# Capítulo 1

# Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações

Evandro Costa, Ryan S.J.d. Baker, Lucas Amorim, Jonathas Magalhães, Tarsis Marinho

***Abstract***

*With the increasing use of Interactive Learning Environments (ILEs) or even Learning Management Systems (LMSs) on the Web to support student learning, an increasingly massive volume of data is being generated by students and instructors participating in a range of interactions. This creates an opportunity for researching online learning, but these data are still only being exploited to a limited degree, compared to the quantity of findings that could be obtained from these data. For example, in order to understand student behaviors and the ways in which students learn, researchers in the emerging field of Educational Data Mining (EDM) have worked to develop data mining methods that are relevant for these types of data sets and research questions. The results of EDM analyses have proven useful for improving educational practices and the design of curriculum, both for classroom and distance use of educational technologies. They have also proven useful for supporting adaptive personalization in online education. In this chapter, the principal concepts and techniques of EDM will be discussed, with examples from relevant educational data sets and with reference to common applications of EDM methods. Such methods will be discussed to address the following data mining tasks: classification and regression, clustering, and association rule mining. Methods specially developed for predicting student knowledge, emotion, and preparation for future learning, will also be discussed.*

***Resumo***

*Com o crescente uso de Ambientes Virtuais de Aprendizagem (AVA) na Web e outras tecnologias para apoio ao processo de ensino e aprendizagem, um grande volume de dados tem sido gerado a partir das diferentes modalidades de interação no sistema envolvendo principalmente estudantes e professores. Entretanto, boa parte desses dados não têm sido*

*analisados, o que se constitui numa lacuna importante, dada a quantidade de informação valiosa que se pode potencialmente extrair de tais dados. Por exemplo, em busca de melhor compreender o comportamento dos estudantes e a forma como eles aprendem, o trabalho realizado por pesquisadores em Mineração de Dados Educacionais tem investido no uso e na melhoria de conhecidas técnicas de mineração de dados para obter conhecimentos relevantes a partir desses dados. Tais conhecimentos podem servir de subsídio para a melhoria das práticas em educação a distância ou presencial, além de ser uma importante ferramenta para viabilizar a personalização do ensino. O objetivo deste capítulo é introduzir os principais conceitos, técnicas e algoritmos de mineração de dados com aplicações a conjuntos de dados educacionais. Neste sentido, serão discutidos métodos sobre as seguintes tarefas: Classificação e regressão, agrupamento de dados, mineração de regras de associação. Particularmente, serão discutidos métodos especialmente desenvolvidos para predição dos conhecimentos, emoções, e aprendizagem futura do estudante.*

## 1.1. Introdução

### 1.1.1. Motivações

Diariamente, motivada principalmente pelas novas tecnologias de coleta e armazenamento de dados e pelo advento da Web, uma vasta quantidade de dados é produzida para os mais diversos setores, a exemplo de Saúde, Educação, Negócios. Portanto, parte ponderável desses dados necessita urgentemente ser analisada.

Em particular, verifica-se que muitas instituições educacionais que fazem uso de Ambientes Virtuais de Aprendizagem (AVAs), ou outras tecnologias de apoio ao processo de ensino e aprendizagem dos estudantes, estão produzindo um grande volume de dados. Assim, considerando a existência de recursos computacionais suficientes para tais instituições, surge então um desafio macro que é o de como explorar adequadamente esses dados, visando obter informação valiosa, considerando principalmente requisitos de qualidade de consistência e corretude, de rápido tempo de obtenção e o seu caráter oportuno. Por exemplo, isso remete a questões importantes, tais como: o que especificamente fazer com esses dados educacionais? Como reverter estas informações extraídas em benefícios para as instituições, para estudantes e professores envolvidos no contexto de um curso? Como cada um destes atores podem identificar e utilizar as informações escondidas nos dados coletados de tal modo que consigam tirar proveitos delas?

Um cenário particular e importante do que foi descrito acima é o da Universidade Aberta do Brasil[1], na qual se geram grandes volumes de dados, através do uso de AVAs. Portanto, trata-se de um exemplo de instituição que se beneficiaria (ou talvez já se beneficia) muito da utilização apropriada de análise de dados, ou mais especialmente de mineração de dados educacionais.

### 1.1.2. Mineração de Dados e Descoberta de Conhecimento

A Mineração de Dados (MD, do inglês, Data Mining, DM), pode vista como uma etapa principal de um processo mais amplo conhecido como descoberta de conhecimento em

[1] http://www.uab.capes.gov.br/

bases de dados (identificado numa área que em inglês se denomina Knowledge Discovery in Databases, KDD). Em KDD verifica-se ainda a inclusão de mais duas grandes etapas: pré-processamento de dados (preparação de dados, abrangendo mecanismos para captação, organização e tratamento dos dados) e pós-processamento dos resultados obtidos na mineração de dados. Neste sentido, de acordo com Fayyad et al. [Fayyad et al. 1996], "KDD é o processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis". Trata-se, portanto, de uma definição abrangente, na qual KDD é descrito como um processo geral de descoberta de conhecimento composto pelas três grandes etapas mencionadas. Os padrões mencionados devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para uma possível tomada de decisão.

No entanto, há uma falta de consenso entre os autores sobre uma definição para o termo Mineração de Dados, dificultando a consolidação de uma definição única. Há inclusive autores que consideram Data Mining como sinônimo de KDD, referindo-se a ambas como uma disciplina que objetiva a extração automática de padrões interessantes e implícitos de grandes coleções de dados [Klösgen and Zytkow 2002]. Doravante neste texto, por simplicidade, não distinguiremos mais estes dois termos.

Mineração de Dados é uma área interdisciplinar, mobilizando principalmente conhecimentos de análise estatística de dados, aprendizagem de máquina, reconhecimento de padrões e visualização de dados [Cabena et al. 1998].

Para descobrir conhecimento que seja relevante, é importante estabelecer metas bem definidas. Segundo Fayyad et al. [Fayyad et al. 1996], no processo de descoberta de conhecimento as metas são definidas em função dos objetivos na utilização do sistema, podendo ser de dois tipos básicos: verificação ou descoberta. Quando a meta é do tipo verificação, o sistema está limitado a verificar hipóteses definidas pelo usuário, enquanto que na descoberta o sistema encontra novos padrões de forma autônoma. A meta do tipo descoberta, em geral, está relacionada com as seguintes tarefas de mineração de dados: predição e descrição, sendo o foco do presente texto.

Tarefas Preditivas objetivam predizer o valor de um determinado atributo (variável) baseado nos valores de outros atributos. O atributo a ser predito é comumente conhecido como a variável preditiva, dependente ou alvo, enquanto que os atributos usados para fazer a predição são conhecidos com as variáveis preditoras, independentes ou explicativas. De modo mais abstrato, a predição se utiliza de uma tupla de variáveis para predizer outras variáveis ou valores desconhecidos [Fayyad et al. 1996].

Tarefas Descritivas procuram encontrar padrões (correlações, tendências, grupos, trajetórias e anomalias) que descrevam os dados.

As metas de predição e descrição são alcançadas abordando alguma das seguintes tarefas e métodos de mineração de dados: classificação, regressão, agrupamento, sumarização, modelagem de dependência e identificação de mudanças e desvios.

### 1.1.3. Mineração de Dados Educacionais

A área emergente de Mineração de Dados Educacionais procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como AVAs, Sistemas Tutores Inteligentes (STIs), entre outros. Com tais métodos visa-se, por exemplo, entender melhor o estudante no seu processo de aprendizagem, analisando-se sua interação com o ambiente. Assim, há a necessidade, por exemplo, de adequação dos algoritmos de mineração de dados existentes para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados. Por outro lado, há uma necessidade significativa e urgente no provimento de ambientes computacionais apropriados para mineração de dados educacionais, oferecendo facilidades de uso para cada um dos atores envolvidos, notadamente ao professor.

#### 1.1.3.1. Origens de EDM

Apesar de algumas iniciativas primeiras com workshops específicos dentro das conferências sobre Artificial Intelligence in Education (AIEd) e sobre Intelligent Tutoring Systems (ITS), foi somente em 2005, em Pittsburgh, EUA, que foi organizado o primeiro Workshop on Educational Data Mining, como parte do 20th National Conference on Artificial Intelligence (AAAI 2005). Daí em diante, houve mais algumas realizações deste workshop entre 2006 e 2007. Seguindo-se, em 2008 lança-se, em Montreal, Canadá, a primeira conferência em EDM: First International Conference on Educational Data Mining, evento este que se estabeleceu e ganhou regularidade de realização anual, estando agora em 2012 na sua quinta edição. Em 2009, esta sociedade investiu na criação de um periódico e publicou o seu primeiro volume do JEDM - Journal of Educational Data Mining. Em 2011 constituiu-se a sociedade científica para EDM (International Educational Data Mining Society [2]). Enfim, a área de EDM está bem consolidada internacionalmente, mas, ainda dando os seus primeiros passos no Brasil, ficando a produção por conta de algumas poucas iniciativas de pesquisas isoladas.

#### 1.1.3.2. Tarefas para EDM

Há diversas tarefas envolvidas em EDM, notadamente as que decorrem diretamente da análise de dados gerados nas interações dos estudantes com os ambientes de aprendizagem. Dessa análise surgem demandas para responder questões relacionadas a como melhorar a aprendizagem do estudante, como desenvolver ambientes educacionais mais eficazes que contribuam efetivamente para os estudantes aprenderem mais e em menos tempo?

Em outra perspectiva, pretende-se saber quais métodos de mineração de se adequam às necessidades presentes na área de EDM? Quais ajustes devem ser feitos nas técnicas de forma a suprir a necessidade de EDM? Do ponto de vista computacional,

[2] Ver detalhes em `http://www.educationaldatamining.org/`

alguns desafios práticos que se apresentam em vários contextos educacionais estão relacionados, por exemplo, a falta de padronização dos dados, o que acaba exigindo grande esforço de pré-processamento [Baker 2011]. Além disso, há a necessidade de adequação dos algoritmos clássicos de mineração de dados para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados [Baker 2010a].

Entre as tarefas e métodos de mineração de dados educacionais a serem discutidos no presente texto, incluem-se: classificação e regressão, agrupamento de dados, mineração de regras de associação. Outras abordagens, entretanto, serão comentadas.

A tarefa de classificação diz respeito ao processo de encontrar um modelo que descreve e distingue classes de dados ou conceitos. Os modelos são derivados com base nas análises de coleções de dados, denominadas conjuntos de treinamentos, os quais correspondem a objetos de dados para os quais os rótulos de classes são conhecidos. O modelo é usado para predizer o rótulo de classe de objetos para os quais o rótulo de classe é desconhecido. Ele associa um item de dado a uma ou várias classes predefinidas. Os modelos derivados podem ser representados em várias formas, tais como: árvore de decisão, regras de classificação, funções matemáticas, redes neurais [Han and Kamber 2000].

Enquanto na classificação a predição é feita para um atributo classificador que assume valores discretos, em modelos de regressão a variável alvo é continua, ou seja, associa um item de dado a uma ou mais variáveis de predição de valores reais. Por sua vez, a análise de agrupamento de dados procura associar um item de dado com um ou vários agrupamentos determinados pelos dados, valendo-se principalmente de medidas de similaridades. Já a abordagem de mineração de regras de associação busca encontrar possíveis relações interessantes entre atributos de uma base de dados. Estas abordagens serão discutidas em detalhes na próxima seção deste capítulo, onde se discutem outras abordagens.

#### 1.1.3.3. Estruturação do Texto

No restante deste texto introduz-se conceitos, métodos e ferramentas utilizados em mineração de dados educacionais, além de uma discussão algumas aplicações em EDM. Na seção seguinte são apresentados os principais métodos utilizados em EDM que têm sido empregados em aplicações reais. Prossegue-se na Seção 1.3 com uma explanação sobre aquisição e preparação de dados. Na Seção 1.4 estão elencadas algumas das principais aplicações de EDM. Na Seção 1.5 estão descritas algumas ferramentas que têm sido utilizadas em EDM. A Seção 1.6 apresenta algumas considerações sobre o capítulo.

## 1.2. Tarefas e Algoritmos de Mineração de Dados Educacionais

O objetivo desta seção é apresentar as principais técnicas utilizadas na mineração de dados educacionais, focalizando tarefas e algoritmos envolvidos, dando ao leitor facilidades para entender as técnicas, percebendo o que cada uma faz e em quais situações são utilizadas.

Em sua grande parte, as técnicas utilizadas na área de EDM são provindas da área de mineração de dados [Baker 2011]. Entretanto, na maioria das vezes há a necessidade

de adaptá-las devido às particularidades existentes em ambientes educacionais e seus dados.

As técnicas estão apresentadas conforme sua categorização nas sub-áreas de EDM, seguindo-se o que consta na taxonomia proposta por Baker et al. [Baker 2011], tal como segue:

- Predição
  - Classificação
  - Regressão
- Agrupamento
- Mineração de Relações
  - Mineração de Regras de Associação
  - Mineração de Correlações
  - Mineração de Padrões Sequenciais
  - Mineração de Causas
- Destilação de dados para facilitar decisões humanas
- Descobertas com modelos

Dos métodos destacados na taxonomia acima, alguns dos mais demandados estão descritos em detalhes, quais sejam: Predição, Agrupamento e Mineração de Relações (apenas Regras de Associação) e os demais estão sucintamente discutidos.

### 1.2.1. Predição

Na tarefa de predição, a meta é desenvolver modelos que façam inferência sobre aspectos específicos dos dados (variáveis preditivas) por meio da análise e associação dos diversos aspectos encontrados nos dados (variáveis preditoras). Um modelo preditivo pode ser entendido como uma função $f(X,\beta) \approx Y$, onde $X$ é um conjunto de variáveis preditoras, $\beta$ são parâmetros desconhecidos e $Y$ é a variável preditiva $Y$. Em outras palavras, deseja-se estimar o valor de $Y$ por meio da descoberta de $\beta$ utilizando-se $X$. No processo de predição, é fundamental que boa parte dos dados sejam rotulados manualmente, ou seja, a aprendizagem do modelo ocorrerá de forma supervisionada e dar-se-á utilizando um conjunto de treinamento com valores previamente conhecidos de $Y$.

Segundo Baker et al. [Baker 2011], há dois benefícios relacionados à utilização da predição em EDM. Primeiro, os métodos de predição podem ser utilizados para estudar quais aspectos de um modelo são importantes para predição. Esta estratégia é frequentemente utilizada em pesquisas que tentam, de forma direta, predizer os benefícios educacionais de determinadas técnicas e ferramentas para um conjunto de estudantes, isso sem considerar os fatores intermediários, como apresentado em [Romero et al. 2008]. Segundo, os métodos de predição auxiliam a predizer o valor das variáveis utilizadas em um

modelo. O intuito de utilizar essa abordagem é verificar quais dados são mais importantes para o modelo pois analisar todos os dados de um grande banco de dados para gerar um modelo é inviável, do ponto de vista financeiro e de tempo [Baker 2011]. Dessa forma, o modelo pode ser construído utilizando parte dos dados e então ser aplicado para modelar dados mais extensos [Baker et al. 2008]. Esse tipo de técnica pode auxiliar no desenvolvimento e uso de atividades instrucionais, pois consegue-se estimar os benefícios educacionais antes mesmo da atividade ser aplicada aos alunos.

Em EDM, são utilizados mais frequentemente dois tipos de técnicas de predição: classificação e regressão. Na classificação a variável preditiva é binária ou categórica e na regressão a variável preditiva é contínua. Em ambos os casos, as variáveis preditoras podem ser categóricas ou contínuas.

A Figura 1.1 representa o funcionamento de um modelo classificador, que tem como entrada um conjunto de treinamento, que consiste de um conjunto de amostras (ou instâncias) de dados onde a classe já é conhecida (ver Tabela 1.1a). A partir desse conjunto de dados, o processo de aprendizagem induz um modelo classificador que em seguida é testado junto a um conjunto de testes, que consiste de um conjunto de amostras cujas classes são ocultadas (ver Tabela 1.1b) e precisam ser preditas a partir do modelo.

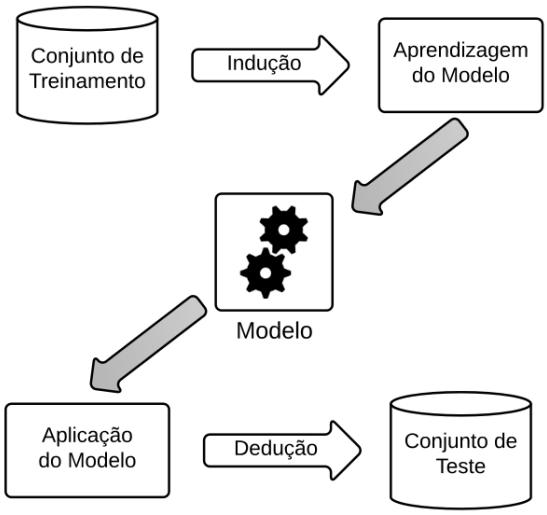


**Figura 1.1. Representação de um modelo classificador.**

| At1 | At2 | At3 | Classe |
|---|---|---|---|
| 0.5 | Falso | Azul | Sim |
| 0.3 | Falso | Branco | Não |
| 0.8 | Verdadeiro | Azul | Não |
| 0.6 | Falso | Verde | Sim |

(a)

| At1 | At2 | At3 | Classe |
|---|---|---|---|
| 0.9 | Falso | Branco | ? |
| 0.2 | Verdadeiro | Azul | ? |
| 0.1 | Verdadeiro | Verde | ? |
| 0.3 | Falso | Verde | ? |

(b)

**Tabela 1.1. Exemplo de conjunto de treinamento (a) e conjunto de teste (b), com atributo contínuo (At1) e discretos (At2 e At3).**

Na classificação, os algoritmos mais utilizados são árvores de decisão e máquina de vetores de suporte. A seguir são apresentados alguns trabalhos que aplicam técnicas

de classificação. Em [Damez et al. 2005], é utilizado uma árvore de decisão fuzzy para modelagem de usuário no intuito de distinguir usuários experientes de leigos. É utilizado um agente para aprender as características cognitivas das interações dos usuários e classificá-los. Feng et al. [Feng and Koedinger 2005] buscam por fontes de erro em predizer o conhecimento de um estudante. Eles utilizam regressão *stepwise* para predizer quais métricas explicam o erro na predição de resultados de exames.

Em relação à regressão, os algoritmos mais populares são regressão linear, redes neurais e máquinas de vetores de suporte para regressão. Como exemplo de utilização de técnicas de regressão em EDM, podemos citar o trabalho de Beck & Woolf [Beck and Woolf 2000] que utilizam regressão linear para predizer variáveis observáveis. O modelo é acoplado num agente de aprendizagem dentro de um STI. O agente aprende a predizer a probabilidade da próxima resposta do estudante estar correta e em quanto tempo o estudante gerará a resposta.

A seguir são apresentados alguns algoritmos de predição, relativamente há modelos de classificação, descrevendo o método de indução de árvore de decisão e o de máquina de vetor de suporte, além do modelo de regressão, apresentado através da abordagem de regressão linear.

#### 1.2.1.1. Árvore de Decisão

Árvores de decisão são modelos estatísticos que utilizam treinamento supervisionado para classificação e predição dos dados. Ou seja, no conjunto de treinamento as variáveis preditivas $Y$ são conhecidas. Uma árvore de decisão possui uma estrutura de árvore, onde cada nó interno (não-folha), pode ser entendido como um atributo de teste, e cada nó-folha (nó-terminal) possui um rótulo de classe [Han and Kamber 2000]. O nó de mais alto nível numa árvore de decisão é chamado de nó-raiz. Um exemplo de árvore de decisão pode ser visto na Figura 1.2.

Após aprendido os parâmetros do modelo, a árvore de decisão irá classificar uma instância de acordo com o caminho que satisfazer as condições desde o nó-raiz até o nó-folha, ao final do processo a instância será rotulada de acordo com o nó-folha. Os algoritmos mais populares de árvore de decisão são o C4.5 [Quinlan 1993], C5.0 [RuleQuest ] e o CART [Breiman 1984].

O algoritmo C4.5, tem como entrada um conjunto de treinamento de dados já classificados (rotulados) e gera um modelo, a partir de um conjunto de dados semelhante, porém não rotulado (conjunto de teste), na forma de uma Árvore de Decisão, utilizando o conceito de entropia da informação. A cada nó da árvore, é selecionado um atributo dos dados que divide o conjunto de amostras de forma mais efetiva em subconjuntos destinados a uma classe ou outra. O critério utilizado para a divisão é o ganho de informação (diferença na entropia) que resulta da escolha do atributo divisor. O atributo com maior ganho de informação é escolhido para tomar a decisão [Quinlan 1993]. O algoritmo C5.0 é uma melhoria do algoritmo C4.5 que promete regras mais precisas, árvores de decisão menores e outras melhorias relacionadas à eficiência e ao custo computacional do algoritmo em si.

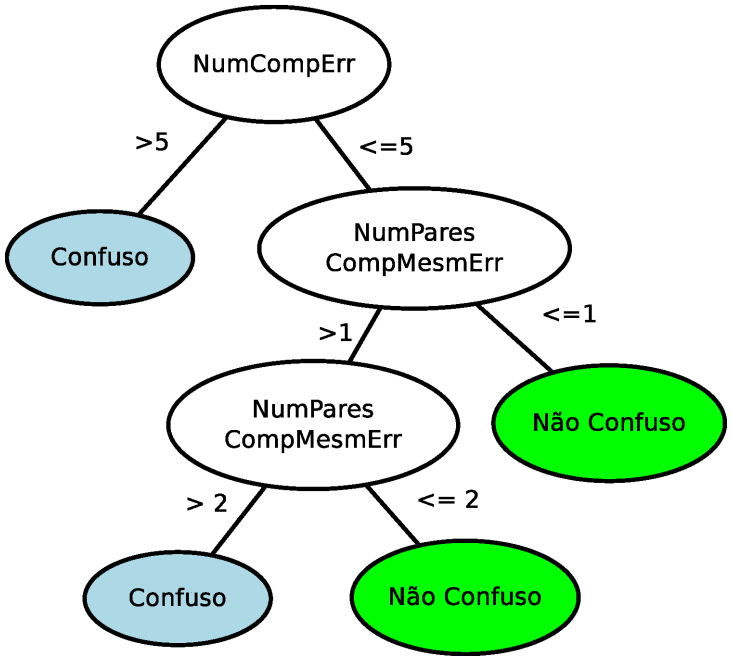


**Figura 1.2. Árvore de Decisão que classifica aprendizes de programação entre Confusos e Não Confusos de acordo com os atributos "Número de Compilação Com Erros" e "Número de Pares de Compilações com o Mesmo Erro".**

### 1.2.1.2. Máquina de Vetores de Suporte

Máquina de Vetores de Suporte (do inglês, *Support Vector Machine* (SVM)) é um algoritmo supervisionado utilizado para a tarefa de classificação que utiliza um hiperplano como separador de classes [Tan et al. 2005]. Este hiperplano é descoberto usando os *vetores de suporte* (conjunto de treinamento) e funciona como um suporte para o limite da decisão ao classificar.

Para dar uma explicação intuitiva do funcionamento da técnica de SVM, considere os dados de treinamento apresentados na Figura 1.3. Suponha que os dados sejam relativos a uma turma com informações dos alunos, representados por círculos, como assiduidade e número de postagens num fórum de discussão (variáveis preditoras). Além disso os dados rotulam cada aluno conforme seu desempenho na disciplina (variável preditiva), alunos que passaram da disciplina (círculos brancos) e alunos que não atingiram a nota mínima (círculos cinzas). Intuitivamente, a meta do SVM é descobrir qual a melhor forma de separar os dois grupos de alunos.

Nota-se que existe um número infinito de hiperplanos (linha tracejada) que podem separar as classes apresentadas (círculos brancos e círculos cinzas). Então o objetivo do SVM é encontrar qual o melhor hiperplano, ou seja aquele que maximize a distância entre as instâncias das classes vizinhas. Um exemplo de melhor hiperplano para os dados apresentados na Figura 1.3 encontrado pelo SVM é apresentado na Figura 1.4.

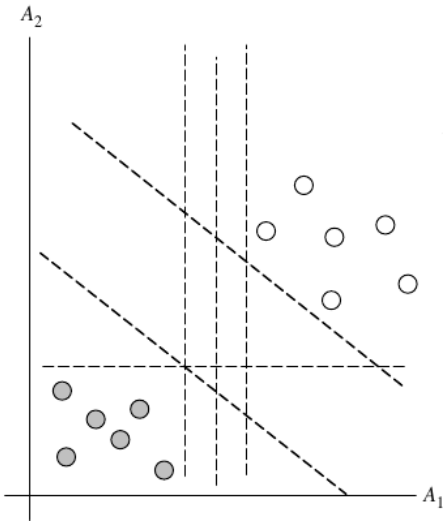


**Figura 1.3. Dados de treinamento, onde existe um número infinito de hiperplanos que podem separar as classes. Os alunos que passaram da disciplina (círculos brancos) e alunos que não atingiram a nota mínima (círculos cinzas). Extraída e adaptada de [Han and Kamber 2000].**

#### 1.2.1.3. Regressão Linear

Regressão linear é uma técnica de predição que envolve uma variável preditiva $y$ e uma única variável preditora $x$, onde $y$ é modelado em função linear de $x$ [Han and Kamber 2000]:

$$y = b + wx, \tag{1}$$

onde assume-se que a variância de $y$ é constante e $b$ e $w$ são coeficientes de regressão. Estes coeficientes podem ser resolvidos pelo método dos mínimos quadrados, que estima qual a reta que melhor representa os dados, ou seja, aquela que minimizar o erro entre os dados atuais e a estimativa da reta.

Como exemplo de regressão linear, suponha uma turma de matemática onde os alunos possuam a sua disposição STI. Ao final do curso, obtém-se dados pareados sobre o tempo de utilização do tutor e das notas finais dos alunos (Tabela 1.2).

Os dados da Tabela 1.2 são apresentados graficamente na Figura 1.5. Analisando-a é possível notar que embora os pontos no gráfico não correspondam a uma linha reta, o padrão tende a uma relação linear entre o tempo de utilização do tutor e a nota final na disciplina.

### 1.2.2. Agrupamento

Em agrupamento, o objetivo é dividir o conjunto de dados em grupos, de forma que os objetos contidos nos dados fiquem agrupados naturalmente de acordo com a semelhança entre eles. Os algoritmos de agrupamento são técnicas de aprendizado não-supervisionado, logo os grupos ou categorias, e até mesmo suas quantidades, não são conhecidos inicialmente. A equação abaixo nos diz quantos grupos de $k$ objetos são possíveis dentro de um conjunto de dados contendo $n$ objetos.

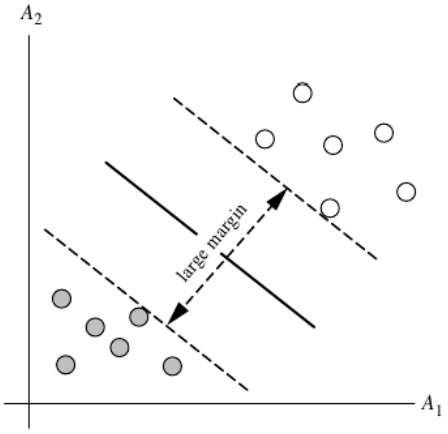


**Figura 1.4. O algoritmo de SVM encontra o melhor hiperplano que separa as classes. Extraída e adaptada de [Han and Kamber 2000].**

**Tabela 1.2. Dados da turma.**

| Tempo de utilização (Média semanal) | Nota final |
|---|---|
| 2 | 6 |
| 3 | 6 |
| 4 | 6.5 |
| 4 | 7 |
| 5 | 7.5 |
| 6 | 7.5 |
| 6 | 7.5 |
| 6 | 8 |
| 7 | 8.5 |
| 8 | 8.5 |
| 8 | 9 |
| 10 | 9.5 |

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \left(\frac{k}{i}\right)(k-i)^n. \quad (2)$$

Se tomarmos, por exemplo, $n = 25$ e $k = 5$, temos que $N(n,k) = 2.436.648.974.110.751$. Esta tarefa é portanto muito complexa e considerada um problema NP-Hard. [3] [4]

Os algoritmos de agrupamento podem tanto começar sem nenhuma hipótese a priori sobre os grupos nos dados (tal como o algoritmo k-means com reinício aleatório), ou começar de uma hipótese específica, gerada possivelmente por pesquisa anterior em outros dados. Um algoritmo de agrupamento pode gerar grupos do tipo *hard*, onde cada

[3]Um problema é considerado NP-Hard quando pode-se assumir que sua complexidade é, pelo menos, tão grande quanto a do problema NP mais complexo.

[4]NP é o conjunto de problemas de decisão onde as instâncias para as quais a resposta é sim podem ser reconhecidas em tempo polinomial por uma máquina de Turing não determinística.

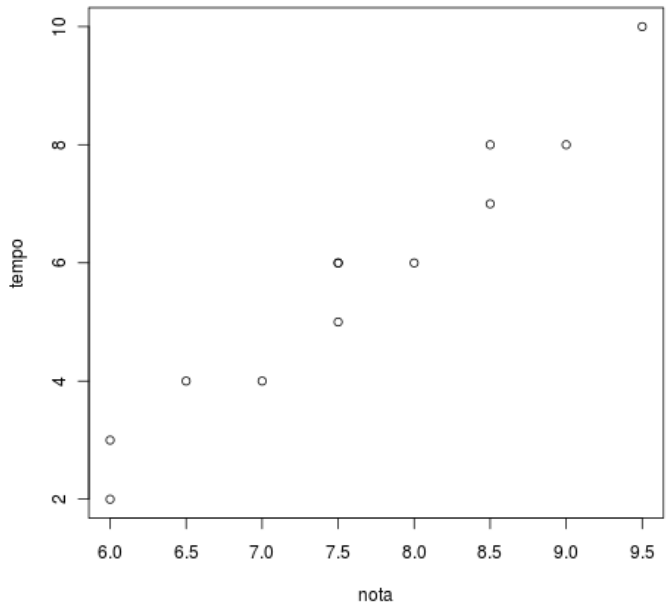


**Figura 1.5. Gráfico dos dados da Tabela 1.2, onde *nota* (Nota final) e *tempo* (Tempo de utilização).**

elemento pertence a somente um grupo (ex.: algoritmo k-means), ou do tipo soft (também conhecido como fuzzy), onde um elemento pode pertencer a mais de um grupo (ex.: Modelos de Mistura Gaussiana).

Diversos trabalhos na área de Mineração de Dados Educacionais fazem uso das técnicas de agrupamento, em [Moreno et al. 2012], um Algoritmo Genético é utilizado para realizar agrupamento inter-homogêneo e intra-heterogêneo de estudantes para fins de atividades de aprendizagem colaborativa. Diversas características dos estudantes são consideradas, tais como uma estimativa de seu nível de conhecimento e estimativas de suas habilidades de comunicação e de liderança.

Em [Xu ], um método estatístico (mistura de distribuições de probabilidade) de agrupamento do tipo fuzzy, chamado de Análise Latente de Classes, é utilizado para agrupar professores de acordo com os seus comportamentos ao utilizar uma biblioteca digital para auxílio à preparação de conteúdo instrucional. Em [Amershi 2009], diferentes tipos de estudantes são identificados em ambientes de aprendizagem a partir de dados oriundos de logs da interface e do rastreamento do movimento dos olhos, esses dados servem de entrada para um algoritmo de agrupamento do tipo K-Means que se encarrega de agrupar os estudantes de acordo com suas similaridades comportamentais.

Em [Talavera 2004], também é utilizado um método estatístico para identificar padrões de comportamento de estudantes em um cenário de colaboração num ambiente de aprendizagem. Em [Shen 2003], os estudantes são agrupados de acordo com suas preferências com o intuito de melhor adaptar os sistemas de ensino a distância de acordo com suas necessidades. Ainda em [Shen 2003], são construídos modelos representativos

de cada grupo, que é por fim utilizado para identificar as melhores práticas de ensino e sugestão de material de acordo com as preferências dos membros de cada grupo.

Em seguida, alguns algoritmos de agrupamento bastante usados em EDM são mostrados.

#### 1.2.2.1. Algoritmo K-Means

O algoritmo K-Means é largamente utilizado para a tarefa de agrupamento. Em sua forma mais comum [Lloyd 1982], algumas vezes referida por Algoritmo de Lloyd, tem o seguinte funcionamento:

O número *k* de grupos que se deseja encontrar precisa ser informado de antemão. Em seguida, *k* pontos são escolhidos aleatoriamente para representar os centróides dos grupos, com isso, um conjunto de elementos, usualmente vetores, é particionado de forma que cada elemento é atribuído à partição, ou grupo, de centróide mais próximo, de acordo com a distância euclidiana comum. A cada iteração do algoritmo, os *k* centróides, ou "médias", e daí vem o nome *means*, são recalculados de acordo com os elementos presentes no grupo e em seguida todos os elementos são realocados para a partição cujo o novo centróide se encontra mais próximo (ver Figura 1.6).

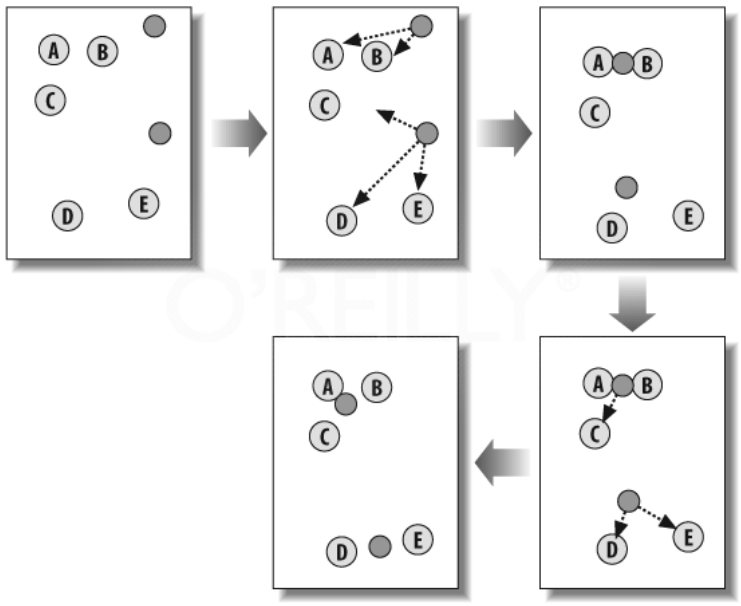


**Figura 1.6. Funcionamento do Algoritmo K-Means, passo-a-passo. (Adaptado de [Segaran 2007])**

Esse processo é repetido até que os elementos sejam atribuídos aos mesmos grupos das interações anteriores por repetidas interações, de forma que os centróides permaneçam estáveis daí em diante.

Tal como aponta [Witten and Frank 2005], este método de agrupamento é simples e efetivo. É fácil provar que o processo minimiza a distância quadrática total de cada ponto do grupo ao seu centróide, e assim que a iteração estabiliza, cada ponto estará atribuído

ao centróide mais próximo e portanto o efeito generalizado é a minimização da distância quadrática total de todos os pontos aos seus centros. No entanto, não há garantias de que se tenha encontrado um mínimo global, portanto, o que se faz, geralmente é iniciar o algoritmo várias vezes, com diferentes pontos de partidas (posições dos centróides) e escolher aquele resultado com menor distância quadrática total.

#### 1.2.2.2. Algoritmo Genético

Algoritmos Genéticos (AGs) são modelos computacionais de busca e otimização de soluções [Passos and Goldsmith 2005], inicialmente propostos por [Holland 1992], com forte inspiração na teoria da evolução das espécies de Charles Darwin [5]. Nos algoritmos genéticos, as soluções, que representam diferentes pontos no espaço de busca, são representadas por cromossomos artificiais dos quais os genes representam as características daquela solução. Uma série de analogias podem ser feitas entre os algoritmos genéticos e a evolução das espécies, entre elas: O Meio Ambiente pode ser entendido como o problema, representado pela função de avaliação. O indivídiduo em adaptação no ambiente é a solução do problema e é representado pelo seu cromossomo, que num AG pode ser uma string, um vetor, uma lista e etc.

O funcionamento do AG busca reproduzir um ambiente natural, onde somente os indivíduos mais aptos prosperam e reproduzem, transmitindo seu código genético para as próximas gerações, tal como descrito a seguir.

Inicialmente, um conjunto de soluções ou cromossomos, chamado de população, é criado de forma aleatória, constituindo-se então na população inicial que tem cada uma de suas soluções aferidas pela função de avaliação e associadas a um certo valor de aptidão. Baseado no princípio da seleção natural, as soluções mais aptas são selecionadas e submetidas aos operadores genéticos. Cada gene, ou característica, tem uma pequena probabilidade de sofrer mutação e cada solução uma outra probabilidade de sofrer cruzamento, o que poderá, ou não, melhorar a aptidão do indivíduo. Ao fim de cada ciclo, a aptidão dos indivíduos, ou seja, das soluções, é medida pela função de avaliação. Esse processo continua por um determinado número de ciclos ou até que a condição de parada seja satisfeita.

Num problema de agrupamento, geralmente, a solução (o conjunto de grupos proposto) é representada em uma matriz, onde cada coluna é um grupo de estudante, por exemplo, e cada matriz é, portanto, uma maneira de se agrupar aqueles estudantes. Nesta maneira de representação das soluções, o AG pode ser visto como um algoritmo de agrupamento do tipo *hard*.

No fluxograma da Figura 1.7, apresentamos a estrutura de um algoritmo genético, onde podemos notar que, basicamente, o algoritmo é composto de um laço principal (ciclo), que representa as gerações de indivíduos, no qual são executados os elementos básicos deste algoritmo: A função de avaliação, a seleção dos indivíduos a compor a nova população, os operadores genéticos (cruzamento e mutação), e a substituição da popula-

[5] Na teoria da evolução das espécies os indivíduos mais aptos tem maiores chances de sobrevivência e portanto de gerar descendentes, perpetuando seu código genético.

ção antiga pela nova população gerada com indivíduos mais aptos.

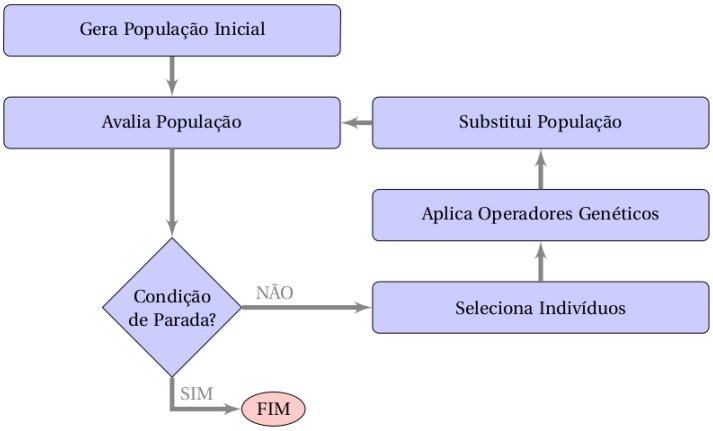


**Figura 1.7. Fluxograma de um algoritmo genético**

#### 1.2.2.3. Outras Abordagens

Além das abordagens descritas acima, existem outras abordagens também bastante usadas em EDM, incluindo métodos estatísticos como Análise de Fatores e Modelos de Mistura Gaussiana.

Os Modelos de Mistura Gaussiana constituem um método estatístico de agrupamento do tipo *fuzzy*. A base dos algoritmos de agrupamento estatístico é o modelo de misturas finitas, onde uma mistura é uma conjunto de $k$ distribuições de probabilidade, representando os $k$ grupos, que governam os valores de atributos dos membros daquele grupo. Em outras palavras, cada distribuição dá a probabilidade de uma instância em particular ter um certo conjunto de atributos se ela for um membro daquele grupo. Cada instância em particular pertence a um e somente um dos grupos, mas não se sabe a qual deles [Witten and Frank 2005].

Com esta fundamentação, os Modelos de Mistura Gaussiana são capazes de produzir agrupamentos com grupos que se sobrepõem ou até mesmo deixar algumas instâncias do conjunto de dados sem estar atribuídas a nenhum grupo. Esta flexibilidade pode ser interessante para algumas aplicações. Uma das desvantagens desta abordagem é um custo computacional relativamente alto.

### 1.2.3. Mineração de Relações

Em mineração de relações, o objetivo é descobrir possíveis relações entre variáveis de um banco de dados com muitas variáveis. Isto pode ser feito investigando quais variáveis estão mais fortemente relacionadas com uma determinada variável de interesse (ex.: a nota do aluno), ou investigando relações fortes entre quaisquer duas variáveis.

Existem quatro tipos de mineração de relações: i) regras de associação, ii) correlações, iii) padrões sequenciais e iv) causas. A seguir apenas a abordagem de regra de associação será descrita, as demais são variações da proposta de descoberta de associações que são sucintamente apresentadas em [Baker and Yacef 2009].

#### 1.2.3.1. Regras de associação

Mineração de regras de associação introduzida em [Agrawal et al. 1993], é uma das mais importantes técnicas de mineração de dados, tendo como objetivo central derivar regras de conhecimento, referindo-se a relacionamentos entre objetos de um conjunto de dados, visando exibir características e tendências. Isto é, procura-se associação entre itens do tipo "uma transação que contém os itens $X$ também possui o conjunto de itens $Y$ " ($X \rightarrow Y$), sendo $X \cap Y = \emptyset$. Assim, a regra tem a forma "Se $X$, então $Y$", onde $X$ é denominado de corpo da regra e $Y$ de cabeça da regra. A cada regra derivada pelo algoritmo, verifica-se a sua validade e importância. Para isso, faz-se uso de duas medidas básicas: o suporte e a confiança, comparando-as com os seus respectivos limiares estabelecidos (suporte mínimo e confiança mínima).

A medida de suporte diz respeito à ocorrência relativa da regra de associação detectada dentro do conjunto de dados de transações, sendo calculada pelo quociente entre o número de transações que sustentam a regra e o número total de transações. Já a medida de confiança de uma regra de associação refere-se ao grau com o qual a regra é verdadeira entre os registros individuais, sendo calculada pelo quociente entre o número de transações sustentando a regra e o número de transações sustentando apenas o corpo da regra.

Um exemplo de uso dessa técnica em EDM é a mineração de regras em um banco de dados de notas de alunos em disciplinas. Neste contexto seria possível derivar regras como "90% dos alunos que têm bom desempenho nas disciplinas de Lógica e Matemática são bem sucedidos também em Programação".

Os algoritmos clássicos de mineração de regras de associacão derivam regras apenas conjuntivas, limitando-se a utilização do operador lógico AND. Desde meados dos anos 90 que vários algoritmos clássicos de derivação de regras de associação têm sido desenvolvidos, por exemplo: quantitative association rule [Srikant and Agrawal 1996], generalized association rule [Srikant and Agrawal 1995], sequential patterns [Mannila et al. 1997] e association rules extended with negation [Tsur et al. 1998]. A literatura de algoritmos de regras de associação é ampla, mas no presente texto vamos ressaltar apenas o clássico algoritmo Apriori. Apriori [Agrawal et al. 1993] consolidou-se como o primeiro algoritmo de mineração de regras de associação assumido como eficiente. Esse algoritmo combina uma estratégia de busca denominada Breadth-first search (BFS) com uma estrutura de árvore para contagem de ocorrência de candidatos.

### 1.2.4. Destilação de dados para facilitar decisões humanas

Nesta área, o principal objetivo é apresentar os dados de forma mais legível e visual para facilitar a compreensão humana e assim apoiar decisões importantes baseadas nos dados. É uma área de interesse crescente dentro da mineração de dados educacional. Pois, a análise dos dados realizada por agente humanos, em sua maioria, só pode ser realizada se os dados forem apresentados de forma apropriada. O principal método dessa área da mineração de dados educacionais é o de visualização da informação. No entanto, a maioria dos métodos de visualização, normalmente, usados dentro da mineração de dados educacional são frequentemente diferentes do que aqueles mais frequentemente usados em problemas

de visualização da informação [Hershkovitz and Nachmias 2008, Kay et al. 2006]. Alguns exemplos dessas particularidades são destacadas por Baker [Baker 2010a]: Os dados são organizados em termos da estrutura do material de aprendizagem (habilidades, problemas, unidades, aulas) e da estrutura de contexto de aprendizagem (alunos, professores, pares de colaboração, classes e escolas).

A destilação dos dados para facilitar decisões humanas tem dois propósitos principais [Baker 2010b]: a) Identificação - os dados são apresentados de forma que humanos possam identificar os padrões mais facilmente, que são difíceis de expressar formalmente; b) Classificação - a destilação de dados pode ser usada também para apoiar a modelos de predição. Neste caso, parte dos dados são exibidos para serem rotulados por humanos. Esses rótulos são utilizados como base para a construção desses modelos.

Segundo Baker[Baker et al. 2006], uma área chave para destilação de dados para facilitar decisões humanas é a metodologia de repetição de texto. Essa metodologia consiste em apresentar pequenas partes da base de dados em formato de texto, após receberem rótulos por agentes humanos. Ainda segundo Baker[Baker 2010a], a repetição de texto tem sido utilizadas para, por exemplo: o desenvolvimento de modelos de predição para usuários que tentam trapacear o sistema em vários ambientes de aprendizagem [Baker et al. 2006, Baker et al. 2010].

A identificação de padrões de aprendizagem e diferenças individuais dos estudantes a partir da visualização é um método chave para exploração de bases de dados educacionais [Baker 2010a]. Como o exemplo apresentado por Baker [Baker 2010a] dentro do domínio do modelo do estudante, como pode ser visto na Figura 1.8.

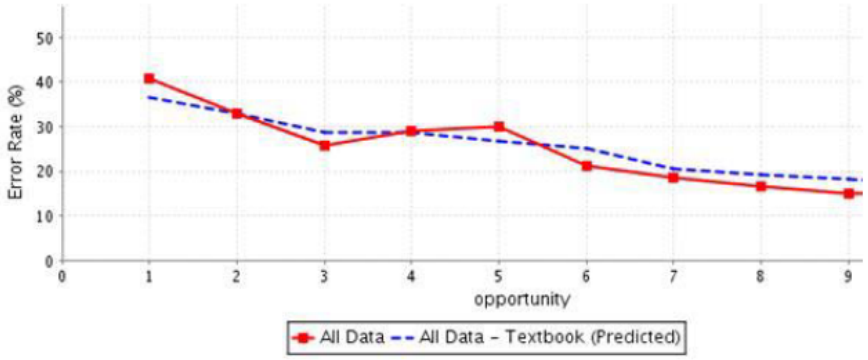


**Figura 1.8. Curva de aprendizagem do desempenho do estudante ao longo do tempo em um tutor (Adaptada de [Koedinger et al. 2010a])**

É apresentado na Figura 1.8 uma curva de aprendizagem clássica, onde o eixo X representa o número de oportunidades (*opportunity*) e o eixo Y o desempenho do estudante, o percentual de respostas corretas ou o tempo para responder (*error rate (%)*). Pode-se perceber na Figura 1.8 que a curva tem uma suave queda, o que indica que o estudante está aprendendo.

### 1.2.5. Descoberta com modelos

Em descoberta com modelos, parte-se de um modelo gerado por um método de predição, tal como classificação, ou por um método de agrupamento, ou ainda manualmente, por

meio de engenharia de conhecimento. Em seguida, esse modelo é utilizado como componente, ou ponto de partida, em outra análise com técnicas de predição ou mineração de relações.

## 1.3. Preparação e Aquisição dos Dados

Nesta seção serão apresentadas as técnicas de preparação (pré-processamento) de dados mais comumente utilizadas em EDM e, em seguida, serão apresentadas as possíveis fontes de dados educacionais para mineração.

### 1.3.1. Preparação de dados

Como em todo problema de Mineração de Dados, em EDM se faz necessária uma etapa de preparação de dados, de forma adequá-los a análise que se pretende efetuar. Um grande problema, no entanto, é que com tantas fontes de dados diferentes em EDM, existe uma falta de padronização na maneira como os dados são coletados e armazenados, que por si só, constitui um dos desafios da área.

Segundo [García 2007], a maioria das tarefas de preparação necessárias em mineração de dados tradicionais, tais como limpeza de dados, identificação de usuário, identificação de sessão, identificação de transação, transformação de dados e enriquecimento, integração e redução de dados não são necessárias em certos AVAs que armazenam dados com o propósito de realizar análises posteriores.

Ainda de acordo com [García 2007], as tarefas típicas de preparação de dados para a área de EDM são: Discretização de Dados (onde valores numéricos são transformados em categorias), derivação de novos atributos e seleção de atributos (novos atributos são criados a partir dos existentes e somente um subconjunto é escolhido), criação de tabelas de sumarização (estas tabelas integram toda a informação desejada para ser minerada no nível apropriado, por exemplo, no nível de estudante), transformação do formato de dados (os dados são transformados para se adequar ao formato requerido pelos algoritmos e frameworks de mineração de dados). Em seguida, veremos uma descrição dessas tarefas.

#### 1.3.1.1. Discretização de Dados

A tarefa de discretização, que é geralmente aplicada quando se pretende realizar classificação ou associação, constitue em transformar valores numéricos contínuos em $n$ intervalos que serão as $n$ categorias que se deseja obter. Basicamente, todo o esforço está em decidir quantas categorias serão necessárias e onde ficarão os limites dos intervalos, no entanto, existem métodos de discretização não-supervisionada, onde não é necessário decidir esses parâmetros.

#### 1.3.1.2. Derivação de Novos Atributos

Esta técnica permite que novos atributos sejam derivados apartir dos atributos originais com o intuito de facilitar a extração de determinada informação de forma mais eficaz. O novo conjunto de atributos pode substituir ou ser agregado aos atributos originais. Con-

sidere, por exemplo, um conjunto de fotografias onde deseja-se classificar de acordo com a presença ou não de uma face humana. Os dados brutos, contendo informações apenas a nível de pixels, podem não ser interessantes para essa tarefa, mas novos atributos, num nível mais alto, podem ser gerados a partir de outros algoritmos que detectem a presença, ou não, de certas bordas na imagem. Esses novos atributos podem servir de entrada a um conjunto maior de técnicas de classificação [Tan et al. 2005].

#### 1.3.1.3. Seleção de Atributos

A tarefa de seleção de atributos é muito importante quando se trabalha com conjunto de dados com alta dimensionalidade, ou seja, com grande número de atributos, o que aumenta o custo computacional de várias técnicas de mineração de dados. Em classificação, por exemplo, comumente trabalha-se apenas com um subconjunto do conjunto de atributos original. Os atributos contidos no subconjunto são escolhidos de acordo com a informação que deseja extrair. A escolha pode ser manual, por um especialista no domínio dos dados, ou automática, por algum algoritmo de seleção automática de atributos.

#### 1.3.1.4. Criação de Tabelas de Sumarização

As tabelas de sumarização são muito utilizadas quando se trabalha com um banco de dados relacional constituído de várias tabelas e precisa-se apenas de uma parte desses dados. Para isso, cria-se uma nova tabela apenas com a informação desejada, já transformada, para a técnica de mineração que se deseja aplicar. Essa tarefa é utilizada em EDM, por exemplo, quando se realiza classificação em dados do Moodle [Romero et al. 2008].

#### 1.3.1.5. Transformação do Formato dos Dados

Essa tarefa é necessária quando os dados de entrada da técnica de mineração que se pretende utilizar precisam estar em um formato específico e diferente do formato atual dos dados.

### 1.3.2. Aquisição dos Dados

Nos últimos anos, com o crescente uso de AVAs, softwares educacionais e outras tecnologias que amparam o ensino por meio do computador, uma grande quantidade de dados tem sido gerada. No Brasil, especificamente, a Universidade Aberta do Brasil tem grande importância neste cenário, devido ao grande número de cursos de ensino a distância utilizando AVAs e outros softwares educacionais, e pode ser um grande aliado da Mineração de Dados Educacionais. Um grande problema, no entanto, é que com tantas fontes de dados diferentes, existe uma falta de padronização na maneira como os dados são coletados e armazenados, ocasionando um grande esforço de pré-processamento de dados, que por si só, constitui um dos desafios da área.

No entanto, algumas fontes de dados educacionais disponíveis na Web, como por

exemplo do repositório PSLC DataShop[6] [Koedinger et al. 2010b].

## 1.4. Principais Aplicações de EDM

A mineração de dados educacional tem sido utilizada diversas áreas, as principais áreas de aplicação são [Baker and Yacef 2009]:

- Modelagem do estudante;
- Modelagem do domínio;
- Suporte pedagógico;
- Descoberta científica.

Nesta seção serão apresentadas como aplicações que utilizam a mineração de dados educacionais podem auxiliar educadores em diferentes áreas, ou mesmo algum sistema computacional, a exemplo de um STI.

### 1.4.1. Modelagem do estudante

Os modelos do estudante armazenam informação sobre características dos alunos, tais como conhecimento, motivação, atitudes, personalidade, além de questões sociais. As técnicas de EDM podem ser utilizadas para dar uma maior acurácia no modelo de estudante e proporcionar uma maior personalização e adaptação dos serviços oferecidos por um AVA.

Modelar as diferenças existentes entre os estudantes possibilita acompanhar o aprendizado de forma individualizada, melhorando significativamente o aprendizado do estudante. Utilizando métodos de EDM é possível modelar atributos do estudante em sistemas de tempo real. Por exemplo, em [Baker et al. 2008], os autores utilizam EDM para detectar comportamentos inadequados dos estudantes em STI. Eles verificam se o estudante está "trapaceando o sistema" (do inglês *gaming the system*), e.g. o estudante pede diversas dicas somente para descobrir a resposta de um determinado problema. Em [D'Mello et al. 2008] é verificado se um estudante está entediado ou frustrado em utilizar o sistema, isto é feito por meio da análise de atributos extraídos da interação dos estudantes com o sistema como por exmplo, informação temporal e informação das respostas. Alguns trabalhos na literatura buscam identificar quais fatores fazem um estudante ser reprovado ou desistir de uma disciplina na universidade, e.g. [Kotsiantis 2009, Dekker et al. 2009, Romero et al. 2008, Superby et al. 2009].

### 1.4.2. Modelagem do domínio

Uma importante área de utilização de EDM é na descoberta de modelos que representem a estrutura de um domínio. Por meio da combinação de arcabouços de modelagem de psicometria com algoritmos de espaço de busca, alguns trabalhos têm conseguido desenvolver abordagens automáticas de descoberta, que a partir de dados conseguem identificar com acurácia modelos de estrutura de domínio.

[6]Repositório para armazenamento e análise de dados educacionais criado pelo Pittsburgh Science of Learning Center: https://pslcdatashop.web.cmu.edu/

### 1.4.3. Suporte pedagógico

O estudo do suporte pedagógico, tanto em softwares de apoio à aprendizagem quanto em outros domínios, como aprendizagem colaborativa, tem o objetivo de descbrir que tipos de suporte pedagógico são mais eficientes na média ou para grupos específicos de estudantes. E neste último caso, torna-se uma tarefa ainda mais complexa devido as particularidades de cada estudante [Baker and Yacef 2009].

### 1.4.4. Descoberta científica

Uma quarta área de aplicação de EDM é na descoberta e confirmação de teorias científicas educacionais conhecidas e na busca de melhor compreender os fatores chave que impactam no processo de aprendizagem, sempre procurando desenvolver melhores sistemas de apoio ao ensino e à aprendizagem.

## 1.5. Ferramentas para EDM

Esta seção tem como objetivo apresentar as principais ferramentas utilizadas na área para realizar o processo de descoberta de conhecimento no contexto educacional.

Existem diversas ferramentas de mineração, comerciais e acadêmicas, disponíveis que proveem algoritmos de mineração, algoritmos de pré-processamento, técnicas de visualização, entre outros, como: DBMiner, Clementine, IBM Intelligent Miner, Weka [Hall et al. 2009] e Rapidminer [Mierswa et al. 2006]. Apesar dos esforços da comunidade de mineração de dados educacionais em propor e construir ferramentas de mineração que levem em conta as particularidades da mineração no contexto educacional, duas dessas ferramentas são muito utilizadas na literatura: Weka e Rapidminer. Por este motivo, essas ferramentas foram escolhidas e serão apresentadas nas subseções a seguir.

### 1.5.1. WEKA

Weka é uma coleção do estado da arte de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento [Hall et al. 2009]. É uma ferramenta de código aberto e foi desenvolvido na Universidade de Waikato na Nova Zelândia. Weka possui uma variedade de algoritmos de aprendizagem, que incluem ferramentas de pré-processamento. Além disso, oferece suporte a todo processo de mineração, que inclui suporte a preparação dos dados de entrada, avaliação estatística da aprendizagem, visualização dos dados de entrada e os resultados. Todas as funcionalidades disponíveis podem ser acessadas através de uma interface comum, apresentada na Figura 1.9.

A interface oferecida pelo Weka permite que os algoritmos de aprendizagem e as diversas ferramentas para transformação possam ser aplicados as bases de dados sem que seja necessário escrever nenhum código. O Weka inclui métodos para os problemas padrões de mineração de dados, como: regressão, classificação, agrupamento, regras de associação e seleção de atributos [Hall et al. 2009]. Todos os algoritmos aceitam o formato padrão estabelecido para o Weka, o ARFF. ARFF é um formato de entrada específico da ferramenta e tem a forma de uma tabela relacional simples. O ARFF pode ser lido de um arquivo e/ou construído a partir de uma base de dados.

Na tela principal apresentada na Figura 1.9, o Weka disponibiliza quatro opções o

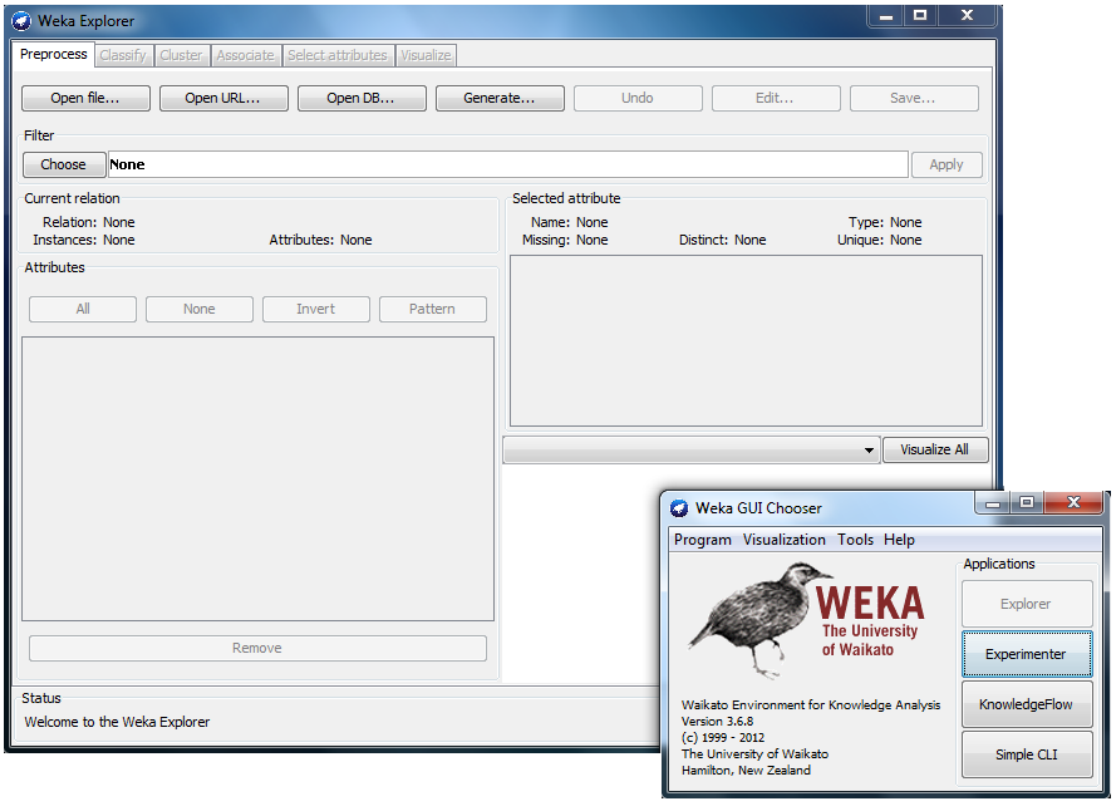


**Figura 1.9. Interface gráfica inicial do Weka e a Interface gráfica Explorer**

usuário: *Explorer, Knowledge Flow, Experimenter e CLI*.

A primeira opção é o *Explorer*. Esta é a opção mais simples para se utilizar o Weka. Ela oferece uma interface que possibilita ao usuário acessar funcionalidades oferecidas pelo Weka e que podem ser acessadas por meio da interface gráfica e suas opções. O usuário pode, por exemplo, escolher ler um arquivo ARFF e construir uma modelo, utilizando algum dos algoritmos disponíveis. A interface possibilita utilizar quaisquer um dos algoritmos implementados pela ferramenta, apresentando dicas e os resultados de cada aprendizagem realizada com a base de dados escolhida.

Outra opção disponibilizada na interface gráfica apresentada na Figura 1.9 é o KnowledgeFlow. Como apresentado na Figura 1.10, esta opção oferece uma interface gráfica que permite ao usuário construir um fluxo para o processamento dos seus dados. O Weka oferece alguns algoritmos incrementais e podem ser usados para processar um conjunto de dados muito grande [Hall et al. 2009]. Essa interface permite que o usuário escolha entre caixas que representam esses algoritmos, arraste-os e estabeleça a configuração desejada. Isso permite que o usuário construa um fluxo para o processamento do conjunto de dados escolhido por meio da conexão desses componentes. Esse componentes podem estar representando as fontes de dados, ferramentas de pré-processamento, algoritmos de aprendizagem, métodos de avaliação e visualização.

A terceira opção disponibilizada pelo Weka é o *Experimenter*. Essa opção oferece uma interface gráfica que possibilita aos usuários um auxílio em uma questão prática:

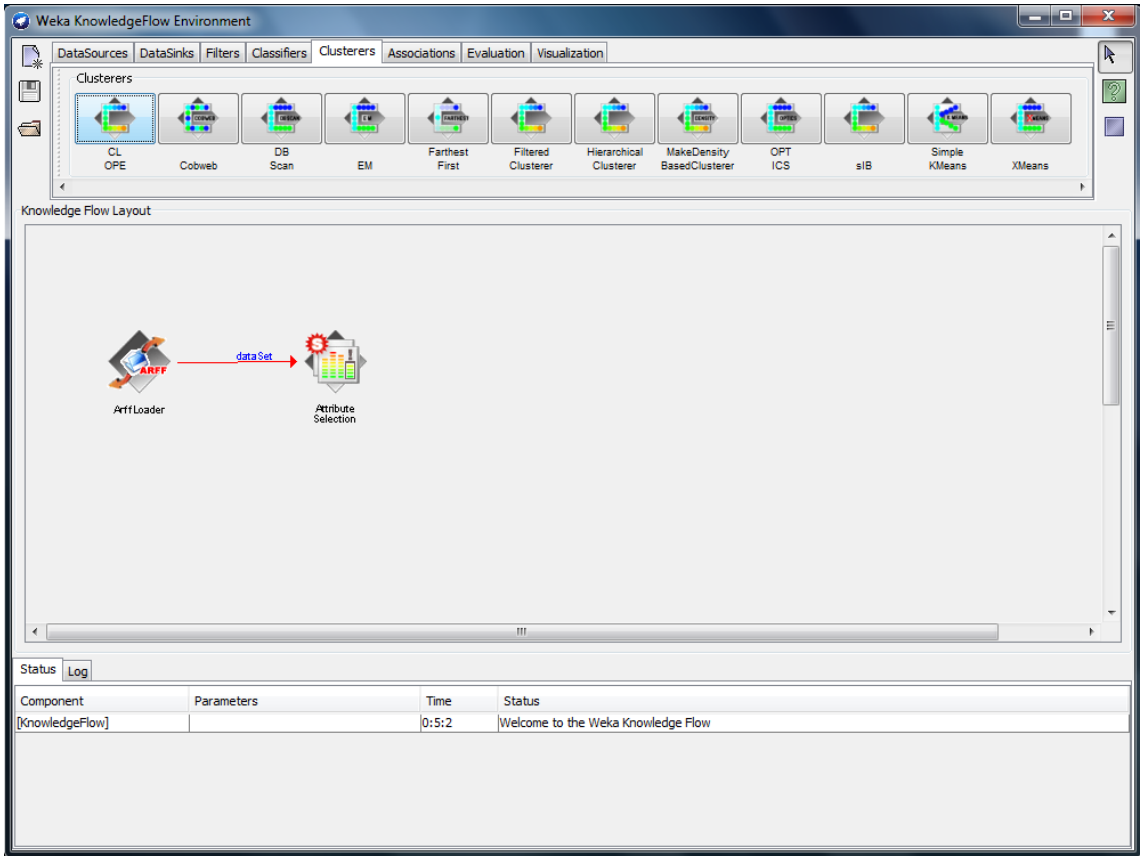


**Figura 1.10. Interface gráfica KnowledgeFlow do Weka**

descobrir quais métodos e parâmetros funcionam melhor para um determinado problema. O usuário pode fazer isso de forma interativa ao aplicar algum algoritmo em sua base de dados. Entretanto, essa interface permite ao usuário automatizar esse processo, tornando mais fácil executar diferentes algoritmos e filtros com diferentes parâmetros.

A última opção oferece a funcionalidade mais básica da ferramenta, onde o Weka pode ser acessado utilizando linhas de comando. Essa opção possibilita o acesso a todas as funcionalidades do sistema. Além disso, o Weka oferece a API Java, que permite a construção de aplicações que utilizem todas as funcionalidades disponibilizadas pela ferramenta. A API também oferece suporte a construção do arquivo ARFF que é o formato específico aceito pelos algoritmos implementados pela ferramenta.

O Weka fornece, ainda, uma vasta documentação on-line, além do material produzido pelo projeto, como o livro *Data Mining Practical Machine Learning Tools and Techniques* [Hall et al. 2009].

### 1.5.2. RapidMiner

O RapidMiner é um sistema de código aberto para a mineração de dados. É um aplicativo distribuído de forma independente para análise de dados, mineração de texto e de dados, além disso, permite a integração com outros produtos desenvolvidos pelo mesmo projeto. Na Figura 1.11 é apresentada uma de suas telas, onde a fonte de dados carregada pode ser

visualizada.

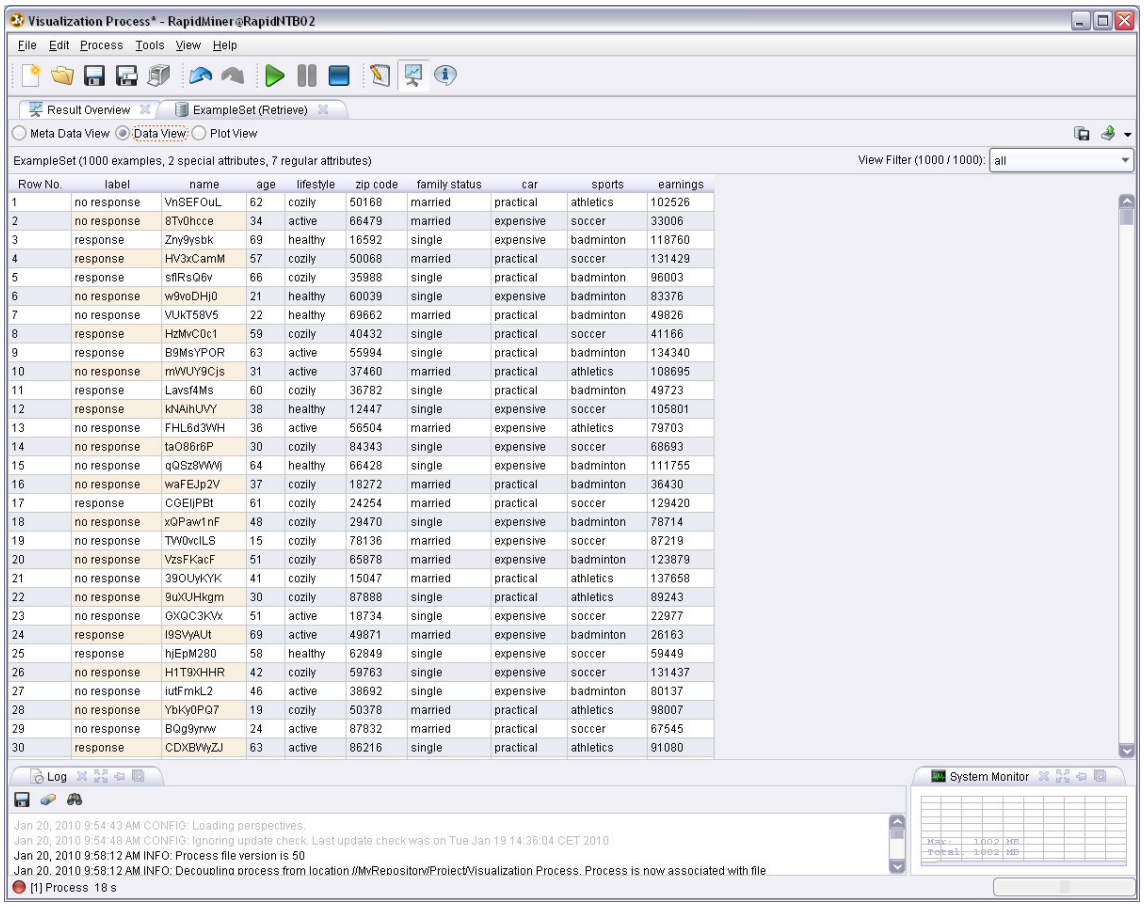


**Figura 1.11. Interface gráfica do RapidMiner (Extraída de: [Rapid-i 2012])**

Segundo seus desenvolvedores[Rapid-i 2012], o *RapidMiner* oferece as seguintes vantagens:

- Integração de dados, ETL Analítica, Análise de Dados e Relatório em uma única suíte;
- Poderosa, mas intuitiva interface gráfica para o usuário para o design do processo de análise;
- Repositório para processo, dados e manipulação de meta-dados;
- Única solução com transformação de meta-dados
- Única solução com suporte ao reconhecimento de erros *on-the-fly* e soluções rápidas;
- Completa e flexível: Centenas de carregamento de dados, transformação de dados, modelagem dos dados e métodos de visualização;

Algumas características interessantes dessa ferramenta são: o fato de ter seu código aberto e, por isso, disponível gratuitamente, funcionar na maioria das principais plataformas e sistema operacional. Além de, assim como o Weka, disponibilizar o acesso as suas funcionalidades por meio de uma interface gráfica intuitiva, linhas de comando e API Java, possibilitando a construção de aplicações que a utilizem por meio de um mecanismo simples. Além disso, o RapidMiner possui a biblioteca de algoritmos de aprendizagem do Weka totalmente integrada e possibilita o acesso a diferentes fontes de dados, como: Excel, Acess, Oracle, Microsoft SQL Server, MySQL, Postgres, Arquivos de texto, entre outros.

Além das das características citadas, outras características são destacadas por desenvolvedores [Rapid-i 2012] como diferenciais da ferramenta, são algumas delas: solução mais abrangente disponível: mais de 500 operadores de integração e transformação dos dados, mineração, avaliação, visualização; e conceito visualização multi-camadas de dados garante a manipulação de dados mais eficiente, entre outras.

O RapidMiner também oferece uma extensa documentação que inclui: Tutorias em videos, guia de instalação e um manual da ferramenta. O manual é muito bem construído e aborda, além de um passo-a-passo de como utilizar a ferramenta, uma introdução aos conceitos fundamentais e necessários sobre mineração de dados.

## 1.6. Considerações Finais

Neste capítulo foi apresentado um apanhado de conceitos, técnicas, tecnologias e aplicações em mineração de dados educacionais, visando disponibilizar um texto básico para orientar o leitor neste campo emergente, olhado na perspectiva da pesquisa em Inteligência Artificial e Educação. Particularmente, foi dado ao leitor uma descrição das principais iniciativas em EDM como campo de pesquisa, tendo atualmente o amparo de uma sociedade científica, a qual patrocina uma conferência internacional e um periódico.

Quanto à parte técnica do texto, enfocou-se os principais métodos utilizados em EDM, oferecendo-se uma boa amplitude e um nível básico de descrição de tais métodos, mas sempre remetendo o leitor a referências apropriadas para possíveis leituras e aprofundamentos no tema. Além disso, foram apresentadas algumas ferramentas de mineração de dados que estão sendo também utilizadas em contextos educacionais. Finalmente algumas aplicações foram elencadas, prestando-se a ilustrar o potencial do campo de mineração de dados educacionais.

## Referências

[Agrawal et al. 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD '93, pages 207–216, New York, NY, USA. ACM.

[Amershi 2009] Amershi, S., C. C. (2009). Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of Educational Data Mining*, 1(1):18–71.

[Baker 2011] Baker, R.S.J.d., I. S. d. C. A. (2011). Mineração de dados educacionais:

Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(2).

[Baker 2010a] Baker, R. (2010a). *Data Mining for Education*. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), Elsevier, Oxford, UK.

[Baker 2010b] Baker, R. (2010b). *Mining Data for Student Models*. In Nkmabou, R., Mizoguchi, R., & Bourdeau, J. (Eds.) Advances in Intelligent Tutoring Systems, pp. 323-338. Secaucus, NJ: Springer.

[Baker and Yacef 2009] Baker, R. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

[Baker et al. 2008] Baker, R. S., Corbett, A. T., Roll, I., and Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3):287–314.

[Baker et al. 2006] Baker, R. S. J. D., Corbett, A. T., and Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the Workshop on Educational Data Mining (held at the 8th International Conference on Intelligent Tutoring Systems – ITS 2006). Jhongli*, pages 29–36.

[Baker et al. 2010] Baker, R. S. J. D., Mitrović, A., and Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. In *Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 267–278, Berlin, Heidelberg. Springer-Verlag.

[Beck and Woolf 2000] Beck, J. and Woolf, B. P. (2000). High-level student modeling with machine learning. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, ITS '00, pages 584–593, London, UK, UK. Springer-Verlag.

[Breiman 1984] Breiman, L. (1984). *Classification and regression trees*. The Wadsworth and Brooks-Cole statistics-probability series. Chapman & Hall.

[Cabena et al. 1998] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Damez et al. 2005] Damez, M., Marsala, C., Dang, T., and Bouchon-Meunier, B. (2005). Fuzzy decision tree for user modeling from human-computer interactions. In *Proceedings of the International conference on human system learning: Who is in control?*, pages 287–302.

[Dekker et al. 2009] Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. In *Proceedings of the International Conference on Educational Data Mining*, pages 41–50.

[D'Mello et al. 2008] D'Mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1-2):45–80.

[Fayyad et al. 1996] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Advances in knowledge discovery and data mining. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *A*, chapter From data mining to knowledge discovery: an overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Feng and Koedinger 2005] Feng, M., H. N. and Koedinger, K. (2005). Looking for sources of error in predicting students' knowledge. In *Educational Data Mining: Papers From the 2005 AAAI Workshop*, pages 54–61, Menlo Park. AAAI Press.

[García 2007] García, E., R. C. V. S. C. T. (2007). Drawbacks and solutions of applying association rule mining in learning management systems. *International Workshop on Applying Data Mining in e-Learning (ADML'07)*, pages 13–22.

[Hall et al. 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[Han and Kamber 2000] Han, J. and Kamber, M. (2000). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Hershkovitz and Nachmias 2008] Hershkovitz, A. and Nachmias, R. (2008). Developing a log-based motivation measuring tool. In de Baker, R. S. J., Barnes, T., and Beck, J. E., editors, *EDM*, pages 226–233. www.educationaldatamining.org.

[Holland 1992] Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press.

[Kay et al. 2006] Kay, J., Maisonneuve, N., Yacef, K., and Reimann, P. (2006). The big five and visualisations of team work activity. In *Proceedings of the 8th international conference on Intelligent Tutoring Systems*, ITS'06, pages 197–206, Berlin, Heidelberg. Springer-Verlag.

[Klösgen and Zytkow 2002] Klösgen, W. and Zytkow, J. M., editors (2002). *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., New York, NY, USA.

[Koedinger et al. 2010a] Koedinger, K. R., Baker, R. S. J., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. (2010a). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, pages 43–55.

[Koedinger et al. 2010b] Koedinger, K. R., Baker, R. S. J., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. (2010b). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, pages 43–55.

[Kotsiantis 2009] Kotsiantis, S. (2009). Educational data mining: a case study for predicting dropout-prone students. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 1(2):101–111.

[Lloyd 1982] Lloyd, S. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129 – 137.

[Mannila et al. 1997] Mannila, H., Toivonen, H., and Inkeri Verkamo, A. (1997). Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289.

[Mierswa et al. 2006] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In Ungar, L., Craven, M., Gunopulos, D., and Eliassi-Rad, T., editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA. ACM.

[Moreno et al. 2012] Moreno, J., Ovalle, D. A., and Vicari, R. M. (2012). A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. *Comput. Educ.*, 58(1):560–569.

[Passos and Goldsmith 2005] Passos, E. L. and Goldsmith, R. (2005). *Data Mining: Um guia Prático*. Campus, 1 edition.

[Quinlan 1993] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Rapid-i 2012] Rapid-i (2012). Disponível em http://rapid-i.com/content/view/181/190/lang,en/. Acessado em 10 de Outubro de 2012.

[Romero et al. 2008] Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187191, 2008. 49 Data Mining 2009*.

[RuleQuest ] RuleQuest. Data mininig tools see5 and c5.0.

[Segaran 2007] Segaran, T. (2007). *Programming collective intelligence*. O'Reilly, first edition.

[Shen 2003] Shen, R., H. P. Y. F. Y. Q. . H. J. (2003). Data mining and case-based reasoning for distance learning. *Journal of Distance Education Technologies*, 1(3):46–58.

[Srikant and Agrawal 1995] Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 407–419, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Srikant and Agrawal 1996] Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, SIGMOD '96, pages 1–12, New York, NY, USA. ACM.

[Superby et al. 2009] Superby, J., Vandamme, J.-P., and Meskens, N. (2009). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pages 37–44.

[Talavera 2004] Talavera, L., G. E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*, pages 17–23.

[Tan et al. 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Tsur et al. 1998] Tsur, D., Ullman, J. D., Abiteboul, S., Clifton, C., Motwani, R., Nestorov, S., and Rosenthal, A. (1998). Query flocks: a generalization of association-rule mining. *SIGMOD Rec.*, 27(2):1–12.

[Witten and Frank 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd edition.

[Xu ] Xu, B., R. M. Understanding teacher users of a digital library service: A clustering approach.