Aplicação de mineração de dados a resultados do teste de competência em leitura de palavras

José Cassiano Grassi Gunji
Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) – São Paulo, SP – Brasil
cassiano.gunji@gmail.com

Maurício Amaral de Almeida
Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) – São Paulo, SP – Brasil
madealmeida@gmail.com

Marcelo Duduchi
Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) – São Paulo, SP – Brasil
mduduchi@terra.com.br

Resumo: O Teste de Competência em Leitura de Palavras é uma ferramenta útil para se medir o nível de desenvolvimento das habilidades de leitura de estudantes em fase de alfabetização. Neste artigo, os resultados obtidos de uma aplicação do TCLP são submetidos a técnicas de Mineração de Dados. Algumas características não evidentes do grupo de teste são identificas, assim como algumas propriedades do TCLP aplicado a esse grupo. Um método complementar de avaliação dos resultados do teste é proposto.

Palavras chave: Mineração de Dados, Inteligência Artificial, Aprendizado de Máquina, Alfabetização, Educação.

1 Introdução

O Teste de Competência em Leitura de Palavras (TCLP) é uma ferramenta útil para se medir o nível de desenvolvimento das habilidades de leitura em estudantes em alfabetização. Ele apresenta resultados significativos quando aplicado a alunos variando do terceiro ano do Ensino Infantil até o oitavo ano do Ensino Fundamental [1]. Apesar disso, o TCLP é mais usualmente empregado na avaliação de alunos do primeiro ao quarto ano do Ensino Fundamental [2]. Com a aplicação deste teste, é possível a medição do nível e do estágio da alfabetização do aluno e como ele se compara com a média dos alunos em seu nível escolar. Este artigo investiga a possibilidade de se obter mais informações da aplicação do TCLP com o auxílio de técnicas de Mineração de Dados (MD), uma especialização da Inteligência Artificial (IA). Para tanto é apropriada uma breve discussão sobre estes dois assuntos de natureza interdisciplinar, o TCLP e a MD.

1.1 O Teste de Competência em Leitura de Palavras

O TCLP avalia o desenvolvimento da leitura ao longo das etapas de aprendizado. Trata-se de um teste de papel e lápis [1] mas possui uma versão eletrônica aplicada via Internet [2]. O teste é composto de 8 itens de treino e 70 itens de teste reunidos num caderno de aplicação. Cada item é composto de uma figura e um elemento escrito. Esse elemento escrito pode ser uma palavra correta ou uma pseudopalavra. Pseudopalavras são seqüências de caracteres que compõe um todo pronunciável, mas que não possui um significado. A tarefa do examinado é circular os itens corretos e cruzar (assinalar com um “X”) os itens incorretos. Há 7 tipos de itens distribuídos

aleatoriamente ao longo do teste, com dez itens de teste para cada tipo. Dois dos tipos são compostos por palavras corretas. Os demais tipos são compostos de pseudopalavras, cada um representando um tipo diferente de erro. Cada erro tem a finalidade de diagnosticar um tipo diferente de falha no processo de aprendizado. A Figura 1 mostra alguns exemplos de itens de teste que compõe o TCLP.

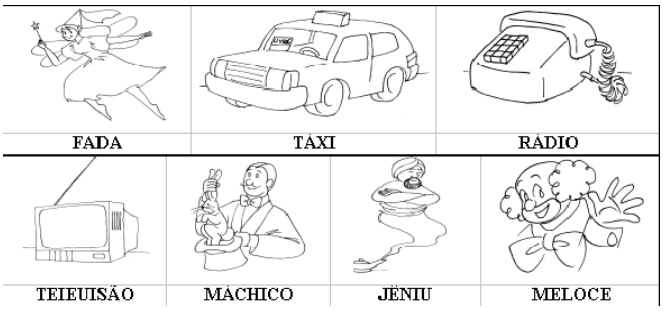

Figura 1: Exemplos de pares figura-palavra que compõe os itens de teste do TCLP (CAPOVILLA, F.; CAPOVILLA, A.; VIGGIANO *et al*, 2005).

Os tipos com itens corretos são: 1) *palavras corretas regulares* (CR), como FADA sob a figura de uma fada; 2) *palavras corretas irregulares* (CI), como TÁXI, sob a figura de um táxi. Os cinco tipos com itens incorretos são: 3) *palavras semanticamente incorretas*, que diferem das figuras às quais estão associadas, ou seja, *vizinhas semânticas* (TS), como RÁDIO, sob a figura de um telefone; 4) *pseudopalavras estranhas* (PE), como MELOCE sob a figura de um palhaço; 5) *pseudopalavras homófonas* (PH), como JÊNIU sob a figura de um gênio; 6) Pseudopalavras pseudo-homógrafas com trocas fonológicas, ou seja, vizinhas fonológicas (TF), como MÁCHICO sob a figura de um mágico; 7) Pseudopalavras pseudo-homógrafas com trocas visuais, ou seja, vizinhas visuais (TV), como TEIEUISÃO, sob a figura de uma televisão (CAPOVILLA, F.; CAPOVILLA, A.; VIGGIANO *et al*, 2005; MACEDO; CAPOVILLA; NIKAEDO *et al*, 2005).

O TCLP é acompanhado de tabelas de normatização para avaliar o grau de desvio entre o padrão de leitura de um examinado e o padrão de leitura normal de seu grupo de referência de acordo com o nível de escolaridade.

1.1.1 As Fases do Processo de Alfabetização

O processo de alfabetização se dá em três estágios [3, 4]. O primeiro é o *logográfico*, em que o aluno trata a palavra escrita como se fosse uma representação pictoideográfica e visual; o segundo é o *alfabético*, em que, com o desenvolvimento da rota fonológica, o aluno aprende a fazer a decodificação grafo-fonêmica; e o *ortográfico*, em que, com o desenvolvimento da rota lexical, o aluno aprende a fazer a leitura visual direta de palavras de alta freqüência. Nota-se que, uma vez que o aluno passa de uma fase à seguinte, as fases anteriores não são abandonadas. Elas apenas ocorrem em menor freqüência e importância. Assim, as estratégias não são mutuamente excludentes, e podem coexistir simultaneamente no leitor e no escritor competentes. Por exemplo, materiais como algarismos matemáticos e sinais de trânsito tendem a ser lidos pela estratégia logográfica. Já palavras novas precisam ser lidas pela estratégia fonológica. Finalmente, palavras conhecidas e familiares, ou de composição morfológica evidente, podem ser lidas mais rapidamente pela estratégia lexical de reconhecimento visual direto.

Capovilla, Varanda e Capovilla [3] explicam que o TCLP foi desenvolvido com o propósito de identificar em qual estágio de alfabetização um aluno se encontra. Assim, o erro de rejeitar pares com palavras corretas regulares (ci) pode indicar dificuldade com o processamento lexical ou falta dele. O erro de deixar de rejeitar pseudopalavras homófonas (ph) também pode indicar dificuldade no processamento lexical, porém em nível mais acentuado, com uso exclusivo da rota fonológica. O erro de deixar de rejeitar pseudopalavras com trocas fonológicas (tf) pode indicar

que a criança está tentando ler exclusivamente pela rota fonológica, ou seja, pela decodificação grafofonêmica estrita, sem fazer uso da rota lexical, mas com o agravante de dificuldades com o processamento fonológico. O erro de deixar de rejeitar palavras semanticamente incorretas (ts) indica que a criança não está fazendo acesso ao léxico semântico. O erro de deixar de rejeitar pseudopalavras com trocas visuais (tv) pode indicar dificuldade com o processamento fonológico e recurso à estratégia de leitura logográfica. Finalmente, o erro de deixar de rejeitar pseudopalavras estranhas (pe) pode indicar sérios problemas de leitura (com ausência de processamento lexical, fonológico e, mesmo, logográfico) ou de atenção.

Capovilla, Joly, Ferracini, *et al* [4] ressaltam que é fundamental conhecer as estratégias de leitura pois, nos distúrbios de leitura, pode haver alterações específicas em uma ou mais dessas estratégias. Por exemplo, dois tipos clássicos de dislexia são a dislexia fonológica e a dislexia morfêmica, também chamada de dislexia de superfície ou semântica. Na dislexia fonológica há dificuldades na leitura pela rota fonológica, porém, a leitura visual-direta pela rota lexical está preservada. Logo, há dificuldades na leitura de pseudopalavras e palavras desconhecidas, mas a leitura de palavras familiares é adequada. Já na dislexia morfêmica há dificuldades na leitura pela rota lexical, sendo a leitura feita principalmente pela rota fonológica. Logo, há dificuldades na leitura de palavras irregulares e longas, com regularizações.

1.2 A Mineração de Dados

Rezende [5, 6] explica que MD é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas da Inteligência Artificial (IA) e que sua finalidade é a "extração de conhecimento previamente desconhecido, implícito e potencialmente útil a partir de dados" [7 *apud* 5, pg. 2]. Em outras palavras: "O foco central de MD é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relação entre dados" [5, pg. 2].

O processo de MD pode ser dividido em três grandes etapas: Pré-processamento, extração de padrões e pós-processamento. Também se pode incluir nessa divisão uma fase anterior ao processo de MD, que se refere ao conhecimento do domínio e a identificação do problema, e uma fase posterior ao processo, que se refere à utilização do conhecimento obtido. A Figura 2 ilustra estas etapas.

Figura 2: Fases do processo de Mineração de Dados [5].

Inicia-se o processo de MD com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados. A seguir, é feita uma seleção de dados a partir dessas fontes, de acordo com os objetivos do processo. Os conjuntos de dados resultantes dessa seleção são pré-processados, ou seja, recebem um tratamento para poderem ser submetidos aos métodos e ferramentas na etapa de extração de padrões.

A etapa de extração de padrões tem como produto usual a obtenção de preditores que, para um dado problema, fornecem uma decisão. Outro produto usual da MD é a obtenção de uma descrição de características intrínsecas nos dados. Essa descrição pode revelar propriedades não evidentes no conjunto de dados ou em subconjuntos dele.

Na etapa seguinte, a de pós-processamento, o conhecimento é avaliado quanto a sua qualidade e/ou utilidade para que, em caso positivo, seja utilizado para apoio a algum processo de tomada de decisão.

> *É importante notar que, por ser um processo eminentemente iterativo, as etapas da Mineração de Dados não são estanques, ou seja, a correlação entre as técnicas e métodos utilizados nas várias etapas é considerável, a ponto da ocorrência de pequenas mudanças em uma delas afetar substancialmente o sucesso de todo o processo. Portanto, os resultados de uma determinada etapa podem acarretar mudanças a quaisquer das etapas posteriores ou, ainda, o recomeço de todo o processo [5, pg. 10].*

A escolha da tarefa de MD é feita de acordo com os objetivos desejáveis para a solução a ser encontrada. As tarefas possíveis de um algoritmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas. As atividades de predição consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em um modelo capaz de reconhecer a classe de um novo exemplo. Os dois principais tipos de tarefas para predição são classificação e regressão. As atividades descritivas consistem na descoberta de comportamentos notáveis e recorrentes no conjunto de dados. Com este resultado, pode-se observar propriedades não evidentes nos dados brutos. Os dois tipos de tarefas para descrição são o agrupamento e as regras de associação.

A escolha do algoritmo é realizada de forma subordinada à linguagem de representação dos padrões a serem encontrados. Pode-se utilizar algoritmos indutores de árvores de decisão ou regras de produção, por exemplo, se o objetivo é realizar uma classificação.

A extração de padrões consiste da aplicação dos algoritmos de mineração escolhidos para a extração dos padrões embutidos nos dados.

O conhecimento extraído da aplicação do algoritmo de MD pode ser utilizado na resolução de problemas na vida real, seja por meio de um Sistema Inteligente ou de um ser humano como apoio a algum processo de decisão.

Rezende [6, 5] ainda observa que há essencialmente dois estilos para se fazer MD: *top-down* e *bottom-up*. No estilo *top-down*, o processo é iniciado com alguma hipótese a ser verificada. Nesse caso, em geral, é desenvolvido um modelo e este é então avaliado para se determinar se a hipótese é válida ou não. No estilo *bottom-up*, não é especificada uma hipótese para validação, apenas são extraídos padrões dos dados. Ainda neste estilo, a abordagem pode ser supervisionada, que é quando se tem alguma idéia do que se está procurando, como também pode ser não-supervisionada, que é quando não se tem idéia do que se está procurando.

2 Material e Método

Neste artigo foi utilizado o resultado da aplicação do TCLP a um conjunto de alunos do 1º e 2º ano do Ensino Fundamental, ambos da mesma escola. Os dados são compostos das respostas de cada aluno a cada questão do teste e estão registrados em uma planilha eletrônica.

A preparação dos dados iniciou-se ainda na própria planilha. O primeiro passo foi a exclusão das respostas aos 8 itens de treino. Usando seus recursos, foi calculada a somatória de acertos em cada categoria de pergunta. As ocorrências de ausência de dados e de dados incoerentes foram removidas. Por fim, a tabela foi formatada no padrão esperado pelo sistema Tanagra [8]. Após a preparação, ficaram disponíveis 85 exemplos úteis na tabela de resultados.

O sistema Tanagra foi escolhido, pois seu uso é livre e gratuito para fins de ensino e pesquisa. Além disso, ele reúne em uma única plataforma um acervo respeitável de recursos de estatística e algoritmos de MD. Por ser um *software* de código aberto, permite que alterações sejam feitas nos algoritmos ou que a plataforma seja utilizada na implementação de novos algoritmos. Outra característica interessante é que os algoritmos implementados neste sistema são bem conhecidos e bastante difundidos nos meios acadêmicos.

O processo de extração de padrões foi feito interativamente. Vários algoritmos foram experimentados para as duas tarefas de MD: predição e descrição. A abordagem foi sempre *bottom-up*, tanto na modalidade supervisionada quanto na não-supervisionada. Os resultados mais significativos estão expostos no tópico a seguir.

3 Resultados

Neste tópico é utilizada a nomenclatura dos tipos de pares figura-palavra do TCLP em sua forma abreviada, a qual está registrada na Tabela 1.

Tipo	Abreviação
Correta Regular	cr
Correta Irregular	ci
Vizinha Semântica	ts
Vizinha Visual	tv
Vizinha Fonológica	tf
Pseudopalavra Homófona	ph
Pseudopalavra Estranha	pe

Tabela 1: Abreviações dos tipos de pares figura-palavra utilizadas nos resultados. Fonte: o autor.

3.1 Tarefa de Classificação

A primeira tarefa executada foi a classificação dos dados. Foi empregado o algoritmo C4.5 utilizando-se as colunas dos totais de acertos em cada categoria de pergunta. A série de cada aluno foi utilizada como o objetivo da classificação. Vale notar que o resultado obtido com a interpretação canônica dos resultados do TCLP é a série de cada sujeito do teste. A matriz de confusão obtida está representada na Tabela 2 e a árvore de decisão no Quadro 1.

	primeira	segunda	Sum
primeira	15	17	32
segunda	9	44	53
Sum	**24**	**61**	**85**

Tabela 2: Matriz de confusão da primeira aplicação do algoritmo C4.5, usando validação cruzada de cinco vias. Fonte: o autor.

1. Se (Total ci < 9) e (Total ph < 6) então SERIE = **primeira** (81,25% de 16 exemplos)
2. Se (Total ci < 9) e (Total ph >= 6) e (Total cr < 10) então SERIE = **primeira** (70% de 10 exemplos)
3. Se (Total ci < 9) e (Total ph >= 6) e (Total cr >= 10) então SERIE = **segunda** (71,43% de 7 exemplos)
4. Se (Total ci >= 9) então SERIE = **segunda** (80,55% de 52 exemplos)

Quadro 1: Árvore de decisão da primeira aplicação do algoritmo C4.5. Fonte: o autor.

A matriz de confusão reproduzida na Tabela 2 é obtida aplicando-se a técnica de validação cruzada (*cross-validation*) de cinco vias. Para tanto, divide-se os dados em cinco partes. Quatro delas são usadas para se criar a árvore de decisão e a quinta é usada para se testar a árvore obtida. O processo é repetido para todas as cinco partes. Nas linhas da tabela encontram-se os alunos que estavam efetivamente em cada série, e nas colunas as séries que a árvore de decisão atribui a cada aluno. Na diagonal principal da matriz encontram-se os casos em que a árvore classifica corretamente os alunos [8, 6].

A primeira informação que se pode observar na árvore de decisão registrada no Quadro 1 é a de que errar apenas uma das questões do tipo Correta Irregular é característica dos alunos da segunda série (ver linha 4). Na linha 3 observa-se que caso o aluno tenha errado pelo menos duas questões do tipo Correta Irregular, mas tenha acertado todas as questões do tipo Correta Regular e pelo menos 6 questões do tipo Pseudopalavra Homófona, também deve ser classificado como sendo da segunda série. O restante da amostra, o algoritmo classifica como sendo da primeira série.

O mesmo algoritmo de aprendizado de máquina foi aplicado novamente. Mas dessa vez, além dos totais de acertos em cada tipo de pergunta, foi fornecido ao algoritmo os acertos às questões individuais. A Matriz de Confusão resultante está reproduzida na Tabela 3 e a árvore de decisão no Quadro 2. A Matriz de Confusão é obtida novamente pela aplicação de validação cruzada de cinco vias.

	primeira	segunda	Sum
primeira	13	19	32
segunda	11	42	53
Sum	**24**	**61**	**85**

Tabela 3: Matriz de Confusão da segunda aplicação do algoritmo C4.5, usando validação cruzada de cinco vias. Fonte: o autor.

1. Se (Resul70 ci = 0) então SERIE = **primeira** (81,25% de 16 exemplos)
2. Se (Resul70 ci = 1) e (Resul53 tf = 0) então SERIE = **primeira** (71.43% de 7 exemplos)
3. Se (Resul70 ci = 1) e (Resul53 tf = 1) e (Total ph < 8) e (Total ci < 9) então SERIE = **primeira** (69,23% de 13 exemplos)
4. Se (Resul70 ci = 1) e (Resul53 tf = 1) e (Total ph < 8) e (Total ci >= 9) então SERIE = **segunda** (81,48% de 27 exemplos)
5. Se (Resul70 ci = 1) e (Resul53 tf = 1) e (Total ph >= 8) então SERIE = **segunda** (81,48% de 27 exemplos)

Quadro 2: Árvore de decisão da segunda aplicação do algoritmo C4.5. Fonte: o autor.

Um fato notável com o exame da árvore de decisão representada na Tabela 3 é o de que as respostas a duas questões específicas do teste já são capazes de classificar alguns alunos. A linha 1 indica que quem errou a questão 70, que é do tipo Correta Irregular, pertence à primeira série. Caso tenha acertado a questão 70, mas errado a questão 53, que é do tipo Vizinha Fonológica, também pertence à primeira série. Vale a pena investigar as causas reais para este comportamento. Estas questões podem ser realmente eficientes para a classificação, mas podem também representar conhecimento oferecido apenas aos alunos da segunda série. Também podem envolver conceitos estranhos ao grupo social ao qual as crianças pertençam. Também é possível que este comportamento seja simplesmente uma aberração estatística que esteja recebendo destaque indevido pelo algoritmo.

Continuando a análise da mesma árvore de decisão, e notando-se que deste ponto em diante, o algoritmo admite que todos os alunos acertaram as questões 53 e 70, passa-se a classificar alunos da segunda série. Estes acertaram pelo menos 8 questões do tipo Pseudopalavra Homófona (linha 5). Caso contrário, acertaram pelo menos 9 questões do tipo Correta Irregular (linha 4). Os demais alunos são classificados como sendo da primeira série.

Um fato que deve ser notado neste ponto é que o algoritmo C4.5 evidencia que pequenas porções do TCLP já são suficientes para classificar os alunos desta amostra. Não é necessário usar todos os dados disponíveis.

3.2 Tarefa de Agrupamento

Algoritmos de agrupamento separam os dados em grupos com características semelhantes. Entretanto, muitas vezes não ficam óbvias quais são as características que os algoritmos usam para fazer os agrupamentos. Descobrir se os agrupamentos fazem sentido e por

quê é uma tarefa da fase de pós-processamento. A seguir estão redigidos os resultados encontrados.

Foram experimentados vários algoritmos com diversas configurações. A combinação que gerou resultados mais facilmente identificáveis e interessantes foi o algoritmo K-Means separando os dados em 4 agrupamentos.

Agrupamento	Descrição	Tamanho
Agrupamento n°1	c_kmeans_1	3
Agrupamento n°2	c_kmeans_2	22
Agrupamento n°3	c_kmeans_3	1
Agrupamento n°4	c_kmeans_4	59

Tabela 4: Característica dos agrupamentos obtidos com o algoritmo K-Means. Fonte: o autor

Observando-se a Tabela 4, que relaciona os agrupamentos encontrados e o número de membros em cada um, nota-se que há dois agrupamentos principais, o n°2 e o n°4. Os demais agrupamentos podem ser exceções, casos especiais ou aberrações estatísticas. Isto será discutido mais à frente. O primeiro passo foi visualizar graficamente os agrupamentos em busca de algum comportamento que os diferenciem uns dos outros.

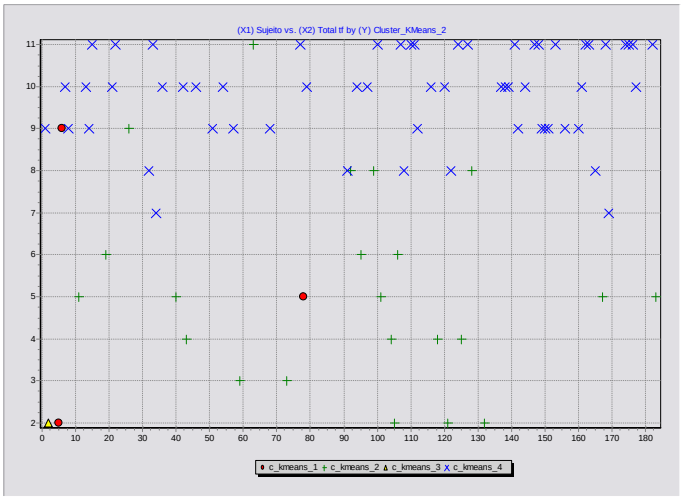

Figura 3: Representação dos agrupamentos em função da pontuação em Vizinha Fonológica. Fonte: o autor.

Os agrupamentos c_kmeans_2 e c_kmeans_4 apresentam bom desempenho em quase todos os testes. As diferenças surgem nos testes do tipo Vizinha Fonológica (tf) e Vizinha Visual (tv), nos quais o agrupamento c_kmeans_2 apresenta um desempenho inferior. O comportamento dos agrupamentos em relação aos testes do tipo Vizinha Fonológica estão representados na Figura 3. O comportamento em relação aos testes do tipo Vizinha Visual é semelhante e não foi representado neste artigo. Nota-se que o eixo da abscissa deste diagrama é o índice dos sujeitos do teste. Dos índices 1 até 79 foram relacionados alunos da primeira série. Dos índices 91 até 183 foram relacionados alunos da segunda série. Assim, fica evidente que os agrupamentos c_kmeans_2 e c_kmeans_4 envolvem alunos das duas séries indistintamente. O erro de deixar de rejeitar pseudopalavras com Vizinhas Fonológicas (tf) pode indicar que a criança está tentando ler exclusivamente pela rota fonológica, ou seja, pela decodificação grafofonêmica estrita, sem fazer uso da rota lexical (etapa ortográfica), mas com o agravante de dificuldades com o processamento fonológico. Esta conclusão é reforçada pela constatação que este agrupamento também errou ao deixar de rejeitar pseudopalavras com Vizinhas Visuais (tv), o que pode indicar dificuldades com o processamento fonológico e recurso à estratégia de leitura logográfica [3].

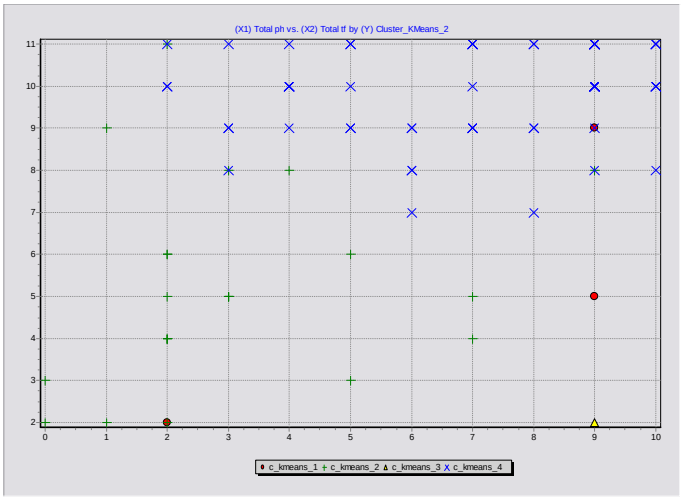

Figura 4: Representação dos agrupamentos em função da pontuação em Pseudopalavra Homófona e Vizinha Fonológica. Fonte: o autor.

Observando-se a distribuição dos agrupamentos com relação aos testes do tipo Pseudopalavra Homófona, do mesmo modo que foi feito até agora, nota-se que o agrupamento c_kmeans_2 também apresenta um desempenho ruim. Entretanto, o agrupamento c_kmeans_4, que vinha apresentando desempenho excelente em todos os testes, passa a apresentar um desempenho mais distribuído. Mas ao se observar a distribuição dos agrupamentos com relação aos testes de Pseudopalavra Homófona (ph) e Vizinha Fonológica (tf) simultaneamente, como mostrado na Figura 4, nota-se que existe uma relação entre os dois: O agrupamento c_kmeans_4 obteve boas notas em ph, em tf ou em ambos ao mesmo tempo. O erro de se deixar de rejeitar pseudopalavras homófonas pode indicar dificuldade no processamento lexical, a última etapa do processo de alfabetização [3]. Já o agrupamento c_kmeans_2, por outro lado, não foi bem em ph, em tf ou em ambos. Como já foi concluído no parágrafo anterior que este agrupamento apresenta dificuldades com a rota alfabética, é razoável esperar que tenha dificuldades também com a rota léxica, pois esta só se desenvolve depois da primeira.

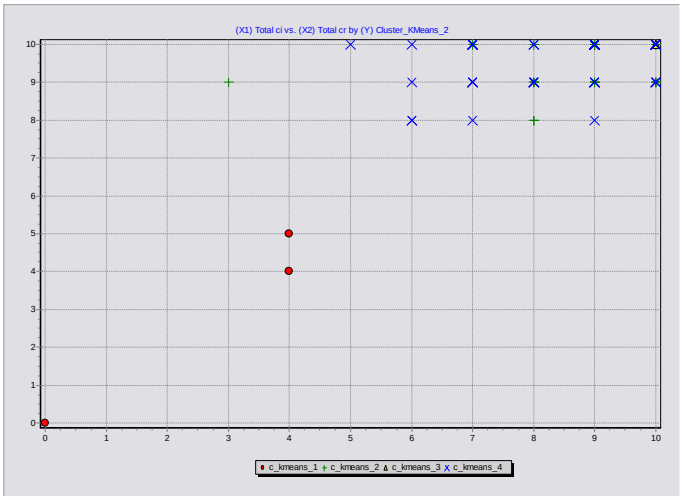

Figura 5: Representação dos agrupamentos em função da pontuação em Correta Regular e Correta Irregular. Fonte: o autor.

A mesma abordagem foi dada à interpretação dos agrupamentos com poucos membros, ou seja, observou-se sua distribuição em relação a cada tipo de teste em relação aos demais agrupamentos. Foi observado que uma importante característica do agrupamento c_kmeans_1 é ter obtido as piores notas nos testes do tipo Correta Regular e Correta Irregular. Na Figura 5, que

representa a distribuição dos agrupamentos em relação aos totais em cr e ci, observa-se que o agrupamento c_kmeans_1 reúne todos os indivíduos que obtiveram notas ruins em ambos os testes. Este agrupamento é composto pelos indivíduos 5, 6 e 78.

Sujeito	Total cr	Total ci	Total ts	Total tf	Total tv	Total ph	Total pe
2	10	10	0	2	1	9	0

Tabela 5: Notas do sujeito 2, único integrante do agrupamento c_kmeans_3. Fonte: o autor.

A característica mais evidente do agrupamento c_kmeans_3 é o fato deste ser composto por um único membro. Suas notas estão relacionadas na Tabela 5. Este aluno obteve notas máximas nos testes do tipo Correta Regular (cr) e Correta Irregular (ci), mas errou todos os testes do tipo Vizinha Semântica (ts). Mais do que isso, é o único sujeito a errar todos os testes deste tipo. Este erro indica que a criança não está fazendo acesso ao léxico semântico [3]. Mas um comportamento assim extremo também pode sugerir uma dificuldade exagerada na transição entre as etapas alfabética e ortográfica, sintoma característico da dislexia morfêmica [4, 9]. Esta conclusão é reforçada pelo fato deste aluno também ter mostrado desempenho ruim nas questões do tipo Vizinha Fonológica (tf), indicando que faz uso apenas da rota alfabética. Também é possível que este aluno não tenha entendido como fazer o teste e esteja ignorando as figuras, apenas assinalando as palavras escritas de maneira correta. Aparentemente, o sujeito 2 domina a etapa alfabética, entretanto, seu desempenho ruim em testes do tipo Vizinha Visual (tv) contradiz esta suposição, indicando que ele apresenta dificuldades com a rota alfabética e esteja usando a rota logográfica. Seu bom desempenho em Pseudopalavras Homófonas (ph) sugere que já domina a rota léxica (a última), mas seu mau desempenho em Pseudopalavras Estranhas (pe) indica sérios problemas de leitura, com ausência de processamento lexical, fonológico e mesmo logográfico. Conclusões tão contraditórias indicam algum problema sério, como distúrbios de atenção e dislexia entre muitas outras dificuldades no aprendizado. Também podem indicar falha no processo de aplicação do teste e na correção e transcrição dos resultados. Em todos os casos, o comportamento do sujeito 2 deve ser investigado mais profundamente.

3.2.1 Classificação dos Agrupamentos Encontrados

O algoritmo C4.5 foi aplicado novamente, mas desta vez, classificando os sujeitos entre os 4 agrupamentos encontrados no item anterior. A árvore de decisão obtida está registrada no Quadro 3 e sua matriz de confusão na Tabela 6.

	c_kmeans_1	c_kmeans_2	c_kmeans_3	c_kmeans_4	Sum
c_kmeans_1	6	0	0	1	7
c_kmeans_2	1	11	0	2	14
c_kmeans_3	1	0	14	1	16
c_kmeans_4	1	3	1	43	48
Sum	**9**	**14**	**15**	**47**	**85**

Tabela 6: Matriz de confusão do algoritmo C4.5 aplicado aos agrupamentos obtidos pela aplicação do algoritmo K-Means, usando validação cruzada de cinco vias. Fonte: o autor.

1. Se (Total tf < 7) então agrupamento = **c_kmeans_2** (85% de 20 exemplos)
2. Se (7 <= Total tf < 9) e (Total cr < 10) então agrupamento = **c_kmeans_2** (60% de 5 exemplos)
3. Se (7 <= Total tf < 9) e (Total cr = 10) então agrupamento = **c_kmeans_4** (100% de 5 exemplos)
4. Se (Total tf >= 9) então agrupamento = **c_kmeans_4** (94,55% de 55 exemplos)

Quadro 3: Árvore de decisão do algoritmo C4.5 aplicado aos agrupamentos obtidos pela aplicação do algoritmo K-Means. Fonte: o autor.

Analisando-se a árvore de decisão representada na Tabela 6, nota-se que um aluno pode ser classificado como pertencendo ao agrupamento c_kmeans_4 se tiver acertado pelo menos 9 das questões do tipo Vizinha Fonológica (tf). Esta informação pode ser observada na linha 4. Também pode ser classificado neste mesmo grupo se tiver acertos em Vizinha Fonológica entre 7

e 9, mas que tenha ao mesmo tempo acertado todas as questões do tipo Correta Regular (cr), conclui-se observando a linha 3. Levando-se em conta a linha 1, classifica-se um aluno como pertencendo ao agrupamento c_kmeans_2 caso tenha acertado menos de 7 questões do tipo Vizinha Fonológica (tf). Os demais (linha 2), também são do mesmo agrupamento.

4 Conclusões

O TCLP é aplicado com o intuito de se verificar o desempenho de um aluno frente ao desempenho de outros alunos em seu mesmo nível de escolaridade. Esta comparação é feita com o uso de tabelas normatizadas, mediante um tratamento estatístico. A primeira aplicação do algoritmo de classificação forneceu uma alternativa a este tratamento e evidenciou características de desempenho de certos alunos que, por si só, são capazes de classificá-los, sem a necessidade de se examinar seu desempenho no teste inteiro. Isto foi evidenciado com as aplicações do algoritmo de classificação C4.5. As árvores de decisão obtidas nestas aplicações devem ser testadas com um novo conjunto de resultados deste teste, aplicado a um novo grupo de alunos. Deste modo, sua acuidade e utilidade podem ser estabelecidas. Caso sejam satisfatórias, este novo método de avaliação do teste pode ser adotado como ferramenta auxiliar na interpretação dos resultados do TCLP.

A seguir, a aplicação do algoritmo de agrupamento separou os alunos em dois grandes grupos: O grupo dos alunos que estavam no estágio alfabético e o grupo que estava no estágio ortográfico. Esta conclusão ainda deve ser confirmada pelo exame de um especialista no domínio, sendo recomendados estudos mais aprofundados. Entretanto, este artigo já foi capaz de identificar o comportamento homogêneo dos alunos dentro de seus respectivos agrupamentos. A nova aplicação do algoritmo de classificação resultou num modelo que rapidamente prediz qual a classe a que pertence um novo indivíduo. Com esta ferramenta, problemas de aprendizado podem ser rapidamente detectados e remediados. Isto fica evidente quando o algoritmo de agrupamento destacou indivíduos com desempenho incomum e interessante. Encontrar estes indivíduos manualmente nos dados brutos demanda trabalho árduo e profundo conhecimento do domínio.

Finalmente, houve indicações de que o uso de MD no tratamento dos resultados do TCLP expõe informações que não são evidentes nos dados e que não são originalmente tratadas pelo teste. A interpretação sugerida pelos autores deste artigo é a de que estas informações referem-se principalmente ao estágio de alfabetização dos alunos, mas admite a possibilidade de que estas informações estejam expondo outras características interessantes dos alunos, como a ocorrência de dislexia. De qualquer modo, fica claro que o uso de técnicas de Mineração de Dados pode ampliar e enriquecer as informações obtidas com a aplicação do Teste de Competência em Leitura de Palavras.

5 Referências

[1] CAPOVILLA, F.; CAPOVILLA, A. G. S.; VIGGIANO, K. *et al*., **Silent reading by deaf and hearing readers:** logographic, alphabetical and lexical processes. *Estud. psicol. (Natal)*, Jan./Apr. 2005, vol.10, no.1, p.15-23. Disponível na Internet: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-294X2005000100003&lng=en&nrm=iso>. Acesso em: 21 abr. 2008. ISSN 1413-294X.

[2] MACEDO, E. C. de; CAPOVILLA, F. C.; NIKAEDO, C. C. *et al*. **Teleavaliação da habilidade de leitura no ensino infantil fundamental**. *Psicol. esc. educ.* Jun. 2005, vol.9, no.1, p.37-46. Disponível na Internet: <http://pepsic.bvs-

psi.org.br/scielo.php?script=sci_arttext&pid=S1413-85572005000100012&lng=pt&nrm=iso>. ISSN 1413-8557.

[3] CAPOVILLA, F. C., VARANDA, C. e CAPOVILLA, A. G. S. **Teste de competência de leitura de palavras e pseudopalavras**: normatização e validação. *Psic*. dez. 2006, vol.7, no.2, p.47-59. Disponível na Internet: <http://pepsic.bvs-psi.org.br/scielo.php?script=sci_arttext&pid=S1676-73142006000200007&lng=pt&nrm=iso>. Acesso em: 21 abr. 2008. ISSN 1676-7314.

[4] CAPOVILLA, A. G. S.; JOLY, M. C. R. A.; FERRACINI, F., *et al.,* **Estratégias de leitura e desempenho em escrita no início da alfabetização**: estratégias de leitura e alfabetização. In: *Psicologia Escolar e Educacional*. dez. 2004, vol. 8, no.2, p.189-197. Disponível na Internet: <http://pepsic.bvs-psi.org.br/scielo.php?script=sci_arttext&pid=S1413-85572004000200007&lng=pt&nrm=iso>. Acesso em: 21 abr. 2008. ISSN 1413-8557.

[5] REZENDE, S. O., **Mineração de dados**. In: *Encontro Nacional de Inteligência Artificial*, 5., São Leopoldo – RS, 2005. Disponível na Internet: <http://www.addlabs.uff.br/enia_site/dw/mineracaodedados.pdf>. Acesso em: 21 abr. 2008.

[6] REZENDE, S. O., **Sistemas Inteligentes**: fundamentos e aplicações. Barueri, SP: Manole, 2003. ISBN 85-204-1683-7.

[7] WITTEN, I. H.; FRANK E. **Data mining**: pratical machine learning tools and techniques with Java implementations. Morgan Kauffmann Publishers: 1999 *apud* REZENDE, S. O., **Mineração de dados**. In: *Encontro Nacional de Inteligência Artificial*, 5., São Leopoldo – RS, 2005. Disponível na Internet: <http://www.addlabs.uff.br/enia_site/dw/mineracaodedados.pdf>. Acesso em: 21 abr. 2008.

[8] RAKOTOMALA, R., **Tanagra** : un logiciel gratuit pour l'enseignement et la recherche, in *Actes de EGC* '2005, RNTI-E-3, vol. 2, p.697-702, 2005.

[9] **DISLEXIA**. in *Wikipedia*. Disponível em <http://pt.wikipedia.org/wiki/Dislexia>. Acesso em: 29 nov. 2007.