UNIVERSIDADE FEDERAL FLUMINENSE

MARIANA TASCA FONTENELLE LÔBO

Uma Proposta de Medida de Relevância de Atributos Multivalorados para Classificação

NITERÓI

2008

UNIVERSIDADE FEDERAL FLUMINENSE

MARIANA TASCA FONTENELLE LÔBO

Uma Proposta de Medida de Relevância de Atributos Multivalorados para Classificação

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Orientador:
Bianca Zadrozny

Co-orientador:
Alexandre Plastino de Carvalho

NITERÓI

2008

Uma Proposta de Medida de Relevância de Atributos Multivalorados para Classificação

Mariana Tasca Fontenelle Lôbo

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre.

Aprovada por:

Profa. Ph.D. Bianca Zadrozny / IC-UFF (Presidente)

Prof. D.Sc. Alexandre Plastino de Carvalho / IC-UFF

Profa. Ph.D. Ana Cristina Bicharra Garcia / IC-UFF

Prof. D.Sc. Luiz Henrique de Campos Merschmann / UFOP

Prof. Ph.D. André Ponce de Leon F. de Carvalho / USP

Niterói, 30 de setembro de 2008.

Dedico este trabalho à minha família, que esteve de mãos dadas comigo durante todo esse percurso.

Agradecimentos

São muitas as pessoas que contribuíram para a conclusão dessa jornada, tão difícil pra mim, por ter que conciliar meu trabalho em tempo integral aos estudos do mestrado.

Em primeiro lugar, gostaria de agradecer ao meu amor, Eberth, pela força, dedicação, companheirismo, amor e paciência. Obrigada por não me deixar fraquejar e por compartilhar comigo cada momento dessa etapa. Obrigada por compreender minha ausência nos momentos em que não pude estar com você por causa dos estudos.

Agradeço também, de forma especial, à minha família. Obrigada pelo incentivo e por estarem de mãos dadas comigo durante essa caminhada. À minha mãe e (literalmente) mestra, que me mostrou que não existe tempo para enfrentar qualquer desafio e mergulhou num mestrado em sua área enquanto eu fazia o meu. Parabéns pela conquista e obrigada pelo exemplo.

Aos professores Bianca Zadrozny e Alexandre Plastino, meus orientadores, pela dedicação e apoio. Obrigada por acreditarem no meu trabalho.

Aos meus amigos do Serpro, que foram muito importantes nessa fase: Ronald Dreux e Paulo Ubirajara, obrigada por compreenderem minha ausência no trabalho. Paula, obrigada pela força e incentivo diário. Luciane, obrigada pelo apoio e consultoria. Sua ajuda foi essencial. Beatriz, obrigada por ouvir minhas intermináveis explicações sobre a pesquisa. Aos outros colegas de trabalho, muito obrigada pela torcida e por estarem sempre ao meu lado.

Aos colegas da UFF, obrigada pelo apoio durante toda essa fase. Em especial ao Jacques, pela atenção e ajuda.

Agradeço a Deus, por ter me ajudado a manter minha serenidade e fé necessárias para a conquista desse objetivo.

Resumo

Uma etapa importante do processo de descoberta de conhecimento em bases de dados é a seleção de atributos, que tem como objetivo a escolha de um subconjunto adequado de atributos que represente a informação importante contida nos dados. A maioria dos métodos existentes para seleção de atributos leva em consideração apenas atributos de tipos simples, como categóricos e numéricos. Esses métodos não se aplicam a atributos multivalorados, que são caracterizados por poderem assumir vários valores simultaneamente para uma mesma instância da base. Porém, em muitas bases reais esses atributos estão presentes. Por exemplo, os tipos de livros que uma pessoa possui podem ser representados através de um atributo multivalorado.

Este trabalho propõe uma medida de relevância para atributos multivalorados, que tem como objetivo medir a sua importância para a classificação. A medida proposta leva em consideração a capacidade que cada atributo tem de determinar a classe da instância.

Para avaliar a medida proposta, foram realizados experimentos com diversas bases de dados, submetidas a classificadores multi-relacionais. Nesses experimentos, foi possível observar que os valores de acurácia obtidos refletem, na maioria dos casos, os valores da medida de relevância proposta. Assim, pôde-se concluir que a medida proposta é um bom indicador da importância do atributo multivalorado para a classificação.

Palavras-chave: mineração multi-relacional, seleção de atributos, atributos multivalorados, medidas de relevância, classificação.

Abstract

An important step in the knowledge discovery in databases (KDD) process is the attribute selection step, which aims at choosing a subset of attributes that can represent the important information within the data. Most of the existing attribute selection methods can only handle simple attribute types, such as categorical and numerical. In particular, these methods cannot be applied to multi-valued attributes, which are attributes that take multiple values simultaneously for the same instance in the database. In many real databases, however, multi-valued attributes are present – e.g. the types of books owned by a person may be represented by a multi-valued attribute.

This dissertation proposes a relevance measure for multi-valued attributes, which aims at measuring their importance for classification. The proposed measure takes into account the ability that the attribute has in determining the instance class.

In order to evaluate the proposed measure, experiments were conducted with several databases submitted to multi-relational classifiers. The experiments show that the resulting accuracy values follow, in most cases, the values of the proposed relevance measure. So we can conclude that the proposed measure is a good indicator of the relevance of multi-valued attributes for classification.

Keywords: multi-relational data mining, attribute selection, multi-valued attributes, relevance measures, classification.

Palavras-chave

1. Mineração multi-relacional
2. Seleção de atributos
3. Atributos multivalorados
4. Medidas de relevância
5. Classificação

Glossário

AL	:	*Average Linking*;
IBGE	:	Instituto Brasileiro de Geografia e Estatística;
IC	:	Instituto de Computação;
ILP	:	*Inductive Logic Programming*;
KDD	:	*Knowledge Discovery in Database*;
kNN	:	*k-Nearest Neighbor*;
MN	:	Monovalorado;
MV	:	Multivalorado;
OLAP	:	*Online Analytical Processing*;
POF	:	Pesquisa de Orçamentos Familiares;
UFF	:	Universidade Federal Fluminense;

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Ao longo dos últimos anos, incentivadas pela redução do custo de armazenamento de dados, muitas empresas – principalmente as grandes organizações – vêm armazenando suas transações diárias em grandes bancos de dados. Este procedimento tem gerado uma imensa massa de informações detalhadas e históricas que, se bem aproveitadas, podem ser de grande valia para o auxílio à tomada de decisões estratégicas. Os processos de descoberta de conhecimento (*Knowledge Discovery in Databases* – KDD) surgiram em função da necessidade de exploração desses dados, transformando-os, através de vários tipos de investigações, em conhecimento útil [1].

O processo de KDD compreende seis etapas: seleção dos dados, limpeza, enriquecimento, transformação ou codificação dos dados, mineração dos dados e, por fim, exibição das informações descobertas [2]. As etapas que antecedem a mineração de dados, na ordem exposta anteriormente, fazem parte do pré-processamento, ou seja, momento em que os dados são preparados para serem investigados. A etapa de mineração de dados pode ter objetivos distintos: classificar elementos, descobrir padrões implícitos na base, dividir as instâncias em grupos com características semelhantes, entre outros. Este trabalho concentra-se na tarefa de classificação, que tem como objetivo estimar a classe à qual pertence uma nova instância a partir dos valores de seus atributos.

Uma das formas de melhorar o desempenho do processo de classificação é realizar uma seleção dos atributos disponíveis, descartando aqueles que não contribuem e que podem até mesmo prejudicar o desempenho dessa tarefa. A seleção de atributos faz parte do pré-processamento da mineração, e pode ser definida como um processo de escolha de um subconjunto adequado de atributos, que representa a informação importante contida nos dados, segundo algum critério [3]. Esta fase é fortemente recomendada principalmente se a base a ser investigada possui um número muito grande de atributos, pois o processamento

da maioria dos algoritmos de mineração pode exigir um grande esforço computacional se uma quantidade grande de atributos for utilizada. Com a utilização de técnicas de seleção de atributos é possível: (a) melhorar o desempenho dos classificadores, eliminando os atributos que não agregam valor à investigação da base e aqueles que deterioram os resultados, (b) simplificar os modelos de classificação, reduzindo o custo computacional de execução desses modelos, (c) reduzir o tamanho da base e (d) simplificar o modelo gerado, fornecendo um melhor entendimento dos resultados obtidos.

Existem várias técnicas para realizar a seleção de atributos, algumas baseadas em medidas de relevância (abordagem *filter*), outras baseadas na utilização do próprio classificador (abordagem *wrapper*) e ainda aquelas que são realizadas dentro do algoritmo de aprendizado (abordagem *embedded*) [4].

Atendendo aos contextos convencionais de mineração de dados, onde a base a ser investigada é representada por uma única tabela ou arquivo seqüencial, a maioria dos algoritmos e medidas de relevância disponíveis hoje na literatura para realizar a seleção de atributos leva em consideração apenas atributos de tipos simples, como categóricos e numéricos. Porém, muitas bases de dados de domínios reais encontradas atualmente possuem atributos multivalorados, que são caracterizados por poderem assumir vários valores simultaneamente. Atributos com essa característica podem contribuir ou não para a tarefa de classificação, dependendo do domínio de aplicação que está sendo investigado. Saber, por exemplo, quais tipos de livros uma pessoa compra (atributo multivalorado) pode ser importante para descobrir se ela tem ou não filhos; mas pode não trazer nenhuma informação útil para descobrir se a pessoa utiliza ou não cartão de crédito (para esta última classe seria importante, talvez, saber se o indivíduo possui muitos gastos). Assim, torna-se importante tratar esse tipo de atributo para que seja possível descobrir sua importância para a classificação.

Não existem na literatura muitas propostas para realizar a seleção de atributos multivalorados ou medidas para determinar sua importância. Em [5], é utilizada uma técnica que transforma os k valores do domínio do atributo multivalorado em atributos binários, permitindo, assim, que algoritmos de seleção convencionais possam ser utilizados. Mas essa técnica traz um problema para a mineração, pois aumenta a dimensionalidade do espaço original.

Na tentativa de contribuir para o processo de seleção de atributos multivalorados, o presente trabalho propõe uma medida de relevância para esse tipo de atributo, levando em consideração sua capacidade de determinar corretamente a classe de uma instância.

Pretende-se que a medida proposta possa ser utilizada em algoritmos de seleção de atributos.

As avaliações da medida proposta são feitas com base na acurácia de um classificador multi-relacional contido na ferramenta Relational Weka [6], desenvolvida a partir da conhecida ferramenta Weka [7]. São realizados testes com bases de dados reais – algumas obtidas de repositórios públicos disponíveis na internet e outras obtidas do Instituto Brasileiro de Geografia e Estatística (IBGE) – e bases sintéticas, geradas a partir dessas mesmas bases reais. Os seguintes aspectos são considerados para a avaliação da medida proposta: (a) análise da qualidade dos atributos multivalorados isoladamente, sem considerar a influência dos outros atributos da base; (b) análise da contribuição do atributo multivalorado para a classificação quando o mesmo é combinado com os outros atributos da base; (c) análise do comportamento da medida proposta com a variação das probabilidades de distribuição dos valores do atributo entre as classes e (d) análise da utilidade da medida para realizar a comparação de dois atributos multivalorados de uma mesma base, na tentativa de descobrir qual deles traz maior contribuição para a classificação.

A mineração de dados a partir de uma base com várias tabelas é conhecida como mineração relacional (também chamada de multi-relacional apenas para enfatizar a utilização de múltiplas tabelas) [8]. Apesar de ser uma área nova dentro do processo de KDD, já existem na literatura várias propostas de algoritmos para realizar a exploração dessas bases [9, 10, 11, 12]. Como a representação dos atributos multivalorados em bases de dados normalmente é feita através de uma tabela separada, para evitar redundâncias dentro da tabela principal [2], pesquisas com esse tipo de atributo, objeto de estudo deste trabalho, estão conseqüentemente inseridas no contexto da mineração multi-relacional.

Este trabalho está organizado conforme especificado a seguir. O Capítulo 2 contém uma revisão bibliográfica sobre classificação multi-relacional: algoritmo k-NN e medidas de distância. No Capítulo 3, é apresentada uma breve revisão sobre seleção de atributos e tipos de medidas de avaliação de atributos. A medida de relevância de atributos multivalorados proposta neste trabalho é apresentada no Capítulo 4. No Capítulo 5, apresentam-se as bases de dados utilizadas, o algoritmo de geração do atributo multivalorado sintético, os resultados obtidos nos experimentos e a avaliação da medida de relevância proposta. Por fim, no Capítulo 6, são apresentadas as conclusões deste trabalho e propostas de trabalhos futuros.

Capítulo 2

Classificação com Atributos Multivalorados

As pesquisas relacionadas à mineração de dados normalmente envolvem algoritmos que extraem padrões de bases de dados onde todos os atributos a serem investigados fazem parte de uma única tabela ou arquivo seqüencial. Tal característica limita a investigação a esse formato de dados, quando utilizados os algoritmos convencionais de mineração. Para que a mineração de dados pudesse ser realizada em bancos de dados relacionais – modelo normalmente utilizado pelas empresas, envolvendo inúmeras tabelas relacionadas através de chaves estrangeiras [2] – surgiram algumas linhas de pesquisa relacionadas à mineração relacional, também chamada de multi-relacional, simplesmente para reforçar o fato de que múltiplas tabelas são envolvidas na mineração [8].

Atributos multivalorados – que são caracterizados por poderem assumir vários valores simultaneamente para uma mesma instância – são implementados em tabelas separadas relacionadas à tabela principal e, portanto, sua utilização em processos de mineração de dados exige a utilização de técnicas multi-relacionais.

2.1 Mineração Multi-Relacional

Tradicionalmente, os algoritmos desenvolvidos em pesquisas relacionadas à mineração de dados utilizam uma única tabela (ou relação) no processo de extração do conhecimento. Alguns autores denominam essa abordagem de aprendizado proposicional, no qual a base é formada por um conjunto de instâncias, cada instância é representada por um conjunto fixo de atributos, e cada atributo possui um único valor para determinada instância. Esses algoritmos ou modelos são chamados proposicionais porque uma instância é caracterizada

pela conjunção de proposições da forma "atributo θ valor", onde θ é um operador relacional, tal como $<, \leq, >, \geq, =, \neq$ [13].

Porém, em contextos que envolvem bases de dados reais, os dados encontram-se, na maioria das vezes, no modelo relacional, onde múltiplas tabelas armazenam as informações de forma normalizada e são interligadas através de chaves estrangeiras. Para que seja possível extrair conhecimento desse tipo de base de dados através de mineração, existem duas alternativas: utilizar algoritmos de mineração multi-relacional, que trabalham diretamente com múltiplas tabelas, ou realizar a proposicionalização, ou seja, a transformação do problema multi-relacional no problema proposicional [13]. Essa transformação, por sua vez, pode ser feita de duas formas: (a) unir as tabelas envolvidas no contexto através de operações de JOIN, agrupando todos os atributos numa única tabela; (b) transformar o modelo relacional em uma única relação (tabela) através da criação de novos atributos, na tabela principal, que sumariza ou agrega as informações das outras tabelas.

Ambas as alternativas citadas podem trazer uma série de problemas: a junção de todas as tabelas da base pode resultar em uma tabela muito grande, com uma quantidade excessiva de atributos, de difícil manipulação. Além disso, a junção das tabelas vai gerar uma série de registros repetidos para um mesmo indivíduo, causado pelos relacionamentos *N:1* e *N:M*, o que pode acarretar problemas estatísticos [14]. A criação de atributos na tabela principal com informações agregadas das outras tabelas, geradas a partir de operações tais como soma (SUM), média (MEAN) ou contador (COUNT), pode gerar uma grande perda de informações. Se a tabela principal armazena dados de pessoas e o objetivo é classificá-las por perfil, por exemplo, o fato de conhecer os títulos dos livros comprados por essas pessoas é de grande importância quando comparado apenas com a quantidade de livros que cada uma compra (informação agregada). Existem, na literatura, algumas propostas de agregação de atributos categóricos com menor perda de informação. Em [15], é apresentado um *framework* que realiza a proposicionalização em bases relacionais utilizando agregações convencionais em atributos numéricos (SUM, MEAN, COUNT) e agregações mais sofisticadas em atributos categóricos.

Existem várias propostas na área de mineração multi-relacional que trabalham com as tabelas no modelo de dados original. A maioria delas está relacionada à Programação Lógica Indutiva (*Inductive Logic Programming* - ILP) [16, 17], que combina indução (a capacidade de gerar modelos genéricos a partir de instâncias específicas) e programação lógica, que utiliza lógica de primeira ordem para representar as relações entre os objetos e implementa o raciocínio dedutivo [13]. Existem, ainda, outras propostas não baseadas em

ILP que apresentaram bons resultados, como em [18], onde apresenta-se um *framework* ilustrado pelo algoritmo Warmr, uma generalização do algoritmo Apriori para modelos relacionais; em [19], onde apresenta-se um *framework* para árvores de decisão multi-relacionais; e em [20], onde explora-se uma abordagem baseada em redes Bayesianas.

Para a tarefa de classificação em mineração de dados, uma instância é caracterizada por um conjunto de atributos. Quando esses atributos podem assumir um único valor para determinada instância, são chamados de atributos monovalorados. Uma pessoa, por exemplo, pode ter um único valor para idade, sexo, cidade onde nasceu, entre outras informações.

Porém, existem certos atributos que podem assumir uma série de valores para uma mesma instância, e são chamados, então, de atributos multivalorados. São exemplos de atributos multivalorados: os tipos de livros que uma pessoa lê, as áreas de pesquisa de um professor etc.

Num banco de dados relacional, que caracteriza-se pelo armazenamento dos dados em múltiplas tabelas, os atributos monovalorados normalmente são armazenados na tabela principal, juntamente com a identificação das instâncias relacionadas a eles. Já os atributos multivalorados são implementados através da construção de outras tabelas, relacionadas à tabela principal através de chaves estrangeiras.

Na próxima seção, será tratado o problema de classificação em bases de dados com atributos multivalorados.

2.2 Classificação em Bases com Atributos Multivalorados

Uma das principais tarefas de mineração de dados é a classificação, que tem como objetivo identificar, entre um conjunto pré-definido de classes, aquela à qual pertence uma instância, a partir de seus atributos. Um tipo comum de técnica de classificação, denominado *eager*, envolve duas fases: na primeira delas, um modelo é construído a partir de uma base de treinamento, onde as instâncias têm sua classe conhecida. Uma base de teste é usada posteriormente para avaliar o modelo. A partir desse modelo construído, uma nova instância que tem sua classe desconhecida pode ser classificada com base no seu conjunto de atributos. A fase de classificação torna-se mais rápida, já que a base de dados não precisará mais ser consultada quando da classificação de uma nova instância. Para classificar uma instância, basta que os valores dos seus atributos sejam avaliados

pelas regras do modelo. Nessa abordagem, a fase de geração do modelo pode ter alto custo computacional e deverá ser realizada toda vez que a base for significativamente atualizada.

Existe um outro tipo importante de técnica de classificação, conhecida como *lazy*, que não envolve a criação prévia de um modelo a partir de uma base de treinamento. Nesse caso, o processamento dos dados só ocorre quando existe uma nova instância a ser classificada e suas características são comparadas com as características das outras instâncias da base, já conhecidas. A expressão *lazy* (preguiçoso) é utilizada pelo fato de esse tipo de algoritmo adiar o processamento dos dados até que uma requisição seja realizada. Assim, não existe o custo da geração de um modelo; porém, a etapa de classificação é mais demorada, já que a cada nova instância a ser classificada, a base é consultada. Normalmente, os algoritmos *lazy* utilizam medidas que definem os critérios de comparação entre as instâncias. As próximas subseções apresentam o algoritmo k-NN (uma técnica de classificação na categoria *lazy*) e algumas medidas de distância que podem ser aplicadas nesse tipo de algoritmo.

2.2.1 Algoritmo k-NN

Um dos algoritmos mais conhecidos de classificação *lazy* é o k-NN (*k Nearest Neighbours*). Essa técnica foi proposta na década de 50, mas só se popularizou como método de classificação nas áreas de mineração de dados e aprendizado relacional no início dos anos 90 [21].

Esse algoritmo tem como idéia básica classificar uma nova instância através da comparação dela com instâncias conhecidas da base. A classe da nova instância é determinada pela classe mais freqüente entre as k instâncias mais similares à instância que está se tentando rotular. O valor de k é um parâmetro de entrada.

Para comparar a similaridade entre as instâncias da base são utilizadas medidas de distância. As k instâncias selecionadas para a classificação de uma nova instância são aquelas consideradas mais próximas (similares) à nova instância, de acordo com a medida de distância definida. Uma medida comumente utilizada nesse tipo de classificador, quando os atributos são numéricos, é a distância Euclidiana, a qual define que a distância entre duas instâncias, $X = \{x_1, x_2, ..., x_n\}$ e $Y = \{y_1, y_2, ..., y_n\}$, é dada por

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{2.1}$$

Caso os atributos a serem avaliados possuam valores em diferentes escalas, é importante que seja feita a normalização dos valores para que a contribuição de cada atributo seja considerada de forma igualitária. Uma das formas de normalização é colocar todos os atributos variando entre 0 e 1, utilizando a seguinte equação:

$$norm(v_i) = \frac{v_i - min\{v_i\}}{max\{v_i\} - min\{v_i\}}, \tag{2.2}$$

onde $norm(v_i)$ é a normalização do valor original v_i do atributo i e $min\{v_i\}$ e $max\{v_i\}$ são, respectivamente, o menor e o maior valores do atributo i na base de dados de treinamento.

Quando os atributos são categóricos (ou nominais), uma forma simples de calcular a distância entre eles é considerar a diferença $(x_i - y_i)$ da Fórmula 2.1 igual a 1 (um), quando os valores forem diferentes, e igual a 0 (zero), quando os valores forem iguais.

Já nos casos em que o atributo a ser considerado na classificação representar um conjunto de itens (atributo multivalorado), distâncias especiais para tratamento de comparação de conjuntos devem ser utilizadas. A próxima subseção apresenta algumas medidas de distâncias para esse tipo de atributo.

2.2.2 Medidas de Distância de Atributos Multivalorados

Para calcular a distância entre instâncias de uma base, faz-se necessário definir uma medida, calculada por uma função a partir das características dessas instâncias. Quando o atributo a ser considerado é multivalorado, é necessário definir uma forma de comparar os dois conjuntos de itens relacionados a um par de indivíduos. Antes de explorar medidas de distâncias de conjuntos propostas na literatura, é importante registrar algumas terminologias e definições utilizadas para caracterizá-las. Uma função $d : S \times S \rightarrow \Re$ é considerada uma métrica se e somente se atender às seguintes condições [22]:

$$\text{Não negativa}: \quad \forall x, y \in S : d(x, y) \geq 0 \tag{2.3a}$$

$$\text{Reflexividade}: \quad \forall x, y \in S : d(x, y) = 0 \Rightarrow x = y \tag{2.3b}$$

$$\text{Simétrica}: \quad \forall x, y \in S : d(x, y) = d(y, x) \tag{2.3c}$$

$$\text{Desigualdade triangular}: \quad \forall x, y, z \in S : d(x, z) \leq d(x, y) + d(y, z) \tag{2.3d}$$

Nem todas as medidas propostas na literatura são consideradas métricas, pois nem sempre as condições 2.3a a 2.3d são atendidas. Quando uma medida satisfaz todas as

condições com exceção da reflexividade, é chamada *pseudo-métrica*; e se a mesma satisfaz todas as condições com exceção da desigualdade triangular, é chamada *semi-métrica*. Existem, na literatura, várias propostas de medidas de distâncias entre conjuntos. Entre elas, destacam-se as medidas *Average Linking* (AL), Tanimoto e RIBL, que foram utilizadas para calcular a distância entre atributos multivalorados nos experimentos deste trabalho. Nas próximas subseções, são descritas com detalhes essas três medidas.

2.2.2.1 *Average Linking*

A medida *Average Linking* (AL), entre dois conjuntos A e B, tenta levar em consideração a informação completa disponível na comparação dos dois conjuntos (ou atributos multivalorados), já que considera todos os pares possíveis formados a partir de A e B. Sua definição pode ser dada como a média das distâncias desses pares [23]:

$$D_{AL}(A,B) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d(a_i,b_j)}{|A||B|}, \tag{2.4}$$

onde $A = \{a_1, a_2, ..., a_n\}$ e $B = \{b_1, b_2, ..., b_m\}$ representam os conjuntos que estão sendo comparados e a distância entre os valores dos atributos é dada por:

$$d(a_i,b_j) = \begin{cases} |a_i - b_j|, & \text{se o valor for contínuo,} \\ 0, & \text{se o valor for discreto e } a_i = b_j, \\ 1, & \text{se o valor for discreto e } a_i \neq b_j. \end{cases} \tag{2.5}$$

Essa medida não satisfaz a propriedade da reflexividade, ou seja, na maioria das vezes a distância de um conjunto para ele mesmo pode ser maior que zero. Esses resultados podem gerar situações contraditórias, já que um conjunto pode ser considerado mais próximo a um outro do que a ele mesmo. Como não satisfaz a propriedade da reflexividade, *Average Linking* não é considerada uma métrica.

2.2.2.2 Tanimoto

A distância Tanimoto entre os conjuntos A e B pode ser definida como:

$$D_T(A,B) = \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{2.6}$$

Essa medida é recomendada quando não se sabe o grau de distância ou similaridade entre os elementos dos conjuntos (ou valores dos atributos multivalorados). Como a idéia da medida Tanimoto é baseada na interseção entre os conjuntos, dois elementos, quando comparados, devem ser considerados iguais ou não. Se forem considerados iguais, são contabilizados na quantidade de interseções da Equação 2.6. Nota-se que essa medida é ideal para ser aplicada na comparação de conjuntos de itens categóricos.

Porém, existem extensões da medida Tanimoto na literatura que propõem uma forma de aplicá-la com graus diferenciados de similaridades [24]. Nesse caso, dois elementos são considerados idênticos se a distância entre eles for menor que um limite definido pelo usuário. A distância entre dois valores de atributos fica, então, definida como:

$$d(a_i, b_j) = \begin{cases} \frac{|a_i - b_j|}{|a_i + b_j|}, & \text{se o valor for contínuo,} \\ 0, & \text{se o valor for discreto e } a_i = b_j, \\ 1, & \text{se o valor for discreto e } a_i \neq b_j. \end{cases} \tag{2.7}$$

Assim, se a distância entre dois valores calculada pela Equação 2.7 for menor que o limite definido pelo usuário, esses valores são considerados iguais e, conseqüentemente, são contabilizados no número de interseções da Equação 2.6.

Essa medida é menos sensível a ruídos porque cada elemento de um conjunto pode ser igualado a um elemento do outro no máximo uma vez. Em medidas como *Average Linking*, cada elemento é comparado com todos os outros. Assim, se um elemento estiver errado (for um ruído) ele poderá alterar bastante o valor da medida, enquanto que na Tanimoto isso não acontece.

Como a medida Tanimoto viola a desigualdade triangular, não pode ser considerada uma métrica.

2.2.2.3 RIBL

No modelo relacional, uma instância é representada por um conjunto de atributos e pode estar relacionada com outras informações em vários níveis. A Figura 2.1 descreve um exemplo desse tipo de representação [8]. O nível 0 contém a instância alvo da classificação e armazena, além da sua identificação (*IdPessoa*), outros atributos monovalorados. O predicado *membro(IdPessoa, idade, sexo, salário, tipo de membro)* descreve a estrutura do nível 0. O nível 1 é representado pelos predicados *carro(IdPessoa, tipo, velocidade máxima, fabricante)* e *casa(IdPessoa, distrito, ano de construção, tamanho)*. Nesse exemplo, pode-se, de forma simplificada, dizer que o nível 1 representa atributos multivalorados, pois indivíduos podem possuir mais de um carro e mais de uma casa. O nível 2 é representado pelas características do distrito onde as casas estão localizadas e, nesse caso específico, não representa atributos multivalorados, já que cada casa só pode estar em um único distrito.

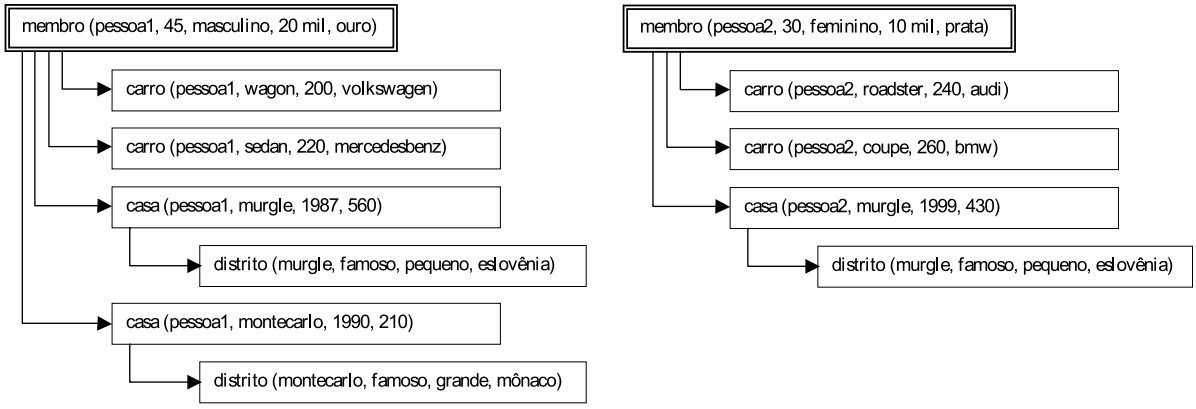

Figura 2.1: Representação dos atributos e relacionamentos de duas instâncias da base: *pessoa1* e *pessoa2*

Para duas instâncias $x = (x_1, x_2, ..., x_n)$ e $y = (y_1, y_2, ..., y_n)$, a distância entre elas é calculada como:

$$dist(x, y) = \sum_{i=1}^{n} d(x_i, y_i)/n, \tag{2.8}$$

onde n representa o número de atributos relacionados às instâncias e a diferença entre os valores dos atributos é definida pela Fórmula 2.5.

A idéia básica da medida RIBL é a seguinte. Para calcular a distância entre duas instâncias, inicialmente seus atributos do nível 0 são comparados. Assim, nessa primeira etapa, são calculadas as distâncias dos atributos *IdPessoa*, *idade*, *sexo*, *salário* e *tipo de membro* entre as instâncias correspondentes aos indivíduos *pessoa1* e *pessoa2* da Figura

2.1. Nesse nível, os identificadores das instâncias são tratados como atributos categóricos, ou seja, $d(pessoa1, pessoa2) = 1$.

Em seguida, os atributos do nível 1 são considerados, utilizando a mesma regra de comparação da Equação 2.8. Nessa etapa, então, são consideradas as distâncias entre as casas e entre os carros, levando-se em consideração cada uma de suas características. No nível 1, $d(pessoa1, pessoa2)$ = ((distância entre os conjuntos de carros de *pessoa1* e *pessoa2*) + (distância entre os conjuntos de casas de *pessoa1* e *pessoa2*))/2. Assim, para calcular a distância entre as instâncias nesse nível, são considerados novamente os atributos *IdPessoa*, *idade*, *sexo*, *salário* e *tipo de membro*, porém com um novo valor para $d(pessoa1, pessoa2)$.

Esse procedimento se repete nos níveis seguintes até que seja atingido o nível máximo definido pelo usuário.

Quando o cálculo da distância em determinado nível deve levar em consideração atributos que representam conjuntos (que é o caso de carros e casas, no nível 1 do exemplo da Figura 2.1), essa distância é calculada da seguinte maneira: considera-se o menor conjunto (ou o primeiro, caso eles sejam do mesmo tamanho) e, para cada item desse conjunto, é calculada a distância para o elemento mais próximo do outro conjunto. Essas distâncias são somadas posteriormente. Para tratar as diferenças de cardinalidade entre os conjuntos, é feita uma normalização utilizando o tamanho do maior conjunto. Ou seja, a distância entre os conjuntos pode ser dada por:

$$D_{RIBL}(A,B) = \begin{cases} \frac{\sum_{i=1}^{n} min_{j=1}^{m}(d(a_i,b_j))}{|B|}, & |A| < |B| \\ \\ \frac{\sum_{j=1}^{m} min_{i=1}^{n}(d(a_i,b_j))}{|A|}, & |A| \geq |B| \end{cases} \tag{2.9}$$

onde $A = \{a_1, a_2, ..., a_n\}$ e $B = \{b_1, b_2, ..., b_m\}$ representam os conjuntos que estão sendo comparados.

Por atender apenas à propriedade da reflexividade, RIBL não pode ser considerada uma métrica. Esse método tem foco no grupo de distâncias mínimas de um conjunto em relação ao outro, gerando uma medida global da similaridade dos conjuntos a partir dos seus elementos mais semelhantes. Essa característica pode ser problemática se existir um ruído no conjunto A, por exemplo, que seja muito mais próximo do conjunto B do que todos os outros elementos de A. Nesse caso, essa distância mínima irá influenciar diretamente no resultado final da medida, gerando uma distorção.

RIBL é mais do que uma medida de distância entre conjuntos, pois trata bases relacionais com diversos níveis, onde cada nível pode ser composto de vários atributos. RIBL é, na verdade, um algoritmo genérico de aprendizado *lazy* para classificação relacional. Quando aplicado em bases relacionais de apenas dois níveis, onde o nível mais detalhado é representado por atributos multivalorados, este método se reduz a uma medida de distância de conjuntos. Já as medidas Tanimoto e *Average Linking*, descritas anteriormente, apresentam apenas critérios de comparação entre conjuntos (ou atributos multivalorados) num único nível.

É importante observar, comparando as Fórmulas 2.5 e 2.7, que o tratamento dos atributos numéricos é feito de forma diferente na medida Tanimoto quando comparado com *Average Linking* e RIBL. Essa diferenciação pode causar diferenças nos valores dessas medidas, mesmo se estiverem sendo considerados apenas atributos do nível 0 (monovalorados), ou seja, mesmo que a base seja composta por uma única tabela e não estejam sendo utilizados atributos multivalorados na classificação.

Não existe uma regra para dizer qual medida de distância apresenta melhores resultados na classificação. Tudo depende do contexto que está sendo investigado e do tipo de informação que se está tentando extrair. Em alguns casos pode ser importante considerar as distâncias entre os elementos mais similares, em outros casos, uma medida mais global, que leva em consideração todos os elementos do conjunto, pode ser mais importante [23].

Uma das formas de melhorar o processo de classificação é realizar uma seleção dos atributos disponíveis, descartando aqueles que não contribuem para a investigação em questão. O próximo capítulo trata de medidas de avaliação de atributos, que têm como objetivo identificar quais características da base podem ser realmente úteis para a mineração de dados do contexto que está sendo analisado.

Capítulo 3

Medida de Relevância de Atributos Multivalorados

Técnicas de seleção de atributos são estudadas há pelo menos quatro décadas com o objetivo de tentar descobrir, em um grande conjunto de atributos de uma base, aqueles com maior importância para a classificação e também para outras técnicas de mineração de dados. Com o passar dos anos, essa preocupação é cada vez maior, pois atualmente a facilidade de armazenamento de dados permite a geração de grandes volumes de informações, tanto em termos de quantidade de registros quanto em termos de quantidade de atributos.

São várias as vantagens de se realizar a seleção de atributos antes de iniciar o processo de mineração de dados: (a) melhora nos resultados da mineração – que no caso da classificação pode ser evidenciada pelo aumento da acurácia – já que com a seleção podem ser removidos dados irrelevantes e ruídos, (b) redução do tempo computacional, (c) melhora no entendimento do domínio, entre outras.

Conforme será visto neste capítulo, muitas técnicas de seleção de atributos se baseiam em medidas de relevância. No entanto, não foi encontrada na literatura nenhuma medida de relevância específica para atributos multivalorados. A maioria das técnicas de seleção de atributos leva em consideração apenas atributos de tipos convencionais, como numéricos e categóricos. Em [5], é proposta uma técnica que transforma os k valores do domínio de um atributo multivalorado em atributos binários. Assim, se o k-ézimo valor do domínio aparece no conjunto de determinada instância, o atributo k mostra o valor "*true*"; caso contrário, o atributo k mostra o valor "*false*". Essa técnica traz um problema para a mineração, pois aumenta a dimensionalidade do espaço original. A quantidade de novos atributos gerados será proporcional ao tamanho do domínio do atributo multivalorado.

Neste capítulo, após uma introdução sobre seleção de atributos e tipos de medidas de relevância de atributos, é apresentada a principal contribuição desta dissertação: uma proposta de medida de relevância para atributos multivalorados.

3.1 Tipos de Algoritmos de Seleção de Atributos

De acordo com a maneira utilizada para descobrir quais atributos da base são os mais importantes para a mineração, as técnicas de seleção de atributos podem ser agrupadas em três abordagens principais: *embedded*, *filter* e *wrapper* [25].

Na abordagem ***embedded***, a seleção de atributos é realizada dentro do próprio algoritmo de aprendizado. Essa abordagem é muito utilizada pelos algoritmos de aprendizado que seguem a estratégia *eager* (gulosa). Durante a fase de treinamento, os algoritmos *eager* substituem os exemplos de treinamento pelo conceito induzido, geralmente na forma de um conjunto de regras ou uma árvore de decisão. Na maioria dos casos, o próprio algoritmo de aprendizado se encarrega de selecionar os atributos que farão parte das regras ou da árvore de decisão. Exemplos de algoritmos de aprendizado que seguem a abordagem *embedded* são CN2, C4.5 e ID3.

Na abordagem ***filter***, a seleção ocorre na etapa de pré-processamento e é totalmente independente do algoritmo de aprendizado que será utilizado na fase de mineração. Nesse tipo de seleção, são utilizadas medidas para selecionar um grupo de atributos considerado relevante, segundo algum critério, enquanto os outros são descartados. Assim, apenas o grupo de atributos selecionados será considerado na fase de mineração.

Na abordagem ***wrapper***, a seleção também ocorre numa etapa de pré-processamento, mas nesse caso o algoritmo de aprendizado que será utilizado na mineração é considerado como "ferramenta" para a seleção. Esse tipo de seleção gera um subconjunto de atributos candidatos e executa o algoritmo de aprendizado considerando apenas esse subconjunto. A acurácia da classificação é utilizada para avaliar o subconjunto de atributos em questão. Essa operação é repetida até que um critério de parada definido pelo usuário seja satisfeito. Assim, vários subconjuntos são avaliados, e aquele que resultou na melhor acurácia do classificador é escolhido como o melhor subconjunto.

A abordagem *wrapper* é considerada eficaz do ponto de vista da acurácia, já que o próprio algoritmo que será utilizado na mineração "escolhe" o subconjunto de atributos relevantes, de acordo com a acurácia obtida. Porém, o custo computacional desse método é muito alto, pois o algoritmo de aprendizado deve ser executado várias vezes.

O foco deste trabalho está relacionado a medidas de relevância de atributos criadas para algoritmos de seleção que utilizam a abordagem *filter*, na etapa de pré-processamento. A próxima seção descreve algumas categorias de medidas de relevância de atributos.

3.2 Medidas de Relevância de Atributos

Assim como atributos monovalorados, os multivalorados (ou conjuntos de itens) podem ou não ter importância para a classificação de uma instância. Conhecer, por exemplo, o conjunto de livros que consumidores de uma livraria compram pode ajudar a descobrir o sexo deles ou se os mesmos possuem ou não filhos. Porém, poderia não fazer diferença conhecer os títulos dos livros para descobrir se esses consumidores possuem cartão de crédito. Nesse caso, talvez, fosse mais importante saber se esses consumidores gastam muito com a compra de livros. Assim, a importância de um atributo multivalorado para a classificação depende do domínio da aplicação a ser investigado e, principalmente, da classe que está se tentando descobrir. Em uma mesma base, um atributo pode ser importante para determinada classe e não apresentar qualquer ganho de informação para outra. E mais, em técnicas de classificação *lazy*, se os conjuntos de itens relacionados às instâncias não forem relevantes, esse atributo multivalorado pode atrapalhar a classificação, pois o fato de duas instâncias da mesma classe apresentarem interseção nula entre seus conjuntos de itens, pode distanciá-las e, conseqüentemente, as mesmas poderão ser rotuladas com classes distintas. Nesse contexto, torna-se importante estudar medidas que possam avaliar a relevância dos atributos para a classificação.

Antes de propor medidas para descobrir se um atributo é importante, é necessário definir que tipo de importância está sendo avaliada, ou seja, o que significa dizer que um atributo é importante. Essa definição pode ser considerada, de forma geral, como:

Definição 3.1 [3]: Um atributo é dito importante se, quando removido, a medida de importância considerada em relação aos atributos restantes é deteriorada.

Existem várias formas de avaliar a importância de atributos, através de medidas segundo características distintas. Algumas delas são descritas a seguir.

3.2.1 Medidas de Distância

Medidas de distância determinam a divergência ou separabilidade entre atributos. Em um problema de duas classes, um atributo X_i é considerado melhor que um atributo X_j

se X_i apresenta uma diferença maior que X_j entre a média de distâncias das instâncias de mesma classe e de classes diferentes. Ou seja, atributos considerados bons são aqueles que aproximam instâncias da mesma classe e separam instâncias de classes distintas. Em atributos de tipos convencionais são usadas, por exemplo, distâncias Euclidianas.

3.2.2 Medidas de Dependência

Medidas de dependência podem mostrar o quanto a classe depende de determinado atributo. Essas medidas estão relacionadas à probabilidade de uma classe ocorrer, dado um determinado valor de atributo. Uma definição de medida de dependência pode ser dada por:

Definição 3.2 [25]: Seja Y o atributo classe. Um atributo X é importante se e somente se existem valores x e y para os quais $P(X = x) > 0$ e $P(Y = y|X = x) \neq P(Y = y)$.

Ou seja, X é importante se a estimativa para a classe Y pode ser modificada em função de determinado valor de X, o que quer dizer que Y é condicionalmente dependente de X. Essa definição leva em consideração apenas o atributo em questão e a classe, ou seja, não são avaliadas as influências de outros atributos. Assim, essa definição é falha se o atributo isoladamente não for preditivo da classe, ou seja, quando $P(Y = y|X = x) = P(Y = y)$. Existem definições alternativas que levam em consideração a influência dos outros atributos do subconjunto que está sendo avaliado, mas essas são mais complicadas de implementar na prática pela dificuldade de estimar as probabilidades conjuntas. Como o foco deste trabalho é descobrir a importância de atributos multivalorados individualmente, essas outras definições não serão tratadas em detalhes.

3.2.3 Medidas de Consistência

Um conjunto de atributos é dito inconsistente se existem dois exemplos com os mesmos valores de atributos, mas pertencentes a classes diferentes. Intuitivamente podemos dizer que é incoerente calcular medidas de inconsistência para atributos isolados, principalmente se os domínios desses atributos forem pequenos, pois fatalmente existirão valores relacionados a instâncias de classes distintas. Como este trabalho tem como foco a descoberta da importância de atributos isoladamente, as medidas de consistência não serão detalhadas.

3.2.4 Medidas de Precisão

Medidas de precisão estão intimamente ligadas ao algoritmo utilizado para a classificação e, conseqüentemente, relacionadas à abordagem *wrapper*, mencionada no início deste capítulo. Nesse tipo de análise, atributos considerados bons são aqueles que aumentam a precisão do classificador.

Definição 3.3 [26]: Dados uma amostra de dados S, um algoritmo de aprendizado I e um subconjunto de atributos F, no qual $\{X_i\} \not\subseteq F$. Um atributo X_i é incrementalmente útil para I em relação a F se a precisão da hipótese produzida por I considerando o conjunto de atributos $\{X_i\} \cup F$ é melhor que a precisão alcançada utilizando-se apenas o subconjunto de atributos F.

Essas medidas podem ser observadas através da análise dos resultados da classificação com e sem o conjunto de atributos em questão. Vale lembrar que o fato de a classificação melhorar com a inclusão de determinado atributo não quer dizer que ele por si só é importante. A combinação desse atributo com os outros que fazem parte do subconjunto utilizado na classificação deve ser considerada. Um atributo individualmente pode ser importante, mas pode não fazer parte do subconjunto ótimo para a classificação de determinada base quando a medida de precisão é considerada. O seguinte exemplo, extraído de [25], mostra essa questão.

Considere o conjunto de instâncias {E1, E2,..., E8} descrito por três atributos, X_1, X_2 e X_3, o qual tem como universo de possíveis exemplos $\{0,1\}^3$ (Tabela 3.1). Seja a função objetivo $f(\vec{x}) = (X_1 \wedge X_2) \vee X_3$. Sob qualquer uma das definições de importância anteriormente apresentadas, todos os três atributos são considerados importantes.

Tabela 3.1: Exemplos de atributos importantes

Instâncias	X_1	X_2	X_3	Classe
E1	1	1	1	1
E2	1	1	0	1
E3	1	0	1	1
E4	1	0	0	0
E5	0	1	1	1
E6	0	1	0	0
E7	0	0	1	1
E8	0	0	0	0

No exemplo da Tabela 3.1, existe um único subconjunto de atributos ótimo em relação à medida de precisão: $\{X_3\}$, pois a precisão obtida para prever a classe com esse atributo é 7/8, ou seja, precisão máxima entre todos os subconjuntos possíveis de atributos.

Qualquer outro atributo ou subconjunto resultaria em uma precisão de no máximo 5/8. Ou seja, dizer que um atributo é importante não garante que ele faça parte do conjunto ótimo, que resulta na melhor precisão do algoritmo utilizado.

Conforme observado nas seções anteriores, dizer que um atributo é relevante não garante que ele seja realmente útil para a classificação. Podemos dizer que todo atributo útil é relevante, mas nem todo atributo relevante é útil. Porém, medidas de relevância são consideradas como um bom indicativo de que o uso de determinado atributo pode trazer melhores resultados na classificação.

A próxima seção apresenta uma proposta de medida de relevância de atributos multivalorados, que pode ser aplicada em algoritmos de seleção de atributos que utilizam a técnica *filter*, ou seja, na qual a seleção é feita durante o pré-processamento. O objetivo dessa medida é tentar indicar se determinado atributo multivalorado será útil ou não para a classificação da base em questão e, ainda, avaliar qual é o melhor atributo entre dois ou mais atributos multivalorados da mesma base.

3.3 A Medida Proposta

Conforme visto na seção anterior, a relevância de atributos para a tarefa de classificação pode ser avaliada sob várias perspectivas. Este trabalho tem como objetivo propor e avaliar uma medida de relevância de atributos multivalorados baseada na medida de dependência, ou seja, no quanto as classes dependem de determinado atributo. Em outras palavras, a medida proposta apresenta um indicador da capacidade de determinado atributo definir a classe em questão.

O escopo deste trabalho refere-se apenas a problemas de classificação binária. A proposta da medida de relevância é calcular a diferença das probabilidades de cada uma das classes ocorrer, dados os valores do atributo multivalorado. Assim, se essa diferença for grande, significa que o atributo tem a capacidade de descrever a classe, ou seja, significa que a classe é dependente desse atributo. Quando a diferença das probabilidades é pequena, significa que o atributo em questão não define a classe e, conseqüentemente, não é considerado importante para a classificação.

Seja x um valor do domínio de um atributo multivalorado X e C uma classe. A probabilidade da classe C ocorrer, dado o valor x, é definida pela regra de Bayes apresentada

na Fórmula 3.1.

$$P(C|x) = P(x \ \wedge \ C)/P(x) \tag{3.1}$$

Essa probabilidade pode ser estimada a partir da base de dados, contando-se o número de instâncias pertencentes à classe C para as quais o atributo multivalorado assume o valor x, dividido pelo número total de instâncias (de qualquer classe) que possuem o valor x no atributo multivalorado.

A partir da Fórmula 3.1, são definidos, nas Fórmulas 3.2a e 3.2b, dois vetores de probabilidades, P_A e P_B, para as classes A e B, respectivamente.

$$P_A[i] = P(A|x_i), \tag{3.2a}$$

$$P_B[i] = P(B|x_i), \tag{3.2b}$$

onde $x_1, x_2, ..., x_n$ representam os valores do domínio do atributo multivalorado.

A partir da diferença entre esses vetores, define-se, na Fórmula 3.3, um terceiro vetor de diferenças, P_D.

$$P_D[i] = |P_A[i] - P_B[i]|. \tag{3.3}$$

A medida de relevância de atributos multivalorados é definida, na Fórmula 3.4, calculando-se a média ponderada do vetor de diferenças. Essa ponderação é importante para dar maior peso aos valores que aparecem muitas vezes na base e menor peso àqueles que aparecem pouco, dando pouca contribuição.

$$I(X) = \frac{\sum_{i=1}^{n} P_D[i].count(x_i)}{\sum_{i=1}^{n} count(x_i)}, \tag{3.4}$$

onde $count(x_i)$ representa a quantidade de vezes que o valor x_i aparece na base.

O resultado dessa medida igual a 1 (um) reflete a relevância máxima de um atributo multivalorado. Esse valor vai ocorrer quando, para todo valor x_i, $P_D[i]$ for igual a 1. Em outras palavras, para qualquer um dos valores do atributo multivalorado, a probabilidade de uma das classes ocorrer será de 100% enquanto que a probabilidade da outra classe

ocorrer será de 0%. Nesse contexto, pode-se dizer que conhecer o valor desse atributo para uma determinada instância é de fato importante para descobrir a sua classe.

Por outro lado, se o valor da medida é igual a zero, o que acontece quando o numerador da Fórmula 3.4 é zero, significa que as probabilidades de cada uma das classes ocorrer, para cada valor x_i desse atributo, são iguais, ou seja, conhecer esse atributo não faz qualquer diferença para a classificação.

A Figura 3.1 apresenta o pseudo-código do cálculo da medida de relevância proposta para atributos multivalorados. O cálculo dos vetores $vCount$, $vCountA$ e $vCountB$, representados nas linhas 02, 03 e 04 do algoritmo, respectivamente, é realizado a partir de uma passagem por todas as tuplas da base.

```
procedure CalculaMedida(X, domínio(X), base de dados)
01.  n := quantidade de itens do domínio do atributo multivalorado;
02.  vCount := vetor que contém, para cada valor x_i, a quantidade de instâncias que
possuem o item x_i;
03.  vCountA := vetor que contém, para cada valor x_i, a quantidade de instâncias
da classe A que possuem o item x_i;
04.  vCountB := vetor que contém, para cada valor x_i, a quantidade de instâncias
da classe B que possuem o item x_i;
05.  for i := 1 to n do begin
06.     P_A[i] := vCountA[i]/vCount[i];
07.     P_B[i] := vCountB[i]/vCount[i];
08.     P_D[i] := |P_A[i] - P_B[i]|;
09.     vSoma := vSoma + (P_D[i] * vCount[i]);
10.     vSomaCount := vSomaCount + vCount[i];
11.  end;
12.  I(X) := vSoma/vSomaCount;
```

Figura 3.1: Pseudo-código do cálculo da medida de relevância

A seguir, um exemplo ilustra a execução do algoritmo.

Dado um atributo multivalorado X e o seu domínio, representado pelo conjunto {a, b, c, d, e, f, g}. Suponha que a base seja formada por 5 instâncias, duas da classe A e três da classe B, cujo atributo multivalorado X é representado pelos conjuntos indicados na Tabela 3.2.

Tabela 3.2: Representação dos conjuntos de cada instância da base exemplo

Instâncias da classe A	Instâncias da classe B
T_1 = {a, b, c}	T_3 = {d, e, f, g}
T_2 = {a, c, d, e}	T_4 = {d, e, f}
	T_5 = {c, e, g}

As ocorrências de cada um dos itens do domínio de X podem ser assim resumidas, na

Tabela 3.3.

Tabela 3.3: Ocorrência dos valores do atributo X na base exemplo

Item	Ocorrências na base	Ocorrências na classe A	Ocorrências na classe B
"a"	2	2	0
"b"	1	1	0
"c"	3	2	1
"d"	3	1	2
"e"	4	1	3
"f"	2	0	2
"g"	2	0	2

Calculando as probabilidades das classes para cada item de X, obtêm-se: P_A = [1, 1, 2/3, 1/3, 1/4, 0, 0], P_B = [0, 0, 1/3, 2/3, 3/4, 1, 1] e P_D = [1, 1, 1/3, 1/3, 2/4, 1, 1].

Utilizando os valores do vetor P_D e as quantidades de ocorrências de cada item na base, pode-se calcular o valor da medida de importância de X, aplicando-se a Fórmula 3.4, conforme indicado a seguir.

$$I(X) = \frac{(1*2)+(1*1)+(1/3*3)+(1/3*3)+(2/4*4)+(1*2)+(1*2)}{2+1+3+3+4+2+2} = 11/17 = 0,647.$$

Para avaliar a medida proposta, foram realizados experimentos com diversas bases, submetidas a classificadores multi-relacionais. As acurácias resultantes da classificação foram observadas em conjunto com as medidas de relevância calculadas para os atributos utilizados na classificação. Medidas de relevância que indicam melhores atributos devem estar compatíveis com acurácias melhores e vice-versa. No próximo Capítulo são descritas as bases de dados utilizadas e os experimentos realizados.

Capítulo 4

Experimentos e Resultados

Este capítulo tem como objetivo descrever as bases de dados utilizadas nos experimentos e mostrar os resultados obtidos. A medida de relevância proposta foi aplicada a cada base e seus resultados foram comparados com a acurácia do classificador, para comprovar sua validade. Quando a medida indica que um atributo multivalorado é relevante, espera-se que a acurácia do classificador melhore com a utilização desse atributo. Por outro lado, quando a medida indica que um atributo multivalorado não é relevante, a tendência é que a acurácia do classificador não se altere (caso o atributo multivalorado seja irrelevante) ou piore (caso o atributo represente uma informação confusa para a classificação).

Foram utilizadas bases de dados reais e posteriormente bases híbridas, onde os atributos convencionais são os mesmos das bases reais e os atributos multivalorados foram gerados sinteticamente a partir de critérios definidos. As bases híbridas foram criadas com o objetivo de verificar, de maneira controlada, como a medida proposta se comporta para atributos multivalorados com características distintas.

Para realizar a classificação nas bases de dados relacionais, este trabalho utiliza a ferramenta Relational Weka [6], desenvolvida a partir da conhecida ferramenta Weka [7]. Conforme descrito no Capítulo 2, foram usadas três medidas de distância na realização da classificação *lazy*: RIBL, Tanimoto e *Average Linking* (AL), aplicadas ao algoritmo k-NN. Em todas as execuções foi realizada validação cruzada com 10 partições e o parâmetro k do k-NN foi fixado com o valor igual a três (k=3).

4.1 Bases de Dados Reais

Foram utilizadas nos experimentos quatro bases de dados reais, todas com classes binárias. Essas bases foram escolhidas por conterem uma ou mais tabelas que implementam um atributo multivalorado. As bases KDD Cup 2000 e Ebooks foram utilizadas em trabalhos anteriores de mineração multi-relacional [27]. Não foram encontradas referências, na literatura, das bases IBGE POF (1999 e 2002) sendo utilizadas em trabalhos de mineração de dados.

4.1.1 KDD Cup 2000

A base KDD Cup 2000 descreve transações de vendas coletadas no *site* Gazelle.com através do software *Blue Martini*. Esse *site* vende produtos de cuidados para as pernas. Os dados foram coletados diretamente do *site* (`http://cobweb.ecn.purdue.edu/KDDCUP/data/`), com autorização requerida, onde já havia sido feito um tratamento mínimo de limpeza. A classe utilizada para a base KDD é a classe original definida na especificação da base, que tem como objetivo separar as instâncias de acordo com o valor gasto em compras no *site*. A Figura 4.1 apresenta a distribuição das instâncias da base KDD para a classe em questão.

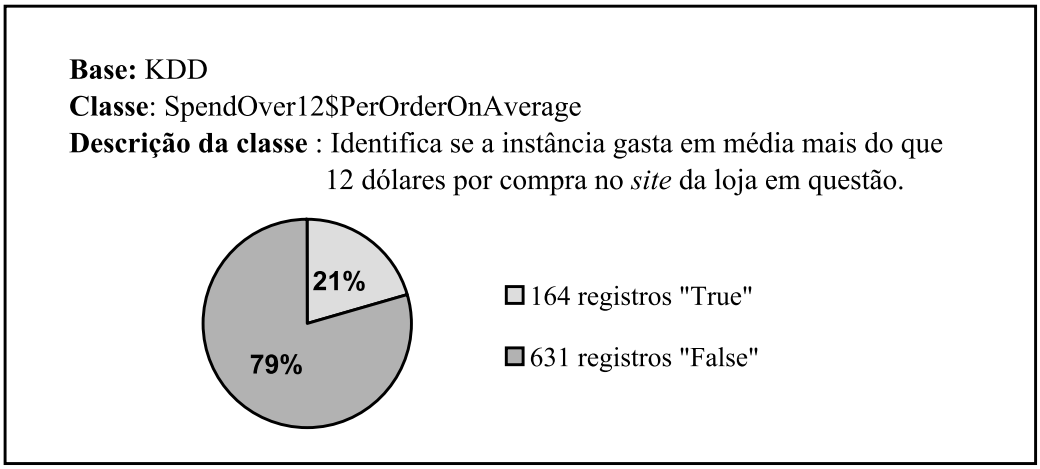

Figura 4.1: Distribuição das instâncias da base KDD para a classe "Spend Over 12 Per Order On Average"

A Figura 4.2 apresenta o relacionamento entre as tabelas da base. Cada cliente pode conter um conjunto de produtos e um conjunto de coleções. As instâncias da base a serem classificadas são representadas pelos registros da tabela de Clientes. São utilizados dois atributos multivalorados nessa base: um deles é representado pelo conjunto de produtos adquiridos pelos clientes (tabela Produtos) e o outro é representado pelo conjunto de

rótulos das coleções dos produtos adquiridos pelos clientes (tabela Coleção). Os atributos monovalorados são armazenados na tabela Cliente. Pela análise do contexto, pode-se inferir que o atributo "Coleção" é uma agregação do atributo "Produto". Porém, a base de dados obtida não especifica essa hierarquia explicitamente.

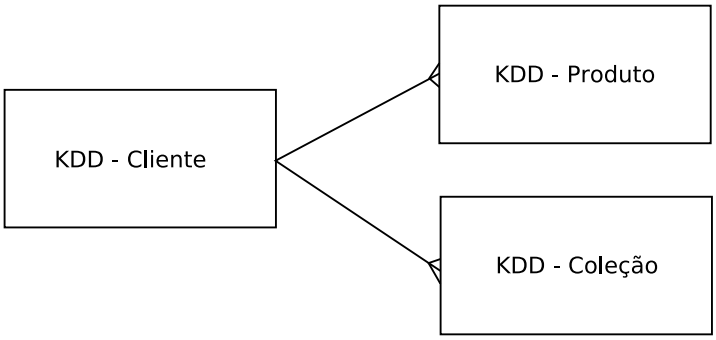

Figura 4.2: Base KDD - Tabelas do modelo de dados

A Tabela 4.1 apresenta as quantidades de registros de cada tabela da base KDD e as quantidades de atributos, por tipo.

Tabela 4.1: Base KDD - Quantidades de registros das tabelas e quantidades de atributos, por tipo.

KDD - Cliente
Quantidade de registros: 795
Quantidade de atributos categóricos: 49
Quantidade de atributos numéricos: 6

KDD - Produto
Quantidade de registros: 1044
Quantidade de atributos categóricos: 2 (IDCliente e IDProduto)

KDD - Coleção
Quantidade de registros: 810
Quantidade de atributos categóricos: 2 (IDCliente e IDColeção)

4.1.2 EBooks

A base EBooks contém dados sobre vendas de livros pela internet realizadas por uma empresa coreana emergente. A descrição detalhada dessa base pode ser obtida em [27]. Como na especificação original da base não existia uma classe definida, foram escolhidos dois atributos para realizarem o papel de classes nos experimentos deste trabalho: sexo

e filhos. As Figuras 4.3 e 4.4 apresentam a distribuição das instâncias da base EBooks para as classes "*Sex*" e "*Kids*", respectivamente. A documentação da base não identificou o significado dos valores {1,2} para a classe *Sex* e {0,1} para a classe *Kids*.

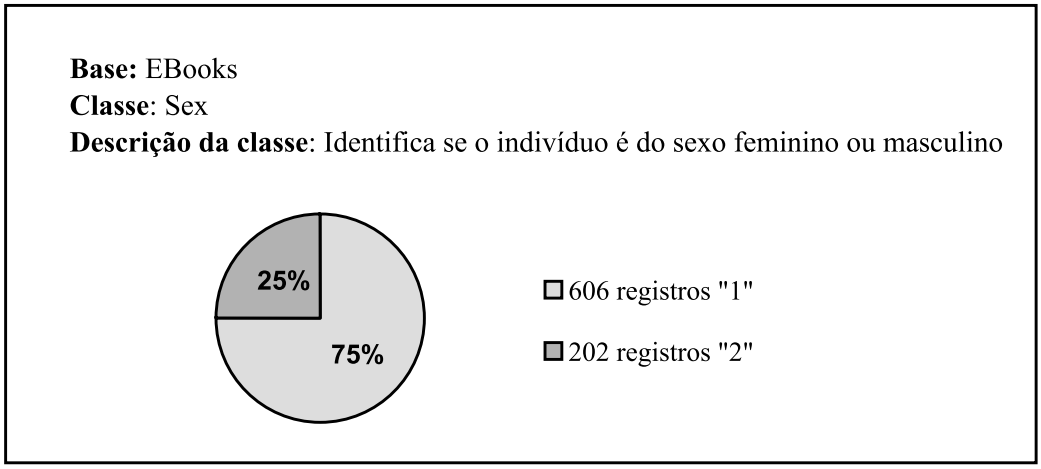

Figura 4.3: Distribuição das instâncias da base EBooks para a classe "*Sex*"

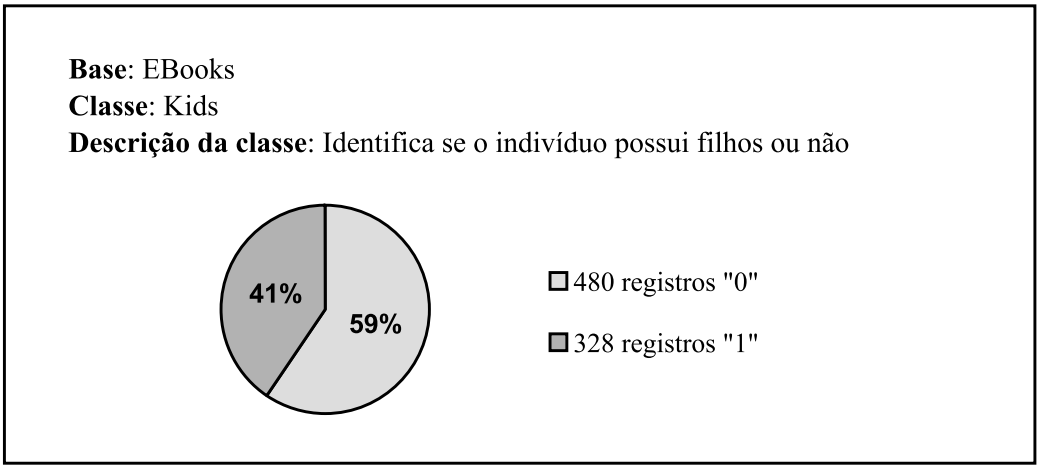

Figura 4.4: Distribuição das instâncias da base EBooks para a classe "*Kids*"

A Figura 4.5 apresenta o relacionamento entre as tabelas da base. Cada consumidor pode conter um conjunto de livros e um conjunto de categorias de livros. As instâncias da base a serem classificadas são representadas pelos registros da tabela de consumidores (*Customers*). São utilizados dois atributos multivalorados nessa base: um deles é representado pelo conjunto de livros adquiridos pelos clientes (tabela *Book*) e o outro é representado pelo conjunto de categorias dos livros adquiridos pelos clientes (tabela *Category*). Os atributos monovalorados são armazenados na tabela *Customers*. Pela análise do contexto, pode-se inferir que o atributo "*Category*" é uma agregação do atributo "*Book*". Porém, a base de dados obtida não especifica essa hierarquia explicitamente.

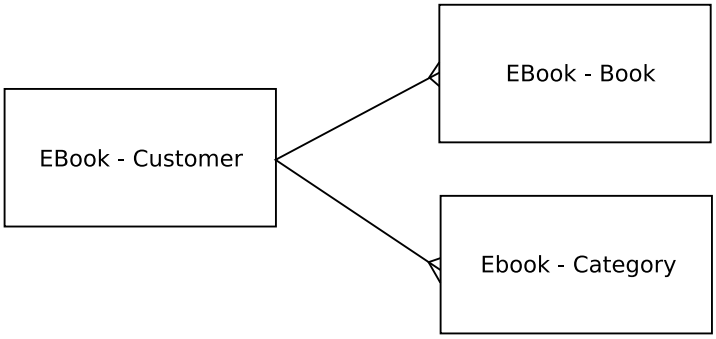

Figura 4.5: Base EBooks - Tabelas do modelo de dados

A Tabela 4.2 apresenta as quantidades de registros de cada tabela da base EBooks e as quantidades de atributos, por tipo.

Tabela 4.2: Base EBooks - Quantidades de registros das tabelas e quantidades de atributos, por tipo.

EBooks - Customer
Quantidade de registros: 808
Quantidade de atributos categóricos: 6
Quantidade de atributos numéricos: 0

EBooks - Book
Quantidade de registros: 17.213
Quantidade de atributos categóricos: 2 (IDCustomer e IDBook)

EBooks - Category
Quantidade de registros: 1.294
Quantidade de atributos categóricos: 2 (IDCustomer e IDCategory)

4.1.3 IBGE POF 2002

A base IBGE POF 2002 descreve uma pesquisa de orçamentos familiares, realizada pelo IBGE em 2002. Os indivíduos a serem classificados são moradores de diversas regiões do Brasil. Esses moradores possuem características individuais e possuem uma lista de produtos adquiridos, que representam o atributo multivalorado. A base foi adquirida do IBGE. Como na especificação original da base não existia uma classe definida, foram escolhidos dois atributos para realizarem o papel de classes nos experimentos deste trabalho: cartão de crédito e sexo. As Figuras 4.6 e 4.7 apresentam a distribuição das instâncias da base IBGE 2002 para as classes "Cartão de crédito" e "Sexo", respectivamente.

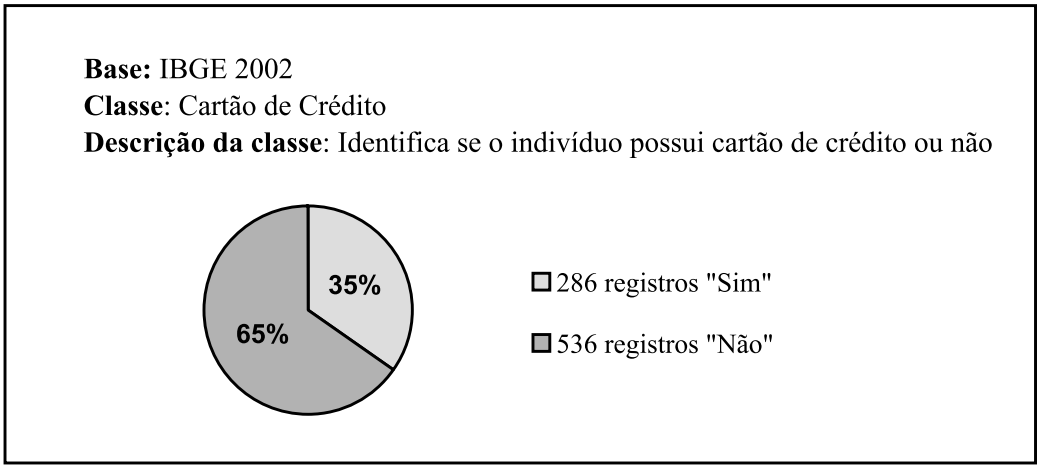

Figura 4.6: Distribuição das instâncias da base IBGE 2002 para a classe "Cartão de crédito"

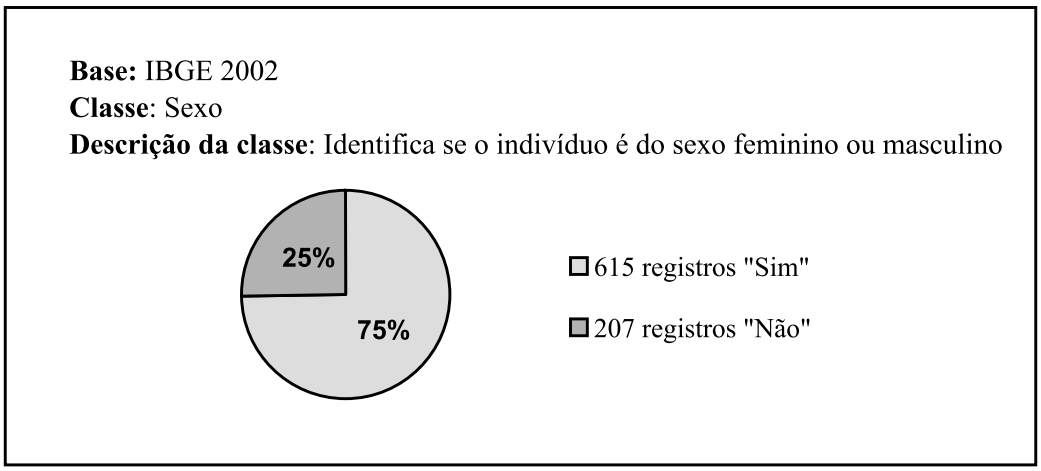

Figura 4.7: Distribuição das instâncias da base IBGE 2002 para a classe "Sexo"

A Figura 4.8 apresenta o relacionamento entre as tabelas da base. Cada morador pode conter um conjunto de produtos e um conjunto de categorias de produtos. As instâncias da base a serem classificadas são representadas pelos registros da tabela Morador. São utilizados dois atributos multivalorados nessa base: um deles é representado pelo conjunto de produtos adquiridos pelos moradores (tabela Produto) e o outro é representado pelo conjunto de categorias dos produtos adquiridos pelos moradores (tabela Categoria). Os atributos monovalorados são armazenados na tabela Morador. Nas especificações dessa base fica claro que a categoria é uma agregação de produtos. No entanto, o modelo representado pela Figura 4.8 – que liga a tabela de categorias à tabela de moradores, e não à tabela de produtos – mostra a forma como as tabelas estão sendo tratadas nos experimentos deste trabalho, sem nenhuma influência direta entre os atributos "Produto" e "Categoria".

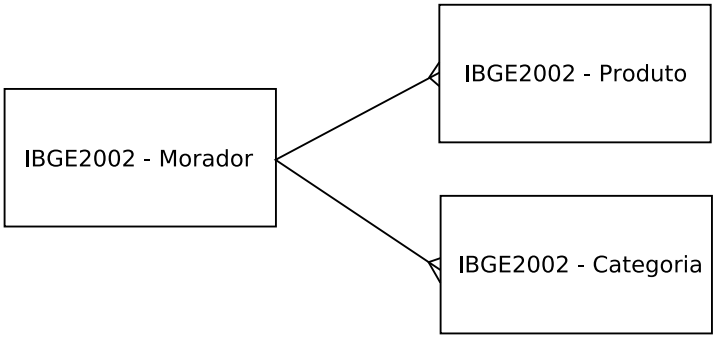

Figura 4.8: Base IBGE 2002 - Tabelas do modelo de dados

A Tabela 4.3 apresenta as quantidades de registros de cada tabela da base IBGE 2002 e as quantidades de atributos, por tipo.

Tabela 4.3: Base IBGE 2002 - Quantidades de registros das tabelas e quantidades de atributos, por tipo.

IBGE 2002 - Morador
Quantidade de registros: 822
Quantidade de atributos categóricos: 7
Quantidade de atributos numéricos: 2

IBGE 2002 - Produto
Quantidade de registros: 2.434
Quantidade de atributos categóricos: 2
(IDMorador e IDProduto)

IBGE 2002 - Categoria
Quantidade de registros: 1.471
Quantidade de atributos categóricos: 2
(IDMorador e IDCategoria)

4.1.4 IBGE POF 1999

A base IBGE POF 1999 descreve uma pesquisa de orçamentos familiares, realizada pelo IBGE em 1999. Os indivíduos a serem classificados são famílias de diversas regiões do Brasil. Essas famílias possuem características próprias e uma lista de produtos adquiridos, que representam o atributo multivalorado. A base foi adquirida do IBGE. Como na especificação original da base não existia uma classe definida, foram escolhidos dois atributos para realizarem o papel de classes nos experimentos deste trabalho: faixa de renda e tamanho da família. As Figuras 4.9 e 4.10 apresentam a distribuição das instâncias da base IBGE 1999 para as classes “Faixa de renda” e “Tamanho da família”, respectivamente.

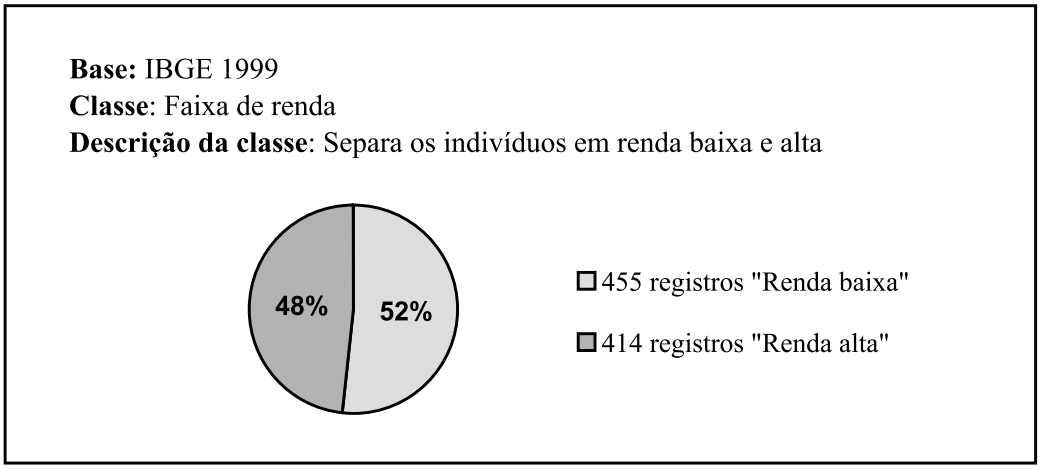

Figura 4.9: Distribuição das instâncias da base IBGE 1999 para a classe "Faixa de renda"

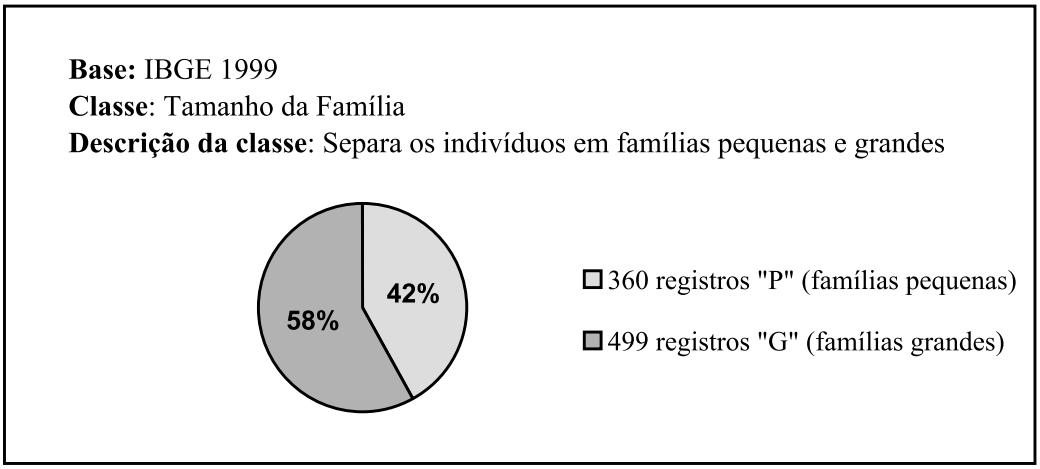

Figura 4.10: Distribuição das instâncias da base IBGE 1999 para a classe "Tamanho da família"

A Figura 4.11 apresenta o relacionamento entre as tabelas da base. Cada família pode conter um conjunto de produtos. As instâncias da base a serem classificadas são representadas pelos registros da tabela Família. Nessa base só existe um atributo multivalorado, representado pelo conjunto de produtos adquiridos pelas famílias (tabela Produto). Os atributos monovalorados são armazenados na tabela Família.

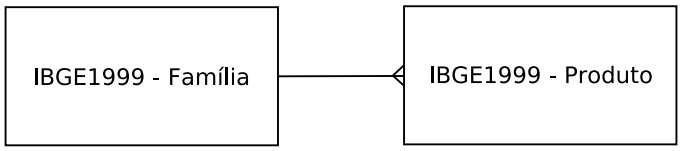

Figura 4.11: Base IBGE 1999 - Tabelas do modelo de dados

A Tabela 4.4 apresenta as quantidades de registros de cada tabela da base IBGE 1999 e as quantidades de atributos, por tipo.

Tabela 4.4: Base IBGE 1999 - Quantidades de registros das tabelas e quantidades de atributos, por tipo.

IBGE 1999 - Família
Quantidade de registros: 859
Quantidade de atributos categóricos: 3
Quantidade de atributos numéricos: 1

IBGE 1999 - Produto
Quantidade de registros: 41.118
Quantidade de atributos categóricos: 2 (IDMorador e IDProduto)

4.2 Bases de Dados Reais - Resultados e Análises

Nesta seção são apresentados os resultados obtidos nos experimentos com as bases reais e uma análise desses valores.

Para cada base e classe avaliada, são apresentados os seguintes resultados:

- Gráficos e/ou tabelas com os valores de acurácia obtidos através da classificação das bases (para as três medidas de distância utilizadas).

- Valor da medida de relevância proposta para os atributos multivalorados em questão.

Serão feitos dois tipos de avaliação:

A1) Para cada atributo multivalorado avaliado, verificar se o valor da medida de relevância é compatível com a acurácia da classificação quando esse atributo é utilizado isoladamente, ou seja, como o único atributo utilizado pelo classificador. Para um valor alto da medida de relevância (acima de 0,7), a acurácia deve ser alta. Já para um valor baixo da medida de relevância, a acurácia tende a ser igual ou pior do que o percentual da classe mais freqüente.

A2) Comparar os resultados de acurácia da classificação para uma determinada base em duas situações: (a) utilização de um conjunto de atributos que engloba os outros atributos da base (monovalorados) e mais o atributo multivalorado e (b) utilização de um conjunto de atributos que engloba somente os atributos monovalorados, sem a participação do atributo multivalorado. Analisar o comportamento desses dois contextos e verificar se o valor da medida de relevância do atributo multivalorado é coerente: se a medida de relevância for boa, a tendência é que o atributo multivalorado melhore a classificação

(quando comparadas as situações (a) e (b)); caso contrário, a tendência é que a acurácia piore ou fique estável. Vale lembrar que o resultado desse experimento é muito sensível à qualidade dos atributos monovalorados que estão sendo combinados com o atributo multivalorado.

Tabela 4.5: Valores da medida de relevância dos atributos multivalorados e suas respectivas acurácias.

Base/Classe: %Classe Mais Freq.	**Atributo**	**Medida Prob**	**Medida Dist**	**Acurácia**
KDD/Spend: 79%	Produto	0,741	RIBL Tanimoto AL	81,89 81,13 81,00
	Coleção	0,721	RIBL Tanimoto AL	82,39 83,27 82,89
EBooks/Sex: 75%	Book	0,795	RIBL Tanimoto AL	79,33 84,03 84,40
	Category	0,576	RIBL Tanimoto AL	75,00 75,25 75,62
EBooks/Kids: 59%	Book	0,508	RIBL Tanimoto AL	60,64 61,14 60,77
	Category	0,235	RIBL Tanimoto AL	59,40 60,15 59,40
IBGE 2002/Cartão: 65%	Produto	0,348	RIBL Tanimoto AL	62,65 65,94 65,08
	Categoria	0,259	RIBL Tanimoto AL	63,87 64,35 65,81
IBGE 2002/Sexo: 75%	Produto	0,603	RIBL Tanimoto AL	75,79 77,00 76,64
	Categoria	0,508	RIBL Tanimoto AL	74,69 77,00 73,84
IBGE 1999/Renda: 52%	Produto	0,229	RIBL Tanimoto AL	42,26 65,31 54,48
IBGE 1999/Família: 58%	Produto	0,226	RIBL Tanimoto AL	41,91 60,42 50,41

Os resultados da avaliação (A1), descritos na Tabela 4.5, demonstram que a medida de relevância proposta pode ser um bom indicador da importância do atributo multivalorado para a classificação. Na maioria dos casos, valores altos da medida de relevância (acima de 0,7) correspondem a valores altos de acurácia. É importante ressaltar que uma acurácia é alta quando seu percentual é maior do que o percentual da classe mais freqüente. Uma acurácia de 80% não pode ser considerada alta se 80% das instâncias pertencem à mesma

classe. Um algoritmo qualquer poderia simplesmente rotular todas as instâncias com o *label* da classe majoritária e conseguiria, sem nenhuma "inteligência", acertar 80% da classificação.

Por exemplo, para base KDD/Spend a medida de relevância do atributo Produto é alta (0,741) e a acurácia para todas as medidas também é alta (em torno de 81%) e maior do que o percentual da classe mais freqüente (79%). Já para a base IBGE 2002/Cartão, a medida de relevância do atributo Produto é baixa (0,348) e a acurácia para todas as medidas também é baixa (variando de 62% a 65%), sendo que o percentual da classe mais freqüente é 65%. Esse comportamento aconteceu na maioria dos casos. Porém, algumas exceções ocorreram apontando valores altos de acurácia com a utilização de atributos considerados ruins pela medida de relevância. Alguns desses valores podem ser vistos na base IBGE 1999 da Tabela 4.5. Uma possível explicação para este comportamento é que a medida proposta leva em consideração os valores dos conjuntos individualmente, sem avaliar a combinação entre eles. Caso a combinação de valores dentro de um conjunto (atributo multivalorado) seja relevante, a medida proposta pode não enxergar essa importância e considerar este atributo "ruim".

Os resultados da Tabela 4.5 não levam em consideração a influência dos atributos monovalorados da base no atributo multivalorado. Os valores de acurácia obtidos foram gerados a partir da classificação das instâncias com o uso do atributo multivalorado de forma isolada, como o único atributo considerado pelo classificador.

Os gráficos das Figuras 4.12, 4.13 e 4.14 exibem os resultados da avaliação (A2). No título dos gráficos encontram-se: a identificação da base e classe, o atributo multivalorado utilizado e o valor da medida de relevância proposta, identificado pelo rótulo "Medida Prob." (medida de probabilidade). O eixo y identifica as medidas de distâncias utilizadas na classificação e o eixo x identifica o valor da acurácia obtida pelo classificador, para cada uma das medidas. Esses valores podem ser visualizados com precisão como rótulos ao lado das barras. A legenda identifica os três grupos de atributos que foram utilizados pelo classificador: **MV <atributo multivalorado>** indica que apenas o atributo multivalorado em questão foi considerado na classificação; **MN (Bons ou Ruins)** identifica que apenas um grupo de atributos monovalorados foi considerado ("MN Bons" indica que foi utilizado o "melhor" grupo de atributos e "MN Ruins" indica que foi utilizado o "pior" grupo de atributos monovalorados); **MV <atributo multivalorado> + MN (Bons ou Ruins)** identifica que o grupo de atributos considerados pelo classificador engloba o atributo multivalorado e os atributos monovalorados.

Para mostrar a influência da qualidade dos atributos que são combinados com os atributos multivalorados nos valores de acurácia, foram feitas duas análises: na primeira coluna de gráficos o atributo multivalorado foi combinado com um conjunto de atributos monovalorados selecionados aleatoriamente (para a base KDD) ou com o pior grupo de atributos monovalorados (para as demais bases). Na segunda coluna de gráficos o atributo multivalorado foi combinado com o conjunto dos melhores atributos monovalorados da base. Para montar o conjunto dos "piores" atributos e dos "melhores" atributos monovalorados, foi utilizado um algoritmo de seleção executado na ferramenta Weka (*Attribute Evaluator: InfoGainAttributeEval; Search Method: Ranker*). A partir do *ranking* com todos os atributos monovalorados da base, o conjunto dos melhores atributos foi obtido através dos atributos mais cotados do *ranking*, e o conjunto dos piores atributos foi construído com aqueles que apareciam nos últimos lugares do *ranking*. A quantidade de atributos desses conjuntos variou de acordo com a quantidade de atributos disponíveis na base. Maiores detalhes sobre esses conjuntos podem ser obtidos no Apêndice A.

Pode-se observar que, nos casos em que a medida de relevância do atributo multivalorado é alta, a tendência é que ele melhore a acurácia da classificação quando combinado com os atributos monovalorados. Porém, essa melhora aparece de forma mais forte quando os atributos monovalorados não são muito bons. Se esse conjunto já era forte, ou seja, se a acurácia do classificador com a utilização do conjunto de atributos monovalorados isolados já era alta sem o atributo multivalorado, a tendência é que a inclusão desse atributo não faça muita diferença para a classificação.

Por outro lado, se a medida de relevância do atributo multivalorado é baixa, a tendência é que a inclusão desse atributo no conjunto dos atributos monovalorados não faça diferença ou até piore a acurácia da classificação. Quando o atributo multivalorado é ruim, ele pode distanciar as instâncias de mesma classe da base, dificultando o processo de classificação.

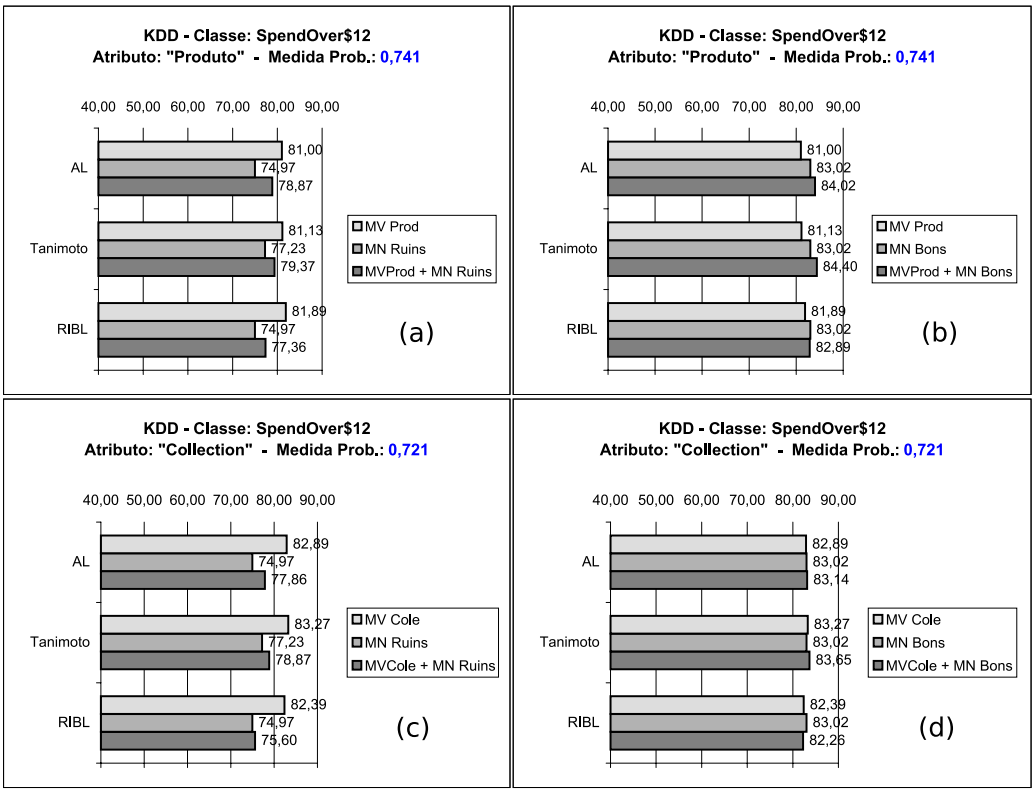

Figura 4.12: Base KDD - Comparação da acurácia da classificação quando o atributo multivalorado (MV) é combinado com atributos monovalorados (MN).

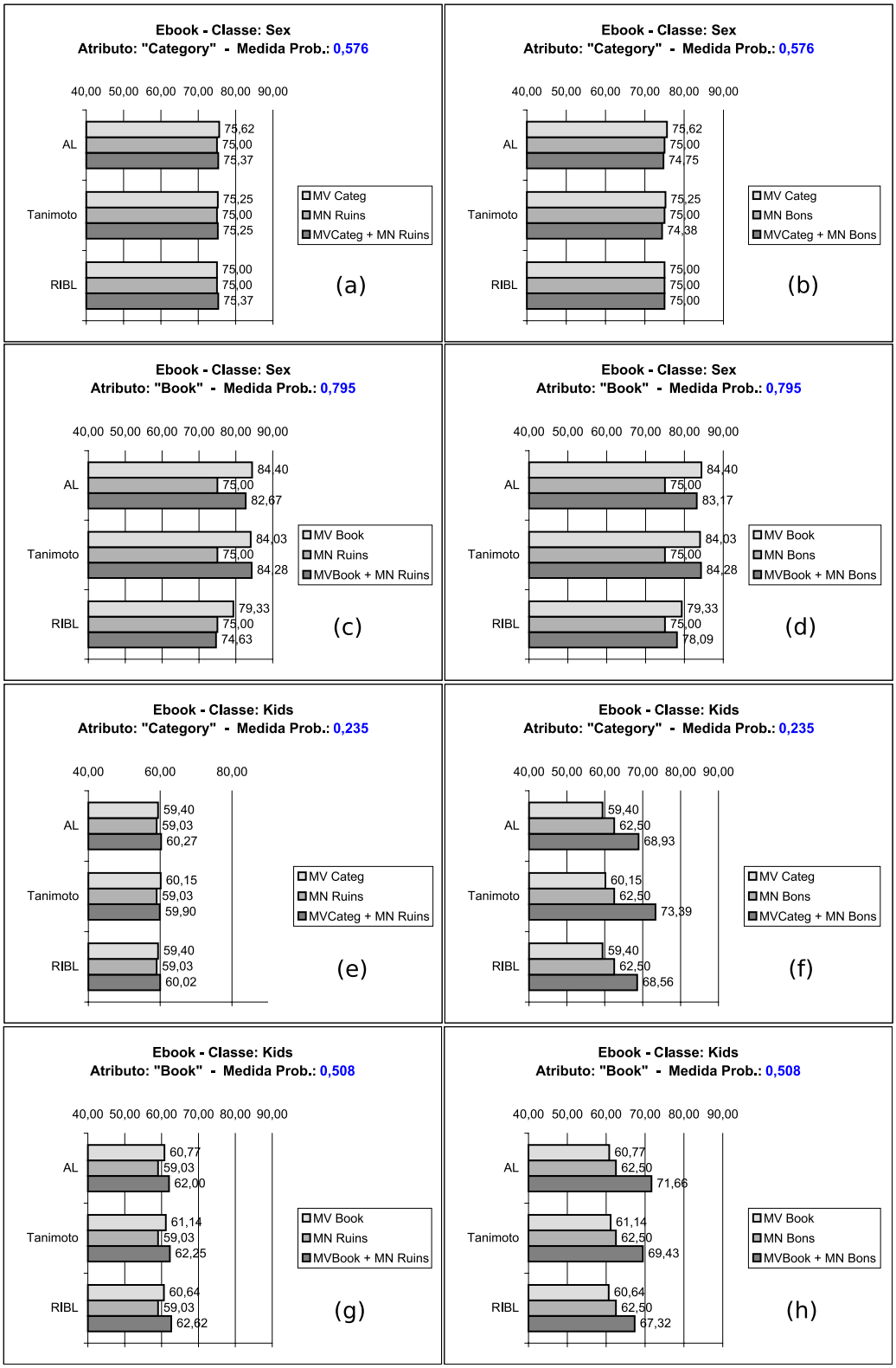

Figura 4.13: Base Ebooks - Comparação da acurácia da classificação quando o atributo multivalorado (MV) é combinado com atributos monovalorados (MN).

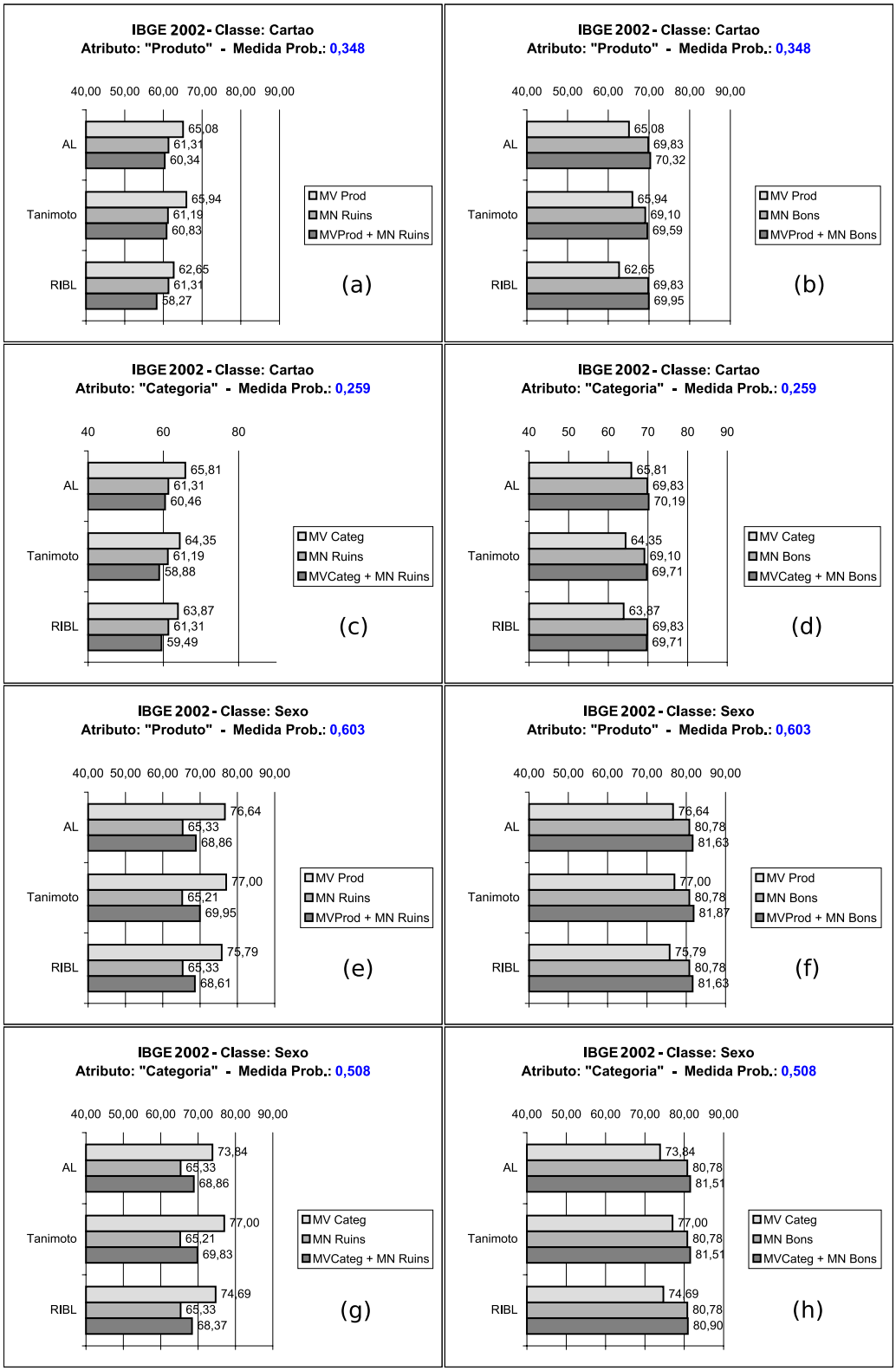

Figura 4.14: Base IBGE 2002 - Comparação da acurácia da classificação quando o atributo multivalorado (MV) é combinado com atributos monovalorados (MN).

Conforme já citado anteriormente, o valor da medida de relevância não garante que a inclusão do atributo multivalorado melhore ou piore a acurácia da classificação. O que existe é uma forte tendência. A influência que um atributo tem no outro pode fazer uma grande diferença, que não pode ser percebida pela medida proposta. Os gráficos (f) e (h) da Figura 4.13 demonstram claramente esse fato. Tanto a acurácia obtida com a utilização dos atributos multivalorados isolados quanto a acurácia obtida com a utilização dos atributos monovalorados eram baixas. Porém, quando combinados, representaram um aumento significativo na qualidade da classificação. Essa influência entre os atributos não será abordada neste trabalho, mas é importante ressaltar que um estudo completo sobre seleção de atributos para a classificação deveria levar em consideração esse ponto.

4.3 Bases de Dados Híbridas

As bases de dados híbridas utilizadas nos experimentos foram geradas a partir das bases reais descritas na seção 4.1. São chamadas híbridas porque a classe e todos os atributos monovalorados (numéricos e categóricos) são reais, enquanto o atributo multivalorado é gerado sinteticamente, para permitir maior controle sobre os experimentos. Os atributos multivalorados da base real original são descartados, permanecendo apenas aqueles gerados sinteticamente.

O procedimento de geração das bases híbridas é descrito a seguir:

1) A partir da base real, é gerado um vetor V, de tamanho n, onde n representa o tamanho do domínio do atributo multivalorado. Esse vetor possui uma estrutura com 4 registros: $V[i].Item$ guarda o identificador do item; $V[i].ProbA$ guarda a probabilidade da classe A ocorrer dado o valor $V[i].Item$; $V[i].ProbB$ guarda a probabilidade da classe B ocorrer dado esse mesmo valor; $V[i].Qtd$ guarda a quantidade de vezes que $V[i].Item$ aparece na base.

Os atributos sintéticos são gerados utilizando o vetor de probabilidades para realizar a distribuição dos valores de atributos, conforme descrito na Figura 4.15. Nas linhas 02 e 03 são calculadas, para cada item, as quantidades de vezes que ele deve ocorrer nas classes A e B, de acordo com o vetor de probabilidades. As variáveis $qtdClasseA$ e $qtdClasseB$ recebem o inteiro mais próximo dos produtos ($V[i].Qtd$ * $V[i].ProbA$) e ($V[i].Qtd$ * $V[i].ProbB$), respectivamente.

Nas linhas 05 e 09 da Figura 4.15 são selecionadas aleatoriamente as instâncias da base que terão esse item nos seus conjuntos. Nesses experimentos, não foram consideradas

repetições de itens para uma mesma instância. Ou seja, as instâncias são escolhidas aleatoriamente, com o cuidado de não serem repetidas para um mesmo item. Assim, o conjunto montado para cada instância não possui itens repetidos.

Seguindo essa lógica, os atributos multivalorados sintéticos possuem o mesmo domínio do atributo multivalorado real, a mesma quantidade de itens para cada valor e a mesma distribuição entre as classes. O que muda são os conjuntos de itens de cada instância. Caso na base real exista algum padrão de combinação de itens do domínio (ex.: instância que possui o item X também possui o item Y), esse padrão pode ser desfeito na geração do atributo sintético, pois os conjuntos são reconstruídos e as interações entre os itens não são consideradas.

```
procedure GeraAtributoSintetico(V)
01.  for i := 1 to n do begin
02.    qtdClasseA := (V[i].Qtd * V[i].ProbA);
03.    qtdClasseB := (V[i].Qtd * V[i].ProbB);
04.    for j := 1 to qtdClasseA do begin
05.      seleciona aleatoriamente uma instância da classe A;
06.      adiciona o item V[i].Item ao conjunto da instância escolhida;
07.    end;
08.    for j := 1 to qtdClasseB do begin
09.      seleciona aleatoriamente uma instância da classe B;
10.      adiciona o item V[i].Item ao conjunto da instância escolhida;
11.    end;
12.  end;
```

Figura 4.15: Pseudo-código da geração do atributo sintético

2) Para cada vetor V originado de uma base real, são gerados 4 novos vetores:

- V_1: possui os mesmos itens e quantidades do vetor original. As probabilidades são recalculadas de forma que o valor absoluto da diferença entre $V[i].ProbA$ e $V[i].ProbB$ seja aumentado em 40%.

- V_2: possui os mesmos itens e quantidades do vetor original. As probabilidades são recalculadas de forma que o valor absoluto da diferença entre $V[i].ProbA$ e $V[i].ProbB$ seja aumentado em 80%.

- V_3: possui os mesmos itens e quantidades do vetor original. As probabilidades são recalculadas de forma que o valor absoluto da diferença entre $V[i].ProbA$ e $V[i].ProbB$ seja diminuído em 40%.

- V_4: possui os mesmos itens e quantidades do vetor original. As probabilidades

são recalculadas de forma que o valor absoluto da diferença entre $V[i].ProbA$ e $V[i].ProbB$ seja diminuído em 80%.

É importante ressaltar que essa variação das diferenças de probabilidades é controlada pelos limites que cada uma das probabilidades pode chegar, ou seja, o menor valor não pode ser inferior a 0 (zero) e o maior valor não pode ultrapassar 1 (um). E ainda, a soma dos valores de $V[i].ProbA$ e $V[i].ProbB$ deve ser sempre igual a 1 (um). Assim, para aumentar as diferenças entre as probabilidades, o menor valor entre $V[i].ProbA$ e $V[i].ProbB$ é diminuído de um fator x enquanto que o maior valor é aumentado nesse mesmo fator, respeitando os limites superiores e inferiores. Para diminuir as diferenças, a operação inversa é realizada, ou seja, o menor valor entre $V[i].ProbA$ e $V[i].ProbB$ é aumentado de x e o maior valor diminuído desse mesmo fator.

Essas limitações na regra de variação das diferenças entre as probabilidades impede que a média de diferenças varie exatamente no percentual aplicado, já que alguns pares $V[i].ProbA$ e $V[i].ProbB$ podem não ser alterados por já estarem nos limites máximos e mínimos.

As regras de montagem dos novos vetores foram criadas com o objetivo de tornar os valores do atributo multivalorado mais ou menos descritivos. Porém, não espera-se que o valor da medida de relevância proposta acompanhe essa mesma variação de percentual, até porque a aplicação do percentual não é homogênea em todos os pares (conforme explicado no parágrafo anterior).

3) A partir do vetor original da base real e dos novos vetores calculados, são geradas 5 bases híbridas:

- $Base0$: o atributo multivalorado é gerado a partir do vetor V, isto é, o vetor calculado diretamente a partir da base real. A medida de relevância tem exatamente o mesmo valor da base original. Mudam apenas os conjuntos de itens gerados.

- $Base1$: o atributo multivalorado é gerado a partir do vetor V_1. O valor da medida de relevância tende a ser maior do que na Base0.

- $Base2$: o atributo multivalorado é gerado a partir do vetor V_2. O valor da medida de relevância tende a ser maior do que na Base1.

- $Base3$: o atributo multivalorado é gerado a partir do vetor V_3. O valor da medida de relevância tende a ser menor do que na Base0.

- *Base*4: o atributo multivalorado é gerado a partir do vetor V_4. O valor da medida de relevância tende a ser menor do que na Base3.

Espera-se, com os resultados da classificação, que as acurácias da *Base*1 e da *Base*2 sejam melhores do que a da *Base*0 e que as acurácias da *Base*3 e da *Base*4 sejam piores que da *Base*0, acompanhando o comportamento do valor da medida de relevância.

4.4 Bases de Dados Híbridas - Resultados e Análises

Nesta seção são apresentados os resultados obtidos com a classificação das bases híbridas, geradas conforme especificado na seção anterior. Analogamente aos experimentos com as bases reais, foram utilizadas três medidas de distâncias na execução da classificação *lazy*: RIBL, Tanimoto e *Average Linking* (AL).

Os resultados estão organizados da mesma forma, ou seja:

- Gráficos com os valores de acurácia obtidos através da classificação das bases (para as três medidas de distância utilizadas).

- Valor da medida de relevância proposta para os atributos multivalorados em questão.

As avaliações (A1) e (A2) feitas com as bases reais foram feitas também com as bases híbridas e resultados semelhantes foram obtidos. Dois novos tipos de avaliação foram projetados especificamente para as bases híbridas, levando-se em conta a possibilidade de variar as probabilidades das classes para cada valor do atributo:

A3) Com as variações realizadas nas bases híbridas, aumentando e diminuindo as diferenças entre as probabilidades das classes, a medida de relevância tem seu valor alterado. Essa avaliação tem como objetivo verificar se a acurácia do classificador acompanha o comportamento da medida de relevância, ou seja, para medidas mais altas espera-se acurácias mais altas e vice-versa.

A4) Verificar se a medida de relevância pode ser usada para comparar dois atributos multivalorados dentro de uma mesma base. Para um atributo com medida de relevância maior, espera-se uma acurácia melhor na classificação. Vale lembrar que as acurácias foram obtidas com a utilização do atributo multivalorado isolado na classificação, sem a influência dos atributos monovalorados.

Os gráficos das Figuras 4.16 e 4.17 mostram os resultados correspondentes à avaliação (A3): a primeira coluna de gráficos exibe as acurácias da classificação com a utilização de cada atributo multivalorado; a coluna da direita exibe os valores da medida de relevância proposta. Para avaliar o comportamento dos valores, o eixo x foi colocado na ordem (intuitiva) do menor para o maior valor da medida de relevância. Os menores valores são encontrados nas bases em que as diferenças entre as probabilidades foram diminuídas ($Base4$: -80% e $Base3$: -40%) e os maiores valores são encontrados nas bases em que as diferenças entre as probabilidades foram aumentadas ($Base1$: +40% e $Base2$: +80%) em relação à base original. O valor da $Base0$, gerada a partir do vetor original, foi colocado no centro.

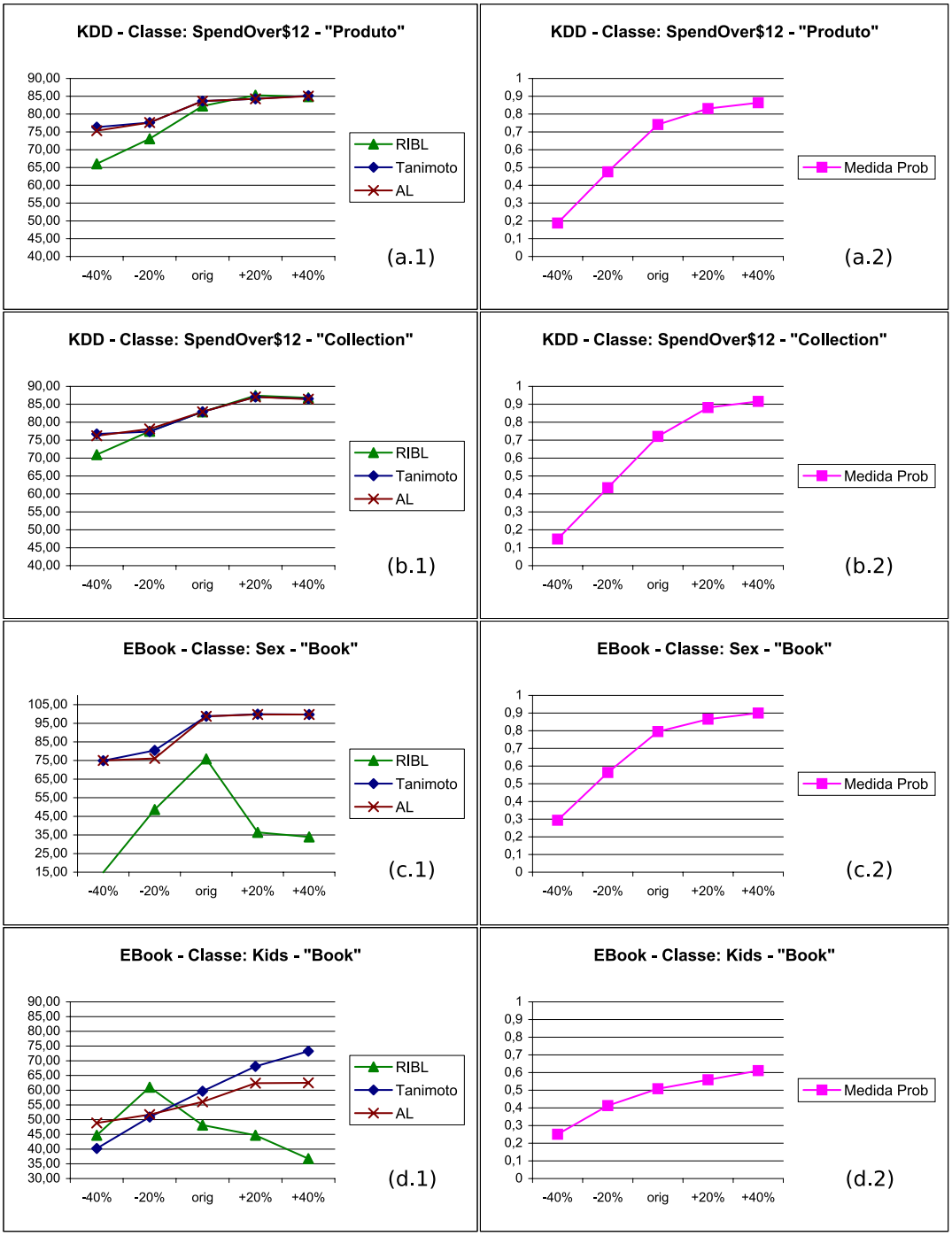

Figura 4.16: Variação da acurácia e da medida de relevância a partir da variação das diferenças entre as probabilidades das classes

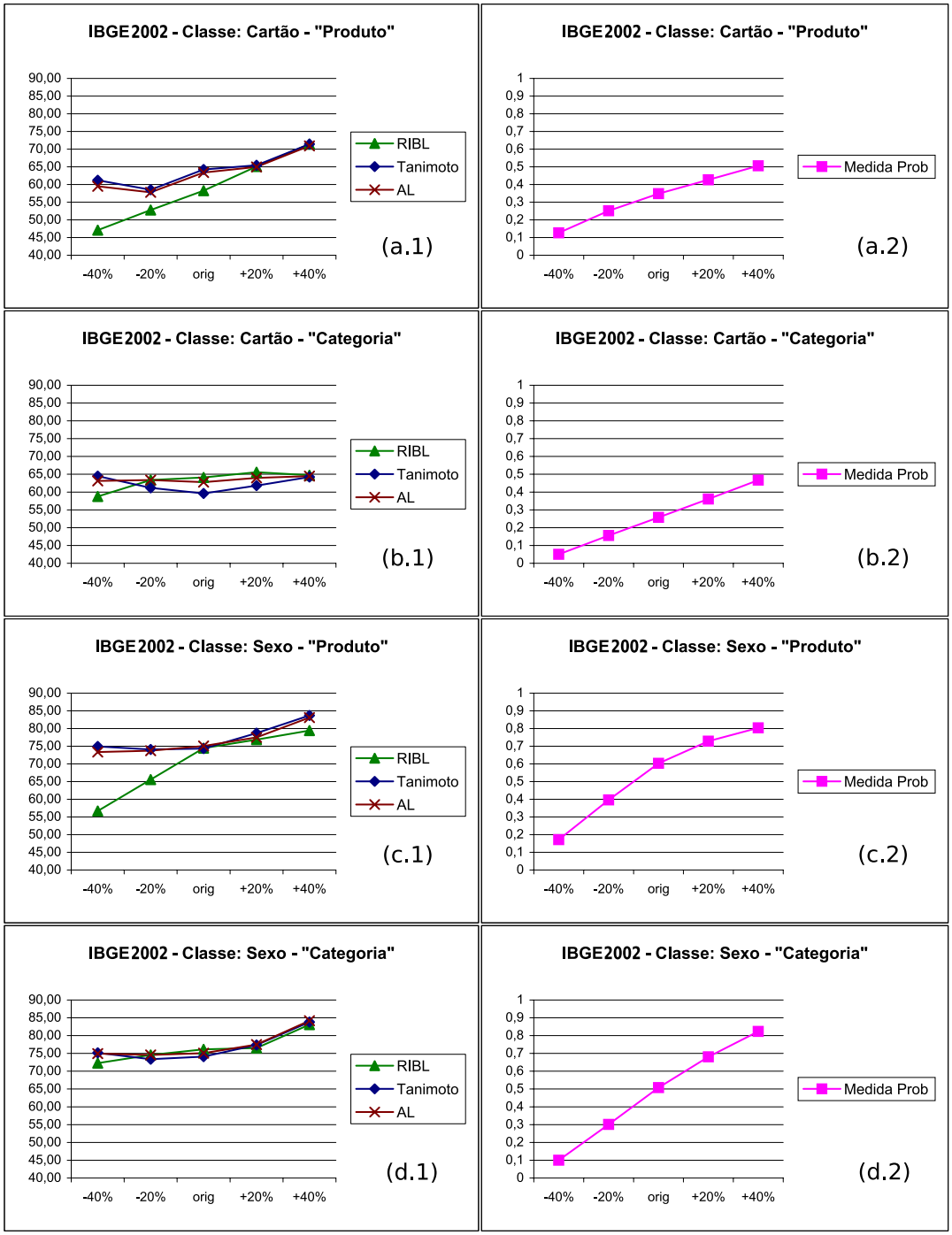

Figura 4.17: Variação da acurácia e da medida de relevância a partir da variação das diferenças entre as probabilidades das classes

Pode-se observar, através do conjunto de gráficos das Figuras 4.16 e 4.17, que na maioria dos casos a tendência de variação das acurácias acompanha exatamente o comportamento da medida de relevância proposta, apesar de ter uma taxa de variação menor. Apenas nos gráficos (c.1) e (d.1) da Figura 4.16 a acurácia gerada a partir da medida RIBL teve um comportamento contraditório. A base de dados representada por esses gráficos apresenta um domínio muito grande do atributo multivalorado (mais de 5000 itens), o que pode gerar, como conseqüência, conjuntos de tamanhos muito variados. Fazendo uma análise detalhada, foi possível observar que a medida RIBL é altamente sensível à diferença entre os tamanhos dos conjuntos que estão sendo comparados. Como ela soma as distâncias mínimas entre as distâncias dos elementos do menor conjunto para o maior, e depois divide pelo tamanho do maior conjunto, têm-se seguinte situação: quanto menor for o conjunto menor, menor será o valor do numerador; quanto maior for o conjunto maior, maior será o denominador, e conseqüentemente menor será a distância.

O exemplo da Tabela 4.6 ilustra a deficiência da medida RIBL para a situação exposta no parágrafo anterior. São utilizados, no exemplo, dois conjuntos A e B.

Tabela 4.6: Exemplo da deficiência da medida RIBL

Situação 1) $A = \{1,2,3\}$, $B = \{3,4,5\}$
RIBL = 0,66
Tanimoto = 0,80
AL = 0,88
Situação 2) $A = \{1,2,3\}$, $B = \{3,4,5,6\}$
RIBL = 0,50
Tanimoto = 0,83
AL = 0,91
Situação 3) $A = \{1,2,3\}$, $B = \{3,4,5,6,7,8,9,10\}$
RIBL = 0,25
Tanimoto = 0,90
AL = 0,96

Nota-se que as interseções nas três situações não mudaram. O que ocorreu foi o aumento do tamanho do segundo conjunto, gerando uma diferença maior entre as quantidades de A e B. Ao contrário da Tanimoto e *Average Linking*, onde a distância entre os conjuntos vai aumentando da situação 1 para a situação 3, a medida RIBL vai diminuindo, o que parece contraditório, já que os conjuntos vão ficando cada vez mais diferentes da situação 1 para a situação 3.

A Tabela 4.7 mostra a diferença entre o tamanho médio do menor e do maior conjunto para todas as bases híbridas. Esses tamanhos médios foram obtidos a partir das cinco bases híbridas geradas de cada base real. A base Ebooks com atributo Produto possui

uma diferença entre os tamanhos do menor e do maior conjunto muito maior do que as outras. Assim, pode-se inferir que o comportamento contraditório da acurácia nos gráficos (c.1) e (d.1) da Figura 4.16 deve-se realmente à incapacidade da medida RIBL de lidar corretamente com esses conjuntos de tamanhos muito diferentes.

Tabela 4.7: Variação dos tamanhos médios dos conjuntos dos atributos multivalorados nas bases híbridas

Base	**Atributo**	**Média Menor conj.**	**Média Maior conj.**	**Média Diferença**
EBooks	Book	7,6	39,4	31,8
	Category	1,0	3,0	2,0
IBGE	Produto	1,0	10,5	9,5
	Categoria	1,0	7,5	6,5
KDD	Produto	1,0	7,0	6,0
	Coleção	1,0	6,3	5,3

Apesar da base KDD real também possuir uma grande diferença entre os tamanhos dos conjuntos, o problema da medida RIBL não ocorreu nesse caso, porque a distribuição das quantidades é menos homogênea. Enquanto na base KDD híbrida existe uma quantidade balanceada entre conjuntos pequenos, médios e grandes, na base real existem muitos conjuntos menores e poucos conjuntos grandes. Ou seja, o problema da diferença grande entre os tamanhos não aparece com muita freqüência.

Um outro exemplo, demonstrado pela Tabela 4.8, pode evidenciar o comportamento distorcido da medida RIBL em comparação com a Tanimoto e *Average Linking*. Na situação (a) não existem interseções entre os conjuntos A e B. O correto seria considerar a distância entre os conjuntos igual a 1, e apenas a medida RIBL não apresenta esse resultado. Nas situações (b) e (c) existe uma interseção parcial entre os conjuntos, de forma que tanto em A quanto em B existem elementos exclusivos de um conjunto, que não existem no outro. A medida RIBL apresenta os mesmos valores de distância para as três situações.

Tabela 4.8: Comportamento das medidas mediante interseções parciais

(a)		**(b)**		**(c)**	
A = {1}		A = {1,2}		A = {1,2,3}	
B = {2,3,4}		B = {2,3,4}		B = {2,3,4}	
RIBL	0,33	RIBL	0,33	RIBL	0,33
Tanimoto	1,00	Tanimoto	0,75	Tanimoto	0,50
AL	1,00	AL	0,83	AL	0,77

A Tabela 4.9 apresenta uma outra situação, onde o conjunto A sempre se apresenta como um subconjunto de B, com variação de quantidades. Nesses casos específicos, a

medida RIBL sempre considera os conjuntos iguais, o que pode ser considerado um problema, dependendo do contexto que está sendo analisado. Se estamos buscando perfis de potenciais consumidores de livros para uma livraria, por exemplo, uma pessoa que compra 100 livros em determinado período não pode ser considerada semelhante à outra que compra apenas um livro, ainda que esse único livro esteja na lista do primeiro consumidor considerado. Para esses casos, a medida *Average Linking* também não tem um comportamento desejável, pois apesar de não considerar os conjuntos como "iguais", ela considera a mesma distância entre eles, independente da quantidade de itens. Apenas a medida Tanimoto apresenta uma variação coerente com as situações expostas.

Tabela 4.9: Comportamento das medidas mediante interseções totais

(a)	
A = {1}	
B = {1,2,3,4}	
RIBL	0,00
Tanimoto	0,75
AL	0,75

(b)	
A = {1,2}	
B = {1,2,3,4}	
RIBL	0,00
Tanimoto	0,50
AL	0,75

(c)	
A = {1,2,3}	
B = {1,2,3,4}	
RIBL	0,00
Tanimoto	0,25
AL	0,75

A análise (A4) é feita através do gráfico da Figura 4.18, onde são comparados dois atributos multivalorados na mesma base, com valores de medidas de relevância distintos. Em cada base real existiam dois atributos multivalorados (com exceção da base IBGE 1999, que só possui um), exemplificados aqui por a e b. Como, para cada base real, foram geradas cinco bases híbridas, foram criados, então, 10 atributos sintéticos – 5 atributos gerados a partir de a, $\{a_1, a_2, a_3, a_4, a_5\}$ e 5 atributos gerados a partir de b, $\{b_1, b_2, b_3, b_4, b_5\}$. A escolha dos atributos a serem comparados nesta análise foi feita a partir desses 10 atributos sintéticos, optando-se pelo par (a_i e b_j) de atributos com maiores diferenças nos valores de suas medidas de relevância.

Em todos os casos, apesar de variações maiores e menores, o atributo que apresentou maior valor da medida de relevância proposta deu origem a uma classificação de melhor acurácia. Assim, pode-se dizer que a medida de relevância proposta também pode ser utilizada para a comparação de atributos de uma mesma base, contribuindo para a seleção dos melhores atributos a serem considerados pelo classificador.

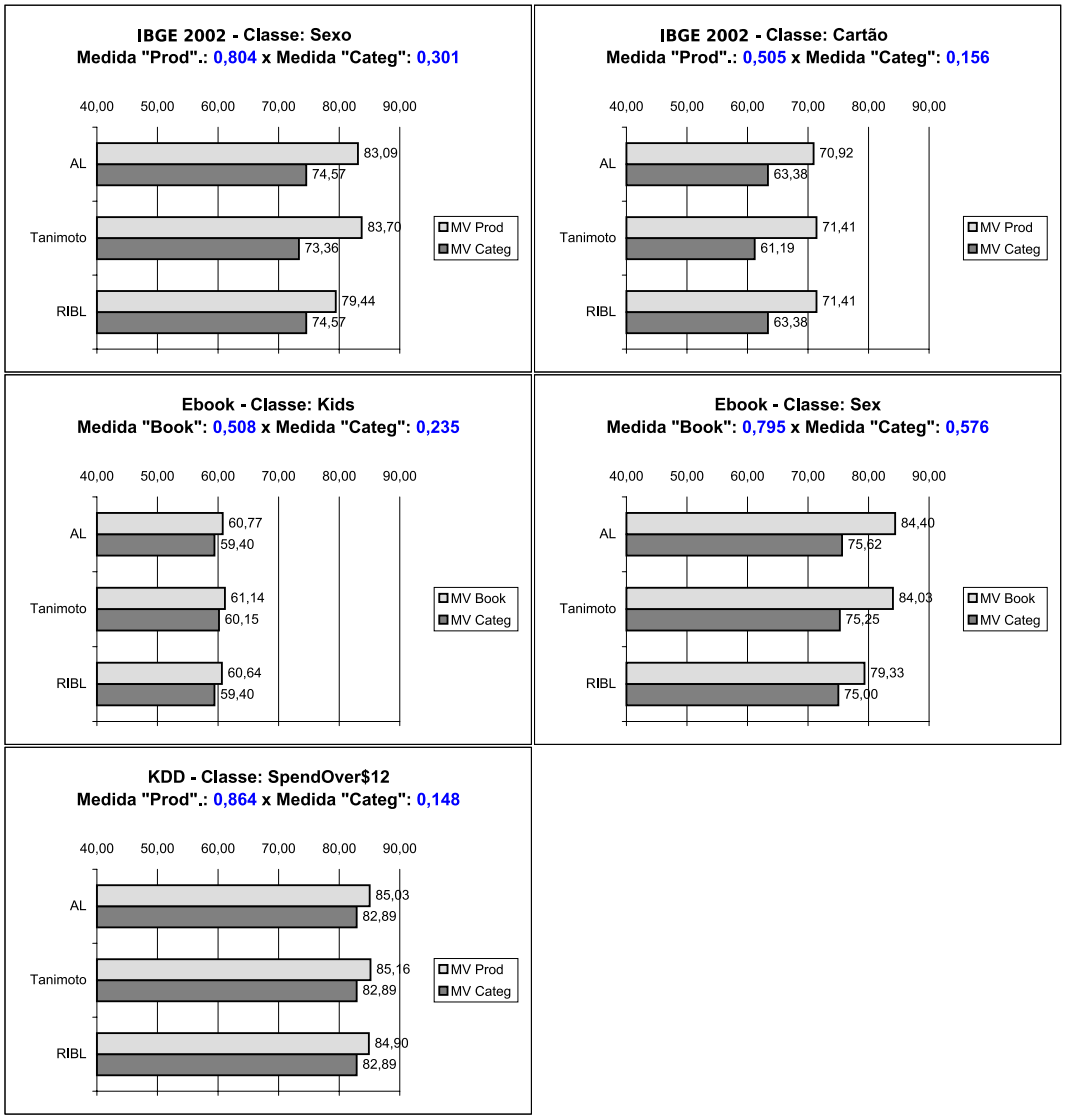

Figura 4.18: Comparação de dois atributos multivalorados da mesma base

Capítulo 5

Conclusões

O trabalho desenvolvido teve como objetivo propor e avaliar uma medida de relevância para atributos multivalorados no contexto de classificação. A medida proposta baseia-se na diferença das probabilidades de cada classe ocorrer, para cada valor do atributo multivalorado em questão. Quanto maior for essa diferença (em média, para todos os valores do atributo), maior será a relevância do atributo multivalorado segundo a medida proposta.

Como não existem na literatura outras medidas de relevância para esse tipo de atributo, o objetivo da avaliação não foi fazer uma análise comparativa com outras medidas, mas sim verificar se a medida proposta pode ser útil para selecionar atributos numa etapa que antecede o processo de classificação. Mais especificamente, foi utilizado um algoritmo de classificação *lazy*, o k-NN, com três diferentes medidas de distância para atributos multivalorados: RIBL, Tanimoto e *Average Linking*.

Foram feitos quatro tipos de testes, em bases reais e híbridas, e foi possível comprovar que na maioria das vezes a medida proposta é um bom indicador da qualidade de um atributo multivalorado. Essa medida poderia ser bastante útil em combinação com algoritmos de seleção de atributos que seguem a abordagem *filter*.

As avaliações realizadas nos experimentos e as conclusões obtidas a partir delas são descritas a seguir.

A1) Para cada atributo multivalorado avaliado, foi verificado que o valor da medida de relevância é compatível com a acurácia da classificação quando esse atributo é utilizado isoladamente, ou seja, como o único atributo utilizado pelo classificador. Para um valor alto da medida de relevância (acima de 0,7), a acurácia foi, na maioria das vezes, maior do que o percentual da classe mais freqüente. Por outro lado, para valores baixos da

medida de relevância, a acurácia se manteve equivalente ou abaixo do valor da classe mais freqüente.

A2) Foram comparados os resultados de acurácia da classificação para uma determinada base em duas situações: (a) utilização de um conjunto de atributos que engloba os outros atributos da base (monovalorados) e mais o atributo multivalorado e (b) utilização de um conjunto de atributos que engloba somente os atributos monovalorados, sem a participação do atributo multivalorado. O valor da medida de relevância do atributo multivalorado, nesses dois contextos, foi coerente: quando a medida de relevância apresentou um valor alto, a inclusão do atributo multivalorado no conjunto de atributos utilizados na classificação (quando comparadas as situações (a) e (b)), melhorou os valores de acurácia do classificador. É importante ressaltar que o resultado desse experimento é muito sensível à qualidade dos atributos monovalorados que estão sendo combinados com o atributo multivalorado.

A3) Com as variações realizadas nas bases híbridas, aumentando e diminuindo as diferenças entre as probabilidades das classes, a medida de relevância tem seu valor alterado. Com essa avaliação, foi possível verificar que a acurácia do classificador acompanha o comportamento da medida de relevância, ou seja, para medidas mais altas foram obtidas acurácias mais altas e vice-versa.

A4) Foi verificado que a medida de relevância pode ser usada para comparar dois atributos multivalorados dentro de uma mesma base. Para um atributo com medida de relevância maior, a acurácia obtida na classificação com o uso desse atributo foi melhor. Vale lembrar que as acurácias foram obtidas com a utilização do atributo multivalorado isolado na classificação, sem a influência dos atributos monovalorados.

Apesar de não ter sido o foco do trabalho, foi possível analisar o comportamento de três medidas de distância de conjuntos, utilizadas em algoritmos do tipo *lazy*: RIBL, Tanimoto e *Average Linking*. A medida RIBL apresenta uma deficiência grave quando utilizada com atributos multivalorados, pois é muito sensível à diferença entre as quantidades dos dois conjuntos comparados, gerando distorções e até situações contraditórias. A medida *Average Linking* teve um bom desempenho, apesar de também apresentar problemas, já que algumas vezes considera diferente de zero a distância entre dois conjuntos idênticos. A medida Tanimoto parece ter as características mais adequadas à comparação de atributos multivalorados. Porém, não podemos afirmar categoricamente qual a melhor de distância para esse tipo de atributo. Cada uma possui características distintas que vão se adequar a diferentes tipos de contextos e objetivos.

Trabalhos Futuros

A medida de relevância de atributos multivalorados proposta neste trabalho leva em consideração a qualidade do atributo de forma isolada. Um estudo mais elaborado poderia ser realizado com o objetivo de construir uma medida que levasse em consideração a influência entre os atributos da base.

Uma outra sugestão de trabalho futuro baseado nesta pesquisa é a construção de uma medida de relevância de atributos multivalorados que possa ser aplicada a bases com mais de duas classes. A medida proposta neste trabalho leva em consideração a diferença entre as probabilidades das duas classes. Uma das formas de aplicar este conceito para várias bases seria através da utilização de um valor de desvio padrão entre as médias de probabilidades de cada classe. Um atributo poderia ser considerado bom quando o valor da medida de relevância fosse maior que o desvio padrão para uma das classes e menor que o desvio padrão para as demais.

APÊNDICE A – Detalhamento das Bases de Dados

Neste apêndice são descritas as tabelas utilizadas nos experimentos, de forma detalhada. Para cada tabela, são listados os seus atributos e respectivos domínios.

A Tabela A.1 apresenta a descrição dos atributos da base KDD Cup 2000. Nos gráficos das Figuras 4.12, 4.13 e 4.14, do Capítulo 4, o grupo de atributos selecionados aleatoriamente, para montar o conjunto de atributos considerado "ruim", é composto pelos seguinte atributos: {WhichDoYouWearMostFrequent, HowDoYouDressForWork, HowManyPairsDoYouPurchase, HowOftenDoYouPurchase, Age, MaritalStatus, OwnOrRentHome, NumberOfVehicles, Gender, PresenceOfChildren}.

O melhor conjunto de atributos, selecionados por um algoritmo de seleção da ferramenta Weka, é composto pelos seguintes atributos: {WhichDoYouWearMostFrequent, HowDoYouDressForWork, HowDidYouHearAboutUs, SendEmail, USState, PresenceOfChildren, EstimatedIncomeCode, HomeMarketValue, AvailableHomeEquity, Occupation}.

Atributo	**Domínio**
WhichDoYouWearMostFrequent	lista com 5 itens de produtos
HowDoYouDressForWork	lista com 5 tipos de estilos de roupas
HowManyPairsDoYouPurchase	lista com 5 faixas de quantidades
YourFavoriteLegwearBrand	lista com 16 itens de produtos
HowDidYouHearAboutUs	lista com 6 itens de fontes de informação
SendEmail	{NULL, True, False}
HowOftenDoYouPurchase	lista com 5 faixas de frequência
USState	lista com 50 siglas de estados americanos
Email	lista com 8 domínios de emails
AccountStatus	{Locked-Out, Active}
TruckOwner	{True, False}
RVOwner	{True, False}

MotorcycleOwner	{True, False}
ValueOfAllVehicles	numérico
Age	numérico
OtherIndivAge	numérico
MaritalStatus	lista com 4 tipos de estado civil
WorkingWoman	{True, False}
MailResponder	{True, False}
BankCardHolder	{True, False}
GasCardHolder	{True, False}
UnknownCardType	{True, False}
TECardHolder	{True, False}
PremiumCardHolder	{True, False}
NumberOfAdults	numérico
EstimatedIncomeCode	lista com 9 faixas salariais
HomeMarketValue	lista com 20 faixas de despesas
NewCarBuyer	{True, NULL}
VehicleLifestyle	lista com 8 estilos de veículos
PresenceOfPool	{True, False}
OwnOrRentHome	{Owner, Renter}
LengthOfResidence	numérico
MailOrderBuyer	{True, False}
YearHomeWasBought	{50s, 60s, 70s, 80s, 90s, ?}
NumberOfVehicles	numérico
CRAIncomeClassification	{1,2,3,4,?}
NumberOfCreditLines	{1,2,3,4,5,6,7,8,9,?}
SpecialityStoreRetail	{True, False}
OilRetailActivity	{True, False}
BankRetailActivity	{True, False}
FinanceRetailActivity	{True, False}
MiscellaneousRetailActivity	{True, False}
UpscaleRetail	{True, False}
UpscaleSpecialityRetail	{True, False}
RetailActivity	{True, False}
DwellingUnitSize	{SINGLE-FAMILY, MULTI-FAMILY}
AvailableHomeEquity	lista com 16 itens
MinorityCensusTract	{True, False}

Gender	{Male, Female}
Occupation	lista com 18 tipos de ocupação
OtherIndivGender	{Male, Female, NULL}
OtherIndivOccupation	lista com 15 itens de ocupação
PresenceOfChildren	{True, False}
UpscaleCardHolder	{True, False}

Tabela A.1: Base KDD - Atributos da tabela *Cliente*

A Tabela A.2 apresenta a descrição dos atributos da base EBooks. Nos gráficos das Figuras 4.12, 4.13 e 4.14, do Capítulo 4, o pior conjunto de atributos para a classe *Sex* é: {kids, mail}, e para a classe *Kids* é: {auto, sex}.

O melhor conjunto de atributos para a classe *Sex* é: {nation, auto}, e para a classe *Kids* é: {mail, nation}.

Quando a classe considerada é *Sex*, o atributo sex não é utilizado. E quando a classe é *Kids*, o atributo kids não é utilizado.

Atributo	**Domínio**
nation	{AD,AI,AL,AT,AU,AW,CA,CN,DE,DZ,GB,GT,ID,KR,TW,US,UZ}
auto	{0,1}
mail	{0,1}
sex	{1,2}
kids	{0,1}

Tabela A.2: Base EBooks - Atributos da tabela *Customer*

A Tabela A.3 apresenta a descrição dos atributos da base IBGE POF 2002. Nos gráficos das Figuras 4.12, 4.13 e 4.14, do Capítulo 4, o pior conjunto de atributos para a classe *Cartão de crédito* é: {sexo, idade, cod-posicao-ocup}, e para a classe *Sexo* é: {idade, cartao, vl-despesa}.

O melhor conjunto de atributos para a classe *Cartão de crédito* é: {cod-ocupacao, vl-despesa, nivel-instruc}, e para a classe *Sexo* é: {cod-ocupacao, cod-posicao-ocup, nivel-instruc}.

Quando a classe considerada é *Cartão de crédito*, o atributo cartão não é utilizado. E quando a classe é *Sexo*, o atributo sexo não é utilizado.

Atributo	Domínio
sexo	{1, 2}
cartao	{1, 2}
idade	numérico
nivel-instruc	{00,02,04,05,06,07,08,09,10,11,13,14,15,16,88}
titular-plano-saude	{1,2}
cod-posicao-ocup	{01,02,03,04,05,06,07}
cod-ocupacao	lista com 187 códigos de ocupação
vl-despesa	numérico

Tabela A.3: Base IBGE2002 - Atributos da tabela *Morador*

A Tabela A.4 apresenta a descrição dos atributos da base IBGE POF 1999. Quando a classe considerada é *Tamanho da Família*, o atributo num-tamfam não é utilizado. E quando a classe é *Faixa de Renda*, o atributo cod-renda não é utilizado.

Atributo	Domínio
cod-cidade	{01,02,03,04,05,06,07,08,09,10}
num-tamfam	numérico
cod-renda	{0,1,2,3,4,5,6,7,8,9,10}

Tabela A.4: Base IBGE1999 - Atributos da tabela *Família*

Referências

[1] FAYYAD U.; SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37–54, 1996.

[2] ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database System.* 5th. ed. USA: Addison-Wesley, 2006.

[3] LIU, H.; MOTODA, H. *Feature Selection for Knowledge Discovery and Data Mining.* USA: Kluwer Academic Publishers, 1998.

[4] LIU, H.; MOTODA, H. Less is more. *Computational Methods of Feature Selection.*, Chapman and Hall/CRC., p. 3–17, 2008.

[5] HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, IEEE Educational Activities Department, USA, v. 15, n. 3, p. 1437–1447, 2003.

[6] WOSNICA, A. Relational weka. *http://cui.unige.ch/woznica/rel_weka/*, 2007.

[7] WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques.* 2nd. ed. USA: Morgan Kaufmann, 2005.

[8] DZEROSKI, S. Multi-relational data mining: an introduction. *SIGKDD Explorations*, ACM, USA, v. 5, n. 1, p. 1–16, 2003.

[9] LAER, V. V.; RAEDT, L. How to upgrade propositional learners to first order logic: A case study. In: *Relational data mining.* Germany: Springer-Verlag, 2001. p. 235–261.

[10] KRAMER, S.; LAVRAC, N.; FLACH, P. Propositionalization approaches to relational data mining. In: *Relational data mining.* Germany: Springer-Verlag, 2001. p. 262–286.

[11] DEHASPE, L.; TOIVONEN, H. Discovery of relational association rules. In: *Relational data mining.* Germany: Springer-Verlag, 2001. p. 189–212.

[12] EMDE, W.; WETTSCHERECK, D. Relational instance based learning. In: *Proceedings of the Thirtheenth International Conference on Machine Learning.* USA: Morgan Kaufnann, 1996. p. 122–130.

[13] LEIVA, H. *MRDTL: A multi-relational decision tree learning algorithm.* Dissertação (Mestrado) — Iowa State University, USA, 2002.

[14] EMDE, W.; WETTSCHERECK, D. Multi-relational data mining using probabilistic relational models: research summary. In: *Proceedings of the First Workshop in Multi-relational Data Mining.* [S.l.: s.n.], 2001.

[15] PERLICH, C. Automated construction of relational attributes acora: A progress report. *Working Paper, New York University*, 2002.

[16] DZEROSKI, S.; LAVRAC, N. *Relational Data Mining*. 1st. ed. Germany: Springer-Verlag, 2001.

[17] LAVRAC, N. Introduction to inductive logic programming. *Computer Science course ComS70301: Learning from Structured Data*, 2002.

[18] KNOBBE, J. et al. Multi-relational data mining. In: *Proceedings of Benelearn 99*. Belgium, The Netherlands: [s.n.], 1999.

[19] KNOBBE, J.; SIEBES, A.; WALLEN, D. M. G. Van der. Multi-relational decision tree induction. In: *Proceedings of the 3rdEuropean Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD*. London, UK: Springer-Verlag, 1999. p. 378–383.

[20] KERSTING, K.; RAEDT, L. D. Bayesian logic programs. In: *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*. German: Albert-Ludwigs University, 2000.

[21] AHA, D. W. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, v. 36, n. 2, p. 267–287, 1992.

[22] BURAGO, D.; BURAGO, Y.; IVANOV, S. *A Course in Metric Geometry*. USA: American Mathematical Society, 2001.

[23] KALOUSIS, A.; WOZNICA, A.; HILARIO, M. A unifying framework for relational distance-based learning. *Technical report*, University of Geneva, 2005.

[24] DUDA, R.; HART, P.; STORK, D. *Pattern Classification and Scene Analysis*. USA: John Willey and Sons, 2001.

[25] LEE, H. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. Tese (Doutorado) — USP - São Carlos, São Paulo, 2005.

[26] CARUANA, R.; FREITAG, D. How useful is relevance? *Working Notes of the AAAI Fall Synposium on Relevance*, p. 25–29, 1994.

[27] PERLICH, C.; PROVOST, F. Distribution-based aggregation for relational learning from identifier attributes. *Machine Learning 62 (1/2) 65-105*, Kluwer Academic Publishers, USA, v. 62, n. 1-2, p. 65–105, 2006.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo