USP - UNIVERSIDADE DE SÃO PAULO
IME - INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DCC - DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

TRATAMENTO DE DADOS AUSENTES PARA ANÁLISE EM GRANDE ESCALA

CLODIS BOSCARIOLI

SÃO PAULO, 2003

CLODIS BOSCARIOLI

TRATAMENTO DE DADOS AUSENTES PARA ANÁLISE EM GRANDE ESCALA

Relatório de estudos apresentado como requisito de avaliação da disciplina MAC-5701 - Tópicos em Ciência da Computação, desenvolvido sob orientação do Professor Dr. Roberto Marcondes Cesar Jr.

Professor Responsável: Dr. Yoshiharu Kohayakawa

SÃO PAULO, 2003

RESUMO

As bases de dados utilizadas para mineração de dados/reconhecimento de padrões são, na grande maioria das vezes, imcompletas e/ou inconsistentes, necessitando passar por um processo de preprocessamento antes da descoberta de padrões propriamente dita.

No pré-processamento, o tratamento de valores ausentes de atributos é uma questão importante para o aprendizado por classificação, uma vez que a ausência de valores pode comprometer a precisão do classificador aprendido.

Neste contexto, o presente trabalho considera o problema de aprendizado quando há ausência de valores nos conjuntos de dados a serem analisados, descrevendo as abordagens e técnicas para tratamento dos mesmos, de forma a obter um conjunto de dados completo (pré-processado), para classificação.

Sumário

Lista de Figuras

Capítulo 1

Introdução

A aplicação de técnicas de mineração de dados em banco de dados é motivada pela crescente necessidade de descobrir novas informações em bancos de dados existentes, na forma de regras ou padrões. Porém, a integração de Banco de Dados e Data Mining não é trivial e várias são as dificuldades inerentes a esse processo, seja na preparação dos dados, na definição dos operadores, na análise do algoritmo a ser adotado ou, mesmo, na interpretação dos conhecimentos gerados. Dentre todas as etapas, é de particular interesse neste estudo a preparação dos dados, na qual são despendidos esforços para garantir a qualidade aos dados, fundamental para uma mineração eficiente.

Os conjuntos de dados do mundo real contém erros e inconsistências. A etapa de pré-processamento é de suma importância, haja vista a qualidade dos dados de entrada incidirem diretamente na qualidade do conhecimento gerado como saída em um processo de descoberta de conhecimento.

Um grande problema encontrado é a ausência de dados nas bases de dados a serem analisadas. A ausência de valores - *missing data* (também conhecido na literatura por valores desconhecidos, dados faltantes ou censurados) - pode ocorrer por uma variedade de motivos, tais como, omissões no processo de entrada dos dados, problemas no processo de conversão entre bases de dados, falhas em mecanismos de leitura, entre outros [14].

Em problemas práticos, não é difícil verificar casos em que o número de dados disponíveis não é o mesmo para todos os atributos *(features)*. Segundo [13], se o número de dados de treinamento for grande o suficiente, pode-se rejeitar alguns deles, mantendo um número menor, o mesmo para todas os atributos. Contudo, nem sempre esta solução é possível, e, para estes casos, os valores ausentes devem ser preditos heuristicamente.

Esta monografia examina o problema de aprendizado quando há ausência de valores nos conjuntos de dados a serem analisados, bem como apresenta várias técnicas para tratamento dos mesmos, estando organizada da seguinte forma:

O Capítulo 2 introduz a etapa de pré-processamento de dados para aprendizado computacional, contextualizando o restante desta pesquisa.

O Capítulo 3 descreve o problema da ausência de dados, abordagens e técnicas para tratamento dos mesmos.

Finalmente, o Capítulo 4 traz as considerações finais sobre a pesquisa por ora realizada, bem como as propostas de trabalhos futuros.

Capítulo

2

Pré-processamento de Dados para Aprendizado Eficiente

Descoberta de Conhecimento em Banco de Dados (KDD) é um processo não trivial que objetiva identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes [4]. Não é porém, sumariação de dados, que pode ser obtida por ferramentas OLAP (*On-line Analytical Processing*), mais sim a obtenção de novos conhecimentos, implícitos na co-relação entre os dados, para auxiliar na tomada de decisões.

A Figura 2.1 ilustra as fases do processo de KDD, onde as fases de seleção dos dados, pré-processamento, limpeza e transformação dos dados estão diretamente ligadas ao problema de tratar anomalias presentes nos dados. A etapa da mineração consiste na aplicação de técnicas de estatística e inteligência artificial em grandes bases de dados, para encontrar tendência ou padrões nos dados, para posterior interpretação por especialistas do domínio.

O pré-processamento dos dados é um processo de KDD semi-automático, freqüentemente tido como sendo uma fase que envolve grande quantidade de conhecimento do domínio. Por semi-automático entende-se a capacidade do analista de dados em identificar

Figura 2.1: Etapas de um Processo de KDD típico, [4].

os problemas presentes nos dados, além da natureza desses, e utilizar os métodos mais apropriados para solucioná-los.

Preparação de dados é utilizada atualmente por analistas de dados para direcionar a descoberta de conhecimento com qualidade e auxiliar no desenvolvimento de sistemas de aplicações em análise de dados efetivos e com bom desempenho. Um dos objetivos da fase de pré-processamento é solucionar problemas nos dados, tais como identificar e tratar dados corrompidos, retirar atributos irrelevantes e tratar valores desconhecidos.

A importância de pré-processamento pode ser destacada por três aspectos fundamentais [15]:

- Os dados do mundo real são impuros;
- Sistemas de data mining de alto desempenho requerem qualidade nos dados;
- Qualidade nos dados de entrada geram qualidade nos padrões obtidos.

Alguns problemas encontrados nos dados extraídos dos sistemas de banco de dados, que evidenciam a necessidade de pré-processamento, descritos na literatura, citados em [1]:

- Grande quantidade de ruídos e inconsistências;
- Excesso de valores desconhecidos;
- Classes desbalanceadas (grande desproporção entre as distribuições das classes).

Desta forma, as principais tarefas realizadas na etapa de pré-processamento são:

- Identificação de inconsistências;

- Identificação de ruídos;
- Verificação de integridade;
- Identificação de atributos duplicados e redundantes;
- Identificação de valores *defaults*;
- Tratamento de valores ausentes;
- Identificação e descrição de valores extremos *(outliers)*;
- Tratamento de conjuntos de dados com classes desbalanceadas;
- Seleção de atributos relevantes.

Dentre todas as tarefas suscitadas acima, o foco deste trabalho, o qual é assunto dos próximos capítulos, é o tratamento de valores ausentes.

Capítulo

3

Tratamento de Dados Ausentes

Este capítulo abordará o problema de dados ausentes em um conjunto de dados, descrevendo as formas de aleatoriedade possíveis dessa ausência, os métodos de tratamento, bem como apresentando as principais técnicas utilizadas para predizer valores desconhecidos para um processo de aprendizagem confiável.

3.1 Descrição do Problema

Um problema relevante na qualidade dos dados de entrada é a ausência de dados, que consiste na não medição de valores de um ou mais atributos para alguns casos.

Apesar da freqüência de dados faltantes em conjuntos de dados, estes são tratados de forma simplista. Contudo, o tratamento de dados faltantes deve ser cuidadosamente pensado para que não sejam introduzidas distorções no conhecimento induzido como saída.

Muitos pacotes de software estatísticos permitem a identificação de dados ausentes, criando um ou mais códigos especiais para preencher, na matriz de dados, os não observados, tais como "não conhecido", "sem resposta". Alguns destes pacotes tipicamente excluem unidades que tenham valores desconhecidos, para quaisquer das variáveis envolvidas em uma análise. Esta estratégia, denominada "análise de casos completos" é, em geral, inapropriada, uma vez que o pesquisador normalmente está interessado em fazer inferências sobre a toda a população alvo [11].

Na maioria dos casos, os atributos de um conjunto de dados não são independentes entre si. Dessa forma, aproximações para os valores desconhecidos podem ser determinadas por meio da identificação de relações entre os atributos. Imputação é um termo utilizado para denotar um procedimento que substitui os valores desconhecidos de um conjunto de dados por valores estimados. Essa abordagem permite que o tratamento de valores desconhecidos seja independente do algoritmo de aprendizado utilizado, o que permite ao analista de dados selecionar o método de tratamento de valores desconhecidos mais apropriado para cada conjunto de dados [1].

De acordo com [12], dados ausentes podem ser preditos usando uma co-relação com outros atributos, isto é, alguma regra é descoberta entre o atributo com valor ausente e outros atributos. Exemplos usando regras para predizer valores:

(a) Se o salário de uma pessoa, que é gerente, está ausente: Observando os salários dos demais gerentes no banco de dados, pode-se ter um indício do valor ausente.

(b) Outro exemplo seria a predição de vendas para um período particular baseado em históricos.

Em [1] há algumas questões que podem ser levantadas no tratamento de dados ausentes:

- Como os valores desconhecidos de um conjunto de dados devem ser tratados?
- O método de imputação escolhido é efetivo para o tratamento desejado?
- O método de tratamento é efetivo mesmo em grandes quantidades de valores desconhecidos?
- Quais as limitações do método de imputação adotado?

Segundo [1], a imputação de dados ausentes pode ser danosa, uma vez que o mais avançado método de imputação somente é capaz de aproximar os valores reais, não conhecidos, dos valores desconhecidos. Os valores preditos são, geralmente, mais bem comportados, uma vez que são preditos em conformidade com os valores dos outros atributos. Nos experimentos conduzidos, quanto mais atributos com valores desconhecidos são introduzidos, e quanto maior a quantidade de valores desconhecidos, mais simples são os classificadores induzidos. Dessa forma, a imputação de valores desconhecidos deve ser cuidadosamente aplicada, sob o risco de simplificar demasiadamente o problema em estudo.

3.2 Padrões de Ausência de Valores

É importante distinguir padrões de ausência de valores, que descrevem quais valores são observados na matriz de dados e que valores estão ausentes, bem como conhecer mecanismos de tratamento, que consideram o tipo de relacionamento existente entre a ausência e os valores das variáveis na matriz de dados.

Seja $Y = (y_{ij})$ uma matriz sem dados ausentes, com linhas $y_i = (y_{i1}, ..., y_{iK})$ onde y_{ij} é o valor da variável Y_j para um i qualquer. Para valores ausentes, defina $M = (m_{ij})$, como sendo uma matriz indicadora de dados ausentes, tal que $m_{ij} = 1$ se y_{ij} é desconhecido e m_{ij} é 0 se y_{ij} está presente. A matriz M então define um padrão de ausência de valores. Alguns métodos para tratamento de dados ausentes aplicam-se a qualquer padrão de ausência de dados, entretanto, outros estão restritos a um padrão em especial.

A Figura 3.1, descrita em [11] apresenta a seguinte classificação para padrões de ausência de valores em um conjunto de dados:

- *Univariate Missing Data* (Quando a falta de valores está confinada a uma única variável)
- *Multivariate Two Patterns* (Quando há, para um conjunto de variáveis $Y_{j+1}, ..., Y_k$, todos os dados observados ou ausentes, para o mesmo conjunto de casos)
- *Monotone*
- *General* (Quando há valores ausentes nos atributos do conjunto de dados, de forma não uniforme)
- *File Matching* (Ocorre quando, na junção de conjuntos de dados, haja variáveis que não puderam ser observadas de forma conjunta)
- *Factor Analysis* (Situação onde os dados de uma variável são totalmente desconhecidos)

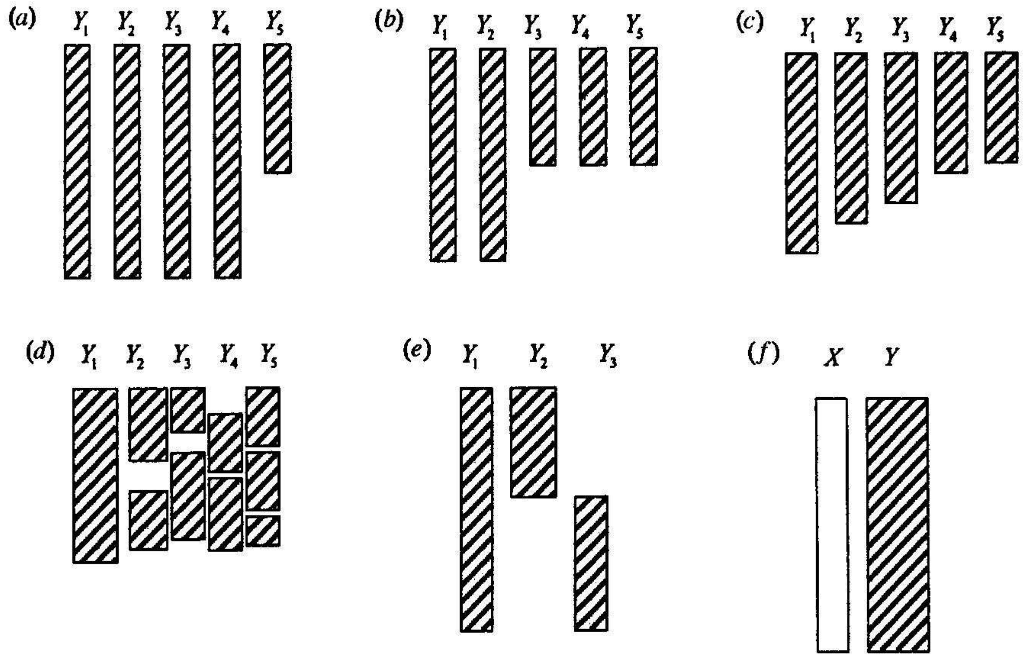

Figura 3.1: As linhas correspondem às observações e as colunas às variáveis. (a) *Univariate Missing Data*; (b) *Multivariate Two Patterns*; (c) *Monotone*; (d) *General*; (e) *File Matching*; (f) *Factor Analysis*
Figura adaptada de [11].

3.3 A Aleatoriedade dos Valores Ausentes

A aleatoriedade dos valores desconhecidos é um fator importante a ser analisado para a escolha do método de tratamento de valores desconhecidos. Na sua forma mais simples, os valores ausentes podem estar distribuídos aleatoriamente nos dados, o que significa que a probabilidade de encontrar um valor desconhecido é a mesma para qualquer valor do atributo.

Por outro lado, os valores desconhecidos podem estar não aleatoriamente distribuídos, o que implica que a probabilidade de encontrar um valor desconhecido pode depender, por exemplo, do valor verdadeiro (não conhecido), do valor ausente.

Segundo [11], o grau de aleatoriedade pode ser classificado como:

1. **MCAR (*Missing Completely at Random*) - Ausentes de forma completamente aleatória**

Seguindo a definição da seção 3.2, o mecanismo de ausência de valores é caracterizado pela distribuição condicional de M dado Y, dito $f(M|Y,\phi)$, onde ϕ denota parâmetros não conhecidos. Se a ausência não depender dos valores de dados de Y, ausentes ou observados, isto é, se

$$f(M|Y,\phi) = f(M|\phi)\; para\, todo\, Y, \phi, \tag{3.1}$$

os dados são considerados MCAR. Note que esta assunção não significa que o padrão é aleatório por si próprio, mas que a ausência de valores não depende dos valores dos dados.

Valores desconhecidos dispostos nesta forma são os que possuem o mais alto grau de aleatoriedade.

É uma suposição muito restrita, a qual dificilmente pode ser encontrada em problemas práticos [1].

2. **MAR (*Missing at Random*) - Ausentes de forma aleatória**

   Seja Y_{obs} os valores observados de Y e seja Y_{mis} os valores ausentes. Uma assunção menos restritiva que MCAR é que a ausência de valor depende somente dos componentes Y_{obs} e não dos compontentes que estão ausentes. Isto é,

$$f(M|Y,\phi) = f(M|Y_{obs},\phi)\; para\, todo\, Y_{mis}, \phi \tag{3.2}$$

   Este mecanismo é chamado de MAR.

   Valores ausentes assim dispostos ocorrem quando os valores conhecidos de um atributo não estão relacionados com os valores desse atributo, mas dependem dos valores de um ou mais atributos do conjunto de dados.

3. **NMAR (*Not Missing at Random*) - Ausentes de forma não aleatória**

   O mecanismo é considerado NMAR se a distribuição de M depende dos valores ausentes na matriz de dados Y.

   Valores ausentes dispostos de forma não aleatória ocorrem quando a probabilidade de um exemplo possuir um valor desconhecido para um atributo depende dos valores do atributo com valores desconhecidos, e possivelmente, dos valores dos demais atributos do conjunto de dados.

3.4 Uma Taxonomia para Tratamento de Valores Ausentes

Os seguintes procedimentos podem ser adotados como alternativa, ao deparar-se com uma base de dados com valores ausentes:

1. **Ignorar ou descartar dados**

   Existem duas abordagens mais utilizadas para descartar dados com valores desconhecidos. A primeira é conhecida como análise de casos completos, que descarta qualquer caso que possua um ou mais valores desconhecidos, é o método default na maioria dos softwares estatísticos. Já a segunda abordagem, conhecida como descarte de casos e/ou atributos, consiste em determinar a extensão dos valores desconhecidos em cada exemplo e em cada atributo, e remover os exemplos e/ou os atributos com grandes quantidades de valores desconhecidos.

   Ambas as abordagens devem ser aplicadas somente quando os valores desconhecidos estão aleatoriamente distribuídos, caso contrário, podem introduzir distorções nos dados.

2. **Estimativa de Parâmetros**

   Procedimentos ML - *Maximum Likelihood* - são utilizados para estimar os parâmetros de um modelo definido para os valores observados nos dados. Procedimentos que utilizam variações do algoritmo EM - *Expectation-Maximization* - [3] podem estimar parâmetros de um modelo na presença de valores ausentes.

3. **Imputação de Dados**

   Classe de procedimentos que visa substituir os valores ausentes por valores estimados. Os métodos pertencentes a essa classe substituem valores desconhecidos por valores estimados por meio de alguma informação extraída do conjunto de dados. Os métodos de imputação podem ser classificadas em:

   (a) **Substituição de casos**

   Método que faz substituição de elementos da amostra. Tipicamente usado em pesquisas de opinião.

   (b) **Imputação pela média/moda**

Método simples, comumente utilizado. Consiste em substituir, por meio dos valores observados, os valores desconhecidos de um certo atributo pela média (para atributos quantitativos), ou pela moda (em atributos qualitativos).

É considerada a melhor estimativa para o valor ausente de um atributo, na ausência de outras informações a respeito dos dados.

(c) **Conhecimento do domínio**

Substituição manual dos valores ausentes por valores inseridos a partir do conhecimento de um especialista do domínio.

Somente é seguro quando o especialista está familiarizado com a aplicação, o conjunto de dados é grande e o número de valores ausentes é pequeno.

(d) **Hot Deck e Cold Deck**

No método Hot Deck, um valor desconhecido é substituído por um valor obtido a partir de uma distribuição estimada pelos dados disponíveis. Tipicamente implementado em dois estágios: no primeiro, o conjunto de dados é particionado em clusters. Após, no segundo estágio, cada exemplo com valores desconhecidos é associado a um dos clusters. Os exemplos completos no cluster são utilizados para estimar os valores desconhecidos. Uma forma de estimar os valores desconhecidos é por meio do cálculo da média ou da moda do atributo, utilizando somente os exemplos membros do cluster.

O método Cold Deck é similar ao Hot Deck.

(e) **Modelo de Predição**

Procedimentos sofisticados para tratamento de dados ausentes que consistem na criação de um modelo preditivo para estimar valores que irão substituir os valores ausentes. O atributo com valores ausentes é utilizado como atributo classe, e os demais atributos são utilizados como entrada para o modelo de predição. Um importante argumento a favor dessa abordagem é que, freqüentemente, os atributos possuem correlações entre si, que podem ser utilizadas para criar um modelo preditivo de classificação ou regressão dependendo se o atributo com valores desconhecidos for qualitativo ou quantitativo, respectivamente.

Limitações:

- Os valores estimados são, geralmente, mais bem comportados que os valores reais, ou seja, são mais consistentes com o conjunto de atributos do que os valores reais seriam.

- A necessidade de correlação entre os atributos.

3.5 Algumas Técnicas Utilizadas no Tratamento de Valores Ausentes

Esta seção objetiva a descrever as principais técnicas que se predispõe a tratar ausência de dados dentre as encontradas, durante esta pesquisa, na literatura.

1. **Substituição pela Média**

   Duas são as técnicas de susbstituição de valores ausentes pela média [8]: *Item Mean Substitution* (IMS) e *Person Mean Substitution* (PMS).

   A abordagem IMS substitui o valor ausente pela média para um atributo particular, calculado por meio de todas as tuplas que contém valor de resposta para este atributo. Já a abordagem PMS substitui o valor ausente com a média de respostas para outros atributos de uma tupla em particular. Esta segunda abordagem pressupõe que os atributos são relacionados, de forma que os valores constantes nos demais atributos são indicadores representativos dos valores potenciais para os dados ausentes.

   Estas duas técnicas são derivadas, das seguintes fórmulas:

$$PMS_i = \frac{\sum_{i=1}^{k} X_i k}{(k - m_i)} \tag{3.3}$$

$$IMS_k = \frac{\sum_{i=1}^{i} X_i k}{(i - m_i)} \tag{3.4}$$

   onde: X_{ik} é uma contagem para a i-ésima tupla, no k-ésimo item; i é o número de tuplas; k o número de atributos; m_i, o número de valores ausentes para os atributos da tupla i; e m_k o número de tuplas com valores ausentes k.

Segundo Downey and King (1995) **apud** [8] sugere que, se o interesse for em estimar confiabilidade, a abordagem PMS foi melhor empregada. Entretanto, se o interesse é meramente checar o relacionamento entre as variáveis, tanto PMS quanto IMS podem ser utilizados, de forma bem aceitável.

2. **Algoritmo C4.5**

   Algoritmo de aprendizado simbólico que induz concceitos por meio de árvores de decisão [9].

   Possui um efetivo algoritmo interno para tratar valores desconhecidos [7].

   Utiliza uma abordagem probabilística para tratar valores ausentes, que podem estar presentes em qualquer atributo, exceto no atributo classe, tanto no conjunto de treinamento quanto no conjunto de teste.

   Dado um conjunto de exemplos de treinamento T, o sistema C4.5 encontra um teste apropriado, com base em um único atributo, que possua um ou mais resultados mutuamente exclusivos $R_1, R_2, ..., R_s$. T é particionado em sub-conjuntos $T_1, T_2, ..., T_s$, tal que T_i contém todos os exemplos em T que satisfazem o teste com resultado R_i. O mesmo algortimo é aplicado para cada sub-conjunto $T_i, i = 1...s$, até que um critério de parada seja satisfeito.

   O C4.5 utiliza a medida "*information gain ratio*" para escolher um bom teste para particionar os exemplos. Se existem valores ausentes em um atributo A, então o sistema C4.5 utiliza o sub-conjunto com todos os valores conhecidos de A para calcular o ganho de informação.

   Uma vez que um teste com base em um atributo A tenha sido escolhido, o sistema C4.5 utiliza uma abordagem probabilística para particionar os exemplos com valores desconhecidos no atributo A. Quando um exemplo em T com valor conhecido é associado a um sub-conjunto T_i, isso pode ser visto como a probabilidade desse exemplo pertencer a T_i ser 1, e para todas as demais partições a probabilidade é 0. Quando um valor não é conhecido, então somente uma declaração probabilística mais fraca pode ser feita. O C4.5 associa a cada exemplo em T_i um peso representando a probabilidade desse exemplo pertencer a T_i. O peso para a partição T_i é a probabilidade desse exemplo pertencer a T_i. Essa probabilidade é estimada por meio da soma dos pesos dos exemplos em T que satisfazem o teste com resultado

R_i, dividido pela soma de pesos dos casos em T com valores conhecidos para o atributo A.

3. **Algoritmo CN2**

   Algoritmo de aprendizado simbólico que induz por meio de conjuntos de regras [2].

   Utiliza abordagem similar à adotada pelo sistema C4.5 para tratamento de dados ausentes. Induz regras da forma

   If $< condicoes >$ then $< classe = C_i > [\#C_1, \#C_2, ..., \#C_N]$

   na qual $< condicoes >$ é uma conjunção de condições que envolvem os atributos do conjunto de dados, e $< classe = C_i >$ é a predição realizada pela regra. O número entre colchetes, $\#C_j$, indica o número de exemplos do conjunto de treinamento cobertos pela regra para cada classe C_j.

   Esses valores podem ser utilizados pelo CN2 para realizar uma classificação probabilística, caso o usuário deseje. Nesse caso, o CN2 responde com as probabilidades de um novo exemplo pertencer a cada uma das classes. Também, esses valores são utilizados pelo algoritmo que avalia regras não ordenadas para verificar qual é a classe mais provável com base em todas as regras que foram disparadas para um novo exemplo. Se mais de uma classe for predita pelas regras disparadas, então os valores $\#C_j$ são somados separadamente para cada classe utilizando somente as regras disparadas. O CN2 responde com a classe mais provável, isto é, aquela que possui o maior valor total de exemplos cobertos pelo conjunto de todas as regras disparadas.

   No caso da indução de regras não ordenadas pelo CN2 se o conjunto de treinamento possui exemplos com valores desconhecidos, então o indutor, após a indução do conjunto de regras, realiza uma segunda passagem sobre o conjunto de regras atualizando os valores $\#C_j$. Para cada exemplo com valores desconhecidos, o CN2 verifica quais regras cobrem esse exemplo. Para que uma regra cubra um exemplo com valores desconhecidos basta este satisfaça as condições da regra e, caso a regra possua condições envolvendo atributos cujos valores não sejam conhecidos, então estas são automaticamente satisfeitas. O exemplo com valores desconhecidos é, então, dividido em frações iguais, as quais são contabilizadas nos valores $\#C_j$.

4. **K-vizinhos mais Próximos com Pesos**

É proposto em [1] a utilização do algoritmo K-vizinhos mais próximos para estimar e substituir valores desconhecidos.

Vantagens:

- Pode ser utilizado tanto para predizer atributos qualitativos, por meio do cálculo da moda entre os K-vizinhos mais próximos, quanto atributos quantitativos, por meio da média dos K-vizinhos mais próximos.

- Não é necessário criar um modelo preditivo para cada atributo com valores desconhecidos, podem ser facilmente adaptados para utilizar qualquer atributo com atributo classe. Essa adaptação é feita por meio da modificação de quais atributos devem ser considerados na função de distância.

- O algoritmo K-vizinhos mais próximos pode ser utilizado mesmo quando existem valores desconhecidos em diversos atributos. Nessa situação, o algoritmo utiliza os valores conhecidos para predizer os ausentes.

Limitações:

- Alto custo de classificação de novos exemplos.

- A busca realizada por esse algoritmo para encontrar os exemplos mais similares necessita de uma passagem completa pelos dados.

5. **Algoritmo EM** (*Expectation-Maximization*)

   A maximização da função de verossimilhança pode ser concretizada por métodos numéricos tais como Newton-Raphson ou scoring de Fisher, mas requerem geralmente um trabalho analítico pesado. Além disso, a implementação destes métodos pode apresentar dificuldades numéricas (requisitos de memória computacional, falta de convergência, instabilidade, etc...), particularmente quando o número de parâmetros é grande. Uma alternativa para o método numérico direto de maximização da função de verossimilhança é conhecido como algoritmo de Esperança-Maximização ou EM [3]. O algoritmo EM é um método geral de estimação de máxima verossimilhançaa para problemas com dados incompletos [5]. Aqui, será descrito, de forma suscinta, o algoritmo EM básico, baseado em [10], [11] e [6] para encontrar parâmetros de máxima verossimilhança em um modelo de dados. Assume-se que os dados $X = \{x_1, ..., x_n\}$ são gerados independentemente da densidade:

$$P(x_i) = \sum_{j=1}^{M} P(x_i \mid \omega_j; \theta_j) P(\omega_j), \tag{3.5}$$

onde cada componente é denotado por ω_j parametrizado por θ_j. Da equação acima, e da assunção de independência, vê-se que a log-verossimilhança dos parâmetros dado um conjunto de dados é

$$\ell(\theta \mid X) = \sum_{i=1}^{N} log \sum_{j=1}^{M} P(x_i \mid \omega_j; \theta_j) P(\omega_j), \tag{3.6}$$

Usando um indicador de variável z, que indica que dados foram gerados para quais componentes e, desta forma, uma função log-verossimilhança para dados completos pode ser escrita como

$$\ell(\theta \mid X, Z) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_(i, j) \; log \; P(x_i \mid z_i; \theta) P(\omega_i; \theta), \tag{3.7}$$

que não envolve um somatório de log.

Uma vez que z não é conhecido, ℓ_c não pode ser utilizado diretamente, sendo necessário trabalhar com a esperança, denotada por $Q(\theta \mid \theta_k)$. De acordo com [3], ela pode ser maximizada iterando as duas etapas seguintes:

$$Etapa\ E : Q(\theta \mid \theta_k = E[\ell_c(\theta \mid X, Z) \mid X, \theta_k]$$

$$Etapa\ M : \theta_k + 1 = \arg\max_{\theta} \; Q(\theta \mid \theta_k) \tag{3.8}$$

A etapa E (esperança) computa a log-verossimilhança esperada dos dados completos e a etapa M encontra os parâmetros que maximizam esta verossimilhança.

Para aplicar este procedimento em dados imcompletos, divide-se o conjunto de dados $X = \{x_1, ..., x_n\}$ em componentes observados X^o e componentes ausentes X^m. De forma análoga, cada vetor x_i é dividido em (x_i^o, x_i^m) onde cada vetor de dados pode ter diferentes componentes ausentes - denotados pelos sobrescritos m_i e o_i.

Para tratar dados ausentes, basta reescrever o algoritmo EM como segue:

$$Etapa\ E : Q(\theta \mid \theta_k = E[\ell_c(\theta \mid X^o, X^m, Z) \mid X^o, \theta_k]$$

$$Etapa\ M : \theta_k + 1 = \arg\max_{\theta}\ Q(\theta \mid \theta_k) \tag{3.9}$$

Comparando esta última equação com a 3.8, vê-se que o indicador de variável Z foi adicionado de X^m, que corresponde aos dados ausentes no conjunto de dados. A etapa E do algoritmo estima os dados ausentes, usando essencialmente a estimativa atual para completar os valores ausentes.

Finalizando, cabe salientar que várias outras técnicas que podem ser aplicadas a conjuntos com dados ausentes foram levantadas nesta pesquisa, e ainda estão sendo estudadas.

Verificou-se que, para a escolha da aplicação de uma técnica em detrimento a outra, deve ser considerado fatores como o tipo dos dados e a forma de como os valores ausentes estão dispostos no conjunto de dados a ser trabalhado.

Capítulo

4

Considerações Finais e Trabalhos Futuros

Esta monografia descreveu o problema de ausência de dados no âmbito de preparação para um processo de classificação e as principais técnicas relatadas na literatura.

A existência no conjunto de dados de outros atributos com informações similares, isto é, com alta correlação, ou poder de predição similar, podem fazer a imputação de valores desconhecidos pouco efetiva, ou até mesmo danosa.

A imputação de valores ausentes pode ser ruim, uma vez que mesmo o mais avançado método de imputação somente é capaz de aproximar os valores reais, não conhecidos, dos valores desconhecidos. Os valores preditos são geralmente mais bem comportados, uma vez que eles são preditos em conformidade com valores desconhecidos introduzidos, e quanto maior for a quantidade de valores desconhecidos, mais simples serão os classificadores induzidos. Dessa forma, a imputação de valores desconhecidos deve ser cuidadosamente aplicada, sob o risco de simplificar sobremaneira o problema em estudo.

Este estudo, cujo tema faz parte do escopo, ainda em delimitação, da pesquisa a ser desenvolvida em minha tese de doutorado, foi bastante relevante, haja vista um levantamento e estudo preliminar que em muito ajudará na organização das idéias para a qualificação.

Como trabalhos futuros, a serem executados nos próximos meses, pode-se citar a descrição e aprofundamento em todas as técnicas pesquisadas, propondo uma taxonomia e obtendo, ao final, um *survey*, a ser submetido para publicação, configurando um primeiro resultado da pesquisa de tese. Ainda, uma explanação sobre este estudo será feita, na forma de seminário, no Grupo de Análise de Dados e Sistemas Gerenciadores de Banco de Dados.

Referências Bibliográficas

[1] A.P.A.E.G. Batista. *Pré-processamento de Dados em Aprendizado de Máquina Supervisionado*. PhD thesis, Instituto de Ciências Matemáticas e de Computação – USP, 2003.

[2] P. Clark and R. Boswell. Rule induction with cn2: Some recent improvements. In *Proc. Fifth European Conference (EWSL1)*, 1991.

[3] P. A. Dempster, N. M. Laird, and B. D. Rubin. maximum likelihood from incomplete data via the em algoritm (with discussion). *Journal of Royal Statistical Society*, 39(1):1–38, November 1977.

[4] U. Fayyad, G. Piatetski-Shapiro, and P. Smyth. Knowledge discovery and data mining: towards a unifying framework. In *Proc. KDD-96*, 1996.

[5] M. M. Garcia. *Imputação de Dados Categorizados usando o Modelo Multinomial*. PhD thesis.

[6] Z. Ghahramani and Jordan I. M. Supervised learning from incomplete data via an em approach. In Advances in Neural Information Processing Systems, 1994.

[7] W. J. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. *RSCTC'2000*, (1):340–347, November 2000.

[8] C.V.K. Populus, R.J.X. Fogg, and Downey R.G. Mean substitution for missing items: Sample size and the effectiveness of the technique. *Technical Report - Kansas State University*, (13):1–20, March 1998.

[9] R. J. Quinlan. *C4.5 Programs for Machine Learning.* Morgan Kaufmann, New York, 1988.

[10] Duda O. R., Hart E. P., and Stork G. D. *Pattern Classification.* John Wiley and Sons, Inc., New York, 2002.

[11] Little J. A. R. and B. D. Rubin. *Statistical Analysis with Missing Data.* John Wiley and Sons, Inc., New York, 2002.

[12] S. Sarawagi and T. V. Raisinghani. Cleaning methods in data warehousing. Technical Report - KR School of Information technology, 1999.

[13] S. Theodoridis and K. Koutroumbas. *Pattern recognition.* Academic Press, San Diego, 1999.

[14] X. Wu and D. Barbará. Learning missing values from summary constraints. *SIGKDD Explorations*, 47(1):21–30, March 2003.

[15] S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied Artificial Intelligence - Special Issue on Data Cleaning and Preprocessing*, 17(5–6):375–381, June 2003.