UTILIZAÇÃO DE ALGORITMOS DE AGRUPAMENTO NA MINERAÇÃO DE DADOS EDUCACIONAIS

Carine G. Webber – CCTI/UCS – cgwebber@ucs.br

Daline Zat – CCTI/UCS – dzat@ucs.br

Maria de Fátima Webber do Prado Lima – CCTI/UCS – mfwplima@ucs.br

Resumo. *A mineração de dados educacionais é uma tarefa importante dado o grande volume de informações produzidas pelos softwares educacionais. Minerar significa buscar padrões relevantes que possam ser usados para aprimorar os processos de aprendizagem. Este estudo analisa ferramentas de mineração em três conjuntos de dados públicos: Geometry, Chinese Tone Study e Álgebra I 2006. Os resultados obtidos foram tabulados e analisados através dos critérios de homogeneidade e separação. A análise identificou que as ferramentas são insuficientes para trabalhar com dados educacionais. Contudo, a ferramenta imunológica foi a que apresentou melhores resultados.*

Palavras chave: *mineração de dados educacionais, algoritmos de agrupamento.*

Use of Clustering Algorithms in the Educational Data Mining

Abstract. *Educational data mining is an important task considering the volume of educational information available. Mining means searching for relevant and useful patterns to improve educational processes. This study analyses mining tools applied to three public datasets named Geometry, Chinese Tone Study and Algebra I 2006. The results were tabulated and analyzed through the criterion of homogeneity and separation. The analysis has concluded that mining tools are not sufficient to work with educational data. Although tests have shown that the immunological tool obtained the best results.*

Keywords*: educational data mining, clustering algorithms.*

1. Introdução

Informações armazenadas em ambientes educacionais constituem fontes riquíssimas de conhecimento que podem ser analisadas através da mineração de dados. A área denominada Mineração de Dados Educacionais (MDE) é um campo de pesquisa que busca extrair informações novas e úteis com o intuito de desenvolver e fortalecer as teorias cognitivas de ensino-aprendizagem. A MDE analisa dados, oriundos de sistemas de aprendizagem interativos, sistemas tutores inteligentes, ambientes de educação à distância e sistemas administrativos referentes a escolas e universidades, com o objetivo de descobrir padrões ou evidências sobre estudantes e formas de aprendizagem (IEEE,

2012). Tal processo de descoberta de conhecimento pode auxiliar professores a conduzirem melhor suas turmas, identificando dificuldades, compreendendo melhor o processo de aprendizagem dos estudantes e a melhorando os métodos de ensino. Como consequência, os professores pode oferecer um *feedback* mais apropriado aos estudantes através de reflexões pertinentes as suas aprendizagens (Romero, 2010).

As pesquisas na área de MDE partem das técnicas comuns de mineração de dados, adaptando e aprimorando-as. Tais técnicas foram classificadas e explicadas por Baker (2010) como sendo de: predição, agrupamento, mineração relacional, descoberta com modelos e destilação de dados para o julgamento humano. Destas categorias, este estudo utilizou as técnicas de *agrupamento*. Elas são capazes de identificar conjuntos de dados similares, que podem corresponder a um padrão ou comportamento típico observado nos dados. A interpretação dos padrões extraídos cabe ao professor e permite que ele identifique dificuldades e reveja decisões didáticas por exemplo.

Neste contexto, este artigo está organizado em 6 seções. A seção 2 introduz os principais conceitos associados aos algoritmos de agrupamento. A seção 3 descreve alguns trabalhos que utilizam técnicas de agrupamento para resolver problemas na MDE. A seção 4 aborda o método de pesquisa utilizado no experimento desenvolvido. Finalmente, a seção 5 apresenta os resultados obtidos no experimento através do método de *agrupamento* aplicado e a seção 6 conclui o artigo.

2. Agrupamento

A tarefa de análise automática de dados consiste em um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou apoiar uma tomada de decisão (Fayyad et al, 1995; Mitchel, 1997). Esse processo envolve a preparação de dados, a aplicação de algoritmos de mineração de dados, a análise e interpretação de padrões, e a avaliação do conhecimento extraído. Os dados devem representar casos, cenários, exemplos ou instâncias representativas do domínio a ser tratado.

O processo de análise automática pode seguir uma abordagem de aprendizagem supervisionada (quando as instâncias são previamente classificadas) ou não supervisionada (ausência de classificação). Em casos onde as instâncias estejam classificadas, pode-se aplicar técnicas para extrair propriedades comuns as instâncias pertencentes a uma mesma classe. Por outro lado, em um conjunto de instâncias não classificadas, as técnicas de agrupamento permitem a partição dos dados em grupos de instâncias similares. Algoritmos de agrupamento podem ser aplicados quando não existam classes bem definidas que possam ser identificadas facilmente *a priori*.

A análise de grupos se vale de modelos estatísticos multivariados visando formar grupos de objetos homogêneos e similares entre si. Mais especificamente as técnicas de agrupamento buscam formar grupos maximizando a similaridade intragrupo e minimizando as similaridades intergrupos (Mitchel, 1997). A análise de grupos permite também identificar objetos que não apresentam similaridades com nenhum grupo.

O ponto de partida do uso de técnicas de agrupamento é a disponibilidade de dados históricos que representem cenários do domínio a ser estudado. De modo geral, algoritmos de agrupamento geram partições dos dados a partir de similaridades entre valores de atributos. A visualização gráfica das partições é um recurso necessário e desejável para a análise dos resultados obtidos. Uma vez geradas, as partições de dados devem ser interpretadas no contexto do domínio, favorecendo *insights* e reflexões que conduzam o usuário a validação de hipóteses e construção de conhecimento. Os grupos

devem ser interpretados em termos de suas características mais relevantes normalmente apresentadas pelo elemento centroide (ponto médio) do grupo.

Os grupos podem ser exclusivos (cada instância pertence a apenas um grupo) ou sobrepostos (uma instância pode pertencer a mais de um grupo). Os métodos exclusivos podem ser baseados em distâncias (as instâncias são particionadas de acordo com medidas de similaridade e de distância) ou hierárquicos (as instâncias são organizadas em níveis hierárquicos partindo de um alto nível geral até o nível de instâncias). Os métodos que permitem sobreposição de instâncias e grupos são probabilísticos (uma instância pode pertencer a cada grupo com uma certa probabilidade) ou *fuzzy* (uma instância pode pertencer a cada grupo em diferentes graus de inclusão). Em geral a escolha do melhor método de agrupamento não é fácil, devendo ser guiada pela natureza dos dados e pelos resultados esperados.

3. Trabalhos Relacionados

Nos últimos anos diversos trabalhos tem explorado os benefícios que o MDE traz ao ambiente educacional. Abdous (2012) realizou um estudo sobre o comportamento de estudantes em um curso de graduação à distância, a fim de analisar a relação entre as questões *online* submetidas com a nota final obtida pelos estudantes. Para isso, utilizou a análise de agrupamentos fornecida pelo software Nvivo9 que analisou todas as interações via *chat* geradas pela ferramenta. Lee (2012) analisou as atividades realizadas por alunos no ambiente MasteringPhysics a fim de verificar as habilidades dos mesmos. Para realizar esta análise utilizou as técnicas de estimativa modal de Bayes e o modelo gaussiano, onde o último modelo se mostrou mais eficiente. Zafra (2012) desenvolveu um algoritmo genético denominado G3P-MI e utilizou-o para determinar a probabilidade de alunos serem aprovados em um curso, analisando as atividades realizadas por eles.

A Conferência Internacional de Mineração de Dados Educacional tem publicado artigos que são relevantes ao contexto deste trabalho. Thai-Nghe (2011) buscou prever o desempenho dos estudantes através do desenvolvimento de um novo modelo de previsão com fatoração tensor, comparando o algoritmo EM com os algoritmos TFMAF (*Tensor Factorization - Moving Average e Forecasting*) e TFF (*Tensor Forecasting of Factorization*) . Li (2011) propôs um método que descobre automaticamente modelos de estudante. O sistema usa um agente de aprendizagem de máquina chamado de SimStudent para adquirir o conhecimento da habilidade e compara o conhecimento adquirido com o modelo humano-gerado e um estudante real. Ele utilizou a medida do erro quadrático para ajuste dos dados dos estudantes para avaliação transversal.

Gong (2011) utilizou técnicas de predição para analisar qual é a melhor forma para modelar o conhecimento dos alunos, considerando a dificuldade do item e as habilidades. Maldonado (2011) usou o método hierárquico aglomerativo (técnica de agrupamento) para analisar as interações homem-máquina através dos *logs* criados pelos grupos. Kardan (2011) considerou o agrupamento não supervisionado através do algoritmo K-média para criar um modelo automático para descobrir e reconhecer testes padrões relevantes durante a interação do estudante com software educacional. Gowda (2011) comparou modelos de aprendizagem utilizando regressão linear para verificar a consistência dos dados com a análise original e a classificação dos dados é realizada através de análise Bayesiana. Trivedi (2011) apresenta a aplicação do algoritmo de agrupamento espectral na mineração de dados educacionais.

Rus (2012) utilizou o algoritmo K-média no ambiente WEKA para analisar e

categorizar os diversos tipos de diálogos existentes em um jogo educacional. Para fornecer subsídios para melhor projetar mecanismos de adaptação e *feedback* em Sistemas de Tutores Inteligentes, Bouchet (2012) utilizou o algoritmo de agrupamento EM para agrupar estudantes que continham conhecimentos prévios diferentes em grupos de acordo com os ganhos de aprendizagem realizados. Já López (2012) utiliza o algoritmo EM para avaliar os dados do fórum do ambiente *Moodle* de uma Universidade para verificar se a participação do aluno nesta ferramenta tem relação com as notas obtidas no final do curso.

Observou-se nos artigos mencionados nesta seção que as técnicas de agrupamento têm sido usadas na MDE. No entanto, nenhum dos trabalhos citados verificou qual técnica de grupamento é mais adequada para ser utilizada na MDE. Nas seções seguintes pretende-se analisar a adequação dos algoritmos de agrupamento.

4. Método

Tendo em vista que o objetivo deste trabalho é verificar o desempenho de algoritmos de agrupamento, foram utilizados três conjuntos de dados públicos com formatos de dados diferenciados retirados do PSCL DataShop: (1) Geometry, composto por dados retirados de um ano de curso de Geometria (Koedinger, 2010); (2) Chinese Tone Study que possui informações sobre os estudantes que estão aprendendo a sua segunda língua (Koedinger, 2010) e; (3) Álgebra I 2006 que armazena uma quantidade relativa de números e equações matemáticas.

Para realizar a análise destes dados, foram escolhidos os algoritmos de agrupamento denominados K-média, EM, Imunológico e hierárquico. O algoritmo K-média (Hantingan, 1975) foi escolhido por ser um dos algoritmos mais conhecidos. Ele forma os grupos visando minimizar a distância entre os elementos do grupo em relação ao centro. O algoritmo EM (Demspter et al, 1977) possui o objetivo de encontrar o melhor ajuste de um modelo para um conjunto de dados através da estimativa da máxima verossimilhança. É um algoritmo estatístico e não precisa da informação do número de grupos. A ferramenta imunológica (Machado, 2011) utiliza conceitos de Sistemas Imunológicos Artificias. Ela forma grupos levando em consideração uma maior homogeneidade entre os elementos do mesmo grupo e uma maior heterogeneidade entre os elementos de grupos diferentes. Os algoritmos hierárquicos (Mello, 2008; Naldi, 2010) agrupam os elementos em uma estrutura de árvore, organizando os grupos em formato hierárquico, resultando assim uma sequência aninhada de partições. Os algoritmos foram testados nas ferramentas WEKA e R.

A execução e análise do experimento previu algumas atividades preliminares. Primeiro os dados passaram por um pré-processamento (limpeza dos dados e normalização). Após foi realizada a adequação das ferramentas ajustando os critérios de avaliação nos algoritmos. Com os algoritmos ajustados, cada algoritmo foi executado diversas vezes para obtenção de dados confiáveis.

Os grupos formados pelos algoritmos selecionados foram avaliados pelos critérios de homogeneidade e separação. O critério de homogeneidade avalia se os dados mais similares possíveis foram colocados no mesmo grupo, por isso quanto menor o resultado obtido mais similar é o grupo. O critério de separação avalia se os grupos estão suficientemente diferentes uns dos outros, ou seja, quanto maior o resultado, maior a distância 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 (Chen et al., 2002; Machado, 2011):

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

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 N_{ci} e N_{cj} são números de instâncias nos grupos do *i-ésimo* e *j-ésimo* e D é a função da distância a ser aplicada (Chen et al., 2002; Machado, 2011):

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

A partir dessas equações consegue-se demonstrar a coerência dos grupos observando que H_{ave} e S_{ave} são fortemente ligados. 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 utiliza como medida a distância euclidiana (Chen et al., 2002), 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;
- x_{il} é o l-ésimo atributo do elemento *i* (Mello, 2008).

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

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 (Metz, 2006; Mello, 2008; Machado, 2011).

5. Resultados e Discussão

O experimento foi realizado considerando-se os três conjunto de dados e as ferramentas Weka, R e Imunológica. Para facilitar a leitura dos dados foram desenvolvidos dois softwares de conversão. Eles lêem os dados de um arquivo original e geram um arquivo específico para cada ferramenta de teste.

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.

5.1 Conjunto de Dados Geometria

Para realizar a análise do conjunto de dados Geometria foram selecionados os 12 atributos considerados mais relevantes: identificação do tipo do problema, número de resoluções do problema, identificação da questão do problema, número de repetições do

problema, resultado da questão (correto ou incorreto), área de conhecimento avaliada, identificação de existência de fórmula na questão, ordem de apresentação da fórmula na questão, existência de um ou mais formatos geométricos sobrepostos na questão, identificação da parte geométrica calculada na questão, identificação do tipo da figura geométrica, identificação da repetição da regra de produção. Para a execução dos algoritmos, o único parâmetro *default* alterado foi o número total de grupos a ser gerado. Foram considerados 8 grupos definidos através de testes preliminares.

Os grupos formados pelo EM ficaram prioritariamente agrupados pelo tipo de problema e subsequentemente pela questão abordada no problema. Os grupos formados pelo algoritmo K-média ficaram agrupados prioritariamente pelos mesmos atributos do EM. O atributo que identifica as respostas de cada questão (correta ou incorreta) mostrou-se mais decisivo na formação dos grupos no algoritmo EM do que no k-média.

Na ferramenta R foram testados três algoritmos. O algoritmo do vizinho mais próximo (*single*) manteve o mesmo comportamento que os algoritmos EM e o K-média. No algoritmo do vizinho mais distante (*complete*) observou-se que o algoritmo formou grupos bem variados. No algoritmo da média das distâncias (*average*) notou-se que não houve repetição dos valores do atributo que identifica o tipo de problema, como ocorreu nos outros algoritmos. A Tabela 1 mostra os valores da homogeneidade de cada grupo nos algoritmos executados na ferramenta WEKA e R.

Tabela 1. Homogeneidade do conjunto de dados geometria (Weka e R)

	Ferramenta Weka						Ferramenta R								
	EM			K-Média			Vizinho mais próximo			Vizinho mais distante			Média das distâncias		
Grupo	NI	%	H	NI	%	H	NI	%	H	NI	%	H	NI	%	H
0	1512	22	1,9444	2141	32	2,2888	3946	58	2,3489	911	13	2,0284	1607	24	2,0679
1	767	11	1,6910	972	14	1,8996	1162	17	2,0532	976	14	2,2969	1124	17	2,3222
2	587	9	1,5670	328	5	1,6221	132	2	1,2392	844	12	1,8949	1215	18	2,2870
3	782	12	1,8336	320	5	1,6187	1409	21	2,3568	1236	18	2,2881	1294	19	2,0833
4	1126	17	2,0919	1675	25	2,0472	21	0	0,4959	1294	19	2,0833	894	13	2,2868
5	652	10	1,7012	784	12	1,7066	32	0	1,2087	894	13	2,2868	21	0	0,4959
6	837	12	1,7875	319	5	1,4117	28	0	1,5607	547	8	2,2189	547	8	2,2189
7	515	8	1,6621	239	4	1,5772	48	1	1,3945	76	1	1,7670	76	1	1,7670
Total	6778	100		6778	100		6778	100		6778	100		6778	100	

Legenda: NI – Nº de Instâncias H-Homogeneidade

Os testes realizados na ferramenta Imunológica consideraram as duas condições de parada disponíveis. A primeira condição de parada testada considerou o número de iterações: 100, 200, 500 e 1000. A tabela 2 exibe um comparativo dos valores da homogeneidade global e separação obtidos em cada execução.

Tabela 2. Homogeneidade e separação X iterações na ferramenta Imunológica

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 segunda condição de parada permite que o usuário defina valores finais esperados para as variáveis de homogeneidade e separação (0,98 para a homogeneidade e 1,6 para a separação). Neste caso, a ferramenta imunológica executou 354 iterações até atingir a condição de parada, atingindo valores de 0,97998 para a homogeneidade e 1,72737 para a separação.

5.2 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. Foram selecionados os atributos: identificação da lição do curso, nível da seção, identificação do problema, número de resoluções do problema, componente que originou a entrada dos dados, identificação do número de repetições por etapa do problema, resultado da questão (correto, incorreto ou *hint*) e mensagem de *feedback* apresentada para o estudante. Com os atributos selecionados e convertidos para cada tipo de ferramenta, iniciou-se a execução dos algoritmos considerando 10 grupos distintos como parametrização inicial. Os demais parâmetros permaneceram com as opções *default*.

No algoritmo EM, os grupos foram agrupados prioritariamente pelo atributo identificação da lição do curso e, em sequência, pelo atributo identificação do problema. Analisando os centróides de cada grupo, identificou-se que os atributos identificação da lição do curso, número de resoluções do problema, identificação do número de repetições por etapa do problema, resultado das questões e mensagem de *feedback* apresentada ao estudante foram os mais relevantes na formação dos grupos. Nos testes da ferramenta R, a execução dos três algoritmos (Vizinho mais próximo, Vizinho mais distante e Média das distâncias) formaram grupos prioritariamente pelos atributos identificação da lição do curso e identificação do problema. A Tabela 3 mostra os resultados de homogeneidade destes algoritmos.

Tabela 3. Homogeneidades do conjunto de dados Língua Chinesa

	Ferramenta Weka						**Ferramenta R**								
	EM			**K-Média**			**Vizinho mais próximo**			**Vizinho mais distante**			**Média das distâncias**		
Grupo	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**
0	3211	7	1,8873	6093	13	1,7000	38437	87	2,0387	6431	15	1,9494	4422	10	1,8940
1	8649	18	1,7063	7401	15	1,7237	2759	6	1,9409	1201	3	1,8266	2773	6	1,8500
2	4584	9	1,6387	5439	11	1,6630	661	1	1,8124	8516	19	1,8935	8563	19	1,8905
3	3345	7	1,6616	6211	13	1,8064	2214	5	1,9632	6591	15	1,8950	3003	7	1,8022
4	6871	14	1,5494	1794	4	1,7107	1	0	0,0000	5930	13	1,8864	6189	14	1,8507
5	7005	14	1,7035	14773	30	1,6236	1	0	0,0000	3328	8	1,7671	6273	14	1,8726
6	2289	5	1,8996	2989	6	1,4947	1	0	0,0000	4879	11	2,0926	7163	16	2,0558
7	3436	7	1,7107	588	1	1,7297	1	0	0,0000	4543	10	2,0390	3035	7	1,9652
8	5425	11	1,5178	2668	6	1,4699	1	0	0,0000	2607	6	1,9919	2601	6	1,9921
9	3628	7	1,7870	487	1	1,7001	1	0	0,0000	51	0	0,9804	55	0	0,9818
Total	48443	100		48443	100		44077	100		44077	100		44077	100	

Legenda: NI – Nº de Instâncias H-Homogeneidade

Os testes realizados com a ferramenta imunológica produziram os resultados da tabela 4. Ela exibe um comparativo dos valores da homogeneidade global e separação

gerados de acordo com os números de iterações (condição de parada).

Tabela 4. Língua Chinesa: homogeneidade e separação X 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

5.3 Conjunto de Dados Álgebra

Para o conjunto de dados Álgebra, os atributos considerados mais relevantes foram: duração para a resolução da questão (em segundos), identificação da sessão, identificação do número de repetições por questão do problema e identificação do resultado da questão. Neste conjunto de dados, o número total de grupos a ser gerado foi alterado para 3 grupos considerando os resultados dos testes preliminares efetuados. Nos algoritmos EM e K-Média, os atributos duração da resolução e 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.

Após a execução dos algoritmos EM, K-Média, Vizinho mais próximo, Vizinho mais distante e Média das distâncias, observou-se que os dados foram agrupados prioritariamente pelo atributo identificação da sessão e, em sequência, pelo atributo duração da resolução da questão. Analisando os atributos do conjunto de dados, identificou-se que os atributos duração, sessão, repetições da etapa e respostas da etapa seriam os mais relevantes na formação dos grupos. A tabela 5 sintetiza os resultados dos testes.

Tabela 5. Homogeneidades do conjunto de dados Algebra

	Ferramenta Weka						Ferramenta R								
	EM			K-Média			Vizinho mais próximo			Vizinho mais distante			Média das distâncias		
Grupo	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**	**NI**	**%**	**H**
0	1641	13	1,2787	8315	66	1,3523	12550	100	1,5778	12459	99	1,5776	12522	100	1,5779
1	3713	30	1,1766	2413	19	1,3207	13	0	1,4821	91	1	1,5698	33	0	1,4958
2	7214	57	1,4815	1840	15	1,3029	5	0	1,3949	18	0	1,4890	13	0	1,4483
Total	12568	100		12568	100		12568	100		12568	100		12568	100	

Legenda: NI – Nº de Instâncias H-Homogeneidade

Para este conjunto de dados, a ferramenta imunológica produziu resultados conforme a tabela 6. Ela apresenta um comparativo dos valores da homogeneidade global e separação gerados de acordo com o número de iterações.

Tabela 6. Álgebra: homogeneidade e separação X 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

6. Resultados e Conclusão

Este trabalho avaliou os grupos formados pelos algoritmos através dos índices de homogeneidade e separação. A tabela 7 exibe um comparativo dos resultados obtidos na execução dos algoritmos para cada conjunto de dados. Para a homogeneidade quando mais próximo de zero for o valor melhor é o resultado. Para a separação, quanto maior o valor melhor é o resultado.

Observando os resultados, uma análise preliminar permite concluir que a ferramenta imunológica alcançou os melhores índices de homogeneidade e separação. 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.

Tabela 7. Melhores resultados para homogeneidade e separação

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

A MDE é um campo de estudo relevante e promissor. Os estudos efetuados identificaram que a aplicação da técnica de agrupamento aos dados educacionais pode contribuir significativamente para o desenvolvimento de software educacional pois ela consegue identificar padrões nos dados. Entretanto, após exaustivos testes e subsequentes análises dos resultados, chegou-se a conclusão que nem todas as ferramentas estão suficientemente preparadas para trabalharem com dados educacionais. A ferramenta R e a ferramenta imunológica trabalham apenas com dados numéricos e não com dados categorizados (muito comuns na área educacional).

Referências Bibliográficas

ABDOUS, M.; HE, W.; YEN, C.-J. Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade. **Educational Technology & Society**, n.15 (3), 2012, p.77–88.

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

BOUCHET,F.; AZEVEDO, R.; KINNEBREW, J.; BISWAS, G. Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering SelfRegulated Learning. In: INT CONF ON EDUCATIONAL DATA MINING , 2012, p.65-72.

CHEN, G., Jaradat, S.A., Bannerjee, N. Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data. **Stat Sinica** 12, p.241-262, 2002.

DEMPSTER, A.P, LAIRD, N.M. RDIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society**, v. 39, p. 1-38, 1976.

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

Press, 1995.

GONG, Y.; BECK, J. Items, Skills, and Transfer Models: Which Really Matters for Student Modeling? In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.81-90.

GOWDA,S.; ROWE,J.; BAKER, R.; CHI, M.; KOEDINGER, K. Improving Models of Slipping, Guessing, and Moment-By-Moment Learning with Estimates of Skill Difficulty. In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.199-208

HARTIGAN, J. A. **Clustering Algorithms**. New York: 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.

KARDAN, S.; CONATI, C. A Framework for Capturing Distinguishing User Interaction Behaviors in Novel Interfaces. In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.159-168

KOEDINGER, K.R., BAKER, R.S.J.d., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., Stamper, J. **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, 2010

LEE, Y. Developing an efficient computational method that estimates the ability of students In: a Web-based learning environment. **Computers & Education**. v. 58, Issue 1, p.579-589, 2012.

LI,N.; COHEN, W.; KOEDINGER, K.; MATSUDA, N. A Machine Learning Approach for Automatic Student Model Discovery. In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.31-40.

LÓPEZ, M.I.; LUNA , J.M; VENTURA, C.; ROMERO, S. Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums. In: INT CONF ONE DUCATIONAL DATA MINING , 2012, p.148-151.

MACHADO L. R. **Desenvolvimento de um Algoritmo Imunológico para Agrupamento de Dados**. Caxias do Sul: Universidade de Caxias do Sul, 2011. 123p.

MALDONADO, R.; YACEF, K.; KAY, J.; KHARRUFA, A , AL-QARAGHULI,A. Analysing Frequent Sequential Patterns of Collaborative Learning Activity Around an Interactive Tabletop. Nominee for Best Paper Award. **In:** INT CONF ON EDUCATIONAL DATA MINING , 2011, p.111-120.

MELLO, C. E. R. **Agrupamento de Regiões: Uma abordagem utilizando acessibilidade.** No Estado do RJ. Rio de Janeiro: UFRJ, 2008. 96f. Dissertação (Mestrado em Ciência em Engenharia de Sistema de Computação).

METZ, J. **Interpretação de clusters gerados por algoritmos de clustering hierárquico.** São Paulo: USP, 2006. 152p. Dissertação.

MICHEL, D. Conditions for Effectively Deriving a Q-Matrix from Data with Non-negative Matrix Factorization. Best Paper Award. In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.41-50.

MITCHELL, T. **Machine Learning**. Boston: McGraw Hill, 1997.

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.

RUS, V.; MOLDOVAN, C.; NIRAULA, N.; GRAESSER, A. Automated Discovery of Speech Act Categories in Educational Games. In: INT CONF ON EDUCATIONAL DATA MINING , 2012, p.25-32.

THAI-NGHE,N.; HORVÁTH,T.; SCHMIDT-THIEME,L. Factorization models for forecasting student performance. In: INT CONF ON EDUCATIONAL DATA MINING , 2011, p.11-20.

ZAFRA, A.; VENTURA, S. Multi-instance genetic programming for predicting student performance in web based educational environments. **Applied Soft Computing**, v.12, Issue 8, p.2693–2706, August 2012.