SEGMENTAÇÃO DO CENSO EDUCACIONAL 2000 UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Josir Cardoso Gomes
Faculdades Ibmec/RJ
josir@jsk.com.br

Ariel Levy
Faculdades Ibmec/RJ
ariellevy@superig.com.br

Gerson Lachtermacher
FCE/UERJ – Ibmec-RJ
glachter@uerj.br

Resumo

O presente trabalho visa demonstrar a viabilidade de utilização de *software* livre (***Mining Tools*** e ***Weka***) em aplicações de Mineração de dados (*Data Mining*), especificamente para classificação de dados. Neste estudo de classificação, algumas técnicas de segmentação estarão sendo avaliadas utilizando-se o critério heurístico indicado por Milligan e Cooper (1985). A base de dados em estudo foi obtida a partir dos indicadores do censo demográfico de 2000 e do censo escolar de 2000, publicada no sítio do INEP, que apresenta o IDH - Índice de Desenvolvimento Humano relacionado com os diversos dados gerados pelo censo educacional. Os aplicativos funcionaram de forma satisfatória e os resultados obtidos no processamento indicaram que o IDH não apresentou uma relação forte com os indicadores de desempenho e investimentos em educação.
Palavras-chave: Mineração de Dados, Preparação de Dados, Software Livre.

Abstract

This article tests the viability of the use of Data Mining Free Software (**Weka**), in the process of data classification, also presenting a new developed free software tool for data preparation called (***Mining Tools*****).** The Milligan e Cooper (1985) criteria is used check the performance of several techniques available using the Demographic Census database (2000), the Educational Census (2000).
Key-words: Data Mining, Data Preparation, Free Software.

1. Introdução

As técnicas de mineração de dados são partes integrantes das mais modernas metodologias de aquisição de conhecimento. (HAN, KAMBER, 2001). Utilizam diversos algoritmos computacionais como forma de buscar novos conhecimentos a partir de um grande volume de dados.

A questão inicial que motivou este estudo foi o agrupamento dos municípios brasileiros sob o ponto de vista educacional como forma de encontrar um padrão entre eles, ou seja, encontrar grupos de municípios que sugerissem padrões de desenvolvimento decorrentes da relação investimento e práticas em educação.

A idéia original foi a aplicação de técnicas de segmentação a partir da base do Censo Educacional desenvolvido pelo INEP – Instituto de Pesquisas em Educação Anísio Teixeira e responder 3 perguntas:

1. Pode-se agrupar os municípios de alguma forma para que se possa encontrar características comuns em termos de indicadores de desempenho e investimentos em educação?
2. Municípios localizados dentro do mesmo estado ou dentro da mesma região geográfica têm características semelhantes?
3. Pode-se encontrar alguma relação entre o IDH e os indicadores de desempenho e investimento em educação ?

Assim, na tentativa de responder a estas três questões, pretendeu-se utilizar técnicas de mineração de dados que tratam da segmentação. Estas técnicas buscam reunir os elementos de um conjunto em

grupos denominados segmentos ou *clusters*, cujos elementos formadores são semelhantes segundo suas características (atributos) e onde cada segmento seja o mais distinto possível em relação aos outros segmentos formados.

Para realizar o estudo, optou-se pelo *software* WEKA (WAIKATO University, 2003) a fim de realizar a segmentação e o *software* Mining Tools (GOMES, 2003) desenvolvido especialmente para este estudo para preparação de dados e para o cálculo da performance do processo de segmentação.

O estudo detalhará o uso da rotina automatizada Mining Tools que irá auxiliar a preparação e normalização de dados e abordará seus critérios no processo de avaliação da qualidade dos segmentos obtidos pelo *software* WEKA facilitando a decisão do número de grupos a serem formados.

A decisão de utilizar o WEKA e construir o Mining Tools partiu da dificuldade da utilização de *softwares* comerciais em função de seus custos elevados. Este fato se constitui em uma barreira para que estudantes, instituições de ensino e pequenas e médias empresas possam utilizar, na prática, as técnicas de mineração de dados.

Inicialmente será apresentada uma revisão teórica sobre a área de mineração de dados abordando as técnicas de segmentação utilizadas neste estudo. Ainda durante a revisão teórica, o conceito de IDH e a construção deste índice serão abordados.

A seguir, serão descritos a metodologia utilizada neste estudo e os resultados encontrados. Além disso, será apresentada uma análise da utilização da função criada por Milligan e Cooper (1985) como índice de qualidade da segmentação obtida. Por fim, serão apresentados tópicos para pesquisas futuras.

2. Revisão Bibliográfica

Nesta seção será apresentada uma revisão bibliográfica da área de mineração de dados com especial atenção aos algoritmos e técnicas de segmentação presentes no software Weka.

2.1. Mineração de dados

Nos últimos anos, as técnicas de mineração de dados passaram a se constituir numa das áreas que mais se expandiram em sistema de informação. Estas técnicas são utilizadas na extração do conhecimento a partir de grandes bases de dados.

A grande expansão da área de Mineração de Dados deve-se ao barateamento no armazenamento digital de informações e na expansão do volume de dados armazenados em organizações públicas e privadas (WHESTPHAL, BLAXTON,1998).

Mineração de dados é uma forma de utilização de grandes bases de dados para auxiliar profissionais na tomada de decisão nas mais diversas indústrias, dentre algumas pode-se citar: varejo, manufaturas, telecomunicações, saúde, seguros e logística.

Apesar do gigantismo apresentado pelos dados, estas bases não se apresentam, freqüentemente, com a devida qualidade. Assim, são comuns os casos de dados faltantes, ou de preenchimento inadequado, dificultando a aquisição de conhecimento. Devido a estes problemas a informação contida nas bases de dados são sub-avaliadas e sub-utilizadas devido à dificuldade de acesso e de análise das mesmas. Assim, torna-se necessária uma minuciosa preparação destas bases, minimizando os efeitos destas falhas. Nesta etapa de preparação dos dados se realiza a "limpeza", removendo ou reparando os dados inconsistentes e ruídos. Ainda durante este processo será realizada a integração dos dados quando várias fontes são combinadas. E ,finalmente, o tratamento dos dados faltantes nos registros e as transformações apropriadas para a mineração de dados, tais como: a normalização; agregação. (HAN, KAMBER, 2001).

Além disso, a análise dos resultados de um estudo de mineração de dados requer um especialista da área em estudo para verificação dos resultados, já que o conhecimento apontado pelos resultados pode não ser evidente, ou, em outros casos, a observação pode produzir uma informação óbvia.

Existe uma variedade de técnicas dentro da área de mineração de dados e suas escolhas decorrerão do tipo de dados, das necessidades de descobertas de tendências ou características desejadas ou do tipo de conhecimento que desejamos obter. Neste ponto, é importante salientar que mineração de dados é uma parte do processo pelo qual podemos descobrir conhecimento (WHESTPHAL, BLAXTON,1998).

2.2. Técnicas de Segmentação e sua relevância

A classificação por segmentação é um processo de divisão de um conjunto de dados (ou objetos) em subconjuntos que detenham um significado e que seus elementos apresentem semelhança em seus atributos. Estes conjuntos são denominados *clusters*, partições ou segmentos. Os objetos são segmentados baseados no princípio de maximizar as similaridades intra-classes e minimizar a similaridade inter-classes (HAN, KAMBER, 2001).

Existem muitas técnicas diferentes para a tarefa de classificação: árvores de classificação, indução de regras, redes neurais, técnicas de segmentação baseadas em proximidades de atributos (*KNN – k-nearest neighbor*), algoritmos genéticos, *rough sets*, lógica *fuzzy*, entre outros. Maiores informações sobre estes algoritmos podem ser obtidas em HAN, KAMBER (2001).

Neste estudo, serão aplicadas as técnicas de classificação baseadas em proximidade. Genericamente, o processo se inicia com a determinação aleatória de um determinado número de centróides equivalentes ao número de segmentos pré-definidos. Em geral ao se iniciar a segmentação, não se conhece os grupos, ou seja, não se tem nenhuma classificação prévia e não se sabe quantos grupos podem ser formados.

O principal problema destes métodos está justamente na determinação do número de grupos ou segmentos que existem em uma base de dados. Geralmente, parte-se da experiência de gerentes que determinam um número de classes esperadas. Para se confirmar esta intuição, realizamos um estudo de análise de sensibilidade, aumentando-se e diminuido-se o número de segmentos.

A fase seguinte é a determinação de centróides aleatórios posicionados pelo software, passando a avaliar a proximidade de cada um dos objetos da base de dados em relação a cada um dos centróides e a sua alocação a um dos segmentos representados pelos centróides. Interativamente os centróides são re-alocados e o processo é repetido até que uma determinada condição de parada seja satisfeita.

Basicamente, os algoritmos se diferenciam em função: de como iniciam os cálculos posicionando o centro das classes; pela forma como as interações são realizadas; e pela fórmula como a proximidade dos centróides é determinada. Os principais algoritmos serão discutidos a seguir.

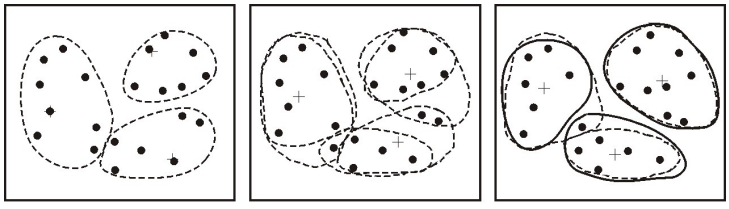
2.2.1. Métodos clássicos de partição: *k-means* e *k-medoids*

É uma classe de algoritmos mais utilizados em aplicativos de mineração de dados. Consistem em obter uma divisão dos dados em k grupos sendo estes grupos representados pelo valor médio (*k-means*) ou pela mediana (*k-medoids*) dos objetos do grupo.

O algoritmo *k-means* tem como parâmetro de entrada, ***k***, e um conjunto de ***n*** objetos que serão classificados em ***k*** segmentos tal que a similaridade entre os elementos do segmento é alta porém a similaridade entre segmentos diferentes é baixa. A similaridade entre os elementos é medida em relação à média apresentada pelos objetos que estão formando o segmento e também pode ser vista como sendo o centro de gravidade do segmento.

De forma randômica o software escolhe ***k*** objetos para representarem os centros dos segmentos, alocando cada um dos segmentos remanescentes a um dos grupos segundo sua similaridade, cujo cálculo podem ter como equações de base a distância ou métrica euclidiana, Manhatan, ou Minkowski (COELHO et al.2003).

Durante as iterações o algoritmo re-posiciona os centros (Figura 1) buscando minimizar a soma das distâncias de seus elementos e assim sucessivamente para todo o conjunto. Aos subgrupos formados denominam-se partições, segmentos ou *clusters*, sendo estes tão compactos geometricamente quanto possíveis para os atributos considerados.



(a) (b) (c)

Figura 1: Segmentação de um conjunto de objetos pelo método *k-means*. Fonte: HAN e KAMBER, 2001

Este método funciona bem na condição de dados compactados em tipo nuvens e que se apresentem bem separadas umas das outras. Apresenta, entretanto a desvantagem de ser necessário poder definir a média dos elementos, o que pode não ser possível em determinados tipos de atributos. Além disso, apresenta a desvantagem de se ter de atribuir um número,k, de grupos como entrada. O método não é indicado para casos em que os segmentos não apresentem a forma convexa, sendo ainda muito prejudicado pela presença de ruídos e dados espúrios, que com certeza irão influenciar as médias. O processo normalmente termina num ponto ótimo local (HAN e KAMBER, 2001), assim para termos um bom resultado seria necessário uma minuciosa investigação variando os pontos de partida.

2.2.2. EM – *Expectation Maximization*

É um algoritmo similar ao *k-means* que gera uma descrição probabilística dos dados em termos de sua média e desvio padrão. Alguns ajustes são necessários na passagem de classes para segmentos, já que este assume uma distribuição de probabilidade, mas não os elementos efetivos de suas participações nas segmentações, assim, utiliza-se estas probabilidades como pesos. A iteração é terminada quando o conjunto de dados distribuídos em classes se afina com os parâmetros de convergência e acuidade descritos mais adiante (”*goodness” of partitioning*”), assim não mais apresenta melhora na qualidade da segmentação (MARKOV,2004).

2.2.3. Métodos de partição hierárquicos

Estes algoritmos trabalham a partir de uma decomposição da base de dados rígida e hierárquica. A abordagem pode ser aglomerativa (de baixo para cima – *bottom up*) ou divisiva (de cima para baixo - *top down*). No caso divisiva, inicia-se com todos os elementos no mesmo grupo, e em partições recursivas, o grupo é dividido em grupos menores, até que determinada condição seja atendida, ou que todos os elementos já esteja devidamente alojados (Figura 2). Estes métodos são prejudicados pelo fato das divisões ou aglomerações serem irreversíveis. Esta rigidez leva a menos custo de computação, por não se preocupar com as possibilidades de combinações de diferentes escolhas. A figura abaixo mostra a forma destas partições e sua lógica:

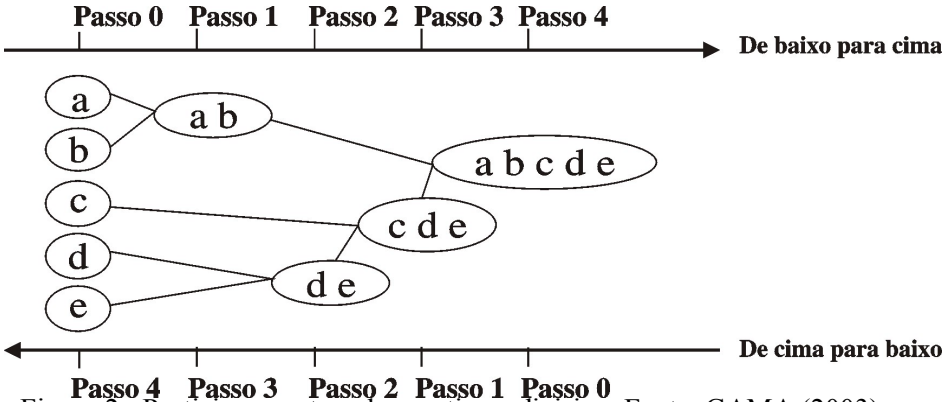


Figura 2: Particionamento aglomerativo e divisivo. Fonte: GAMA (2003)

Apresentam a vantagem de não requerer a determinação prévia do número de segmentos a ser utilizado. Estes algoritmos são comumente usados nas ciências biológicas e sociais desde o século passado sendo vasto o número de trabalhos e referências teóricas (Hartigan *apud* DASGUPTA, 2002). Apesar de permitirem a analise dos dados em vários níveis de sua granularidade simultaneamente, apresentam sua heurística bastante simples (DASGUPTA, 2002).

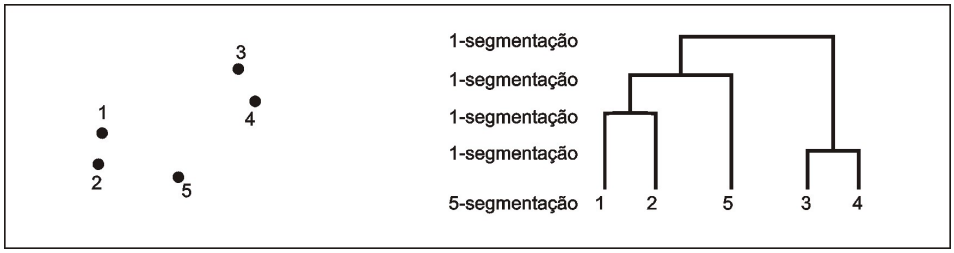


Figura 3: Particionamento hierárquico Fonte: Dasgupta (2002)

No software WEKA, como exemplo destes algoritmos encontra-se o *"Farthest First Traversal Algorithm"* apresentado por Hochbaum e Shmoys em 1985 como uma melhor abordagem na solução dos problemas de alocação de centros *k-means* (DASGUPTA, 2002). A idéia é partir de um ponto qualquer e ao imaginá–lo como um centro, escolher o próximo ponto de centro aquele que apresentar a maior distância e assim proceder sucessivamente até que se obtenha os *k* centros. A partir desta definição os demais elementos são alocados a seus segmentos. Esta solução permite um menor custo de computação(DASGUPTA, 2002).

2.2.4. Métodos de segmentação baseados em modelos

Estes métodos procuram otimizar a alocação dos dados segundo um dado modelo matemático assumindo que os dados são um misto de diversas curvas de distribuições probabilísticas. Assim estes modelos seguem dois tipos de abordagem: uma estatística e outra em pesquisa operacional e redes neurais. Na abordagem estatística, esta baseada numa pré classificação dos dados conforme as observações de suas distribuições, encontra-se características especificas de cada grupo. Ao incorporar a simplicidade e generalidade das características encontradas pelas distribuições probabilísticas este algoritmo agrega qualidade por não depender exclusivamente de cada elemento observado (HAN e KAMBER,2001).

No WEKA o algoritmo que apresenta a abordagem estatística é o COBWEB, que é um método simples de segmentação conceitual e incremental. Os elementos de entrada são descritos por atributos de categoria e dispostos segundo uma árvore de classificação (Figura 4) estando cada nó atrelado a um conjunto de probabilidades condicionais. (HAN e KAMBER,2001; GAMA, 2003).

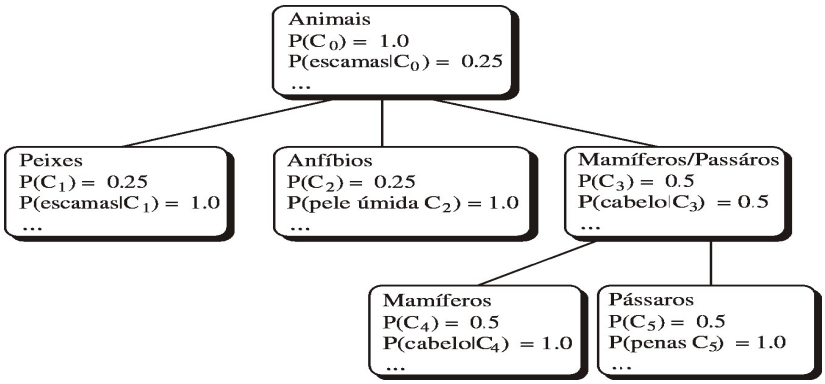


Figura 4: A árvore de classificação. Fonte: Han & Kamber,2001.

Esta árvore difere de uma árvore de decisão, já que cada nó esta associado a um conceito e sua distribuição probabilística, enquanto que na árvore de decisão, os rótulos derivam apenas da lógica. A estratégia de busca é de baixo para cima ou aglomerativa sendo cada elemento visitado apenas uma vez. Portanto, a estrutura pode variar segundo a ordem de visitação, mas os operadores transformarão a estrutura de forma a optar por uma das ações: assinalar o elemento a um nó existente; criar um novo nó; combinar dois nós; dividir um nó em dois.

Durante a classificação, o algoritmo escolherá a ação através da avaliação da função utilidade (HAN e KAMBER,2001; Gama, 2003). Ao gerar uma segmentação hierárquica, onde os "segmentos" são descritos estatisticamente (MARKOV,2003) este algoritmo observa os valores de dois parâmetros,

a saber: Ponto de corte ou convergência denominado de *Cutoff* (padrão=0,002) e Acuidade ou *Acuity* (padrão=1,0). Neste estudo mantiveram-se os valores padrões e as mesmas definições de parâmetros atribuídas durante o estudo com *k-means*. Ainda segundo Markov, quase sempre se encontram diferenças nas segmentações geradas pelos algoritmos *k-means* e Cobweb.

2.3 – A qualidade da segmentação

Um dos maiores problemas dos algoritmos acima citados reside na dificuldade de se encontrar o número ideal de segmentos em que serão divididos os dado. O principal problema de algoritmos baseados em particionamento é a necessidade do parâmetro k, que normalmente é desconhecido. Segundo Kaufman e Rousseeuw (1990, p.38), "é fato que nem todos os valores sugeridos de k levem a partições naturais. A implicação imediata deste fato é que qualquer algoritmo baseado em particionamento normalmente precisa ser executado diversas vezes para diferentes valores de k. O valor mais apropriado para k pode então ser escolhido de acordo com alguma heurística". É fato que, se considerarmos esta heurística de maneira que possa ser quantificada, chega a ser possível escolher o valor ótimo de k de maneira automática: executa-se o algoritmo para todos (ou muitos dos) possíveis valores de k e escolhe-se o que seja mais adequado segundo a heurística. (COELHO et al, 2003).

Esta técnica parte de uma segmentação previamente realizada e qualifica o resultado dos dados segmentados a partir da variância entre os segmentos gerados e a variância interna dos elementos de cada partição.

A função G resume em um índice de qualidade quão distantes (ou dissimilares) estão os segmentos entre si e quão próximos (ou semelhantes) estão os elementos de cada segmento. Assim, quanto maior o resultado da função G, melhor será a qualidade da segmentação realizada. O valor de ***G*** é dado por:

$$G(k) = \frac{(n-k)B}{(k-1)W} \tag{1}$$

onde ***n*** é o número de elementos do conjunto estudado, ***k*** o número de partições gerados, ***B*** é o traço da matriz de covariância entre os segmentos:

$$B = \sum_{i=1}^{I} n_i \sum_{v=1}^{V} (\bar{X}_i^v - \bar{X}^v)^2 \tag{2}$$

onde:

- n_i é o número de elementos de cada partição ***i***
- $\bar{X}_i^v$ é a média dos elementos de cada partição ***i***
- $\bar{X}^v$ é a média de todos os elementos do conjunto.

e ***W*** é o traço da matriz de covariância dos elementos agrupados dentro do mesmo segmentos.

$$W = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{v=1}^{V} (x_{ij}^v - \bar{X}_i^v) \tag{3}$$

Assim, tendo cada objeto ***V*** atributos (ou características) e cada partição i tendo J_i elementos, $x_{ij}{}^v$ representa cada característica ***v*** de cada objeto ***j*** de uma dada partição ***i***.

3. Metodologia

Para um melhor entendimento da base de dados explorada neste estudo serão discutidos a seguir os aspectos dos índices de educação e desenvolvimento humano e em seguida será descrita a metodologia adotada no estudo.

3.1. Censo da educação x IDH

Para exemplificação do uso das ferramentas acima apresentadas, foi utilizada a base do censo da educação disponibilizada pelo INEP (INEP 2003), combinando-a com outras bases disponibilizadas no mesmo sítio, de onde foram obtidos os salários pagos aos professores nos diversos estágios da educação do ensino básico ao ensino médio. Ao estudo foram adicionadas duas colunas referentes aos gastos por aluno por unidade da federação de 1998 e seu percentual médio de crescimento entre 1996 e 2000.

Esta abordagem, buscava relacionar a dotação de verbas em governos anteriores ao atual estágio de desenvolvimento. Na base do INEP já se apresentavam outras características de distribuição populacional e renda de cada município. Assim, supôs-se que com a segmentação desta nova base formada, se poderia obter sucesso na busca de grupos que identificassem tendências ou conhecimentos não tão facilmente observáveis pela simples análise da base.

O IDH - Índice de Desenvolvimento Humano - de certa forma já concentra estas expectativas da área educacional pois seu cálculo é influenciado por três aspectos: Longevidade, Conhecimento e Padrão de Vida. A longevidade é medida pela expectativa de vida no nascimento. O conhecimento é medido por uma combinação da taxa de alfabetização de adultos e uma relação bruta decorrente da população do ensino básico, fundamental, médio e universitário e o padrão de vida é baseado na renda per capta. Esta abordagem fica mais clara na figura abaixo:

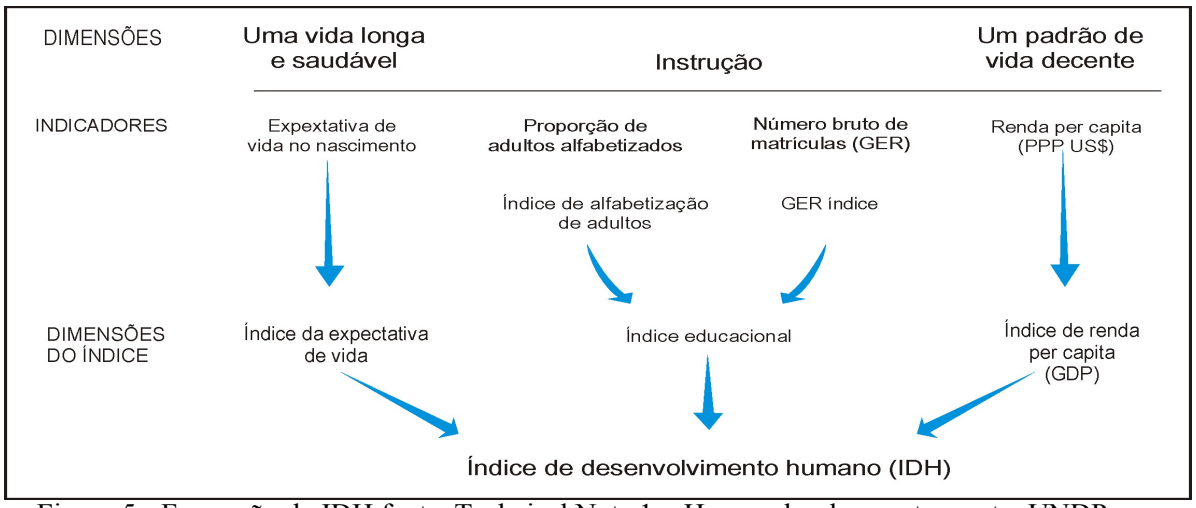


Figura 5 - Formação do IDH fonte: Technical Note 1 – Human development reports- UNDP

3.2. Preparação de Dados

A primeira parte de qualquer trabalho em *Mineração de dados* é a preparação de dados. Esta preparação é necessária para que se possam eliminar dados faltantes ou com erro. No caso da técnica de segmentação é importante, que se normalizem os atributos para que todos fiquem com a mesma ordem de grandeza.

Os dados brutos foram retirados de três planilhas disponibilizadas no sítio do INEP (INEP 2003). A primeira planilha é o próprio Censo Educacional 2000 que contém dados estatísticos e demográficos de todos os 5507 municípios brasileiros. Exceto pelo nome do município, todas as variáveis eram contínuas representando percentuais ou valores absolutos.

Assim, o primeiro processamento foi utilizar o *software* OpenOffice 1.1.0 para abrir a planilha eletrônica e analisar os dados. Após análise inicial, os atributos que representavam números de matrícula, número de docentes e número de estabelecimentos foram substituídos por valores relativos para que não houvesse uma variação muito grande entre cidades pequenas e grandes centros urbanos como São Paulo.

Uma segunda intervenção foi realizada, pois se percebeu que a planilha do Censo não continha dados econômicos, o que tornava a base deficiente em aspectos relevantes tais como o investimento em educação per capta e o salário médio dos professores. Assim, foram mescladas 5 novas colunas a

partir de outras duas planilhas desta vez fornecidas pelo IPEA/DISOC. Ao final do processo tínhamos 61 atributos para cada município.

O próximo passo foi normalizar as colunas para que todas tivessem a mesma ordem de grandeza. A normalização escolhida segundo a equação proposta (HAN e KAMBER, 2001). Esta técnica realiza uma transformação em cada atributo de cada elemento da base de dados a partir da seguinte equação:

$$z_j^v = \frac{x_j^v - \bar{X}^v}{\hat{S}} \tag{4}$$

onde:

$$\hat{S} = \frac{1}{n}\sum_{j=1}^{n}\left(x_j^v - \bar{X}^v\right) \tag{5}$$

isto é, o desvio padrão amostral de cada atributo tal que:

- ***n*** o número de elementos do conjunto
- ***v*** representa o índice de cada atributo, ou seja, equivale a cada coluna quando nos referenciamos à planilha eletrônica;
- $\bar{X}^v$ é média aritmética de um mesmo atributo para todos os elementos e
- x_j^v representa cada elemento da base de dados.

Uma vez tendo a base preparada e normalizada, o próximo passo consistia em executar um *software* de Mineração de dados para montar os partições. Como foi dito anteriormente, os pesquisadores não tinham acesso a nenhum *software* proprietário, assim optou-se pelo WEKA. Este *software* foi desenvolvido na linguagem de programação Java pela Universidade de Weikato da Nova Zelândia (WEIKATO, 2003) e segue a filosofia do *software* livre que "se refere à liberdade dos usuários executarem, copiarem, distribuírem, estudarem, modificarem e aperfeiçoarem o *software*" (STALLMAN, 1996).

Uma das dúvidas no início da pesquisa era verificar se uma ferramenta baseada nesta filosofia de *software* livre estaria ou não pronta o suficiente para realizar experimentos efetivos de mineração de dados. Se houvesse êxito ou não nos processamentos necessários sem perda de qualidade ou de performance para um caso real, uma contribuição importante seria dada a pesquisadores que quisessem utilizar este tipo de ferramenta.

A partir deste ponto, três problemas deveriam ser resolvidos: como realizar a normalização, como importar a planilha em formato XLS (Microsoft Excel) para o *software* Weka, já que este só aceitava o formato ARFF (formato criado pelo Weka para importar os dados) e como automatizar o cálculo da função G (equação 1).

A normalização e o cálculo da função G poderiam ter sido feitos pelo módulo de planilha do Open Office através de macros ou outros recursos disponíveis mas preferiu-se desenvolver um *software* que automatizasse ao máximo o processo. Buscou-se desenvolver um aplicativo que conseguisse:

1. ler a planilha eletrônica sem o auxílio de nenhum *software* proprietário;
2. permitir a seleção das variáveis que seriam normalizadas;
3. realizar a normalização dos dados;
4. Gravar os dados no formato ARFF, formato necessário para que se pudesse interagir com o *software* WEKA.
5. Ler o resultado gerado pelo WEKA para que fosse possível calcular a função G.

O aplicativo foi desenvolvido com a ferramenta Borland Delphi 7 e foi denominado Mining Tools. Seguindo a filosofia de *software* livre, o programa foi disponibilizado na Internet para que qualquer estudante ou pesquisador possa fazer uso dele. Além disso, tanto o nome quanto as mensagens explicativas do *software* foram traduzidos para a língua inglesa para que pesquisadores de outros países também pudessem analisar e fazer uso do *software*.

Em resumo, o seguinte modelo de processamento foi adotado:

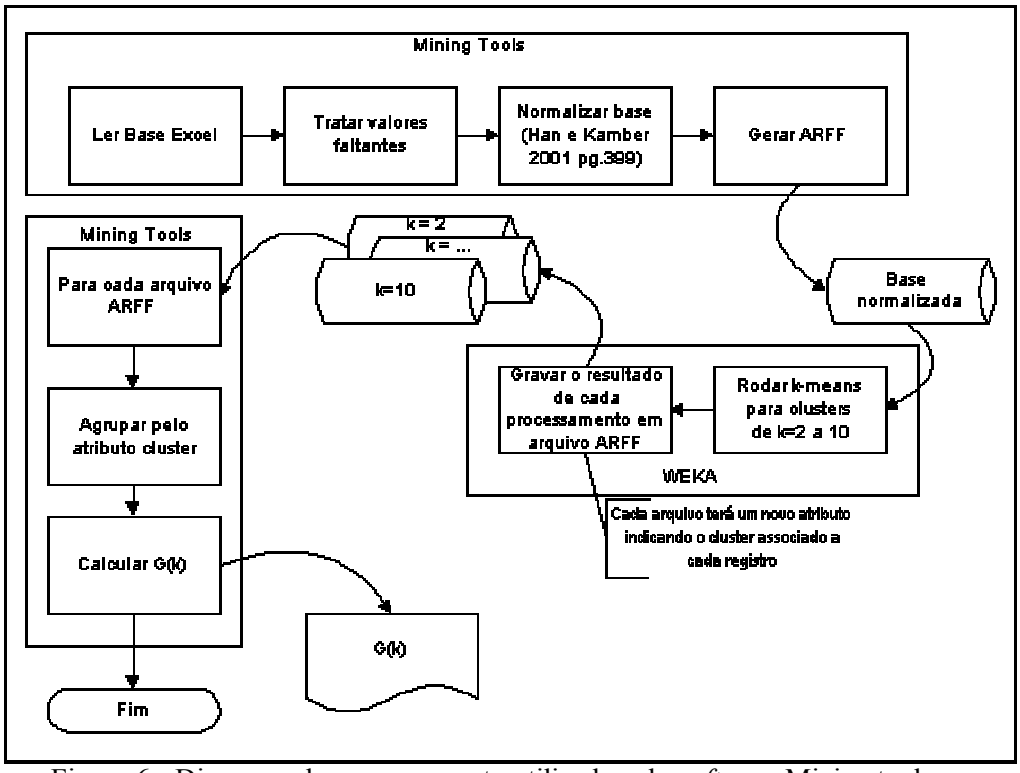


Figura 6 - Diagrama do processamento utilizado pelo *software* Mining tools.

Após o processamento, também foram também foram feitos outros dois experimentos:

- A segmentação natural por unidade da federação ou pela região geográfica teria um resultado melhor que os partições criados pelo algoritmo *k-means*, ou seja, os municípios de cada estado ou de cada região se comportavam da mesma forma ?
- Se fosse feita uma segmentação discretizando os municípios por IDH em 3 e 5 categorias, o índice obtido pela função G seria melhor ?

Inicialmente os resultados obtidos pela segmentação natural por UF, por região e após realizar uma discretização do IDH determinando 2, 3 e 5 partições podem ser vistos no Tabela 1.

k	*G(k)*
UF	**90,91**
Região	**345,39**
IDH(2)	**988,55**
IDH(3)	**512,33**
IDH(5)	**363,89**

Tabela 1: Segmentação por UF, Região e IDH

Ou seja, a melhor segmentação natural seria dividir os municípios em 2 categorias se baseando no índice do IDH. Já os resultados obtidos com o algoritmo *k-means* foram bem próximos aos anteriores:

k	*G(k)*
2	**1166,01**
3	**671,28**
4	**518,95**
5	**440,44**
6	**366,41**
7	**325,14**
8	**289,13**
9	**261,96**
10	**239,99**

Tabela 2: segmentação pelo *k-means*

Além do *k-means*, a versão do Weka utilizada no experimento implementava três outros algoritmos de segmentação: *CobWeb*, *EM* e *Farthest First*. Assim, procurou-se avaliar se outros algoritmos fariam alguma diferença no resultado da segmentação e no cálculo do G(k).

Para responder à segunda questão, o mesmo processamento foi utilizado para os outros algoritmos disponíveis no WEKA e o único algoritmo que mostrou resultados interessantes foi o *Farthest First* (FF). Com ele, o resultado da função G foi bem superior aos outros, como pode-se avaliar na tabela 3:

K	*G(k)*
2	**370,94**
3	**453,09**
4	**350,79**
5	**343,61**
6	**315,84**
7	**276,45**

Tabela 3: Segmentação utilizando algoritmo Farthest First (FF)

Assim, selecionou-se a segmentação realizada pelo algoritmo *k-means* com k=2 para uma análise mais detalhada. Alguns fatos puderam ser constatados:

- Notadamente, o algoritmo separou os municípios que são mais desenvolvidos dos menos desenvolvidos pois somente 3% dos municípios do primeiro segmento tinham um IDH acima da média nacional.
- O primeiro grupo era formado por 2174 municípios e a grande maioria definitivamente era formada por municípios do Norte e Nordeste. Somente 2 municípios do Sul do país (Tunas do Paraná e Laranjal) estavam presentes. Já no Sudeste, quase todos os municípios pertencentes a este

segmento pertenciam ao chamado Vale do Jequitinhonha, área sabidamente de extrema probreza. E do Centro-Oeste somente 9% dos municípios da região apareceram neste grupo.
- O segundo segmento, entretanto, continha elementos em toda a faixa do IDH e continha municípios de todos os estados da federação.

Ao realizar-se a mesma análise para k=3, que seria a segunda melhor segmentação, constatou-se que o segundo segmento do processamento anterior tinha sido dividido em dois e alguns elementos que antes estavam no grupo dos menos desenvolvidos passaram para o novo segmento criado. Da mesma forma, o IDH não teve uma correlação direta com os grupos.

Assim, a partir da análise de segmentação, pode-se concluir que o IDH favorece o estudo do fator educacional nos municípios entretanto, por si só, não consegue categorizar de forma ideal os municípios brasileiros em relação a educação já que, sob o critério da função G, a qualidade da segmentação efetuada ficou bem abaixo da segmentação não-supervisionada do algoritmo FF.

4. Conclusões e considerações finais

Este estudo possibilitou comparações e descobertas interessantes, ainda que passíveis de uma análise mais acurada por especialistas da área educacional, no sentido de melhor extrair as razões e formular as hipóteses para a avaliação do conhecimento gerado. Dentre os resultados do estudo pode-se concluir que a região geográfica e a unidade da federação não são critérios para uniformizar os municípios em relação à educação. O fato de dois municípios do Sul aparecerem no grupo dos menos desenvolvidos, merece um estudo mais detalhado.

Em relação aos algoritmos, verificou-se que o processamento do FF foi bem mais rápido que *k-means* o que condiz coma teoria matemática por trás do algoritmo. Entretanto, à primeira vista, pode-se concluir que o *k-means* é mais preciso e sempre trará um índice G melhor que FF, mas experiências iniciais executadas antes do experimento com outras bases de dados colocaram por terra essa suposição.

Quanto ao uso do WEKA, o *software* realizou as diversas segmentações com rapidez e sem erros. A única deficiência aparente está na documentação que não é de fácil entendimento para profissionais que não conheçam a linguagem de programação Java.

Além disso, o *software* Mining Tools favoreceu o acesso ao WEKA facilitando o tratamento e a preparação dos dados permitindo assim que pesquisadores de outras áreas menos ligadas à informática possam ter uma nova ferramenta de preparação de dados com uma interface mais amigável.

Por fim, pode-se afirmar que o método de Milligan e Cooper (1985) se mostrou uma ferramenta rápida e consistência de avaliação sobre a qualidade da segmentação e o método aqui apresentado pode facilitar o trabalho de outros pesquisadores para problemas similares.

5. Trabalhos futuros

Durante a realização deste trabalho, novas e interessantes questões, puderam ser formuladas ficando aqui como sugestões para futuros trabalhos.

A primeira sugestão seria a utilizar outros *softwares* proprietários de mineração de dados e verificar se os resultados das avaliações através do critério de Milligan e Cooper repetem os resultados obtidos ou se podem trazer novas informações.

Uma outra sugestão seria incorporar outros atributos que não estavam à disposição para este estudo, tais como dados econômicos e de investimento em infra-estrutura além de verificar se os resultados se alteram e se, por exemplo, algum processamento seja gerado com uma qualidade melhor que as encontradas aqui.

Um dos pontos motivadores da pesquisa foi a confecção do software Mining Tools como ferramenta aberta justamente para que, desta forma, outros pesquisadores possam ter a liberdade de utilizar este método em outras bases de dados, validando e melhorando assim os resultados aqui obtidos.

Como última sugestão, seria interessante repetir o experimento com dados de 2003 e verificar se houve alteração entre os municípios e se a escolha entre 2 segmentos iria se manter. O simples fato de municípios trocarem de segmento poderia dar sugestões para que pesquisadores na área de educação avaliassem quais os fatores que fizeram estes municípios melhorarem ou não. Este tipo de estudo

possibilitaria novos conhecimentos sobre as formas e efeitos da administração de recursos destinados à educação.

Referências Bibliográficas

- COELHO, Paulo S.; LACHTERMACHER, Gérson; EBECKEN Nelson F. F.; **Classificação de Dados: uma visão geral**, ENANPAD, 2003.
- COELHO, Paulo S.; ESPENCHITT, Dilson G.; CARVALHO, Júlio L. N.**; Em busca do k ótimo para o algoritmo k-means***; em prelo*
- CRAW, Susan , WIRATUNGA, Nirmalie; **Applications of Datamining - clustering Lab.** Disponível em http://www.scms.rgu.ac.uk/staff/smc/teaching/datamining/clustering/ . Acesso em 13/01/04
- DASGUPTA, Sanjoy. **Performance guarantees for hierarchical partitioning;** In Proceedings of the fifteenth conference on Computational Learning Theory, 2002.
- ??FISHER, Doug; **Knowledge Acquisition Via Incremental Conceptual Clustering;**1987**;** Disponível em http://kiew.cs.uni-dortmund.de:8001/mlnet/instances/81d91eaa-db6c2e1493 Acesso em 31/01/04
- §_____; **Iterative Optimization and Simplification of Hierarchical Clusterings;** 1996; Disponível em http://www-2.cs.cmu.edu/afs/cs/project/jair/pub/volume4/fisher96a-html/html-final.html Acesso em 31/01/04
- GAMA, João; **Métodos de Agrupamento – Clustering.** Disponível em http://www.liacc.up.pt/~pbrazdil/Ensino/ecdas/cluster.pdf . Acesso em 08/12/2003.
- GOMES, J. C. ; **Mine Tools – Utilitário para normalização de dados e interface com o Weka**; Disponível em http://groups.yahoo.com/group/datamining_ibmec/filesAcesso em 08/12/2003.
- HAN, Jiawei ; KAMBER, Micheline; **Data Mining- Concepts and Techniques;** Londres, Academic Press, 2001.
- INEP, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira; I**ndicadores do Censo Demográfico de 2000 e do Censo Escolar de 2000**. Disponível em http://groups.yahoo.com/group/datamining_ibmec/files/bases . Acesso em 17/11/2003.
- KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: an Introduction to Cluster Analysis.** John Wiley & Sons, 1990.
- MARKOV, ZDRAVKO; **Datamining;** Disponível em http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-10.pdf acesso em 12/01/04
- MOBASHER,BAMSHAD; **Mineração de dados Techniques: Classification and Clustering.** Disponível em http://maya.cs.depaul.edu/~classes/cs589/lectures/lecture3. Acesso em 4/12/2003.
- STALLMAN, RICHARD; **The Free Software Definition.** Disponível em http://www.gnu.org/philosophy/free-sw.html. Acesso em 4/12/2003
- UNITED NATIONS; **Human Development Reports 2002.** Disponível em http://hdr.undp.org/reports/global/2002/en/pdf/backtwo.pdf . Acesso em 04/12/2003
- WAIKATO UNIVERSITY; **Weka 3: Machine Learning Software in Java.** Disponível em http://www.cs.waikato.ac.nz/ml/weka/ Acesso em 04/12/2003
- WHESTPHAL, Christopher BLAXTON, Teresa; **Mineração de dados Solutions: Methods and Tools for Solving Real –World Problems**; NewYork; John Wiley & Sons, 1998.