	**80%**	0%	10%	11%	**94%**	29%
No_HInt_or_Other_Message_Text	6%	1%	10%	2%	1%	**100%**	0%	0%	1%	7%
Correct_Now_go_on_to_syllable_2	3%	**57%**	5%	0%	0%	0%	**87%**	**63%**	0%	**53%**
No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_hand_corner_for_help	8%	2%	3%	**82%**	9%	0%	0%	2%	0%	0%
No_this_is_not_correct_Please_click_on_hint_in_the_upper_right_corner_for_help	4%	5%	5%	11%	5%	0%	0%	15%	4%	1%
Focus_on_the_pitch_change_and_listen_to_the_sound_again	0%	1%	1%	1%	1%	0%	0%	1%	0%	5%
The_sound_pitch_goes_down_at_the_beginning_then_up_at_the_middle_ie_it_is_falling-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
This_is_the_third_tone_Its_pitch_goes_down_then_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
You_must_complete_the_steps_in_the_correct_order_Remember_that_you_can_ask_for_a_hint_if_you_need_help	0%	1%	1%	2%	0%	0%	0%	2%	0%	0%

The_sound_pitch_is_going_down_ie_it_is_high-falling_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Im_sorry_but_you_are_not_done_yet_Please_continue_working	0%	0%	0%	1%	1%	0%	0%	1%	0%	0%
Correct_Now_click_on_Done_to_go_to_the_next_one	0%	1%	3%	0%	0%	0%	3%	2%	0%	0%
No_this_is_not_correct_Please_click_on_hint_in_the_top_right_corner_for_help	1%	0%	1%	0%	1%	0%	0%	0%	0%	0%
Correct_Now_go_on_to_syllable_2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_pitch_is_high-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Focus_on_the_frequency_change_and_listen_to_the_sound_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_frequency_is_high-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_pitch_is_nearly_high-flat_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
This_is_the_first_tone_Its_pitch_stays_stable_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Focus_on_the_picth_change_and_listen_to_the_sound_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_forth_tone_Its_pitch_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_is_short_and_its_pitch_is_low-flat_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Please_click_on_the_highlighted_button	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_frequency_is_going_down_ie_it_is_high-falling_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_forth_tone_Its_frequency_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Focus_on_the_pitch_change_and_listen_to_the_sound_againThe_sound_pitch_is_high-rising_Listen_againThis_is_the_second_tone_Its_pitch_goes_up_Please_click_the_correc	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_neutral_tone_It_is_short_with_little_pitch_change_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_fourth_tone_Its_pitch_goes_down_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
The_sound_frequency_goes_down_at_the_beginning_then_up_at_the_middle_ie_it_is_falling-rising_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Atributo\Grupos	0	1	2	3	4	5	6	7	8	9
The_sound_is_short_and_its_frequency_is_low-flat_Listen_again	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_neutral_tone_Its_short_with_no_frequecy_change_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_second_tone_Its_frequency_goes_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
This_is_the_second_tone_Its_pitch_goes_up_Please_click_the_correct_answer	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100 %

Após as tabulações dos atributos, foram analisados os resultados dos cálculos de homogeneidade e separação obtidos pelo algoritmo K-média. O cálculo da homogeneidade global resultou o valor 1,6652 e o cálculo da separação resultou o valor 2,1665. A homogeneidade de cada grupo pode ser observada na tabela 49.

Tabela 49 - Homogeneidades dos grupos formados pelo K-média.

Grupo	Nº Instâncias	%	Homogeneidade
0	6093	13%	1,7000
1	7401	15%	1,7237
2	5439	11%	1,6630
3	6211	13%	1,8064
4	1794	4%	1,7107
5	14773	30%	1,6236
6	2989	6%	1,4947
7	588	1%	1,7297
8	2668	6%	1,4699
9	487	1%	1,7001
Total	48443	100%	

5.4.2.3 Software R

A execução dos algoritmos na ferramenta R iniciou pelo algoritmo conhecido como vizinho mais próximo (*single*), em sequência, pelo algoritmo vizinho mais distante (*complete*)

e, por último, o algoritmo que faz a média das distâncias (*average*). Foram gerados 10 grupos levando em consideração os testes preliminares efetuados para o conjunto de dados.

Para a execução dos algoritmos na ferramenta R foi necessário diminuir o número de instâncias de 48443 para 44077. As instâncias removidas pertencem à lição do curso *lesson18*. A redução do número total de instâncias ocorreu devido a grande quantidade de memória utilizada durante a execução dos algoritmos excedendo o total de 16 GB.

5.4.2.3.1 Vizinho mais próximo

A tabela 50 apresenta os centróides probabilísticos, gerados pelo algoritmo vizinho mais próximo. Observa-se que os grupos foram formados prioritariamente pelo atributo 1 (identificação da lição do curso) seguindo pelo atributo 3 (identificação do problema).

Tabela 50 - Centróides probabilísticos gerados pelo algoritmo vizinho mais próximo.

Grupo	Centróides
0	lesson16,s2,l16_w23,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
1	lesson4,s2,l04_w12_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
2	lesson11,s1,l11_w02,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
3	lesson2,s2,l02_w04_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
4	lesson2,s1,l02_w17_swf,3,radioGroup_UpdateRadioButton,8,CORRECT, Correct_Now_go_on_to_syllable_2
5	lesson2,s2,l02_w09_swf,18,radioGroup2_UpdateRadioButton,6,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
6	lesson13,s2,l13_w02,2,radioGroup_UpdateRadioButton,10,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
7	lesson13,s2,l13_w04,2,radioGroup_UpdateRadioButton,7,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
8	lesson13,s2,l13_w13,2,radioGroup2_UpdateRadioButton,5,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
9	lesson13,s2,l13_w15,2,radioGroup_UpdateRadioButton,6,CORRECT, Correct_Now_go_on_to_syllable_2

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 2,0251 e o cálculo da separação resultou o valor 1,7542.

A tabela 51 exibe a homogeneidade de cada grupo. Observa-se que os 6, 7, 8 e 9 foram constituídos somente por uma instância. Isto significa que o algoritmo não consegue formar 10 grupos como ocorre com outros algoritmos.

Tabela 51 - Homogeneidades dos grupos formados pelo vizinho mais próximo.

Grupo	Nº Instâncias	%	Homogeneidade
0	38437	87%	2,0387
1	2759	6%	1,9409
2	661	1%	1,8124
3	2214	5%	1,9632
4	1	0%	0,0000
5	1	0%	0,0000
6	1	0%	0,0000
7	1	0%	0,0000
8	1	0%	0,0000
9	1	0%	0,0000
Total	44077	100%	

5.4.2.3.2 Vizinho mais distante

A tabela 52 apresenta os centróides probabilísticos, gerados pelo algoritmo vizinho mais distante. Observa-se que os grupos foram formados prioritariamente pelo atributo 1 (identificação da lição do curso) seguindo pelo atributo 3 (identificação do problema).

Tabela 52 - Centróides probabilísticos gerados pelo algoritmo vizinho mais distante.

Grupo	Centróides
0	lesson7,s2,l05_w02_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
1	lesson9,s2,l09_w10,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
2	lesson16,s2,l16_w11,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
3	lesson14,s3,l14_w04,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
4	lesson15,s2,l14_w27,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
5	lesson17,s2,l17_w02,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
6	lesson3,s2,l02_w01_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
7	lesson4,s2,l04_w12_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
8	lesson2,s2,l02_w04_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
9	lesson16,s3,l16_w23,1,done_ButtonPressed,17,INCORRECT, No_HInt_or_Other_Message_Text

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 1,9313 e o cálculo da separação resultou o valor 1,6644. A homogeneidade de cada grupo pode ser observada na tabela 53.

Tabela 53 - Homogeneidades dos grupos formados pelo vizinho mais distante.

Grupo	Nº Instâncias	%	Homogeneidade
0	6431	15%	1,9494
1	1201	3%	1,8266
2	8516	19%	1,8935
3	6591	15%	1,8950
4	5930	13%	1,8864
5	3328	8%	1,7671
6	4879	11%	2,0926
7	4543	10%	2,0390
8	2607	6%	1,9919
9	51	0%	0,9804
Total	44077	100%	

5.4.2.3.3 Média das distâncias

A tabela 54 apresenta os centróides probabilísticos, gerados pelo algoritmo média das distâncias. Observa-se que os grupos foram formados prioritariamente pelo atributo 1 (identificação da lição do curso) seguindo pelo atributo 3 (identificação do problema).

Tabela 54 - Centróides probabilísticos gerados pelo algoritmo média das distâncias.

Grupo	Centróides
0	lesson6,s2,l05_w02_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
1	lesson7,s2,l07_w19,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
2	lesson16,s2,l16_w11,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
3	lesson13,s3,l13_w30,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
4	lesson14,s3,l14_w04,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
5	lesson15,s2,l17_w02,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
6	lesson3,s2,l02_w01_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, No_HInt_or_Other_Message_Text
7	lesson4,s2,l04_w12_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
8	lesson2,s2,l02_w04_swf,1,radioGroup_UpdateRadioButton,1,CORRECT, Correct_Now_click_on_Done_to_do_the_next_one
9	lesson16,s3,l16_w23,1,done_ButtonPressed,13,INCORRECT, No_HInt_or_Other_Message_Text

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 1,9110 e o cálculo da separação resultou o valor 1,7057. A homogeneidade de cada grupo pode ser observada na tabela 55.

Tabela 55 - Homogeneidades dos grupos formados pelo algoritmo média das distâncias.

Grupo	Nº Instâncias	%	Homogeneidade
0	4422	10%	1,8940
1	2773	6%	1,8500
2	8563	19%	1,8905
3	3003	7%	1,8022
4	6189	14%	1,8507
5	6273	14%	1,8726
6	7163	16%	2,0558
7	3035	7%	1,9652
8	2601	6%	1,9921
9	55	0%	0,9818
Total	44077	100%	

5.4.2.4 Sistema Imunológico

Para a execução do algoritmo imunológico modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 10 grupos considerando os resultados dos testes preliminares e selecionando a opção de normalização dos dados. Os demais parâmetros permaneceram com as opções *default*.

Os testes foram efetuados utilizando as duas condições de parada disponíveis para o algoritmo. A condição de parada definida por número de iterações considerou os valores: 100, 200, 500 e 1000. A condição de parada referente aos índices atribuídos para a homogeneidade e separação selecionou os seguintes valores: 0,33 para a homogeneidade e 0,7 para a separação.

As tabelas 56, 57, 58 e 59 apresentam os centróides gerados pela condição de parada referente ao número de iterações.

Tabela 56 - Centróide gerado após 100 iterações do algoritmo imunológico.

Grupos	Centróide
0	1.0159665783585445, 0.6738349322809166, 0.043828948387444965, 0.0, 0.007297887431500173, 0.15410837276833972, 0.06945124881963931, lesson14;
1	1.017105431949983, 0.6598539924167169, 0.04217137254399635, 0.4769232790243064, 0.003076174643430056, 0.05921713179761929, 0.03021405565058044, lesson14;
2	0.0, 0.2942595276862677, 0.006011980948645858, 0.4457367711694252, 0.0018011952948605513, 0.0, 0.02012368357703981, lesson16;
3	0.0, 0.8066125484218006, 0.15275089786141516, 0.1725191110809763, 0.0038326651424511627, 0.06367581548909296, 0.05150909003289974, lesson3;
4	0.8018924357610925, 0.19767580688811348, 0.004119710700416881, 6.272448897112114E-5, 0.007441020240664457, 0.13965060948294344, 0.05213876584154658, lesson13;
5	0.5003799326910402, 0.6914322641236432, 0.03787226847868508, 0.0, 0.008594916350391276, 0.1488592207416956, 0.07550906498438761, lesson14;
6	0.427537793586865, 0.7013512916719805, 0.043814176564523064, 0.48019817099754547, 0.0027828574079118667, 0.0, 0.01695839739750107, lesson14;
7	0.0, 0.29880829762934985, 0.005817402021907258, 0.0, 0.006482538608840194, 0.12647194417091745, 0.05133423714201261, lesson16;
8	0.24916891523179982, 0.5135358355693949, 0.04085986510985724, 0.6220668200038743, 0.017776359168850366, 0.5740596919229393, 0.1399358865092715, lesson14;
9	0.8191116278451592, 0.20145667265986963, 0.0038757741451105336, 0.44902556391425696, 0.008105347578772298, 0.039244420520480446, 0.02679207411522113, lesson16;

Tabela 57 - Centróide gerado após 200 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.4849540743200553, 0.11850518390409859, 0.0, 0.0, 0.0, 0.0, 0.06236583867655515, lesson16;
1	0.0, 0.977189364926942, 0.8120683785812696, 0.33578083519138274, 0.0, 0.0, 0.031035684379196665, lesson2;
2	0.999740683476327, 0.7951742083144099, 0.0, 0.0, 0.0, 0.5038964686327636, 0.12498547650389795, lesson3;
3	0.4879486327243999, 0.5964835187727534, 0.0, 0.0, 0.015050042917947292, 0.5112211997987945, 0.09383529728288979, lesson16;
4	0.4769511199175207, 0.741794240966305, 0.0, 0.0, 0.0, 0.0, 0.06293891830187973, lesson17;
5	1.0334295761172243, 0.8844438051101872, 0.0, 0.3380448757094994, 0.0, 0.0, 0.031145909493326136, lesson18;
6	1.015647612888098, 0.4671339043143631, 0.0, 0.0, 0.0, 0.5028412871957851, 0.12479209365954719, lesson14;
7	1.0341570860595015, 0.7865269136209092, 0.1052924192454859, 0.0, 0.015196545372274358, 0.0, 0.0, lesson13;
8	0.49419429817132426, 0.2633245108662071, 0.0, 0.6829193476719344, 0.0, 0.0, 0.0, lesson13;
9	0.48408195441425955, 0.5473846591248644, 0.0, 0.6782130982156951, 0.0, 0.0, 0.0, lesson14;

Tabela 58 - Centróide gerado após 500 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.5070955604787668, 0.7498832258188216, 0.15756153918068994, 0.0, 0.10634840487751351, 0.5369982863015901, 0.3546587458694296, lesson3;
1	0.505591867049363, 0.08466450550417855, 0.0, 0.3329275008345642, 0.0, 0.0, 0.031021372678160415, lesson16;
2	0.0, 0.970536578686091, 0.6354449377911894, 0.0, 0.0, 0.0, 0.06261570098153468, lesson2;
3	1.03656640122966, 0.036036624815354266, 0.0, 0.0, 0.0, 0.0, 0.06254511290969923, lesson7;
4	0.0, 0.3459675075785057, 0.0, 0.0, 0.0, 0.0, 0.06289375509123274, lesson14;
5	0.0, 0.018170137765124925, 0.0, 0.34049089021526036, 0.0, 0.0, 0.03099304241397435, lesson16;
6	0.0, 0.6526065852606829, 0.0, 0.6320883551537609, 0.0, 0.0, 0.0, lesson3;
7	1.0592306218466696, 0.3360975554187881, 0.0, 0.0, 0.03038752559061144, 0.0, 0.06258996390033024, lesson16;
8	1.082502167778963, 0.7265704339178346, 0.05209347414369155, 0.3511698951128394, 0.0, 0.0, 0.031198564162251856, lesson4;
9	0.4987402735394904, 0.575417774066538, 0.0, 0.332160951044267, 0.0, 0.0, 0.03115246173846023, lesson14;

Tabela 59 - Centróide gerado após 1000 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.0, 0.15651912132698145, 0.009652237464418584, 0.30747157075897885, 0.0033255577376911805, 0.051223559277578094, 0.0312997127161143, lesson16;
1	0.5033819266784056, 0.1714434223337412, 0.0027561268201186265, 0.2086397344604106, 0.0027234624948926436, 0.024374456044828467, 0.0320806669914832, lesson16;
2	1.0130083157160603, 0.31589653184398764, 0.003397249552284345, 0.7026295959647825, 0.022538046200354357, 0.1815254065150614, 0.037030856489033606, lesson14;
3	1.0300626425082746, 0.7899942780243195, 0.09058037608442464, 0.008105870994371793, 0.011004998756541394, 0.3526946165447787, 0.12302284587512494, lesson4;
4	0.3943458694812523, 0.6558013181285206, 0.017323330802738884, 0.5890876875019183, 0.0053210628847744045, 0.0868699582287161, 0.031191131258243625, lesson15;
5	1.0358232227813777, 0.7577872260793936, 0.07132524863179232, 0.43322433333784494, 0.0024547112092640697, 0.04150374856295365, 0.02882855595029414, lesson4;
6	0.5017467488621719, 0.675276021229376, 0.022890548177957798, 0.0, 0.008936817672495639, 0.16334778895942265, 0.07752871052993719, lesson14;
7	0.0, 0.5975099801835565, 0.011325765372558333, 0.10321591794336285, 0.00515290133871681, 0.10184709804946689, 0.05902879184077285, lesson17;
8	0.0579710620339125, 0.9063142970501468, 0.3011006798095927, 0.30648598045973013, 0.004167651427142751, 0.07113643939251345, 0.041809995612748216, lesson2;
9	1.037169607758599, 0.2995846634403754, 0.005147176280577053, 0.11102061018934786, 0.004506251043649367, 0.06001760030881038, 0.04218267016101324, lesson13;

A tabela 60 exibe um comparativo dos valores da homogeneidade global e separação gerados a partir da condição de parada que utiliza o número de iterações. Observando os dados apresentados pelos valores iniciais da homogeneidade e separação, identifica-se que eles são decisivos para o resultado final apresentado após as iterações.

Tabela 60 - Comparativo da homogeneidade e separação versus o número de iterações

Iterações	Homogeneidade Inicial	Homogeneidade Final	Separação Inicial	Separação Final
100	0,3061	0,3056	0,7637	0,7724
200	0,4274	0,4212	0,7569	0,7749
500	0,3409	0,3355	0,8232	0,8487
1000	0,3145	0,3114	0,7546	0,7889

A tabela 61 apresenta os centróides gerados pela condição de parada referente aos índices esperados para a homogeneidade e separação.

Tabela 61 - Centróide gerado após o algoritmo imunológico atingir a homogeneidade global de 0,33 e a separação de 0,7.

Grupos	Centróides
0	0.4981701089436599, 0.12440568980362457, 0.0, 0.0, 0.0, 0.0, 0.06242561788354988, lesson13;
1	0.5011441420563626, 0.7095807600276177, 0.0, 0.0, 0.030432032662591226, 0.0, 0.06257102536516777, lesson17;
2	1.0164961671240733, 0.27617253190661273, 0.0, 0.6573561198824303, 0.0, 0.0, 0.0, lesson16;
3	1.012040071782882, 0.23149914624917012, 0.05263665192947441, 0.0, 0.0, 0.0, 0.0, lesson7;
4	0.0, 0.8453809475570544, 0.052757228470095126, 0.33266104328264, 0.0, 0.0, 0.03111716907059406, lesson3;
5	0.0, 0.2951413754363721, 0.0, 0.0, 0.0, 0.0, 0.062412468527349066, lesson16;
6	1.0222774877293845, 0.5198575714524314, 0.0, 0.0, 0.015130976806199833, 0.0, 0.06237613451014525, lesson14;
7	0.49480795906514374, 0.21882590894631493, 0.0, 0.666128836629105, 0.04556197024676618, 0.0, 0.0, lesson16;
8	0.0, 0.590693860394373, 0.0, 0.0, 0.0, 0.5035533637775164, 0.12436995321533886, lesson14;
9	0.49988445024130335, 0.5303971286642636, 0.0, 0.6648585256787412, 0.0, 0.0, 0.0, lesson14;

O algoritmo imunológico executou 93 iterações até atingir a condição de parada selecionada chegando aos valores de: 0,3299 para a homogeneidade e 0,8263 para a separação.

5.4.2.5 Resultados

Os resultados da homogeneidade global e separação obtidos nos testes são apresentados na tabela 62. Observa-se que o algoritmo imunológico foi o que apresentou melhores resultados para o conjunto de dados Língua Chinesa.

Tabela 62 - Resultados da homogeneidade global e separação para o conjunto de dados Língua Chinesa.

Critério/Algoritmos	EM	K-média	VP	VD	MD	Imunológico
Homogeneidade	1,6805	1,6652	2,0251	1,9313	1,9110	0,3056
Separação	1,9995	2,1665	1,7542	1,6644	1,7057	0,7724

5.4.3 Conjunto de Dados Álgebra

O estudo de caso para o conjunto de dados Álgebra iniciou com a seleção dos atributos considerados mais relevantes, sendo escolhidos os seguintes:

1. Duração para a resolução da questão (em segundos);

2. Identificação da sessão;
3. Identificação do número de repetições por questão do problema;
4. Identificação do resultado da questão.

Com os atributos selecionados e convertidos para cada tipo de ferramenta, inicia-se a execução dos algoritmos. A lista completa dos atributos pode ser encontrada no anexo G.

5.4.3.1 Algoritmo EM

Para a execução do algoritmo EM modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 3 grupos considerando os resultados dos testes preliminares efetuados. Os demais parâmetros permaneceram com as opções *default*.

Os atributos 1 (duração da resolução) e 3 (número de repetições por questão) são do tipo numérico e possuem muitas variações, por isso foram discretizados. Isto é, utilizou-se a opção *discretize* da ferramenta WEKA que através de um algoritmo forma intervalos com os valores dos atributos.

A tabela 63 apresenta os centróides probabilísticos, gerados pelo algoritmo EM. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 2 (identificação da sessão) e, em sequência, pelo atributo 1 (duração da resolução da questão).

Tabela 63 - Centróides probabilísticos gerados pelo algoritmo EM.

Grupos	Centróides
0	'\'(2.75-4.25]\'',ES_02-5,'\'(1.5-4.5]\'',HINT_LEVEL_CHANGE
1	'\'(2.75-4.25]\'',ES_02-5,'\'(-inf-1.5]\'',OK
2	'\'(8.75-30.5]\'',ES_02-6,'\'(1.5-4.5]\'',OK

Analisando os atributos do conjunto de dados, identificou-se que os atributos 1 (duração), 2 (sessão), 3 (repetições da etapa) e 4 (respostas da etapa) seriam os mais relevantes na formação dos grupos. Nessa primeira análise, levou-se em consideração o grande número de valores que cada atributo pode receber.

A tabela 64 apresenta a distribuição em percentual dos valores do atributo 1 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo possui intervalos de tempo das resoluções de cada questão.

Tabela 64 - Distribuição dos valores do atributo 1 do conjunto de dados Álgebra nos grupos formados pelo EM.

Atributo\Grupos	0	1	2
'(-inf-1.25]'	30%	11%	1%
'(1.25-2.25]'	24%	11%	6%
'(2.25-2.75]'	1%	7%	0%
'(2.75-4.25]'	**33%**	**26%**	17%
'(4.25-4.75]'	0%	2%	0%
'(4.75-8.75]'	9%	24%	31%
'(8.75-30.5]'	3%	16%	**37%**
(30.5-inf)'	1%	4%	8%
Total	100%	100%	100%

A tabela 65 apresenta a distribuição em percentual dos valores do atributo 2 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica a sessão dos problemas.

Tabela 65 - Distribuição dos valores do atributo 2 do conjunto de dados Álgebra nos grupos formados pelo EM.

Atributo\Grupos	0	1	2
ES_02-5	**25%**	**30%**	17%
ES_02-6	20%	5%	**32%**
ES_02-7	10%	21%	7%
ES_02-8	20%	20%	15%
ES_02-9	2%	12%	6%
ES_02-10	13%	9%	15%
ES_02-5a	6%	1%	5%
ES_02-7a	2%	1%	2%
ES_02-9a	1%	0%	1%
Total	100%	100%	100%

A tabela 66 apresenta a distribuição em percentual dos valores do atributo 3 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica o intervalo de repetições por questão do problema.

Tabela 66 - Distribuição dos valores do atributo 3 do conjunto de dados Álgebra nos grupos formados pelo EM.

Atributo\Grupos	0	1	2
'(-inf-1.5]'	0%	**87%**	**45%**
'(1.5-4.5]'	**69%**	13%	35%
'(4.5-inf)'	31%	0%	20%
Total	100%	100%	100%

A tabela 67 apresenta a distribuição em percentual dos valores do atributo 4 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica as respostas de cada problema (correta ou incorreta). A solicitação de ajuda dividida em dois níveis (*initial_hint, hint_level_change*) e identificação de inconsistências nos dados inseridos (*bug*).

Tabela 67- Distribuição dos valores do atributo 4 do conjunto de dados Álgebra nos grupos formados pelo EM.

Atributo\Grupos	0	1	2
OK	5%	**88%**	**55%**
BUG	0%	0%	17%
ERROR	3%	1%	18%
INITIAL_HINT	10%	11%	9%
HINT_LEVEL_CHANGE	**82%**	0%	0%
Total	100%	100%	100%

Após as tabulações dos atributos, foram analisados os resultados dos cálculos de homogeneidade e separação obtidos pelo algoritmo EM. O cálculo da homogeneidade global resultou o valor 1,3649 e o cálculo da separação resultou o valor 1,6887. A homogeneidade de cada grupo pode ser observada na tabela 68.

Tabela 68- Homogeneidades dos grupos formados pelo EM.

Grupo	Nº Instâncias	%	Homogeneidade
0	1641	13%	1,2787
1	3713	30%	1,1766
2	7214	57%	1,4815
Total	12568	100%	

5.4.3.2 Algoritmo K-média

Para a execução do algoritmo k-média modificou-se o parâmetro referente ao número total de grupos a ser gerado, formando 3 grupos considerando os resultados dos testes preliminares efetuados. Os demais parâmetros permaneceram com as opções *default*.

Os atributos 1 (duração da resolução) e 3 (número de repetições por questão) são do tipo numérico e possuem muitas variações, por isso foram discretizados. Isto é, utilizou-se a opção *discretize* da ferramenta WEKA que através de um algoritmo forma intervalos com os valores dos atributos.

A tabela 69 apresenta os centróides gerados pelo algoritmo k-média. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 2 (identificação da sessão) e, em sequência, pelo atributo 1 (duração da resolução da questão).

Tabela 69 - Centróides gerados pelo algoritmo K-média.

Grupos	Centróides
0	\'(4.75-8.75]\'',ES_02-5,'\'(-inf-1.5]\'',OK
1	\'(8.75-30.5]\'',ES_02-6,'\'(-inf-1.5]\'',BUG
2	\'(1.25-2.25]\'',ES_02-5,'\'(1.5-4.5]\'',HINT_LEVEL_CHANGE

A tabela 70 apresenta a distribuição em percentual dos valores do atributo 1 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo possui o intervalo de tempo das resoluções de cada questão.

Tabela 70- Distribuição dos valores do atributo 1 do conjunto de dados Álgebra nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2
'(-inf-1.25]'	6%	4%	24%
'(1.25-2.25]'	6%	4%	**33%**
'(2.25-2.75]'	3%	0%	1%
'(2.75-4.25]'	23%	12%	31%
'(4.25-4.75]'	1%	0%	0%
'(4.75-8.75]'	**37%**	4%	5%
'(8.75-30.5]'	18%	**70%**	4%
'(30.5-inf)'	6%	6%	3%
Total	100%	100%	100%

A tabela 71 apresenta a distribuição em percentual dos valores do atributo 2 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica a sessão dos problemas.

Tabela 71 - Distribuição dos valores do atributo 2 do conjunto de dados Álgebra nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2
ES_02-5	**25%**	5%	**30%**
ES_02-6	13%	**62%**	12%
ES_02-7	14%	5%	10%
ES_02-8	19%	7%	22%
ES_02-9	9%	6%	3%
ES_02-10	15%	7%	15%
ES_02-5a	3%	4%	6%
ES_02-7a	2%	2%	2%
ES_02-9a	1%	1%	1%
Total	100%	100%	100%

A tabela 72 apresenta a distribuição em percentual dos valores do atributo 3 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica intervalo de repetições por questão do problema.

Tabela 72 - Distribuição dos valores do atributo 3 do conjunto de dados Álgebra nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2
'(-inf-1.5]'	**64%**	**44%**	0%
'(1.5-4.5]'	23%	30%	**83%**
'(4.5-inf)'	12%	26%	17%
Total	100%	100%	100%

A tabela 73 apresenta a distribuição em percentual dos valores do atributo 4 do conjunto de dados Álgebra considerando os 3 agrupamentos formados. Este atributo identifica as respostas de cada problema (correta ou incorreta). A solicitação de ajuda dividida em dois níveis (*initial_hint, hint_level_change*) e identificação de inconsistências nos dados inseridos (*bug*).

Tabela 73 - Distribuição dos valores do atributo 4 do conjunto de dados Álgebra nos grupos formados pelo K-média.

Atributo\Grupos	0	1	2
OK	**58%**	**79%**	23%
BUG	10%	3%	**40%**
ERROR	11%	8%	20%
INITIAL_HINT	10%	9%	13%
HINT_LEVEL_CHANGE	11%	1%	5%
Total	100%	100%	100%

Após as tabulações dos atributos, foram analisados os resultados dos cálculos de homogeneidade e separação obtidos pelo algoritmo k-média. O cálculo da homogeneidade global resultou o valor 1,3389 e o cálculo da separação resultou o valor 1,7619. A homogeneidade de cada grupo pode ser observada na tabela 74.

Tabela 74- Homogeneidades dos grupos formados pelo k-média.

Grupo	Nº Instâncias	%	Homogeneidade
0	8315	66%	1,3523
1	2413	19%	1,3207
2	1840	15%	1,3029
Total	12568	100%	

5.4.3.3 Software R

A execução dos algoritmos na ferramenta R iniciou pelo algoritmo conhecido como vizinho mais próximo (*single*), em sequencia pelo algoritmo vizinho mais distante (*complete*) e por último o algoritmo que faz a média das distâncias (*average*). Foram gerados 3 grupos levando em consideração os testes preliminares para o conjunto de dados.

5.4.3.3.1 Vizinho mais próximo

A tabela 75 apresenta os centróides probabilísticos, gerados pelo algoritmo vizinho mais próximo. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 2 (identificação da sessão) seguindo pelo atributo 1 (duração da resolução da questão).

Tabela 75 - Centróides probabilísticos gerados pelo algoritmo vizinho mais próximo.

Grupo	Centróides
0	4,ES_02-6,1,OK
1	429,ES_02-5,1,OK
2	519,ES_02-6,1,OK

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 1,5576 e o cálculo da separação resultou o valor 1,2991. A homogeneidade de cada grupo pode ser observada na tabela 76.

Tabela 76 - Homogeneidades dos grupos formados pelo vizinho mais próximo.

Grupo	Nº Instâncias	%	Homogeneidade
0	12550	100%	1,5778
1	13	0%	1,4821
2	5	0%	1,3949
Total	12568	100%	

5.4.3.3.2 Vizinho mais distante

A tabela 77 apresenta os centróides probabilísticos, gerados pelo algoritmo vizinho mais distante. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 2 (identificação da sessão) seguindo pelo atributo 1 (duração da resolução da questão).

Tabela 77 - Centróides probabilísticos gerados pelo algoritmo vizinho mais distante.

Grupo	Centróides
0	4,ES_02-6,1,OK
1	153,ES_02-5,1,OK
2	429,ES_02-6,1,OK

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 1,5774 e o cálculo da separação resultou o valor 1,3458. A homogeneidade de cada grupo pode ser observada na tabela 78.

Tabela 78 - Homogeneidades dos grupos formados pelo vizinho mais distante.

Grupo	Nº Instâncias	%	Homogeneidade
0	12459	99%	1,5776
1	91	1%	1,5698
2	18	0%	1,4890
Total	12568	100%	

5.4.3.3.3 Média das distâncias

A tabela 79 apresenta os centróides probabilísticos, gerados pelo algoritmo média das distâncias. Observa-se que os grupos foram agrupados prioritariamente pelo atributo 2 (identificação da sessão) seguindo pelo atributo 1 (duração da resolução da questão).

Tabela 79 - Centróides probabilísticos gerados pelo algoritmo média das distâncias.

Grupo	Centróides
0	4,ES_02-6,1,OK
1	241,ES_02-5,1,OK
2	429,ES_02-6,1,OK

Após a formação dos centróides, foram gerados os valores da homogeneidade e separação. O cálculo da homogeneidade global resultou o valor 1,5775 e o cálculo da separação resultou o valor 1,2972. A homogeneidade de cada grupo pode ser observada na tabela 80.

Tabela 80 - Homogeneidades dos grupos formados pelo algoritmo média das distâncias.

Grupo	Nº Instâncias	%	Homogeneidade
0	12522	100%	1,5779
1	33	0%	1,4958
2	13	0%	1,4483
Total	12568	100%	

5.4.3.4 Sistema Imunológico

Para a execução do algoritmo imunológico modificou-se os parâmetros referentes ao número total de grupos a ser gerado, formando 3 grupos considerando os resultados dos testes preliminares e selecionando a opção de normalização dos dados. Os demais parâmetros permaneceram com as opções *default*.

Os testes foram efetuados utilizando as duas condições de parada disponíveis para o algoritmo. A condição de parada definida por número de iterações considerou os valores: 100, 200, 500 e 1000. A condição de parada referente aos índices atribuídos para a homogeneidade e separação selecionou os seguintes valores: 0,111 para a homogeneidade e 0,8 para a separação.

As tabelas 81, 82, 83 e 84 apresentam os centróides gerados pela condição de parada referente ao número de iterações.

Tabela 81 - Centróide gerado após 100 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.0069594522113408976, 0.08051114542009064, 1.0157549093487603, ES_02-5;
1	0.005243085468645763, 0.0, 0.0, ES_02-6;
2	0.0034845572355290105, 0.0, 0.7394071977707947, ES_02-5;

Tabela 82 - Centróide gerado após 200 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.012970283477695665, 0.07323910051508149, 0.9558590332523483, ES_02-5;
1	0.026096980961758873, 0.06344750046452938, 0.40109413546877565, ES_02-6;
2	0.01854681983012264, 0.029413917786470036, 0.0, ES_02-6;

Tabela 83 - Centróide gerado após 500 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.016086091386573034, 0.053910662526133475, 0.7913458611062354, ES_02-5;
1	0.028381678068232122, 0.3067498503115807, 0.4104880208583291, ES_02-6;
2	0.019619918449311533, 0.027660122332257647, 0.027394197133478106, ES_02-6;

Tabela 84 - Centróide gerado após 1000 iterações do algoritmo imunológico.

Grupos	Centróide
0	0.01670366237707196, 0.06756569774513396, 0.7670422753025671, ES_02-5;
1	0.02897600565365369, 0.08619512012950406, 0.21808492073043778, ES_02-6;
2	0.01791062949856456, 0.022260362070488905, 0.0, ES_02-5;

A tabela 85 apresenta um comparativo dos valores da homogeneidade global e separação gerados a partir da condição de parada que utiliza o número de iterações. Observando os dados apresentados pelos valores iniciais da homogeneidade e separação identifica-se que eles são decisivos para o resultado final apresentado após as iterações.

Tabela 85 - Comparativo da homogeneidade e separação versus o número de iterações.

Iterações	Homogeneidade Inicial	Homogeneidade Final	Separação Inicial	Separação Final
100	0,0961	0,0958	0,7836	0,7848
200	0,0930	0,0920	0,6163	0,6630
500	0,1204	0,1193	0,6650	0,6898
1000	0,1044	0,1038	0,6121	0,6227

A tabela 86 apresenta os centróides gerados pela condição de parada referente aos índices esperados para a homogeneidade e separação.

Tabela 86 - Centróide gerado após o algoritmo imunológico atingir a homogeneidade global de 0,111 e a separação de 0,65.

Grupos	Centróide
0	0.019864048274544483, 0.03193673154288472, 0.028251530420343715, ES_02-6;
1	0.013346131113543314, 0.1986354214128797, 0.9657095096211663, ES_02-5;
2	0.01734561101241469, 0.03363731212283496, 0.7532663223086763, ES_02-5;

O algoritmo imunológico executou 4584 iterações até atingir a condição de parada selecionada chegando aos valores de: 0,1110 para a homogeneidade e 0,7552 para a separação.

5.4.2.5 Resultados

Os resultados da homogeneidade global e separação obtidos nos testes são apresentados na tabela 87. Observa-se que o algoritmo imunológico foi o que apresentou melhores resultados para o conjunto de dados Álgebra.

Tabela 87 - Resultados da homogeneidade global e separação para o conjunto de dados Álgebra.

Critério/Algoritmos	EM	K-média	VP	VD	MD	Imunológico
Homogeneidade	1,3649	1,3389	1,5776	1,5774	1,5775	0,0920
Separação	1,6887	1,7619	1,2991	1,3458	1,2972	0,6630

5.4.4 Considerações Finais

Após o término das execuções dos algoritmos com todos os grupos formados, pode-se efetuar dois tipos de análises sobre os resultados:

1) Interpretação dos grupos buscando informações sobre o conhecimento dos estudantes e as suas formas de aprendizagem.
2) Qualidade dos grupos avaliando-os de forma quantitativa utilizando critérios de avaliação, tais como: homogeneidade e separação.

Este trabalho focou no segundo tipo de avaliação. Avaliaram-se os grupos formados pelos algoritmos através dos índices de homogeneidade e separação. A tabela 88 exibe um comparativo dos resultados obtidos na execução dos algoritmos para cada conjunto de dados. Observando os resultados identificou-se que o algoritmo imunológico alcançou os melhores índices de homogeneidade e separação, isto é, apresentou os valores mais próximos à zero para a homogeneidade e para a separação os valores ficaram mais distantes de zero.

Tabela 88 - Resultados da homogeneidade e separação para todos os conjuntos de dados.

Conjuntos de Dados	Critérios	Algoritmos					
		EM	K-média	VP	VD	MD	Imunológico
Geometria	**Homogeneidade**	1,8306	1,9756	2,2571	2,1547	2,1850	0,9123
	Separação	2,4558	2,4644	2,4147	2,3512	2,3907	1,5679
Língua Chinesa	**Homogeneidade**	1,6805	1,6652	2,0251	1,9313	1,9110	0,3056
	Separação	1,9995	2,1665	1,7542	1,6644	1,7057	0,7724
Álgebra	**Homogeneidade**	1,3649	1,3389	1,5776	1,5774	1,5775	0,0920
	Separação	1,6887	1,7619	1,2991	1,3458	1,2972	0,6630

Durante a execução do experimento, identificou-se alguns pontos fortes e fracos de cada ferramenta. A tabela 89 apresenta esses pontos que podem auxiliar o desenvolvimento de futuras pesquisas na MDE.

Tabela 89 – Relação de pontos fortes e fracos de cada ferramenta.

Ferramenta	Pontos Fortes	Pontos Fracos
WEKA	• Criação do centróide inicial através de um valor médio. • Cálculo da similaridade dos grupos utilizando métricas específicas para os dados categorizados.	• Durante os testes executados no experimento não foi identificado nenhum ponto fraco.
R	• Agrupamento de dados no formato hierárquico.	• Consumo elevado de memória. • Cálculo da similaridade específico para dados numéricos.
Imunológico	• Geração dos grupos utilizando os critérios de homogeneidade e separação.	• Cálculo da similaridade específico para dados numéricos. • Inicialização não determinística dos centróides no inicio da execução do algoritmo.

6 CONCLUSÃO

6.1 SÍNTESE DO TRABALHO

A MDE é uma disciplina que vem adquirindo destaque dentro da área de MD. Ela é um campo que busca obter novas informações através de dados educacionais, para auxiliar estudantes e professores no processo de ensino-aprendizagem. Este trabalho apresentou a importância e a complexidade deste campo. Os dados educacionais estão em constante evolução dificultando a análise manual. Por isso, técnicas automáticas como o agrupamento de dados são necessárias.

A técnica de agrupamento consiste em formar grupos de dados com grande similaridade intragrupo e uma grande dissimilaridade entre elementos de grupos diferentes, através de métodos como: particional e hierárquico. Os estudos efetuados neste trabalho identificaram que a aplicação da técnica de agrupamento nos dados educacionais pode contribuir significativamente para o desenvolvimento de software educacional. Através da MDE é possível reconhecer padrões de comportamento em grupos.

Movido pelo objetivo de contribuir para o desenvolvimento da MDE, este trabalho apresenta um estudo comparativo dos algoritmos de agrupamento: k-média, EM, sistema imunológico e hierárquicos. As ferramentas WEKA e R foram utilizadas juntamente com três conjuntos de dados públicos: *Geometry, Chinese Tone Study e Algebra I 2006*. Os resultados das execuções foram tabulados e analisados através dos critérios de homogeneidade e separação.

O estudo comparativo foi dividido inicialmente por conjunto de dados seguido pelas execuções dos algoritmos. Para cada algoritmo executado, foram aplicados os critérios de homogeneidade e separação.

Com a análise dos resultados para cada conjunto de dados do experimento, identificou-se que os melhores índices de homogeneidade e separação foram alcançados pelo algoritmo imunológico. Em segundo lugar, houve uma variação entre os algoritmos k-média e EM. Por fim, os algoritmos pertencentes à ferramenta R apresentaram os índices mais elevados de homogeneidade e separação para os conjuntos de dados do experimento.

Após os testes e as análises dos resultados, chegou-se a conclusão que nem todas as ferramentas e algoritmos estão preparados suficientemente para trabalharem com dados educacionais. Pois, a ferramenta R e o algoritmo imunológico estão preparados para trabalhar

somente com dados numérico e não com dados categorizados como é que caso de alguns dados educacionais.

6.2 CONTRIBUIÇÃO DO TRABALHO

O objetivo de contribuir para o desenvolvimento da MDE e identificar os algoritmos mais apropriados foram alcançados ao término deste trabalho. Identificou-se que o algoritmo imunológico foi o mais apropriado para trabalhar com os dados educacionais dentre os algoritmos testados. Os resultados apresentados por este trabalho contribuem para as áreas da Ciência da Computação e MDE.

A área da MDE ganhou uma nova opção para ampliar as suas pesquisas envolvendo reconhecimento de padrões de comportamento em grupos, além de informações sobre os algoritmos existentes, facilitando a escolha dos algoritmos para a análise dos resultados de pesquisas futuras.

A área da Ciência da Computação ganhou novos desafios. A partir dos testes identificou-se que algumas ferramentas e algoritmos precisam ser aprimorados para obter resultados satisfatórios utilizando dados educacionais. Isto gera novas oportunidades para desenvolvedores trabalharem no aprimoramento dos algoritmos ou de ferramentas complementares. Identificou-se a necessidade das seguintes melhorias:

- A ferramenta R demostrou um consumo elevado de memória em comparação as outras ferramentas utilizadas neste experimento. Quando o arquivo de dados possuía um volume considerável de instâncias, o consumo de memória ultrapassou 16 GB. Sendo o ponto mais crítico da ferramenta.
- O algoritmo imunológico inicializa os centróides utilizando o algoritmo k-média. Esta inicialização demostrou uma instabilidade, pois a cada execução o valor é modificado. O ideal seria que os centróides fossem inicializados através de um valor médio gerado após várias execuções internas.
- A ferramenta R e o algoritmo imunológico estão preparados para trabalharem com dados do tipo numérico. Isto é, o cálculo da similaridade dos grupos é feito utilizando o cálculo da distância. O ideal para trabalhar com dados categorizados é utilizar métricas como: Jaccard, Overlap, Dice e Co-seno (Metz,2006; Machado, 2011).

6.3 TRABALHOS FUTUROS

Os resultados obtidos neste trabalho possibilitam novos estudos nas áreas da Ciência da Computação e MDE. O desenvolvimento de novos softwares educacionais utilizando o algoritmo imunológico como referência seria uma das linhas de pesquisa abordadas pelas as áreas. Além da continuidade nos estudos comparativos dos algoritmos de agrupamento utilizando os critérios de avaliação: homogeneidade e separação.

REFERÊNCIAS BIBLIOGRÁFICAS

ARRUDA, G. F. **Uma abordagem de redes complexas para agrupamento de dados.** 2011. 65f. Monografia (Graduação em Engenharia Elétrica com ênfase em Eletrônica) – Escola de Engenharia de São Carlos da Universidade de São Paulo, São Paulo. 2011.

BAKER, R.S.J.d. **Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model.** Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling. p. 76-80. 2007.

BAKER R.S.J.B., YACE K. **The State of Educational Data Mining in 2009: A Review and Future Visions.** *Journal of Educational Data Mining, v. 1, Issue 1, October 2009.* p. 3-17, 2009

BAKER R.S.J.B. **Data mining for Education**. *International Encyclopedia of Education (3rd edition), v. 7. p. 112-118. Oxford, UK: Elsevier.* 2010.

BROWNLEE, J. **Clever Algorithms**: Nature-Inspired Programming Recipes. Austrália: LuLu, 2011. 436 p.

CEN, H., KOEDINGER K., JUNKER B. **Learning Factors Analysis A General Method of Cognitive Model Evaluation and Improvement**. The 8th International Conference on Intelligent Tutoring Systems. 2006.

CHAMBERS, J. **Software for Data Analysis. Programming with R. Series: Statistics and Computing**. 1st ed. 2008. Corr. 2nd printing, 2008, XIV, 498 p.

CHEN, G. et al. **Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data**. Laboratories of Genetics, National Institute on Aging, National Institute of Health, Baltimore, NY, USA. 2002.

DEMPSTER, A.P. et al. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. Journal of the Royal Statistical Society, London, UK, v. 39, p. 1-38, dez. 1976.

DUARTE, F. J.F. **Optimização da Combinação de Agrupamentos baseado na Acumulação de Provas pesadas por Índices de Validação e com uso de Amostragem.** 2008. 391f. Tese (Doutor em Engenharia Eletrotécnica e de Computadores) - Universidade de Trás-os-Montes e Alto Douro. Portugal, Vila Real. 2008.

EVERITT, B. S. **Cluster Analysis.** Edward Arnold, 3 edition. 1993.

FAYYAD, U. M., PIATESKY-SHAPIRO, G. e SMYTH, P. **From Data Mining to Knowledge Discovery**: An Overview, em Advances in Knowledge Discovery and Data Mining. AAAI Press. 1995.

FORREST, S. et al. **Self-Nonself Discrimination in a Computer**. In IEEE Symposium on Research in Security and Privacy, Oakland, USA. p. 202-212, mai.1994.

GIRKE, T. **R & Bioconductor Manual.** University of California Riverside. 2012. Disponível em: <http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual#clustering_km>. Acesso em 13 de junho de 2012.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques.** Morgan Kaufmann. 550p. 2001.

HARTIGAN, J. A. **Clustering Algorithms**. New York: Publisher John Wiley & Sons, 351 p. 1975.

IEEE Task Force of Educational Data Mining. 2012. Disponível em: <http://datamining.it.uts.edu.au/edd/ >. Acesso em: 16 de abril de 2012.

Internacional Educational Data Mining Society**.** 2012. Disponível em: <http://www.educationaldatamining.org/>. Acesso em: 02 de março de 2012.

JAIN, A.; et al. **Data Clustering: A Review**. Journal: ACM Computing Surveys, Redwood , v. 31, n. 3, p. 264-323, set. 1999.

JUNIOR, N. L. C. ***Clusterização* baseada em algoritmos *fuzzi***. 2006. 166f. Dissertação (Pós-graduação em Ciência da Computação) - Universidade Federal de Pernambuco. Recife. 2006.

KAUFFMAN, L. e ROUSSEEUW, P. J. **Finding groups in data – an introduction to cluster analysis.** New York, Wiley Interscience. 1989.

KEOGH, E. **A Gentle Introduction to Machine Learning and Data Mining for the Database Community**. 18º Simpósio Brasileiro de Bancos de Dados - SBBD 2003, Manaus, 2003.

KOEDINGER, K.R., BAKER, R.S.J.d., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

MACHADO L. R. **Desenvolvimento de um Algoritmo Imunológico para Agrupamento de Dados.** 2011. 123f. Monografia (Graduação em Ciência da Computação) - Universidade de Caxias do Sul, Rio Grande do Sul. 2011.

MARK, H., EIBE, F., GEOFFREY, H., PFAHRINGER, B., REUTEMANN, P. e WITTEN, I. H. **The WEKA Data Mining Software: An Update**. SIGKDD Explorations, 11(1). 2009.

MAXIMILIANO, A. S. CORDEIRO, M. T. A. 2008. 45 fls. **Partição de grupos e validação da análise de agrupamento para equipamentos de fiscalização eletrônica de trânsito**. Monografia - Universidade Estadual do Paraná. Paraná. 2008.

MELLO, C. E. R. **Agrupamento de Regiões: Uma abordagem utilizando acessibilidade**. 2008. 96f. Dissertação (Mestrado em Ciência em Engenharia de Sistema de Computação) - Universidade Federal do Rio de Janeiro. Rio de Janeiro. 2008.

METZ, J. **Interpretação de clusters gerados por algoritmos de clustering hierárquico**. 2006. 152 f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Universidade de São Paulo. São Paulo. 2006.

NALDI, M. C. **Técnicas de combinação para agrupamento centralizado e distribuído de dados**. 2010. 276f. Tese (Doutorado em Ciências Matemáticas e de Computação) - Universidade de São Paulo. São Paulo. 2010.

OLIVEIRA, T. B. S. **Clusterização de dados utilizando técnicas de redes complexas e computação bioinspirada.** 2008. 112 f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Universidade de São Paulo, São Paulo. 2008.

PECHENIZKIY, M., CALDERS, T., CONATI, C., VENTURA, S., ROMERO, C. e STAMPER, J. (eds.). **Proceedings of the 4th International Conference on Educational Data Mining.** International Conference on Educational Data Mining (EDM) 2011. Eindhoven, 6-8 July, 2011.

PEREIRA, D. L. **Estudo do PSO na Clusterização de dados**. 2007. 52f. Monografia (Ciência da Computação). Universidade Federal de Lavras. Minas Gerais. 2007.

PRADO LIMA, M.F.W., WEBBER, C.G. e GUIMARÃES, B. B. **Estudo do Desenvolvimento Cognitivo Individual e de Grupos através da Análise Automática.** Revista Renote Novas Tecnologias na Educação, v 9, n.1. 2011.

PREGIBON, D. **Data Mining. Statistical Computing and Graphics.** p-8. 1997.

R. Disponível em: <http://www.r-project.org/>. Acesso em 28 de abril de 2012.

R Development Core Team**. R Data Import/Export**. Version 2.15.0. 2012. Disponível em: < http://cran.r-project.org/doc/manuals/R-data.pdf >. Acesso em 13 de junho de 2012.

REZENDE, S. O., MARCACINI, R. M. e MOURA,M. F. **O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento.** Revista de Sistemas de Informação da FSMA n.7. p. 7-21. 2011.

ROMERO, C. e VENTURA, S. **Educational Data Mining. A survey from 1995 to 2005.** *Expert Systems with Applications, 33*(1). p. 135–146. 2007.

ROMERO, C. e VENTURA, S. **Educational Data Mining: A Review of the State-of-the-Art.** IEEE Transaction n Systems, Man, and Cybernetics, Part C: Applications and Reviews. Issue 6. p. 601-618. 2010.

SANTOS, D. S. ***Bee clustering:* um Algoritmo para Agrupamento de Dados Inspirado em Inteligência de Enxames**. 2009. 89f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul. Porto Alegre, Rio Grande do Sul. 2009.

SCHEUER, O. e MCLAREN, B.M. **Educational Data Mining.** In: Encyclopedia of the Sciences of Learning, Springer. 2011.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., e TAN, P.. **Web usage mining: Discovery and applications of usage patterns from web data**. SIGKDD Explorations, 1(2). p. 12–23. 2000.

TAN, P., STEINBACH, M., KUMAR, V. **Introdução ao Data Mining**. Ciência Moderna. 900 f, 2009.

THAI-NGHE, N., HORVÁTH, T., SCHMIDT-THIEME, L. **Factorization Models for Forecasting Student Performance.** In: Proceedings of the Fourth International Conference on Educational Data Mining. p. 11-20. 2011.

The R Foundation for Statistical Computing e Department of Statistics and Mathematics Wirtschaftsuniversität Wien. **R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments.** 2008.

VENABLES, W. N., SMITH, D. M. e o R Development Core Team**. An Introduction to R Notes on R: A Programming Environment for Data Analysis and Graphics**. Version 2.13.1 .2011.

WITTEN, I. H. e FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann, San Francisco, Segunda Edição. 2000.

XU, R., Student Member, IEEE e WUNSCH II, D. **Survey of Clustering Algorithms**. IEEE Transactions on Neural Networks, vol. 16, no. 3, p. 645-678, Mai 2005.

YADAV, S. K. e PAL, S. **Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification.** World of Computer Science and Information Technology Journal (WCSIT), volume 2, No. 2, p. 51-56, 2012.

ANEXO

Anexo A – Método do cálculo da homogeneidade dos grupos.

```
public List<Homogeneidade> CalculaHomogeneidadeGrupo(Instances centroides,
                  Instances[] grupos, Instances todasInstanciasArquivo) {
       ////lista para armazenar as homogeneidades
       List<Homogeneidade> listaHomogeneidadeGrupos = new
       ArrayList<Homogeneidade>();

       //fórmula da distância euclidiana do pacote do WEKA
       DistanceFunction m_DistanceFunction = new EuclideanDistance();
       //faz a normalização dos dados
       m_DistanceFunction.setInstances(todasInstanciasArquivo);

       Double vlrHomogeneidadeGrupo = 0.0;
       for (int x = 0; x < grupos.length; x++) {
             Double dist = 0.0;
             for (int j = 0; j < grupos[x].numInstances(); j++) {
                   dist += m_DistanceFunction.distance(grupos[x].instance(j),
                                      centroides.instance(x));
             }
             //calcula o valor da homogeneidade do grupo
             vlrHomogeneidadeGrupo = (1.0 / grupos[x].numInstances()) * (dist);

             // cria o objeto e adiciona na lista
             Homogeneidade homogeneidadeGrupo = new Homogneidade();
             homogeneidadeGrupo.setCodigoGrupo(x);
             homogeneidadeGrupo.setNroInstancia(grupos[x].numInstances());
             homogeneidadeGrupo.setVlrHomogeneidade(vlrHomogeneidadeGrupo);
             listaHomogeneidadeGrupos.add(homogeneidadeGrupo);

       }
       return listaHomogeneidadeGrupos;
}
```

Anexo B – Método do cálculo da homogeneidade global.

```
public Double CalculaHomogeneidadeGlobal(List<Homogeneidade> listHomogeneidade) {

       Double somatorioDistancias = 0.0;
       Integer totalInstancias = 0;
       Double homogeneidadeGlobal = 0.0;

       //percorre a lista com todas as homgeneidades de cada grupo;
       for (int i = 0; i < listHomogeneidade.size(); i++) {
             somatorioDistancias +=  listHomogeneidade.get(i).getNroInstancia() *
             listHomogeneidade.get(i).getVlrHomogeneidade();

             totalInstancias += listHomogeneidade.get(i).getNroInstancia();
       }
       /* Formula FInal: 1/numero total instancias do arquivo* somatorio([numero
       instancia de cada cluster Ci * Homogeneidade de cada cluster Ci]
       EX (seu exemplo): 1/6*([3*h1]+[3*H2]) */
       homogeneidadeGlobal = (1.0/totalInstancias)*(somatorioDistancias);
       return homogeneidadeGlobal;
}
```

Anexo C – Método do cálculo da separação

```
public Double calcularSeparacao(Instances centroides, Instances[] grupos,
Instances todasInstanciasArquivo) throws Exception {

	Double somatorioDistancias = 0.0;
	Integer denominador = 0;
	Double separacao = 0.0;

	// fórmula da distância euclidiana do pacote do WEKA
	DistanceFunction m_DistanceFunction = new EuclideanDistance();
	//faz a normalização dos dados
	m_DistanceFunction.setInstances(todasInstanciasArquivo);

	for (int i = 0; i < grupos.length; i++) {

		/* NCi = quantidade de instancias(registros) que este grupo possui */
		Integer nci = grupos[i].numInstances();

		for (int j = i; j < grupos.length; j++) {
			if(i == j) {
				continue;
			}

			/* NCj = quantidade de instancias(registros) que este grupo
			possui */
			Integer ncj = grupos[j].numInstances();

		//[(C1*C2)*Dist(C1,C2)]+[(C1*C3)*Dist(C1,C3)]+[(C2*C3)*Dist(C2,C3)]
			somatorioDistancias += nci*ncj*(
			m_DistanceFunction.distance(centroides.instance(i),centroides.i
			nstance(j)));

			/* (C1*C2)+(C1*C3)+(C2*C3) */
			denominador+=nci*ncj;
		}
	}

/* exemplo:
S =
1/((C1*C2)+(C1*C3)+(C2*C3))*[(C1*C2)*Dist(C1,C2)]+[(C1*C3)*Dist(C1,C3)]+[(C2*C3)*D
ist(C2,C3)] */
	separacao = (1.0/denominador)*(somatorioDistancias);

	return separacao;

}
```

Anexo D – Arquivo com os resultados do agrupamento efetuado pela ferramenta R utilizando o algoritmo vizinho mais próximo para o conjunto de dados Geometria.

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[69] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 1 2 2
[137] 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 3 1 1 2 2 1 2 3 3 4 4 4 4 2 1 1 2 2 2 2 2 2 2 3 1 1 2 1 2 2 5 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[205] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 4 1
[273] 1 2 2 2 2 2 2 2 2 2 3 2 1 2 2 1 2 2 2 2 2 2 2 3 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[341] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 2 2 2
[409] 2 1 2 2 2 1 2 2 2 2 2 2 3 1 1 2 2 2 2 1 1 2 2 4 3 3 4 2 1 2 2 5 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[477] 4 4 7 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 2 2 2 2 2 2 2 2 3 2 2 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[545] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[613] 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 3 1 2 2 1 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[681] 4 4 4 4 4 4 4 4 2 1 2 4 4 4 4 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 7 7 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2
[749] 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 3 4 4 4 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1
[817] 1 1 1 1 1 1 2 1 1 2 2 1 1 2 2 2 3 2 2 2 2 1 2 2 2 1 2 3 4 4 2 1 2 2 5 2 2 2 3 3 2 2 2 2 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[885] 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[953] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 1 1 1
[1021] 1 1 1 1 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 3 1 1 2 2 1 2 3 4 4 2 1 2 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[1089] 4 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 3 3 2 1 1 2 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 3 2 1 2 2 1 2 2 2 2
[1157] 2 2 2 2 3 2 2 2 2 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1225] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2
[1293] 2 3 3 2 1 2 2 1 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 2 1 2 1 1 1 2 2 3 3 4 4 2 1 2 2 2 2 2 2 2 2 2 2 2 3 3 1 1
[1361] 2 1 2 2 2 5 2 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1429] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1497] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 3 1 2 2 1 2 3 4 4 4 4 4 4 4
[1565] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 1 2 1 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1633] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1701] 1 1 1 1 1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1769] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1837] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 2 2 3 1 2 2 2 2 1 2 2 4 3 4 2 1 2 2 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[1905] 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 4 4 2 2 2 2 3 2 2 2 2 2 2 2 2 2 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[1973] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[2041] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 3 3 1 1 2 2 2 2 1 2 3 4 4 4 2 1 2 2 2 2 2 2 5 2 2 1 2 2 2 2 3 2 2 2 4
[2109] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 4 4 4 4 4 4 4 4 4 1 1 2 2 2 2 1 2 2 3 3 4 4 4
[2177] 2 2 1 2 2 2 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

[2245] 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[2313] 1 1 1 1 1 1 1 1 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 2 1 2 2 2 1 2 2 2 2 2 3 3 2 2 2 1 2 2 1 2

[2381] 2 1 2 2 1 2 2 2 1 2 2 2 1 2 2 2 2 3 2 1 2 2 2 1 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 4 4

[2449] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[2517] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 2 2 2 1 1 2

[2585] 1 1 2 2 2 2 2 2 2 2 2 3 3 1 2 2 2 1 2 4 3 4 2 1 2 2 2 5 2 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4

[2653] 4 4 4 4 4 4 4 4 4 1 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[2721] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[2789] 1 1 1 1 1 1 1 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 1

[2857] 2 2 2 2 2 2 2 2 2 3 1 2 2 2 1 2 2 3 4 4 2 1 2 2 2 2 2 2 2 5 2 2 2 2 3 2 1 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 3 4 4 4 4 4 4 4 4 4 4

[2925] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 4 4 4 4 4 4 1 2 2 2 2 2 2 2 3 3 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 3 2 2 1 2

[2993] 2 1 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 3 2 1 2 2 2 1 1 2 2 2 2 2 2 2 2 3 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[3061] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3129] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 2 3 3 1 2 2 1 2 3 4 4 2 1 2 2 5 2 4 4 4 4 4 4 4 4 4

[3197] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 1 1 1

[3265] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3333] 1 1 8 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 8 8 1 1 1 1 1 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2

[3401] 2 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 3 3 1 1 2 2 2 2 1 1 2 2 4 4 3 3 4 4 2 2 1 1 2 2 2 2 5 5 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3469] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3537] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 1 2 2 1 1 2 2 1 2 2 2 2 2 3 3 2 2 2 2 2 2 1 2 2 1 2 3 4 4 2 1 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[3605] 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 4 1 7 7 7 7 1 1 1 1 1 1 1 1 1 2 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3673] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 3 1 2 2 2 1 4 4

[3809] 4 3 3 2 1 2 5 2 2 2 2 2 2 2 2 3 3 2 2 1 2 1 2 5 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 4

[3877] 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[3945] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4013] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 1 1 2

[4081] 2 2 2 3 3 2 2 2 2 2 2 1 2 2 2 2 1 1 2 2 3 3 4 4 4 4 2 1 1 2 2 2 2 5 2 2 2 2 2 2 2 2 2 3 2 1 1 2 2 2 2 1 2 2 1 2 5 2 2 1 1 2 2 1 2 2 2 2

[4149] 3 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4

[4217] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4285] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 2 2 2 2 2 3 3 1 2 2 1 2 2 3 4 4 2 1 2

[4353] 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4421] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 7 7 7 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4489] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 3 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 3

[4557] 2 2 2 2 1 2 2 2 2 2 2 2 2 3 2 1 2 2 1 2 2 2 2 2 2 2 2 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4625] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1

[4693] 1 1 1 1 1 1 1 2 1 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 3 1 1 2 2 2 2 1 2 4 3 4 4 2 1 2 2 2 2 2 2 5 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[4761] 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 7 7 7 1 1 1 1 1 1 1 1 1

[4829] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[4897] 1 1 1 8 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 2 2 2 2 2 3 3 2 1 1 2 2

[4965] 1 2 4 3 4 4 2 1 2 1 2 5 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5033] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5101] 1 1 1 1 1 1 1 1 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 1 1 2 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 3 1 2 2 2 2 1 2 2 3 4

[5169] 4 2 1 2 2 2 2 2 2 2 2 3 3 2 2 2 2 1 2 2 2 2 2 2 2 3 2 1 2 2 1 2 2 2 2 2 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6

[5237] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 7 7 7 7 1 1 1 1 1 1 1 1

[5305] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5373] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 1 2 2 2 2 1 1 2 2 2 1 2 1 2 2

[5441] 2 2 2 2 2 2 3 3 2 1 1 2 2 2 1 2 3 4 4 4 2 1 2 2 2 2 2 2 2 3 3 2 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4

[5509] 4 4 4 4 4 4 4 4 4 1 7 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 6 4 4 4 4 4 4 4 4 4 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[5577] 4 4 4 4 4 4 4 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 1 2 2 1 1 2 2 2 2 2 2 2 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5645] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2 2 2 2 1 1 1 2 2 2 2 2 2

[5713] 2 3 1 2 2 2 1 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5781] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[5849] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 3 3 1 2 2 1 2 2 3 4 4 2 1

[5917] 2 2 5 2 2 2 2 2 1 2 2 2 1 2 2 5 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 2 2 3 3 2 2 2 2 2 2 3 3 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 1 2

[5985] 2 2 2 2 2 2 2 3 2 1 2 2 2 2 2 2 2 2 2 3 2 1 2 2 1 2 2 2 2 2 2 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1

[6053] 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[6121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[6189] 1 1 2 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 1 1

[6257] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[6325] 1 1 1 8 8 8 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2

[6393] 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 1 2 2 2 2 1 1 2 3 4 4 4 2 1 2 2 2 3 2 2 2 2 2 2 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[6461] 4 4 4 4 4 4 4 4 4 4 4 4 4 1 2 5 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[6529] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[6597] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 1 2 2 2 2 2 2 2 3 1 1 2 2 2 1 2 4 3 3 4 2 1 2 2 2 2 2 2 3 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

[6665] 4 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 2 2 1 2 2 2 2 2 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 1 2 2 2

[6733] 2 2 2 2 2 3 1 1 2 2 2 2 1 2 3 4 4 4 2 1 2 2 2 2 2 2 5 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

Anexo E – Lista completa dos atributos do conjunto de dados Geometria.

Atributo	**Conteúdo**	**Coluna selecionada**
Row	Numeração das linhas.	Não selecionada.
Sample Name	Nome do conjunto de dados.	Não selecionada, pois contem somente o valor "All Data".
Anon Student Id	Id do estudante.	Não selecionada.
Session Id	Código do estudante.	Não selecionada.
Time	Resulta a soma da coluna "Problem Start Time" mais a coluna "Duration (sec)". Marca o inicio que o estudante começou a resolver os exercícios e vai sendo incrementado com a duração da resolução dos exercícios.	Não selecionada.
Time Zone	Coluna preenchida somente com o valor "US/Eastern".	Não selecionada.
Duration (sec)	Duração em segundo da resolução do exercício.	Não selecionada, dados incompletos.
Student Response Type	Coluna preenchida somente com o valor "ATTEMPT".	Não selecionada.
Student Response Subtype	Coluna vazia.	Não selecionada.
Tutor Response Type	Coluna preenchida somente com o valor "RESULT".	Não selecionada.
Tutor Response Subtype	Coluna vazia.	Não selecionada.
Level(Unit)	Coluna preenchida somente com o valor "Area".	Não selecionada.
Problem Name	Nome do problema.	**Selecionada.**
Problem View	Número de resoluções de um problema.	**Selecionada.**
Problem Start Time	Marca o inicio da resolução do problema.	Não selecionada.
Step Name	Indica uma etapa do problema a ser resolvido.	**Selecionada.**
Attempt At Step	Identifica quantas vezes a etapa foi repetida.	**Selecionada.**
Outcome	Indica se a etapa foi executada	**Selecionada.**

	corretamente ou não.	
Selection	Indica uma etapa do problema a ser resolvida. Mesmos dados que a coluna Step Name.	Não selecionada.
Action, Input, Feedback Text, Feedback Classification, Help Level e Total Num Hints	Coluna vazia.	Não selecionada.
KC(Geometry)	Coluna somente com o valor "Geometry".	Não selecionada.
KC Category(Geometry)	Coluna vazia.	Não selecionada.
KC(Original)	Áreas do conhecimento a serem avaliadas.	**Selecionada.**
KC Category(Original)	Coluna vazia.	Não selecionada.
KC(Area)	Identifica se a etapa utiliza ou não fórmulas.	**Selecionada.**
KC Category(Area)	Coluna vazia.	Não selecionada.
KC(Textbook)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Textbook)	Coluna vazia.	Não selecionada.
KC(Textbook New)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Textbook New)	Coluna vazia.	Não selecionada.
KC(Decompose)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Decompose)	Coluna vazia.	Não selecionada.
KC(Textbook_New_Decompose)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Textbook_New_Decompose)	Coluna vazia.	Não selecionada.
KC(DecomposeArith)	Especificações da etapa a ser avaliada	Não selecionada, dados incompletos.
KC Category(DecomposeArith)	Coluna vazia.	Não selecionada.
KC(DecompArithDiam)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(DecompArithDiam)	Coluna vazia.	Não selecionada.

KC(xDecmpTrapCheat)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(xDecmpTrapCheat)	Coluna vazia.	Não selecionada.
KC(LFASearchModel0)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel0)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1)	Coluna vazia.	Não selecionada.
KC(LFASearchAICWholeModel0)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchAICWholeModel0)	Coluna vazia.	Não selecionada.
KC(LFASearchAICWholeModel1)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchAICWholeModel1)	Coluna vazia.	Não selecionada.
KC(LFASearchAICWholeModel2)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchAICWholeModel2)	Coluna vazia.	Não selecionada.
KC(textbook2)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(textbook2)	Coluna vazia.	Não selecionada.
KC(LFASearchAICWholeModel3)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchAICWholeModel3)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-renamed&chgd)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-renamed&chgd)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-backward)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.

KC Category(LFASearchModel1-backward)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-backward)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-backward)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-renamed&chgd.2)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-renamed&chgd.2)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-renamed&chgd.3)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-renamed&chgd.3)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1.context-single)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1.context-single)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-context)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-context)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-context)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-context)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-back-context)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-back-context)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-back-context)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-back-context)	Coluna vazia.	Não selecionada.

KC(LFASearchModel1-back-context)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-back-context)	Coluna vazia.	Não selecionada.
KC(LFASearchModel1-renamed)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(LFASearchModel1-renamed)	Coluna vazia.	Não selecionada.
KC(Single-KC)	Coluna preenchida com o valor "Single-KC".	Não selecionada.
KC Category(Single-KC)	Coluna vazia.	Não selecionada.
KC(Unique-step)	Código da etapa.	Não selecionada.
KC Category(Unique-step)	Coluna vazia.	Não selecionada.
KC(Decompose_height)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Decompose_height)	Coluna vazia.	Não selecionada.
KC(Circle-Collapse)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
KC Category(Circle-Collapse)	Coluna vazia.	Não selecionada.
School	Coluna vazia.	Não selecionada.
Class	Coluna vazia.	Não selecionada.
CF(Factor add-or-m)	Especificações da etapa a ser avaliada.	Não selecionada, dados incompletos.
CF(Factor backward)	Ordem da fórmula.	**Selecionada.**
CF(Factor base-formula-p)	Indica se utiliza a fórmula da base.	Não selecionada, dados incompletos.
CF(Factor base-or-height)	Indica se é base ou altura ou não se aplica.	Não selecionada, dados incompletos.
CF(Factor basic-shape)	Indica se é um tipo básico como triangulo, circulo ou não se aplica.	Não selecionada, dados incompletos.
CF(Factor cir-quad)	Indica se é um quadrado ou circulo ou não se aplica.	Não selecionada, dados incompletos.
CF(Factor circle-formula)	Indica que formula é usada: área, circunferência diâmetro ou não se aplica.	Não selecionada, dados incompletos.

CF(Factor circle-given)	Indica o tipo circulo dado.	Não selecionada, dados incompletos.
CF(Factor circle-goal)	Indica o objetivo.	Não selecionada, dados incompletos.
CF(Factor embedd3-tri-reg_prob_fix)	Atributo não identificado.	Não selecionada.
CF(Factor embeddedness)	Indica se existe mais de um formato geométrico, um sobre posto ao outro.	**Selecionada.**
CF(Factor figure-part)	Indica qual parte da figura geométrica será calcula.	**Selecionada.**
CF(Factor figure-type)	Indica o tipo da figura geométrica.	**Selecionada.**
CF(Factor non-standard-orientation-or-shape)	Atributo não identificado.	Não selecionada.
CF(Factor parallelogram)	Fator especifico para o paralelograma.	Não selecionada.
CF(Factor parallelogram-type)	Fator especifico para o paralelograma.	Não selecionada.
CF(Factor repeat)	Indica se a fórmula já foi utilizada antes.	**Selecionada.**
CF(Factor required)	Atributo não identificado.	Não selecionada.
CF(Factor trapezoid-part)	Atributo não identificado.	Não selecionada.

Anexo F – Lista completa dos atributos do conjunto de dados Língua Chinesa.

Atributos	**Definição**	**Coluna Selecionada**
Row	Identificador sequencial da linha do conjunto de dados.	Não selecionada.
Sample Name	Nome do conjunto de dados.	Não selecionada, pois contem somente o valor "All Data".
Anon Student Id	Identificador do estudante.	Não selecionada.
Session Id	Identificação da seção.	Não selecionada.
Time	Data informando o momento do inicio da etapa.	Não selecionada.
Time Zone	Coluna preenchida com o valor "UTC".	Não selecionada.
Duration (sec)	Marca a duração em segundos para a resolução da etapa.	Não selecionada, pois possui dados em branco.
Student Response Type	Identificador de pedido ajuda na etapa.	Não selecionada, pois existe outro atributo com essa informação.
Student Response Subtype	Coluna vazia.	Não selecionada.
Tutor Response Type	Identificador de pedido ajuda na etapa.	Não selecionada, pois existe outro atributo com essa informação.
Tutor Response Subtype	Coluna vazia.	Não selecionada, dados incompletos.
Level(Unit)	Identificador da Lição do "curso".	**Selecionada**
Level(Section)	Nível da seção.	**Selecionada**
Problem Name	Identificador do problema por lição.	**Selecionada**
Problem View	Número de resoluções do problema.	**Selecionada**
Problem Start Time	Marca o início da resolução dos problemas.	Não selecionada.
Step Name	Componente que originou a entrada dos dados.	**Selecionada**
Attempt At Step	Identifica quantas vezes a etapa foi repetida.	**Selecionada**
Outcome	Identificador da resposta da etapa.	**Selecionada**
Selection	Opção selecionada no tutor.	Não selecionada
Action	Ação executada no tutor.	Não selecionada

Input	Valor inserido no tutor.	Não selecionada, dados incompletos.
Feedback Text	Mensagem de feedback apresenta para o estudante.	**Selecionada**
Feedback Classification	Coluna vazia.	Não selecionada, dados em branco.
Help Level	Identificador do tipo de mensagem.	Não selecionada, dados incompletos.
Total Num Hints	Coluna preenchida com o valor "3" ou vazia.	Não selecionada, dados incompletos.
KC(Default)	Identificador dos tipos te tons.	Não selecionada, dados incompletos.
KC Category(Default)	Categoria dos tons. Preenchido com o valor 'tones' ou vazio.	Não selecionada, dados incompletos.
KC(Single-KC)	Coluna preenchida com o valor 'Single-KC'.	Não selecionada.
KC Category(Single-KC)	Coluna vazia.	Não selecionada.
KC(Unique-step)	Identificador do passo.	Não selecionada, dados incompletos.
KC Category(Unique-step)	Coluna vazia.	Não selecionada.
School	Coluna preenchida com o valor 'CMU'. Identifica a escola.	Não selecionada.
Class	Coluna vazia.	Não selecionada.

Anexo G – Lista completa dos atributos do conjunto de dados Álgebra.

Atributos	**Definição**	**Coluna Selecionada**
Row	Identificador da linha.	Não selecionada.
Anon Student Id	Identificador dos estudantes.	Não selecionada.
Session Id	Identificador da seção para cada estudante.	Não selecionada.
Time	Marca o inicio da resolução da lição (coluna Problem Start Time) mais a soma da duração (coluna duration).	Não selecionada.
Time Zone	Coluna vazia.	Não selecionada.
Duration (sec)	Duração para a resolução da etapa.	**Selecionada.**
Student Response Type	Possui o identificador se foi pedido ajuda na etapa.	Não selecionada, pois a informação se repete na coluna Outcome.
Student Response Subtype	Coluna vazia.	Não selecionada.
Tutor Response Type	Possui o identificador da resposta da ajuda.	Não selecionada, pois a informação se repete na coluna Outcome.
Tutor Response Subtype	Coluna vazia.	Não selecionada.
Level(Unit)	Coluna preenchida com o valor "ES_02".	Não selecionada, pois existe somente um valor.
Level(Section)	Identificação da sessão.	**Seleciona.**
Problem Name	Nome do problema.	Não selecionado, pois possui poucas repetições.
Problem View	Número de resoluções de um problema.	Não selecionado, pois possui poucas repetições.
Problem Start Time	Inicio da resolução da etapa.	Não selecionada.
Step Name	Nome da etapa.	Não selecionado, pois possui poucas repetições.
Attempt At Step	Identificador de quantas vezes a etapa foi repetida.	**Seleciona.**
Outcome	Marca se a etapa está correta, incorreta e o tipo de pedido de ajuda.	**Seleciona.**
Selection	Mostra qual será a próxima ação	Não selecionada.
Action	Mostra qual operação matemática o estudante executou.	Não selecionada, dados incompletos.
Input	O valor modificado.	Não selecionada, dados incompletos.
Feedback Text	Feedback das ações.	Não selecionada, dados

		incompletos.
Feedback Classification	Identifica se ocorreu um bug, ou mostra uma informação (ajuda).	Não selecionada, dados incompletos.
Help Level	Nivel da ajuda.	Não selecionada, dados incompletos.
Total Num Hints	Coluna vazia.	Não selecionada.
Condition Name	Coluna preenchida com os valores “control” e “treatment”	Não selecionada.
Condition Type	Coluna Vazia.	Não selecionada.
KC(Default)	Aparenta serem os passos de ajuda ao estudante.	Não selecionada, dados incompletos.
KC Category(Default)	Coluna preenchida com o valor “autocreated” e vazio.	Não selecionada, dados incompletos.
KC(Single-KC)	Coluna preenchido com o valor “Single-KC”.	Não selecionada.
KC Category(Single-KC)	Coluna vazia.	Não selecionada.
School	Escola onde do tutor foi rodado.	Não selecionada, coluna preenchida somente com o valor “CWCTC”.
Class	Salas que utilizaram o tutor.	Não selecionada.