Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciências da Computação

Classificadores para dados simbólicos do tipo intervalo baseados em modelos de regressão

Diego Cesar Florencio de Queiroz

Trabalho de Graduação

Recife
17 de novembro de 2009

Universidade Federal de Pernambuco
Centro de Informática

Diego Cesar Florencio de Queiroz

Classificadores para dados simbólicos do tipo intervalo baseados em modelos de regressão

Trabalho apresentado ao Programa de Graduação em Ciências da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciências da Computação.

Orientadora: *Renata Maria Cardoso Rodrigues de Souza*

Recife
17 de novembro de 2009

Este trabalho é dedicado a Rebeca por aguentar todas as minhas esquisitices e me amar mesmo assim, a minha mãe, por ter me dado a luz e me criado e a professora Renata, por ter me apresentado ao mundo científico e estimulado meu espírito investigativo.

Agradecimentos

Gostaria de agradecer primeiramente a Rebeca, novamente, por ter me apoiado e confiado em minha capacidade e se mantido ao meu lado embora muitas vezes estivéssemos distantes, a minha mãe e minhas irmãs, principalmente a minha irmã Rhayanna que com suas habilidades incríveis de designer me ajudou a fazer os dois pôsteres, à professora Renata por ter sido muito paciente e ter considerado minhas opiniões em nossas discussões, aos meus amigos e colegas de iniciação científica, Diogo e Telmo, pelas várias reuniões em que discutimos os trabalhos uns dos outros. Aos membros do grupo de pesquisa em Análise de Dados Simbólicos do CIn-UFPE por terem sempre me instigado, assim como a professora, a ir além, mais especialmente para Marco Antônio e Carlos Wilson, pois sem a ajuda deles esse trabalho não teria sido formatado de maneira adequada, nem teria sido entregue a tempo.

Gostaria também de agradecer a todos os meus outros amigos, especialmente a turma do RPG (Bruno, Christian, Otacílio, João, Hercílio, Raony, Letícia, Marcelo, Fábio, Denise, Victor, Caius e Kirlian), por me lembrar de sair de casa e me divertir, especialmente quando estava empacado em alguma parte do meu trabalho. E por último agradeço aos meus grandes amigos Diego Távora e Beto, por terem crescido comigo e terem ouvido muitas divagações absurdas de minha parte.

Em suma, muito obrigado a todos que um dia deixaram uma marca na minha vida, pois foram responsáveis por me ajudar a me tornar o que sou.

O homem que moveu a montanha começou carregando pequenas pedras.
—CONFÚCIO

Resumo

Esse trabalho introduz diferentes classificadores de padrões para dados intervalares baseados em modelos de regressão linear e logística. Seis abordagens abordagens foram idealizadas. Essas abordagens diferem na maneira de representar os intervalos e no modelo de regressão utilizada. O primeiro assume que cada intervalo é um par de variáveis quantitativas e ajusta modelos de regressão linear clássica sobre essas variáveis. O segundo assume que cada intervalo é um par de variáveis quantitativas, mas ajusta dois modelos de regressão linear clássica separados sobre essas variáveis e combina os resultados de uma maneira apropriada. O terceiro classificador considera que cada intervalo é representado pelos seus centros e ajusta modelos de regressão logística clássica sobre os centros dos intervalos. O quarto assume que cada intervalo é um par de variáveis quantitativas e ajusta modelos de regressão logística clássica sobre essas variáveis. O quinto considera que cada intervalo é representado pelos seus vértices e modelos de regressão logística clássica são ajustados sobre cada vértice dos intervalos. O último também assume que cada intervalo é um par de variáveis quantitativas, mas ajusta dois modelos de regressão logística clássica separados nessas variáveis e combina os resultados de uma maneira apropriada. Experimentos com conjuntos de dados sintéticos e uma aplicação com um conjunto de dados intervalares real demonstram a funcionalidade e eficiência desses classificadores.

Palavras-chave: Dados Simbólicos Intervalares, Classificação, Regressão Logística, Análise de Dados Simbólicos

Abstract

This work introduces different pattern classifiers for interval data based on the logistic regression methodology. Six approaches are considered. These approaches differ in the way of representing the intervals and in the model of regression used. The first classifier considers that each interval is represented by a pair of quantitative variables and performs a classic linear regression on these variables. The second considers that each interval is a pair of quantitative variables, but performs two separate classic linear regressions on these variables and cambine the results in a appropriate way. The second classifier considers that each interval is represented by the centers of the intervals and performs a classic logistic regression on the centers of the intervals. The third one assumes each interval as a pair of quantitative variables and performs a conjoint classic logistic regression on these variables. The fourth one considers that each interval is represented by its vertices and a classic logistic regression on the vertices of the intervals is applied. The last one also assumes each interval as a pair of quantitative variables, but performs two separate classic logistic regressions on these variables and combine the results in some appropriate way. Experiments with synthetic data sets and an application with a real interval data set demonstrate the usefulness of these classifiers.

Keywords: Interval Symbolic Data, Classification, Logistic Regression, Symbolic Data Analysis;

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Todo novo começo vem do fim de algum outro começo
—SÊNECA (filósofo romano)

A cada dia mais operações ou processos são automatizados, ou seja, para cada nova transação como compras pela internet, operações bancárias, ligações através de aparelhos de telefonia móvel, entre outras, novos registros em enormes bancos de dados são armazenados. Para armazenar tal volume de informação, sistemas gerenciadores de banco de dados estão presentes em praticamente todas as instituições públicas e privadas, de pequeno, médio e grande porte, contendo os mais diferentes dados sobre produtos, clientes, fornecedores, funcionários, etc. Além disso, os avanços na área de aquisição de dados tornam mais fácil a coleta de informações, desde um leitor de RFID (*Radio-frequency Identifiers*) até sistemas de monitoramento remotos que geram grandes volumes de informação.

Contudo, num ambiente em constante mudança tornam-se necessárias novas técnicas e ferramentas de extração e análise de conhecimento que agilizem o processo de decisão empresarial. A construção de Data Warehouses é considerada um dos primeiros passos para simplificar a análise de dados no apoio à tomada de decisão. Porém, na prática, a análise de dados contida na estrutura de data warehouses geralmente não extrapolam da realização de consultas e diante desse fato, diversos estudos têm sido direcionados ao desenvolvimento de tecnologias de extração de conhecimento.

A descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) é uma área de pesquisa em bastante evidência no momento, que visa desenvolver meios automáticos de aquisição de conhecimento em bases de dados. As técnicas utilizadas nesse processo são genéricas e derivam de diferentes áres de conhecimento, tais como: estatística, inteligência artificial e banco de dados. As técnicas estatísticas englobam algoritmos que podem ser aplicados para descobrir estruturas ou associações em um conjunto de dados, realizar previsões, etc. Dentre estas técnicas destacam-se os modelos de regressão que têm como objetivo descrever o comportamento de uma variável dependente (chamada variável de resposta) a partir de informações provenientes de um conjunto de variáveis independentes (chamadas variáveis explicativas). Além disso, através dos modelos de regressão é possível identificar o grau de associação entre as variáveis ou mensurar o impacto que uma variável exerce sobre outra.

Embora as técnicas tradicionais da estatística sejam bastante aplicadas para sumarizar e analisar conjuntos de dados, com o grande crescimento da área de tecnologia da informação - TI, estas técnicas têm se tornado inadequadas para tratar conjuntos de dados representados por informações mais complexas como *intervalos* ou células multi-valoradas. Além disso,

tais técnicas não possuem estruturas adequadas que permitam sintetizar grandes conjuntos de dados, com a menor perda possível de informação dos dados originais. Como uma alternativa para generalizar as atuais técnicas estatísticas para esses tipos de dado mais complexos, surgiu a análise de dados simbólicos (*Symbolic Data Analysis - SDA*).

A análise de dados simbólicos [1] e [2] é uma área, que nasceu da influência simultânea de vários campos de pesquisa como: análise de dados clássica, inteligência artificial, aprendizagem de máquina e banco de dados. O principal objetivo de SDA é desenvolver modelos para o tratamento de dados mais complexos, como intervalos, conjuntos e distribuições de probabilidades ou de pesos.

CAPÍTULO 2

Aprendizagem Supervisionada e Análise de dados Simbólicos

Quando o aluno está pronto, o mestre aparece.
—PROVÉRBIO CHINÊS

2.1 Aprendizagem Supervisionada

Aprendizagem de máquina é uma área da inteligência artificial que tem como objetivo o desenvolvimento de algoritmos que permitam computadores transformar seu comportamento baseados em dados. Um dos maiores focos da aprendizagem de máquina é automaticamente aprender a reconhecer padrões complexos e fazer decisões ou auxiliar o processo decisório sobre dados.

A Aprendizagem de Máquina se divide em Apredizagem não-supervisionada e Aprendizagem supervisionada, porém o foco deste trabalho é a aprendizagem supervisionada.

A Aprendizagem supervisionada é um paradigma de aprendizagem de máquina, que tem como objetivo extrair conhecimento de um conjunto de exemplos. Esse paradigma está diretamente relacionado com a resolução de problemas de classificação e regressão.

Ambos os problemas são similares, pois se utilizam de um conjunto de características que descrevem cada elemento do conjunto de exemplos para construir um modelo que seja capaz de obter a que classe pertence um novo elemento, ou, no caso de regressão, obter um valor numérico baseado nas características desse novo elemento.

Neste trabalho será abordado um problema de classificação, cujo objetivo é definir um modelo decisório que seja capaz de discriminar um novo exemplo. Em um problema de classificação um conjunto de exemplos contém informações sobre características e uma descrição rotulada da classe para cada elemento. Tal rótulo é gerado através do conhecimento de um especialista ou como resultado de um algoritmo de aprendizagem não-supervisionada.

Como visto na figura 2.1 , Algoritmos de sistemas classificadores, também chamados sistemas discriminantes, são caracterizados por duas etapas:

Etapa de aprendizado

A etapa de aprendizado corresponde à etapa em que um modelo decisório aprende, através de um conjunto de exemplos, como classificar novos exemplos à uma das classes pré-existentes no conjunto de exemplos. A construção desse modelo depende das restrições impostas pela metodologia utilizada.

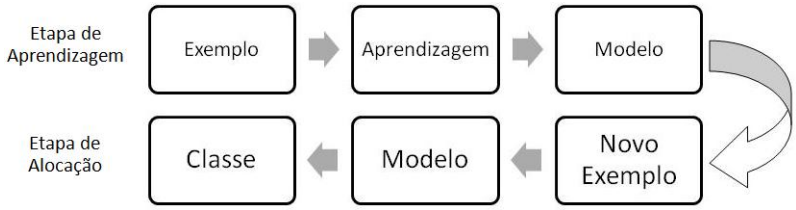

Figura 2.1 Etapas de um sistema de classificação

Etapa de classificação ou alocação

Após a realização da etapa de aprendizado, uma regra é adotada para classificar os novos exemplos, tais regras podem ser baseadas em medidas de proximidade (similaridade, dissimilaridade) ou outros critérios.

Diferentes abordagens de aprendizagem têm sido desenvolvidas e utilizadas na área de classificação supervisionada. Algumas abordagens serão discutidas nesse capítulo: abordagem estatística, redes neurais e aprendizagem simbólica. Como este trabalho utiliza a metodologia de regressão, será dado um maior enfoque à abordagem estatística.

A abordagem estatística consiste em encontrar parâmetros de um modelo de predição para uma amostra de exemplos. A aprendizagem neural está associada à aprendizagem estatística no sentido em que tem como objetivo encontrar pesos dentro de uma estrutura de rede, já a aprendizagem simbólica visa estabelecer regras, através da definição de conceitos.

2.1.1 Abordagem Estatística

Os procedimentos estatísticos de classificação supervisionada são técnicas multivariadas, que tem como objetivo discriminar e alocar objetos. A classificação é um processo empregado para investigação visual ou algébrica das diferenças entre classes de objetos baseando-se em características. A alocação de objetos consiste em obter uma regra para associar um novo objeto a uma das classes pré-definidas.

A aprendizagem estatística é realizada através de duas metodologias distintas: paramétrica (quando há necessidade de fazer suposições sobre a distribuição dos dados) e não-paramétrica. A Regra de Bayes é uma técnica paramétrica bastante utilizada em aplicações estatísticas. Essa regra estima a probabilidade *a posteriori* para classificar um objeto. O Discriminante Logístico é uma técnica de classificação paramétrica obtida através de modelos de regressão linear de forma específica. Os métodos de Kernel K-vizinhos mais próximos são técnicas não lineares e não paramétricas que se utilizam de estimativas de densidade local para classificar os objetos sem fazer suposições sobre as distribuições dos dados. Os métodos Discriminante Linear e Discriminante Quadrático são técnicas não paramétricas que empregam estimativas das matrizes de covariância para encontrar uma função empírica sobre os dados que apresente uma boa separação entre as classes. Esses métodos são robustos, pois não são definidos de forma estrita, ou seja, pequenos desvios das suposições básicas em suas definições não afetam o desempenho dos classificadores, e grandes diferenças devem causar uma ineficácia relativa do método associada à distribuição dos dados.

2.1.2 Abordagem de Redes Neurais

Redes Neurais Artificiais [3], também conhecida como abordagem conexionista de aprendizagem de máquina, são estruturas computacionais, que foram inicialmente inspiradas em propriedades do sistema neural biológico. A estrutura de uma rede neural é composta por três unidades fundamentais: neurônio, arquitetura e algoritmo de aprendizagem.

O neurônio é unidade computacional básica que forma a rede, a arquitetura define como essas unidades estão conectadas entre si e o algoritmo de aprendizagem descreve o método utilizado para mudar o comportamento individual dos neurônios, almejando um comportamento coletivo (comportamento da rede emerge do comportamento de cada unidade). Desse modo, os principais modelos de redes neurais são descritos através de diferentes arquiteturas a técnicas de aprendizagem.

O modelo inicial de neurônio foi proposto por McCulloch e Pitts em 1943 [4] e, desde então inúmeros outros modelos foram surgiram a partir desse trabalho, como Perceptron, Multi-layer Perceptron, redes RBF, etc., além de vários estudos sobre técnicas de aprendizagem, sendo que o algoritmo de retropropagação (*backpropagation*) é um dos mais utilizados.

2.1.3 Abordagem Simbólica

Os trabalhos em aprendizagem de simbólica iniciaram-se há cinco décadas com o objetivo de desenvolver métodos computacionais, que produziriam mecanismos capazes de adquirir conhecimento através de métodos de indução sobre exemplos ou dados. A aprendizagem simbólica é uma abordagem de aprendizagem de máquina que tem sido largamente empregada em inteligência artificial. A construção de métodos de aprendizagem simbólica consiste de uma seqüência de procedimentos baseados em operações lógicas ou binárias que determinam a descrição de um determinado conceito, a partir de um conjunto de exemplos desse conceito provido de um supervisor ou professor e de um conhecimento prévio. A descrição é discriminante, sendo muito importante para a classificação de dados.

Os conjuntos de conceitos podem ser organizados através de uma generalização hierárquica representada por um grafo, como é o caso em uma das técnicas mais utilizadas na abordagem simbólica: Árvores de Decisão.

Árvores de Decisão é uma técnica de abordagem simbólica que tem como objetivo particionar o conjunto de exemplos em subconjuntos, cujos elementos satisfazem às restrições geradas a partir dos atributos inerentes aos mesmos. Uma árvore é um grafo que é constituído de nós e ramos onde cada nó terminal representa uma decisão. Um dos algoritmos mais utilizados na abordagem simbólica é o algoritmo TDIDT (Top-Down Induction Decision Trees), que é utilizado para gerar indutivamente a árvore e realizar o procedimento de *pruning*.

2.2 Análise de Dados Simbólicos

Dados Simbólicos podem surgir de diferentes formas. Por exemplo, considere uma variável $X = \{Cor\}$ e a população de interesse como $\Omega = \{Espécies\ de\ frutas\ produzidas\ num\ pomar\}$. Uma espécie de fruta $k \in \Omega$ pode assumir os seguintes valores: $X(k) = \{Verde, Vermelha\}$ e

uma outra espécie w pode assumir $X(w) = \{Verde, Amarela\}$. Em outro exemplo, após várias medições de controle, um paciente saudável pode ter o valor de sua pressão oscilando no intervalo $[115, 118]$. Um outro paciente, também saudável, poderia ter sua pressão oscilando no intervalo $[114, 116]$. Uma análise clássica utilizando o ponto médio dos intervalos perderia a informação sobre a variação de pressão no estado saudável para cada paciente.

Em outra situação, uma companhia de seguros de saúde possui um banco de dados com centenas (ou milhares) de informações a respeito das consultas de seus segurados, onde cada entrada desse banco armazena: o tipo de especialista consultado, o local do exame, os exames realizados, os medicamentos solicitados, etc. Entretanto, a seguradora pode não estar interessada em uma consulta em especial, mas em todas as consultas realizadas por um dado cliente. Neste caso, todas as consultas realizadas pelo cliente podem ser agregadas, produzindo dados simbólicos. Assim, seria extremamente atípico que o peso (kg) desse determinado cliente, em todas as suas consultas, fosse igual a $70kg$. No entanto, poderíamos observar que seu peso oscilou no intervalo $[68kg, 73kg]$. Em um outro cenário, poderíamos supor que um banco não estaria interessado no valor monetário na conta corrente de um certo indivíduo, mas na variação desse valor ao longo de um ano.

Uma última situação nos leva a refletir que nos dias atuais torna-se cada vez maior a necessidade de analisar grandes bases de dados, mas apesar do grande poder de processamento dos computadores atuais, o esforço computacional necessário para a manipulação dessas grandes massas de dados ainda é um problema. A agregação da informação contida nestas bases de dados em bases menores seria a forma ideal para analisá-los. Além disso os métodos tradicionais de análise de dados foram desenvolvidos numa época onde a quantidade de informação gerada pela sociedade era infinitamente menor que a disponível atualmente.

A Análise de Dados Simbólicos[5] constitui uma extensão de alguns métodos utilizados para a análise de dados clássicos. Bock e Diday [1] apresentam de maneira sólida os principais conceitos de SDA e os principais métodos estatísticos desenvolvidos para a manipulação de dados dessa natureza.

Além dos exemplos citados anteriormente, os dados simbólicos também podem ser obtidos a partir da aplicação de um algoritmo de agrupamento para simplificar grandes conjuntos de dados e descrever, de uma maneira auto-explicativa as classes associadas aos grupos obtidos, também podem ser obtidos como o resultado da descrição de conceitos por especialistas e, por fim, a partir de bancos de dados relacionais para estudar conjuntos de unidades cuja descrição necessita a fusão eventual de várias relações.

2.2.1 Tabelas de Dados Simbólicos

Como já foi descrito através de exemplos, os dados simbólicos podem descrever indivíduos levando em conta, ou não, imprecisão ou incerteza, ou podem descrever itens mais complexos, tais como grupos de indivíduos. Tais dados são apresentados em tabelas de dados simbólicos.

Em tabelas de dados simbólicos, as linhas correspondem aos indivíduos ou grupos de indivíduos e as colunas são as variáveis simbólicas que os descrevem. Um exemplo de tabela de dados simbólicos é apresentado abaixo. Neste exemplo, as linhas são grupos de indivíduos e nas colunas temos três variáveis simbólicas: peso (representado por um intervalo), marca de automóvel (representado por um conjunto de categorias) e seguro (representado por uma

distribuição de pesos).

ID	Peso	Marca de Automóvel	Seguro
1	[68.8, 70.1]	{Ford, Honda}	{(3/4) Sim, (1/4) Não}
2	[60.3, 84.5]	{GM, Fiat, KIA}	{(1/6) Sim, (5/6) Não}
3	[49.4, 55.2]	{Ford, GM}	{(4/5) Sim, (1/5) Não}

Tabela 2.1 Exemplo de uma tabela de dados simbólicos.

2.2.2 Variáveis Simbólicas

Na análise de dados clássicos, as variáveis assumem um único valor ou categoria para um dado indivíduo, entretanto as variáveis simbólicas podem assumir para um dado indivíduo ou grupo de indivíduos: conjuntos de categorias, intervalos, histogramas, etc. As variáveis simbólicas dividem-se basicamente em: variáveis multi-valoradas (ordenadas ou não-ordenadas), variáveis de tipo intervalar e variáveis modais.

Variáveis multi-valoradas não-ordinais

Y é definido como uma variável simbólica multi-valorada se seus valores $Y(k)$ correspondem a subconjuntos finitos do domínio $D : |Y(k)| < \infty$ para todos os indivíduos $k \in E$. Por exemplo, seja Y os bancos existentes em k cidades brasileiras, onde $D = \{Banco\ doBrasil, Bradesco, BancoReal, \ldots, CaixaEconômica\}$. Logo, para o indivíduo $k = Paulista$, $Y(Paulista) = \{Banco\ do\ Brasil, Banco\ Real\}$, já para o indivíduo $k = Barreiros$, $Y(Barreiros) = \{Banco\ do\ Brasil, Caixa\ Econômica\}$.

Variáveis multi-valoradas ordinais

Uma variável simbólica Y é definida como multi-valorada ordinal se D suporta uma relação de ordem $\prec$, tal que, para quaisquer pares de elementos, $a, b \in D$, existe a relação $a \prec b$ ou a relação $b \prec a$. Um caso comum é representado por uma variável qualitativa com domínio finito $D = \{a, b, c, \ldots, h\}$, onde: $a \prec b \prec c \prec \ldots \prec h$. Na prática, $a \prec b$ é interpretado como *a antecede b* ou *a é menor que b*. Assim, para dois indivíduos $k, t \in E$, onde observamos $a = Y(k)$ e $b = Y(t)$ para a variável Y, é possível indicar qual dos indivíduos é estritamente *melhor* que o outro: $a \prec b$ ou $b \prec a$. Por exemplo, $Y = qualidadedoproduto$ e $D = \{excelente, bom, razovel, pobre, insuficiente\}$

Variáveis intervalares

Define-se Y como uma variável simbólica intervalar se $\forall k \in E$, o subconjunto $U := Y(k)$ é um intervalo em $\Re$ ou um intervalo relacionado a uma dada ordem $\prec$ em $D : Y(k) = [\alpha, \beta]$, tal que, $\alpha, \beta \in D, \alpha \leq \beta$ e $\alpha \preceq \beta$. Por exemplo, seja Y o tempo de estudo semanal de um estudante (em horas), os possíveis valores para indivíduos poderíam ser $Y(c) = [0, 2]$ ou $Y(v) = [5, 6]$.

Variáveis modais

A variável simbólica Y também pode ser definida como modal, com domínio D sob o conjunto de objetos simbólicos $E = a, b, \ldots$, é uma variável multi-valorada que, para cada objeto $a \in E$, apresenta não apenas um conjunto de categorias $Y(a) \subseteq D$, mas também uma freqüência, probabilidade ou peso $w(y)$, associado a cada categoria $y \in Y(a)$, que indica o quão freqüente, típico ou relevante a categoria y é para o objeto a. Por exemplo, seja Y a distribuição das agências bancárias em k cidades. Então, para uma cidade t, $Y(t) = \{Banco\ do\ Brasil(0.5), Bradesco\ (0.4), Banco\ \ Real(0,1)\}$.

2.2.3 Técnicas de aprendizagem supervisionada estendidas para dados simbólicos

Várias ferramentas de aprendizagem supervisionada foram estendidas para lidar com dados simbólicos: Ichino, Yaguchi e Diday [6] introduziram um classificador simbólico utilizando uma abordagem baseada em regiões para dados multi-valorados. Nessa abordagem os exemplos de classes são descritos por uma região ou conjunto de regiões obtida através da utilização de uma aproximação de um grafo de vizinhança mútua (*Mutual Neighbourhood Graph - MNG*) e um operador de junção simbólico. Souza [7] propôs uma aproximação de MNG para reduzir a complexidade da etapa de aprendizado sem reduzir o desempenho do classificador em termos de precisão da predição. D'Oliveira, De Carvalho e Souza [8] apresentaram uma abordagem orientada a regiões em que cada região é definida pela casca convexa dos objetos pertencentes a uma classe. Ciampi, Diday, Lebbe, Perinel e Vignes [9] introduziram uma generalização das árvores de decisão binárias para predizer o conjunto dos membros de classe de dados simbólicos. Rossi e Conan-Guez [10] generalizaram as redes neurais MLP para trabalhar com dados intervalares. Mali e Mitra [11] estenderam a rede baseada na função difusa de base radial (*Fuzzy Radial Basis Function - RBF*) para trabalhar no domínio dos dados simbólicos. Appice, D'Amato, Esposito e Malerba [12] introduziram uma abordagem de aprendizado preguiçoso (chamada *Symbolic Objects Nearest Neighbor SO-SNN*), que estende o algoritmo tradicional de classificação k-vizinhos mais próximos ponderados para dados intervalares e modais. Silva e Brito [13] propuseram três abordagens para a análise multivariada de dados intervalares, tendo como foco a análise do discriminante linear.

CAPÍTULO 3

Classificadores Propostos

Glória verdadeira consiste em se fazer aquilo que merece ser escrito; em escrever o que merece ser lido.

—PLÍNIO, O JOVEM (uma carta a Tácito)

Modelos de regressão têm se tornado componentes integrais de qualquer análise de dados cujo objetivo seja descrever o relacionamento entre uma variável resposta e uma ou mais variáveis explicativas. Neste capítulo, será feita uma introdução aos modelos de regressão linear clássica e regressão logística clássica, apresentados trabalhos relacionados e então serão apresentados os 6 classificadores propostos, sendo que 2 deles utilizam modelos de regressão linear e os outros 4 utilizam a metodologia de regressão logística.

3.1 Regressão Linear Clássica

A regressão linear clássica tem como objetivo modelar o relacionamento de uma ou mais variáveis preditoras X e uma variável resposta, contínua, Y, levando em consideração uma variável aleatória ε, que representa o erro presente em medições ou de variáveis não inclusas explicitamente no modelo. Segue abaixo o modelo de regressão linear clássico:

Seja $\Omega = \{(\boldsymbol{x}(i), y(i))(i = 1, \ldots, N)\}$ um conjunto de dados clássicos. Cada padrão i de Ω é descrito por um conjunto de p variáveis contínuas $(\boldsymbol{x}(i) = \{X_{i1}, \ldots, X_{ip}\}$, chamadas variáveis preditoras, uma variável contínua Y, que representa as variáveis resposta, tal que $Y \in \Re$ e $\boldsymbol{\beta} = \beta_0, \ldots, \beta_p$ um vetor de parâmetros desconhecidos.

Então, o modelo de regressão linear clássico é definido por:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \tag{0}$$

Esta expressão permite que E(Y|x) tome valores no intervalo $-\infty$ e $+\infty$. Para ajustar o modelo, é necessário estimar o valor dos coeficientes de regressão $\boldsymbol{\beta}$. Tais parâmetros são estimados tradicionalmente pelo Método dos Mínimos Quadrados. Esse método seleciona parâmetros estimados $\hat{\boldsymbol{\beta}}$ cuja soma das diferenças quadradas com as respostas observadas nos dados é tão pequena quanto possível.

Porém algumas suposições são necessárias: a matriz $\boldsymbol{x}$ que reúne todos os vetores que descrevem cada elemento i do conjunto de dados deve possuir posto completo e a variável que representa o erro (ε) deve seguir uma distribuição normal com variância constante, o que implica que a distribuição condicional da variável Y será, também, normal, com média $E(Y|x)$ e variância constante.

Seja $\boldsymbol{y}$ um vetor que possui todas as variáveis respostas para os i elementos do conjunto de dados, a equação para estimar os parâmetros através do método dos mínimos quadrados é:

$$\hat{\beta} = (x^T x)^{-1} x^T y \tag{1}$$

3.2 Regressão Logística Clássica

O modelo de regressão logística [14] tem sido utilizado na estatística há muitos anos e recentemente tornou-se objeto de estudo da comunidade de aprendizagem de máquina. Essa ferramenta tradicional da estatística surgiu da necessidade de modelar as probabilidades posteriores dos níveis de classe dada sua observação através de funções lineares nas variáveis preditoras.

Seja $\Omega = \{(\boldsymbol{x}(i), y(i))(i = 1, \ldots, N)\}$ um conjunto de dados clássicos. Cada padrão i de Ω é descrito por um conjunto de p variáveis contínuas $(\boldsymbol{x}(i) = \{X_{i1}, \ldots, X_{ip}\}$, chamadas variáveis preditoras, uma variável dicotômica Y, que representa as variáveis resposta, tal que $Y = 0$ ou $Y = 1$ e $\boldsymbol{\beta} = \beta_0, \ldots, \beta_p$ um vetor de parâmetros desconhecidos.

A forma específica do modelo de regressão logística é:

$$E(Y|\boldsymbol{x}) = \pi_{\boldsymbol{x}} = \frac{exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)} \tag{2}$$

Porém alguns aspectos da regressão linear são desejados (ser uma função linear sobre os parâmetros e pode possuir valores contínuos dependendo dos valores da matriz $\boldsymbol{x}$). Tais aspectos são obtidos através da transformação logit, que é dada a seguir:

$$g(x) = ln[\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \tag{3}$$

No caso logístico o erro ε possui uma distribuição com média 0 e variância igual a $\pi(\boldsymbol{x})[1 - \pi(\boldsymbol{x})]$, o que implica na distribuição condicional da variável Y seguir uma distribuição binomial com probabilidade dada pela média condicional $\pi(\boldsymbol{x})$. Além disso, a regressão logística utiliza outra metodologia para estimar os coeficientes de regressão, essa metodologia é o método da máxima verossimilhança [15].

3.3 Trabalhos Relacionados

Alguns trabalhos relacionados que culminaram no desenvolvimento deste trabalho de graduação foram publicados com relação à metodologia utilizada para representação intervalar e a utilização de modelos de regressão para dados simbólicos intervalares, no artigo [16] os autores propuseram um classificador binário baseado na metodologia probit para dados simbólicos intervalares. No artigo [17], os autores utilizam uma abordagem logística e a comparam com a abordagem baseada no modelo probit para classificadores binários, já no artigo [18], dois dos modelos apresentados neste trabalho foram desenvolvidos. Tais artigos foram anexados no apêndice A deste trabalho.

3.4 Introdução aos classificadores propostos

Na etapa de aprendizado, modelos de regressão linear e logística serão ajustados para modelar a probabilidade a posteriori das classes de um conjunto de dados simbólicos de treinamento, e na etapa de alocação ou classificação, novos elementos são classificados como pertencentes à determinada classe de acordo com o argumento calculado ou a probabilidade a posteriori estimada.

Suponha que existam K classes de padrões, rotuladas $1,\ldots,K$. Seja $\Omega = \{(\boldsymbol{x}(i), y(i))(i = 1,\ldots,N)\}$ um conjunto de dados simbólicos de treinamento. Cada padrão i de Ω é descrito por um conjunto de variáveis simbólicas $\{X_1, \ldots, X_p\}$ e uma variável categórica Y que toma um valor pertencente ao conjunto $C = \{1,\ldots,K\}$ das K classes. Uma variável simbólica [19] X é uma correspondência definida de $\Im$ em $\Re$ de modo que para cada i de Ω, $X(i) = [a,b] \subseteq \Im$ onde $\Im = \{[a,b] : a,b \in \Re, a \leq b\}$ é um intervalo.

Aqui, os padrões de treinamento simbólicos $\{\boldsymbol{x}(i), y(i)\}$ têm $\boldsymbol{x}(i) = ([a_{i1},b_{i1}],\ldots,[a_{ip},b_{ip}])$ como um vetor de variáveis intervalares covariantes e $y(i)$ como uma resposta multivariada. Considerando a resposta multivariada, cada uma das categorias é codificada através de uma variável indicadora [20].

Então se o conjunto C possui K classes, existirão K variáveis indicadoras Y_k, $k = 1,\ldots,K$, sendo que $Y_k = 1$ se $Y = k$, caso contrário $Y_k = 0$. Cada variável indicadora Y_k segue uma distribuição de Bernoulli com parâmetro π_k.

K funções lineares ajustadas $f_k(\boldsymbol{x})$ são construídas com o objetivo de modelar os argumentos para as K classes, e K probabilidades posteriores são estimadas para cada classe $\pi_k = Pr(Y = k/\boldsymbol{x}) = Pr(Y_k = 1/\boldsymbol{x})$.

A idéia é utilizar a abordagem um-contra-todos, utilizando a regressão linear clássica e a regressão logística clássica e introduzir análise de regressão linear e logística multiclasse para dados intervalares em que cada classe possui uma ou mais respostas binárias contra as outras classes.

3.5 Classificadores Propostos Baseados em Modelos de Regressão Linear

Nessa seção, serão apresentados os 2 classificadores de dados simbólicos intervalares baseados na metodologia de regressão linear.

3.5.1 Classificador linear baseado no argumento posterior dos limites conjuntamente

O classificador de padrões para dados intervalares linear baseado no argumento posterior simples definido para os limites inferior e superior dos intervalos conjuntamente (APSLC) será apresentado nessa seção. Esse classificador considera os limites inferior e superior como variáveis independentes, ajustando um modelo linear clássico para cada classe.

Uma variável intervalar X é representada por um par de variáveis quantitativas (x_L, x_U) onde $x_{Lj} = a_j$ e $x_{Uj} = b_j$. Assim, cada padrão de treinamento i possui um vetor de $2p$ va-

lores quantitativos $\boldsymbol{x} = \{x_{L1}, x_{U1}, \ldots, x_{Lp}, x_{Up}\}$ como um vetor de covariáveis. Seja $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \ldots, \beta_{k2p})$ um vetor de parâmetros desconhecidos de dimensão $(2p+1)$ para a classe $k, (k = 1, \ldots, K)$ e $\boldsymbol{z} = (1, x_{L1}, x_{U1}, \ldots, x_{Lp}, x_{Up})$ um vetor de entrada com $(2p+1)$ componentes.

Para $k = 1, \ldots, K$ um modelo linear multiclasse para dados intervalares baseados no argumento posterior simples definido para os limites inferior e superior dos intervalos conjuntamente é realizado da seguinte forma:

$$f_k^1(\boldsymbol{x}) = \boldsymbol{\beta}_k \boldsymbol{z} \tag{4}$$

O modelo multiclasse para dados intervalares baseado no argumento posterior definido para os limites inferior e superior dos intervalos conjuntamente é especificado em termos de K transformações lineares em que $\boldsymbol{z}$ é o vetor de entrada com $(2p+1)$ componentes. O vetor de coeficientes $\boldsymbol{\beta}_k, (k = 1, \ldots, K)$, nesse modelo multiclasse é estimado pelo método dos mínimos quadrados.

Seja ω um novo padrão a ser classificado como pertencente a uma de K classes, e seu vetor de covariáveis $\boldsymbol{x}(\omega) = (x_{\omega L1}, x_{\omega U1}, \ldots, x_{\omega Lp}, x_{\omega Up})$ onde $x_{\omega Lj} = a_j$ e $x_{\omega Uj} = b_j$.

A regra de classificação é definida a seguir: ω pertence à classe k se:

$$G^1(\omega) = argmax_{k \in C} f_k^1(\boldsymbol{x}_\omega) \tag{5}$$

3.5.2 Classificador linear baseado na associação dos argumentos posteriores dos limites separadamente

O classificador de padrões para dados intervalares linear baseado na associação dos argumentos posteriores dos limites inferior e superior dos intervalos separadamente(AAPLS) será apresentado nessa seção. A análise nesse método consiste no ajuste de duas regressões lineares binárias para cada classe. As saídas desses ajustes são combinadas com o objetivo de obter um argumento posterior associado.

Esse tipo de análise tem sido um interessante fato na pesquisa com relação a ensembles de classificadores [21] e resultados experimentais demonstram que é útil se os classificadores cometem erros em diferentes regiões do espaço de entrada, utilizando diferentes conjuntos de características, assim como, diferentes conjuntos de dados de treinamento.

Cada padrão de treinamento i possui dois vetores de p valores quantitativos $\boldsymbol{x}_L = (x_{L1}, \ldots, x_{Lp})$ e $(\boldsymbol{x}_{U1}, \ldots, x_{Up})$ como vetores de covariáveis onde $x_{Lj} = a_j$ e $x_{Uj} = b_j$. Seja $\boldsymbol{\alpha}_k = (\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kp})$ e $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})$ dois vetores de parâmetros desconhecidos, ambos de dimensão $(p+1)$ para cada classe k e $\boldsymbol{z}_L = (1, x_L)$ e $\boldsymbol{z}_U = (1, x_U)$ dois vetores de entrada, cada um com $(p+1)$ componentes.

Nesse método, inicialmente dois modelos lineares multiclasse para os limites inferior e superior são planejados separadamente e a média das probabilidades posteriores é computada para derivar uma decisão consensual [22]. Então, para $k = 1, \ldots, K$ um modelo linear multiclasse para dados intervalares baseados na associação dos argumentos posteriores definida para os limites inferior e superior dos intervalos é desenvolvida da seguinte forma:

$$f_k^2(\boldsymbol{x}) = \frac{f_{Lk}(\boldsymbol{x}_L) + f_{Uk}(\boldsymbol{x}_U)}{2} \tag{6}$$

onde

$$f_{Lk}(\boldsymbol{x}_L) = \boldsymbol{\alpha}_k \boldsymbol{z}_L \tag{7}$$

$$f_{Uk}(\boldsymbol{x}_U) = \boldsymbol{\beta}_k \boldsymbol{z}_U \tag{8}$$

O modelo multiclasse para dados intervalares baseados na associação do argumentos posteriores definida para os limites inferior e superior dos intervalos separadamente é especificada em termos de K transformações lineares relacionadas aos limites inferiores em que $\boldsymbol{z}_L$ é o vetor de entrada com $(p+1)$ componentes e K transformações lineares relacionadas aos limites superiores em que $\boldsymbol{z}_U$ é o vetor de entrada com $(p+1)$ componentes. Os vetores de coeficientes $\boldsymbol{\alpha}_k$ e $\boldsymbol{\beta}_k, (k = 1,\ldots,K)$ nesse modelo multiclasse, também são estimados pelo método dos mínimos quadrados.

Seja $\boldsymbol{\omega}$ um novo padrão, que será alocado a uma de K classes, e seu vetor de variáveis $\boldsymbol{x}_\omega = (\boldsymbol{x}_L(\omega), \boldsymbol{x}_L(\omega))$, que podem ser descritas como $\boldsymbol{x}_L(\omega) = (x_{\omega L1},\ldots,x_{\omega Lj},\ldots,x_{\omega Lp})$ e $\boldsymbol{x}_U(\omega) = (x_{\omega U1},\ldots,x_{\omega Uj},\ldots,x_{\omega Up})$ onde $x_{\omega Lj} = a_j$ e $x_{\omega Uj} = b_j$. A regra de classificação é definida a seguir: $\boldsymbol{\omega}$ é alocado na classe k se:

$$G^2(\omega) = argmax_{k\in C} f_k^2(\boldsymbol{x}_\omega) \tag{9}$$

3.6 Classificadores Propostos Baseados em Modelos de Regressão Logística

O modelo de regressão logística é mais flexível em suas suposições que a análise do discriminante. Pois ele não requer a suposição de que as variáveis independentes estejam distribuídas segundo uma distribuição normal, sejam linearmente relacionadas, ou que possuam variância igual dentro de cada grupo. Estando livre das suposições da análise de discriminante, o modelo de regressão logística pode ser utilizado em muitas situações.

A principal diferença entre a regressão logística e o modelo clássico de regressão linear é que a variável resposta na regressão logística é binária ou dicotômica, enquanto na regressão linear é um valor numérico contínuo.

3.6.1 Classificador logístico baseado na probabilidade posterior simples baseada nos centros dos intervalos

O classificador de padrões para dados intervalares logístico baseado na probabilidade posterior simples definida para os centros dos intervalos (LPPSC) será apresentado nessa seção.

Nesse classificador, um intervalo $[a,b]$ é representado pelo seu ponto médio $(a+b)/2$. Cada padrão de treinamento i possui um vetor de p centros $\boldsymbol{x}_c = (x_1^c,\ldots,x_p^c)$ como um vetor de covariáveis onde $x_j^c = (a_j + b_j)/2$. Seja $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1},\ldots,\beta_{kp})$ um vetor de parâmetros

desconhecidos de dimensão $(p+1)$ para a classe k $(k=1,\ldots,K)$ e $\boldsymbol{z}=(1,x_1^c,\ldots,x_j^c,\ldots,x_p^c)$ um vetor de entrada com $(p+1)$ componentes.

Para $k=1,\ldots,K$ o modelo logístico para dados intervalares baseados na probabilidade posterior simples definida para os centros dos intervalos é calculado da seguinte forma:

$$\pi_1(k)=Pr(Y_k=1/\boldsymbol{x}^c)=\frac{exp(\boldsymbol{\beta}_k^T\boldsymbol{z})}{1+exp(\boldsymbol{\beta}_k^T\boldsymbol{z})} \tag{10}$$

A equação (1) pode ser reescrita como:

$$\log\frac{Pr(Y_k=1/\boldsymbol{x}^c)}{Pr(Y_k=0/\boldsymbol{x}^c)}=\boldsymbol{\beta}_k^T\boldsymbol{z} \tag{11}$$

Na etapa de aprendizado, o modelo multiclasse para dados intervalares baseados na probabilidade posterior simples definida para os centros dos intervalos é especificada em termos de K transformações logísticas clássicas em que z é o vetor de entrada com $(p+1)$ componentes. O vetor de coeficientes $\boldsymbol{\beta}_k,(k=1,\ldots,K)$ nesse modelo multiclasse é estimado pelo método da máxima verossimilhança.

Desse modo, seja $\boldsymbol{\omega}$ um novo elemento, que deve ser alocado a uma das K classes do conjunto C, e seu vetor de intervalos $\boldsymbol{x}(\boldsymbol{\omega})=([a_{\omega 1},b_{\omega 1}],\ldots,[a_{\omega p},b_{\omega p}])$. Seja $\boldsymbol{x}^c(\boldsymbol{\omega})=(x_{\omega 1}^c,\ldots,x_{\omega p}^c)$ onde $\boldsymbol{x}_{\omega j}^c=(a_{\omega j}+b_{\omega j})/2$ é um vetor de covariáveis de $\boldsymbol{\omega}$.

Na etapa de alocação, a regra de classificação é definida a seguir: $\boldsymbol{\omega}$ é classificado como pertencente à classe k se:

$$\pi_1(k)=Pr(Y_k=1/\boldsymbol{x}^c(\boldsymbol{\omega}))\geq\pi_1(m)=Pr(Y_m=1/\boldsymbol{x}^c(\boldsymbol{\omega})),\forall\, m\in C \tag{12}$$

3.6.2 Classificador logístico baseado na probabilidade posterior simples definida para os limites conjuntamente

O classificador de padrões para dados intervalares logístico baseado na probabilidade posterior simples definida para os limites inferior e superior dos intervalos conjuntamente (LPPSLC) será apresentado nessa seção. Esse classificador considera os limites inferior e superior como variáveis independentes, ajustando um modelo logístico clássico para cada classe.

Uma variável intervalar X é representada por um par de variáveis quantitativas (x_L,x_U) onde $x_{Lj}=a_j$ e $x_{Uj}=b_j$. Assim, cada padrão de treinamento i possui um vetor de $2p$ valores quantitativos $\boldsymbol{x}=\{x_{L1},x_{U1},\ldots,x_{Lp},x_{Up}\}$ como um vetor de covariáveis. Seja $\boldsymbol{\beta}_k=(\beta_{k0},\beta_{k1},\beta_{k2},\ldots,\beta_{k2p})$ um vetor de parâmetros desconhecidos de dimensão $(2p+1)$ para a classe $k,(k=1,\ldots,K)$ e $\boldsymbol{z}=(1,x_{L1},x_{U1},\ldots,x_{Lp},x_{Up})$ um vetor de entrada com $(2p+1)$ componentes.

Para $k=1,\ldots,K$ um modelo logístico multiclasse para dados intervalares baseados na probabilidade posterior simples definida para os limites inferior e superior dos intervalos conjuntamente é realizada da seguinte forma:

$$\pi_2(k)=Pr(Y_k=1/\boldsymbol{x})=\frac{exp(\boldsymbol{\beta}_k^T\boldsymbol{z})}{1+exp(\boldsymbol{\beta}_k^T\boldsymbol{z})} \tag{13}$$

A equação (13) pode ser reescrita como:

$$\log \frac{Pr(Y_k = 1/\boldsymbol{x})}{Pr(Y_k = 0/\boldsymbol{x})} = \boldsymbol{\beta}_k^T \boldsymbol{z} \tag{14}$$

Como no modelo anterior, o modelo multiclasse para dados intervalares baseado na probabilidade posterior definida para os limites inferior e superior dos intervalos conjuntamente é especificado em termos de K transformações logísticas em que $\boldsymbol{z}$ é o vetor de entrada com $(2p+1)$ componentes. O vetor de coeficientes $\boldsymbol{\beta}_k, (k = 1, \ldots, K)$, nesse modelo multiclasse é, também, estimado pelo método da máxima verossimilhança.

Seja ω um novo padrão a ser classificado como pertencente a uma de K classes do conjunto C, e seu vetor de covariáveis $\boldsymbol{x}(\omega) = (x_{\omega L1}, x_{\omega U1}, \ldots, x_{\omega Lp}, x_{\omega Up})$ onde $x_{\omega Lj} = a_j$ e $x_{\omega Uj} = b_j$.

A regra de classificação é definida a seguir: ω pertence à classe k se:

$$\pi_2(k) = Pr(Y_k = 1/\boldsymbol{x}(\omega)) \geq \pi_2(m) = Pr(Y_m = 1/\boldsymbol{x}(\omega)), \forall\, m \in C \tag{15}$$

3.6.3 Classificador logístico baseado na probabilidade posterior simples definida para os vértices dos padrões

Nessa seção, o classificador de padrões para dados intervalares logístico baseado na probabilidade posterior simples definida para os vértices dos hipercubos (LPPSV) será introduzido. Nesse método, um vetor de intervalo simbólico $\boldsymbol{x} = ([a_1, b_1], \ldots, [a_p], b_p)$ visualizado no espaço de descrição $\Re^p$ por um hipercubo R com 2p vértices, pode ser descrito por uma matriz de todos os vértices do hipercubo no $\Re^p$. Tal metodologia foi utilizada por Chouakria e Diday [23] para sua extensão de análise de componentes principais (*Principal Components Analysis - PCA*) e por Silva e Brito [13] para sua extensão de análise do discriminante linear. Então, para $p = 2$, o vetor $\boldsymbol{x} = ([a_1, b_1], [a_2, b_2])$ pode ser descrito pela matriz:

$$M = \begin{pmatrix} a^1 & \ldots & a^p \\ \vdots & \ddots & \vdots \\ b^1 & \ldots & b^p \end{pmatrix}$$

Então, agora cada padrão de ω tem uma matriz de $2p$ linhas que são vetores p numéricos, correspondendo a todas as possíveis combinações dos limites dos intervalos $[a_j, b_j], (j = 1, \ldots, p)$ e um modelo logístico clássico binário é ajustado sobre o conjunto de padrões de treinamento ω^* com $n \times 2p$ elementos.

Cada padrão de treinamento i resultante do conjunto de padrões de treinamento ω^* tem um vetor de p valores quantitativos $\boldsymbol{x}^v = (x_1^v, \ldots, x_p^v)$ como um vetor de covariáveis onde $x_j^v, (j = 1, \ldots, p)$ é uma coordenada de um vértice de um hipercubo no $\Re^p$. Seja $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})$ um vetor de parâmetros desconhecidos de tamanho $(p+1)$ para a classe $k, (k = 1, \ldots, K)$ e $\boldsymbol{z} = (1, x_1^v, \ldots, x_j^v, \ldots, x_p^v)$ um vetor de entrada com $(p+1)$ componentes.

Para $k = 1, \ldots, K$ um modelo logístico multiclasse para dados intervalares baseados na probabilidade posterior definida para os vértices dos intervalos é ajustado da seguinte forma:

$$\pi_3(k) = Pr(Y_k = 1/\boldsymbol{x}^v) = \frac{exp(\boldsymbol{\beta}_k^T \boldsymbol{z})}{1 + exp(\boldsymbol{\beta}_k^T \boldsymbol{z}} \tag{16}$$

A equação (7) pode ser reescrita como:

$$\log \frac{Pr(Y_k = 1/\boldsymbol{x}^v)}{Pr(Y_k = 0/\boldsymbol{x}^v)} = \boldsymbol{\beta}_k^T \boldsymbol{z} \tag{17}$$

Esse modelo logístico multiclasse para dados intervalares baseados na probabilidade posterior definida para os vértices dos intervalos é especificado, também, em K transformações logísticas em que $\boldsymbol{z}$ é o vetor de entrada com $(p+1)$ componentes. O vetor de coeficientes $\boldsymbol{\beta}_k, (k = 1, \ldots, K)$ nesse modelo multiclasse é, também, estimado pelo método da máxima verossimilhança.

Seja ω um novo padrão, que será classificado como pertencente a uma das K classes do conjunto C, e seu vetor de intervalos $\boldsymbol{x}_\omega = ([a_{\omega 1}, b_{\omega 1}], \ldots, [a_{\omega p}, b_{\omega p}])$. Realizando a representação dos vértices, esse padrão possui $2p$ vetores de covariáveis $\boldsymbol{x}_\omega^v(t) = (x_{\omega 1}^v(t), \ldots, x_{\omega p}^v(t))$ onde $x_{\omega j}^v(t), (t = 1, \ldots, 2^p)$ é uma coordenada de um vértice do hipercubo $\boldsymbol{x}_\omega$ no $\Re^p$.

Na etapa de alocação, a regra de classificação é definida a seguir: ω pertence à classe k se

$$\Pi(k) \geq \Pi(m), \forall\, m \in C \tag{18}$$

$\Pi(k)$ é uma probabilidade definida segundo um dos métodos seguintes:

$$\Pi(k) = \max\{\pi_3(k) = Pr(Y_k = 1/\boldsymbol{x}_\omega^v(t))\}, \forall\, v \in \{1, \ldots, 2^p\} \tag{19}$$

$$\Pi(k) = \min\{\pi_3(k) = Pr(Y_k = 1/\boldsymbol{x}_\omega^v(t))\}, \forall\, v \in \{1, \ldots, 2^p\} \tag{20}$$

$$\Pi(k) = \frac{1}{2^p} \sum_{t=1}^{2^p} \{\pi_3(k) = Pr(Y_k = 1/\boldsymbol{x}_\omega^v(t))\} \tag{21}$$

3.6.4 Classificador logístico baseado na associação da probabilidade posterior definida para os limites

O classificador de padrões para dados intervalares logístico baseado na associação da probabilidade posterior dos limites inferior e superior dos intervalos separadamente(LAPPLS) será apresentado nessa seção. A análise nesse método consiste no ajuste de duas regressões logísticas binárias para cada classe. As saídas desses ajustes são combinadas com o objetivo de obter uma probabilidade posterior associada. Quando os limites inferiores não estão correlacionados com os limites superiores de alguma maneira, o esquema de combinação da regressão logística pode melhorar o desempenho.

Cada padrão de treinamento i possui dois vetores de p valores quantitativos $\boldsymbol{x}_L = (x_{L1}, \ldots, x_{Lp})$ e $(\boldsymbol{x}_{U1}, \ldots, x_{Up})$ como vetores de covariáveis onde $x_{Lj} = a_j$ e $x_{Uj} = b_j$. Seja $\boldsymbol{\alpha}_k = (\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kp})$ e $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})$ dois vetores de parâmetros desconhecidos, ambos de dimensão $(p+1)$ para cada classe k e $\boldsymbol{z}_L = (1, x_L)$ e $\boldsymbol{z}_U = (1, x_U)$ dois vetores de entrada, cada um com $(p+1)$ componentes.

Nesse método, inicialmente dois modelos logísticos multiclasse para os limites inferior e superior são planejados separadamente e a média das probabilidades posteriores é computada

para derivar uma decisão consensual [22]. Então, para $k = 1,\ldots,K$ um modelo logístico multiclasse para dados intervalares baseados na associação da probabilidade posterior definida para os limites inferior e superior dos intervalos é desenvolvida da seguinte forma:

$$\pi_4(k) = \frac{Pr(Y = 1/\boldsymbol{x}_L) + Pr(Y = 1/\boldsymbol{x}_U)}{2} \tag{22}$$

onde

$$Pr(Y = 1/x_L) = \frac{exp(\boldsymbol{\alpha}_k^T \boldsymbol{z}_L)}{1 + exp(\boldsymbol{\alpha}_k^T \boldsymbol{z}_L)} \tag{23}$$

$$Pr(Y = 1/x_U) = \frac{exp(\boldsymbol{\beta}_k^T \boldsymbol{z}_U)}{1 + exp(\boldsymbol{\beta}_k^T \boldsymbol{z}_U)} \tag{24}$$

As equações (23) e (24) podem ser reescritas, respectivamente, como:

$$\log \frac{Pr(Y = k/\boldsymbol{x}_L)}{Pr(Y \neq k/\boldsymbol{x}_L)} = \alpha_{k0} + \alpha_{k1} x_{L1} + \ldots + \alpha_{kp} x_{Lp} \tag{25}$$

$$\log \frac{Pr(Y = k/\boldsymbol{x}_U)}{Pr(Y \neq k/\boldsymbol{x}_U)} = \beta_{k0} + \beta_{k1} x_{U1} + \ldots + \beta_{kp} x_{Up} \tag{26}$$

O modelo multiclasse para dados intervalares baseados na associação da probabilidade posterior definida para os limites inferior e superior dos intervalos separadamente é especificada em termos de K transformações logísticas relacionadas aos limites inferiores em que $\boldsymbol{z}_L$ é o vetor de entrada com $(p+1)$ componentes e K transformações logísticas relacionadas aos limites superiores em que $\boldsymbol{z}_U$ é o vetor de entrada com $(p+1)$ componentes. Os vetores de coeficientes $\boldsymbol{\alpha}_k$ e $\boldsymbol{\beta}_k, (k = 1,\ldots,K)$ nesse modelo multiclasse, também são estimados pelo método da máxima verossimilhança.

Seja ω um novo padrão, que será alocado a uma de K classes do conjunto C, e seus vetores de covariáveis $\boldsymbol{x}_L(\omega) = (x_{\omega L1},\ldots,x_{\omega Lj},\ldots,x_{\omega Lp})$ e $\boldsymbol{x}_U(\omega) = (x_{\omega U1},\ldots,x_{\omega Uj},\ldots,x_{\omega Up})$ onde $x_{\omega Lj} = a_j$ e $x_{\omega Uj} = b_j$. A regra de classificação é definida a seguir: ω é alocado na classe k se

$$\pi_4(k) = Pr(Y_k = 1/\boldsymbol{x}_\omega) \geq \pi_4(k) = Pr(Y_m = 1/\boldsymbol{x}_\omega), \forall\, m \in C \tag{27}$$

3.7 Considerações Finais

Nesse capítulo foram introduzidos 6 classificadores de dados simbólicos de tipo intervalo baseados nas metodologias de regressão linear e regressão logística. Os seis classificadores logísticos diferem entre si pelo modo como a representação intervalar é utilizada e pelo modelo de regressão utilizado.

O primeiro modelo linear (APSLC) ajusta uma equação de regressão linear para cada classe utilizando os limites inferiores e superiores de cada intervalo como variáveis independentes, porém de forma conjunta.

O segundo modelo linear (AAPLS) ajusta duas equações de regressão linear para cada classe utilizando os limites inferiores e superiores como variáveis independentes de maneira separada, combinando os resultados através de uma média.

O primeiro modelo logístico (LPPSC) ajusta uma equação de regressão logística para cada classe utilizando os centros os intervalos como variáveis independentes, para predizer a classe de um novo elemento.

O segundo modelo logístico (LPPSLC) ajusta uma equação de regressão logística para cada classe utilizando os limites inferiores e superiores de cada intervalo como variáveis independentes, porém de forma conjunta.

O terceiro modelo logístico (LPPSV) ajusta uma equação de regressão logística para cada classe utilizando os vértices dos hipercubos que representam os intervalos como variáveis independentes, combinando os resultados através de três métodos (regra dos mínimos, regra dos máximos e regra da média).

O quarto modelo logístico (LAPPLS) ajusta duas equações de regressão logística para cada classe utilizando os limites inferiores e superiores como variáveis independentes de maneira separada, combinando os resultados através de uma média.

É importante ressaltar que todos os modelos utilizam uma transformação nas variáveis intervalares, para variáveis clássicas, sendo que no primeiro modelo logístico, o número de variáveis (dimensão) é mantido, no primeiro e segundo lineares e segundo e quarto logísticos, é o dobro do número inicial, no terceiro modelo, é igual a 2^p, onde p é o número de variáveis inicial.

A diferença entre o primeiro modelo linear e o segundo modelo linear, assim como no segundo modelo logístico e no quarto modelo logístico é que no primeiro modelo linear e no segundo modelo logístico as novas variáveis são ajustadas de maneira conjunta, enquanto que no segundo modelo linear e no quarto modelo logístico as variáveis são ajustadas para o modelo de regressão de forma separada, sendo uma regressão para os limites inferiores e outra regressão para os limites superiores, para cada classe.

CAPÍTULO 4

Experimentos e Resultados

Uma teoria é algo que ninguém acredita, exceto a pessoa que a fez . Um experimento é algo que todos acreditam, exceto a pessoa que o fez .

—ALBERT EINSTEIN (prêmio Nobel de Física em 1921)

Para avaliar o desempenho dos classificadores propostos, foram realizados experimentos com dois conjuntos de dados simbólicos intervalares sintéticos, tais conjuntos foram gerados com baixo e moderado graus de sobreposição. Também foram realizados experimentos com um conjunto de dados simbólicos intervalares real. Neste capítulo, serão apresentados os resultados dos classificadores propostos quando utilizados sobre os conjuntos de dados sintéticos e real.

A precisão dos classificadores para os conjuntos de dados sintéticos é aferida através da taxa de erro de classificação, que é estimada na estrutura de uma simulação de Monte Carlo onde o conjunto de teste e de treinamento são selecionados aleatoriamente de cada conjunto de dados sintético.

Métodos de Monte Carlo são uma classe de algoritmos computacionais que se baseiam na repetição de amostragem aleatória para computar seus resultados. Neste trabalho, os experimentos com dados sintéticos utilizam a abordagem de um método de Monte Carlo, onde conjuntos de dados são gerados e classificados 500 vezes, para então, a taxa de erro média ser obtida.

O conjunto de treinamento corresponde a 75% do conjunto de dados original e o conjunto de teste corresponde a 25% do conjunto de dados original. No conjunto de dados real, foi utilizada a abordagem "leave-one-out".

4.1 Experimentos com conjuntos de dados sintéticos

Inicialmente, dois conjuntos de dados clássicos quantitativos no $\Re^2$ são gerados (Figuras 4.1 e 4.2). Esses conjuntos de dados têm 450 pontos dispersos entre três classes de tamanhos diferentes: uma classe com forma de elipse e tamanho 200 e duas classes de formato esférico, possuindo tamanhos 150 e 100 respectivamente.

Cada classe nesses conjuntos de dados intervalares sintéticos foi gerada através de duas distribuições normais independentes. Cada ponto (z_1, z_2) de cada um desses conjuntos de dados clássicos quantitativos é utilizado como "semente"para um vetor de intervalos (retângulo), como definido a seguir:

$$([z_1 - \gamma_1/2, z_1 + \gamma_1/2], [z_2 - \gamma_2/2, z_2 + \gamma_2/2]) \tag{19}$$

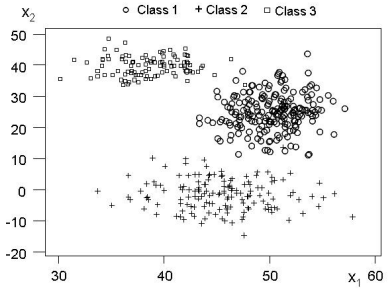

Figura 4.1 Conjunto de dados clássicos quantitativos 1

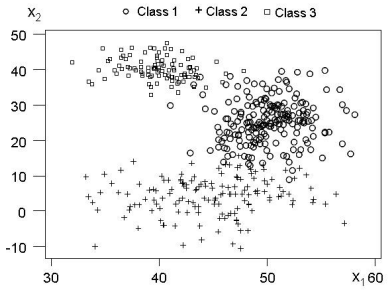

Figura 4.2 Conjunto de dados clássicos quantitativos 2

onde os parâmetros $\gamma_1 e \gamma_2$ são selecionados aleatoriamente de um mesmo intervalo pré-definido.

O conjunto de dados simbólicos intervalares sintéticos 1 foi construído através do conjunto de dados clássicos quantitativos 1, de acordo com os seguintes parâmetros (doravante chamados *configuração 1*):

Classe 1: $\mu_1 = 50$, $\mu_2 = 25$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 36$;
Classe 2: $\mu_1 = 45$, $\mu_2 = -2$, $\sigma_1^2 = 25$ e $\sigma_2^2 = 25$;
Classe 3: $\mu_1 = 38$, $\mu_2 = 40$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 9$;

A figura 4.3 mostra o conjunto de dados intervalares sintéticos 1, no qual γ_1 e γ_2 foram selecionados aleatoriamente do intervalo $[1,10]$. Esse conjunto de dados intervalar mostra baixo grau de sobreposição.

O conjunto de dados simbólicos intervalares sintéticos 2 foi construído a partir do conjunto de dados clássicos quantitativos 2 de acordo com os seguintes parâmetros (doravante chamados *configuração 2*):

Classe 1: $\mu_1 = 50$, $\mu_2 = 25$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 36$;
Classe 2: $\mu_1 = 45$, $\mu_2 = 5$, $\sigma_1^2 = 25$ e $\sigma_2^2 = 25$;
Classe 3: $\mu_1 = 45$, $\mu_2 = 40$, $\sigma_1^2 = 9$ e $\sigma_2^2 = 9$;

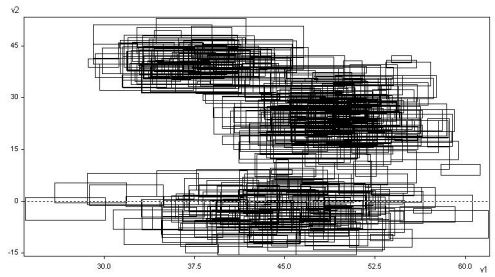

Figura 4.3 Conjunto de dados simbólicos intervalares 1

A figura 4.4 mostra o conjunto de dados intervalares sintéticos 2, no qual γ_1 e γ_2 foram selecionados aleatoriamente do mesmo intervalo $[1,10]$. Esse conjunto de dados intervalar mostra um grau moderado de sobreposição.

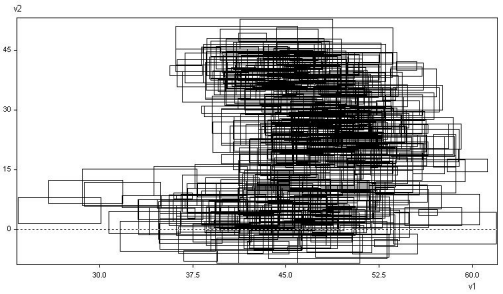

Figura 4.4 Conjunto de dados simbólicos intervalares 2

Na estrutura da simulação de Monte Carlo, 500 replicações desse processo de geração são realizados. Os parâmetros γ_1 e γ_2 são escolhidos aleatoriamente 100 vezes de cada um dos intervalos: $[1,10]$, $[1,20]$, $[1,30]$, $[1,40]$ e $[1,50]$.

Além disso, esse processo foi repetido para as sementes dos conjuntos de dados clássicos quantitativos (configurações 1 e 2). Os classificadores APSLC, AAPLS, LPPSC, LPPSLC, LPPSV e LAPPLS foram aplicados aos conjuntos de dados simbólicos intervalares 1 e 2. A taxa de erro de classificação é calculada para o conjunto de teste e a taxa de erro estimada de classificação corresponde à média das taxas de erro encontradas nas 100 replicações do conjunto de teste para cada intervalo de γ_1 e γ_2.

As tabelas 4.1 e 4.2 mostram os valores das médias e desvios padrão (em parênteses) da taxa de erro de classificação obtidas com os classificadores lineares (tabela 4.1) e logísticos (tabela 4.2) para a configuração 1 para cada γ_1 e γ_2 escolhidos dos intervalos $[1,10]$, $[1,20]$, $[1,30]$, $[1,40]$ e $[1,50]$. Dados intervalares nessa configuração mostram baixo grau de dificuldade.

Entre os classificadores lineares, o melhor desempenho foi alcançado pelo classificador de padrões para dados intervalares baseado na associação da probabilidade posterior dos limites

inferior e superior dos intervalos separadamente (AAPLS), o seu desempenho melhorou com o aumento do tamanho dos intervalos, enquanto o classificador APSLC se manteve relativamente estável em relação ao tamanho dos intervalos.

Entre os classificadores logísticos, o melhor desempenho foi alcançado pelo classificador de padrões para dados intervalares logístico baseado na associação da probabilidade posterior dos limites inferior e superior dos intervalos separadamente(LAPPLS). Além disso, analogamente ao classificador linear AAPLS, o desempenho do LAPPLS melhorou com o aumento do tamanho dos intervalos. Como esperado, o pior desempenho foi obtido pelo classificador de padrões para dados intervalares logístico baseado na probabilidade posterior simples definida para os centros dos intervalos (LPPSC).

É interessante ressaltar que a taxa de erro para o método LPPSC não foi afetada pelo tamanho dos intervalos. Esse resultado pode ser explicado pelo fato de que a estrutura dos pontos que pertencem aos conjuntos de dados 1 e 2 são preservados durante as replicações realizadas na estrutura da simulação de Monte Carlo.

Com respeito ao classificador para dados intervalares baseados na probabilidade posterior simples definida para os vértices dos hipercubos (LPPSV), a taxa de erro piorou quando o tamanho do intervalo cresceu para todas as regras de classificação. A segunda melhor taxa de erro foi alcançada pelo método LPPSLC. Além disso, é importante perceber que o desempenho do classificador LPPSLC foi afetado sutilmente pelo tamanho dos intervalos.

Tamanho para γ_i, $i = 1,2$	APSLC	AAPLS
$[1,10]$	3.64	3.48
	(0.01797759)	(0.01787968)
$[1,20]$	3.67	2.84
	(0.01929960)	(0.01738624)
$[1,30]$	3.64	2.26
	(0.01691604)	(0.01460347)
$[1,40]$	3.69	1.90
	(0.01924696)	(0.01505071)
$[1,50]$	3.68	1.84
	(0.01833132)	(0.01621484)

Tabela 4.1 A média (%) e o desvio padrão (entre parênteses) da taxa de erro para o conjunto de dados simbólicos intervalares sintéticos 1 (classificadores lineares)

Tamanho para γ_i, $i = 1,2$	LPPSC	LPPSLC	LPPSV			LAPPLS
			Regra do mínimo	Regra do máximo	Regra da média	
$[1,10]$	57.42	1.66	1.69	1.72	3.10	1.29
	(0.0303)	(0.0131)	(0.0109)	(0.0135)	(0.0144)	(0.0096)
$[1,20]$	57.89	1.72	4.45	3.05	9.41	1.58
	(0.0321)	(0.0121)	(0.0215)	(0.0145)	(0.0269)	(0.0118)
$[1,30]$	58.24	1.35	12.20	5.51	18.24	1.20
	(0.0355)	(0.0111)	(0.261)	(0.0217)	(0.0307)	(0.0118)
$[1,40]$	57.87	1.67	19.38	7.60	22.95	0.93
	(0.0308)	(0.0132)	(0.0261)	(0.0290)	(0.0347)	(0.0102)
$[1,50]$	58.03	1.33	25.31	9.81	25.90	0.87
	(0.0339)	(0.0108)	(0.0218)	(0.0298)	(0.0343)	(0.0099)

Tabela 4.2 A média (%) e o desvio padrão (entre parênteses) da taxa de erro para o conjunto de dados simbólicos intervalares sintéticos 1 (classificadores logísticos)

As tabelas 4.3 e 4.4 mostram os valores das médias e desvios padrão (em parêteses) das taxas de erro de classificação obtidas pelos classificadores lineares e logísticos aplicados na configuração 2 assim como para cada γ_1 e γ_2 escolhidos dos intervalos $[1,10]$, $[1,20]$, $[1,30]$, $[1,40]$ e $[1,50]$. Dados intervalares nessa configuração apresentam grau moderado de dificuldade de classificação.

Tamanho para γ_i, $i = 1,2$	APSLC	AAPLS
$[1,10]$	12.51	11.33
	(0.03064123)	(0.02980246)
$[1,20]$	12.95	10.37
	(0.03353984)	(0.02808535)
$[1,30]$	12.84	9.20
	(0.03066527)	(0.02717949)
$[1,40]$	12.72	8.16
	(0.03348891)	(0.02921632)
$[1,50]$	12.43	8.07
	(0.03193513)	(0.02899221)

Tabela 4.3 A média (%) e o desvio padrão (entre parênteses) da taxa de erro para o conjunto de dados simbólicos intervalares sintéticos 2 (classificadores lineares)

Entre os classificadores lineares, o melhor desempenho foi novamente alcançado pelo classificador AAPLS. No segundo conjunto ambos os classificadores lineares apresentaram o mesmo comportamento em relação ao comportamento apresentado no primeiro conjunto.

Entre os classificadores logísticos, o melhor desempenho também foi alcançado pelo classificador LAPPLS. Contudo, o desempenho do classificador baseado na metodologia LAPPLS foi apenas ligeiramente superior àquele alcançado pela metodologia LPPSLC. Além disso, o desempenho de ambos é afetado levemente quando o tamanho dos intervalos aumenta. Esse fato sugere que esses métodos são estáveis quanto ao tamanho dos intervalos.

Como esperado, o pior desempenho foi novamente obtido pelo classificador LPPSC. Porém é interessante ressaltar que as médias e desvios foram os mesmos para as configurações 1 e 2.

Tamanho para γ_i, $i = 1,2$	LPPSC	LPPSLC	LPPSV			LAPPLS
			Regra do mínimo	Regra do máximo	Regra da média	
$[1,10]$	57.42	5.60	7.74	7.22	10.53	5.10
	(0.0303)	(0.0220)	(0.0225)	(0.0244)	(0.0314)	(0.0195)
$[1,20]$	57.89	5.73	15.36	12.42	20.54	4.96
	(0.0321)	(0.0191)	(0.0280)	(0.0287)	(0.0344)	(0.0182)
$[1,30]$	58.24	5.78	25.16	17.83	27.42	4.78
	(0.0355)	(0.0250)	(0.299)	(0.0311)	(0.0343)	(0.0211)
$[1,40]$	57.87	5.63	32.11	20.44	29.00	4.96
	(0.0308)	(0.0236)	(0.0296)	(0.0374)	(0.0358)	(0.0224)
$[1,50]$	58.03	5.49	38.01	23.37	31.00	5.08
	(0.0339)	(0.0246)	(0.0315)	(0.0368)	(0.0379)	(0.0229)

Tabela 4.4 A média (%) e o desvio padrão (entre parênteses) da taxa de erro para o conjunto de dados simbólicos intervalares sintéticos 2

Esse resultado é explicado pelo fato de que os conjuntos de dados quantitativos 1 e 2 são muito parecidos.

Com respeito ao classificador LPPSV, como na configuração 1, a taxa de erro foi degradada a medida que o tamanho dos intervalos aumentou para todas as regras de classificação.

Os resultados das tabelas 4.1 e 4.2 demonstraram a importância de levar em consideração a informação presente nos limites inferiores e superiores dos intervalos tanto de forma conjunta como separadamente em modelos baseados em regressão linear e logística para classificação de dados simbólicos intervalares. Para comparar os métodos APSLC e AAPLS e os métodos LPPSLC e LAPPLS, dois testes t de Student foram executados a um nível de significância de 5%.

As tabelas 4.5 e 4.6 mostram as hipóteses nula e alternativas e os valores observados da estatística dos testes seguindo uma distribuição t de Student com 198 graus de liberdade. Nessa tabela μ_1 e μ_2 são, respectivamente, a taxa de erro média para as abordagens APSLC e AAPLS (Tabela 4.5) e LPPSLC e LAPPLS (tabela 4.6).

Tamanho para γ_i, $i = 1,2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_2 < \mu_1$			
	Conjunto de dados 1	Decisão	Conjunto de dados 2	Decisão
$\gamma \in [1,10]$	-79.22	Rejeitar H_0	-276.26	Rejeitar H_0
$\gamma \in [1,20]$	-321.15	Rejeitar H_0	-590.92	Rejeitar H_0
$\gamma \in [1,30]$	-617.91	Rejeitar H_0	-890.51	Rejeitar H_0
$\gamma \in [1,40]$	-734.66	Rejeitar H_0	-1027.85	Rejeitar H_0
$\gamma \in [1,50]$	-752.85	Rejeitar H_0	-1012.90	Rejeitar H_0

Tabela 4.5 Estatísticas do teste t de Student comparando os métodos APSLC e AAPLS

Desses resultados nós podemos rejeitar a hipótese de que o desempenho médio (medido pela taxa erro) do método APSLC é igual ao desempenho médio da abordagem AAPLS e que o desempenho médio do método LPPSLC é igual ao desempenho médio da abordagem LAPPS. Sendo as abordagens AAPLS e LAPPLS superiores.

Tamanho para γ_i, $i = 1,2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_2 < \mu_1$			
	Conjunto de dados 1	Decisão	Conjunto de dados 2	Decisão
$\gamma \in [1,10]$	-227.81	Rejeitar H_0	-170.07	Rejeitar H_0
$\gamma \in [1,20]$	-82.83	Rejeitar H_0	-291.85	Rejeitar H_0
$\gamma \in [1,30]$	-92.98	Rejeitar H_0	-305.67	Rejeitar H_0
$\gamma \in [1,40]$	-443.59	Rejeitar H_0	-205.91	Rejeitar H_0
$\gamma \in [1,50]$	-313.97	Rejeitar H_0	-121.99	Rejeitar H_0

Tabela 4.6 Estatísticas do teste t de Student comparando os métodos LPPSLC e LAPPLS

4.2 O método Leave-One-Out

O método Leave-One-Out, é um caso particular da técnica N-Fold Cross-Validation, que é uma técnica de validação clássica da área estatística e de aprendizagem de máquina. O método Leave-One-Out é geralmente utilizado para estimar a taxa de erro sobre conjuntos de dados de tamanho pequeno. O número de conjuntos construídos (*Folds*) é igual ao número de elementos do conjunto original, ou seja, $N = K$, onde K é igual ao número de indivíduos no conjunto de dados original. Após a divisão, um dos K subconjuntos, ou seja, um elemento, é selecionado aleatoriamente para o conjunto de teste, e os outros são utilizados no conjunto de treinamento, após a classificação é calculada a taxa de erro. O processo é repetido até que todos os elementos tenham sido utilizados no conjunto de teste e então a média das taxas de erro obtida para cada iteração do algoritmo é calculada.

4.3 Experimentos com conjuntos de dados reais

O conjunto de dados intervalares sobre carros tem sido aplicado para validar diferentes métodos de classificação supervisionada e não-supervisionada na literatura da Análise de dados Simbólicos, como por exemplo [13],[24], [25], [26], [27], [28]. Neste trabalho, os classificadores APSLC, AAPLS, LPPSC, LPPSLC, LPPSV e LAPPLS foram aplicados a esse conjunto de dados e os resultados da taxa média de erro obtidos pelo processo de validação cruzada utilizando o método Leave-One-Out. Além disso, uma comparação entre esses classificadores e um classificador linear introduzido na literatura de SDA é considerado.

O conjunto de dados intervalares sobre carros possui 33 modelos de carros, cada um descrito por 8 variáveis intervalares e uma variável categórica que define a classe de cada modelo. Contudo, quatro variáveis (Passo, Comprimento, Largura e Altura) têm valores de seus limites inferiores iguais aos valores de seus limites inferiores para quase todos os 33 modelos de carros. Além disso, existem alguns pares de padrões do conjunto de dados que possuem valores idênticos para essas variáveis. Esse fato afetou a convergência do classificador linear APSLC e do classificador logístico LPPSLC , que consideram a análise conjunta dos limites inferiores e superiores.

Então, duas aplicações do classificador APSLC e do classificador LPPSLC foram realizadas. Na primeira aplicação os limites superiores para as variáveis Passo, Comprimento, Largura

e Altura foram removidos e a taxa de erro obtida pelo classificador APSLC foi 39.39% e a taxa de erro obtida pelo classificador LPPSLC foi 48.48%. Na segunda aplicação, os limites inferiores para as variáveis Passo, Comprimento, Largura e Altura foram removidos e a taxa de erro obtida para o classificador APSLC foi 33.33%, já para o classificador LPPSLC a taxa de erro obtida foi 57.57

Com respeito aos classificadores AAPLS, LPPSC, LPPSV e LAPPLS, cada aplicação foi realizada utilizando o conjunto de dados original. Os valores para a taxa de erro obtidos foram: 36.36% para o classificador LPPSC, 30.3%, 36.4% e 30.3% para o classificador LPPSV considerando as regras do mínimo, máximo e média, respectivamente, 27.2% para o classificador LAPPLS e, também, 27.2% para o classificador AAPLS.

Os resultados mostram que o desempenho do classificador LAPPLS foi melhor que o dos outros classificadores. Com relação à comparação com o classificador linear baseado na abordagem distribucional introduzido em [13], o classificador LAPPLS apresentou um nível de precisão tão bom quanto o demonstrado por [13].

Elemento	Preço	Motor	Velocidade Máxima	...	Altura	Categoria
Alfa 145	[27806; 33596]	[1370; 1910]	[185; 211]	...	[143; 143]	Utilitário
Alfa 156	[41593; 62291]	[1598; 2492]	[200; 227]	...	[142; 142]	Berlina
Alfa 166	[64499; 88760]	[1970; 2959]	[204; 211]	...	[142; 142]	Luxo
Aston Martin	[260500; 460000]	[5935; 5935]	[298; 306]	...	[124; 132]	Esporte
Audi A3	[40230; 68838]	[1595; 1781]	[189; 238]	...	[142; 142]	Utilitário
Audi A6	[68216; 140205]	[1781; 4172]	[216; 250]	...	[145; 145]	Berlina
Audi A8	[123849; 171417]	[2771; 4172]	[232; 250]	...	[144; 144]	Luxo
Bmw 3	[45407; 76392]	[1796; 2979]	[201; 247]	...	[142; 142]	Berlina
Bmw 5	[70292; 198792]	[2171; 4398]	[226; 250]	...	[144; 144]	Luxo
Bmw 7	[104892; 276792]	[2793; 5397]	[228; 240]	...	[143; 143]	Luxo
Ferrari	[240292; 391692]	[3586; 5474]	[295; 298]	...	[130; 130]	Esporte
Punto	[19229; 30885]	[1242; 1910]	[155; 170]	...	[148; 148]	Utilitário
Fiesta	[19242; 24742]	[1242; 1753]	[155; 170]	...	[132; 132]	Utilitário
Focus B	[27492; 34092]	[1596; 1753]	[185; 193]	...	[143; 143]	Berlina
Honda NSK S	[205242; 215242]	[2977; 3179]	[260; 270]	...	[129; 129]	Esporte
Lamborghini S	[413000; 423000]	[5992; 5992]	[335; 335]	...	[111; 111]	Esporte
Lancia YU	[19837; 29034]	[1242; 1242]	[158; 174]	...	[144; 144]	Utilitário
Lancia KL	[58806; 81306]	[1998; 2959]	[212; 220]	...	[146; 146]	Luxo
Maserati GTS	[155000; 159500]	[3217; 3217]	[280; 290]	...	[131; 131]	Esporte
Mercedes SLS	[132800; 262500]	[2799 : 5987]	[232; 250]	...	[129; 129]	Esporte
Mercedes Classe CB	[55902; 115248]	[1998; 3199]	[210; 250]	...	[143; 143]	Berlina
Mercedes Classe EL	[69243; 389405]	[1998; 5439]	[222; 250]	...	[144; 144]	Luxo
Mercedes Classe SL	[128202; 394342]	[3199; 5786]	210 : 240	...	144 : 144	Luxo
Nissan Micra U	[18492; 24192]	[998; 1348]	150 : 164	...	144 : 144	Utilitário
Corsa U	[19212; 30612]	[973; 1796]	[155; 202]	...	[155; 202]	Utilitário
Vectra B	[36492; 49092]	[1598; 2171]	[193; 207]	...	[143; 143]	Berlina
Porsche S	[147704; 246412]	[3387; 3600]	[280; 305]	...	[130; 131]	Esporte
Twingo U	[16992; 23492]	[1149; 1149]	[151; 168]	...	[142; 142]	Utilitário
Rover 75 U	[21492; 33042]	[1119; 1994]	[160; 185]	...	[142; 142]	Utilitário
Rover 75 B	[50490; 65399]	[1796; 2497]	[195; 210]	...	[143; 143]	Berlina
Skoda Fabia	[19519; 32686]	[1397; 1896]	[157; 183]	...	[145; 145]	Utilitário
Skoda Octavia	[27419; 48679]	[1585; 1896]	[190; 191]	...	[143; 143]	Berlina
Passat	[39676; 63455]	[1595; 2496]	[192; 220]	...	[146; 146]	Luxo

Tabela 4.7 Conjunto de dados intervalares sobre carros.

CAPÍTULO 5

Conclusão e Trabalhos Futuros

Nem todo fim é o objetivo. O fim de uma melodia não é seu objetivo, e ainda assim se uma melodia não chegou ao seu fim, ela não chegou ao seu objetivo.

—FRIEDRICH NIETZSCHE

5.1 Conclusão

A principal contribuição desse trabalho foi a introdução de seis classificadores de padrões baseados na metodologia de regressão linear e logística para dados simbólicos do tipo intervalo. Os dados de entrada são classes de padrões descritos por vetores de características, para os quais cada valor de característica é um intervalo.

Todos os classificadores apresentados começam com a construção de funções lineares para modelar a probabilidade posterior das classes, sendo que dois deles baseiam-se na metodologia de regressão linear e os outros 4 baseiam-se numa distribuição logística, após isso, essas probabilidades são utilizadas para classificar novos padrões em uma dessas classes. Três diferentes representações da informação contida nos dados intervalares são consideradas nesse trabalho e os classificadores diferem na maneira em que utilizam essas representações. São elas: intervalos representados pelos centros, intervalos representados por um par (conjunto ou separado) de variáveis quantitativas e, cada intervalo também é representado pelos seus vértices.

Experimentos com conjuntos de dados simbólicos intervalares sintéticos, variando de casos fáceis a moderados de sobreposição de classes e uma aplicação com o conjunto de dados intervalares sobre carros foi realizada. A precisão dos resultados fornecidos pelos classificadores foi analisada pela taxa média de erro de classificação e o melhor resultado foi alcançado com os classificadores que realizam duas regressões (lineares e logísticas) separadas sobre os limites inferior e superior dos intervalos, respectivamente (APPLS e LAPPLS) e utiliza a média das probabilidades posteriores para obter uma probabilidade posterior associada.

Com relação à comparação desse método com um classificador baseado na metodologia de discriminante linear já introduzido na literatura da Análise de Dados Simbólicos, foi observado que para o conjunto de dados intervalares sobre carros, os classificadores provêem praticamente os mesmos resultados de taxa de erro.

5.2 Trabalhos futuros

A abordagem todos contra todos utilizada para decompor o problema multiclasse em K problemas de classificação binários, onde K é o número de classes do problema multiclasse inicial, apresenta um problema quando o número de classes cresce, pois a medida que isso acontece, mais classificadores binários são necessários para resolver cada parte decomposta. Essa desvantagem poderia ser resolvida utilizando um método de regressão linear multivariada, que permitiria o ajuste simultâneo para todas as classes em um único modelo de regressão.

Também é interessante pesquisar por que os classificadores baseados nos limites inferiores e superiores separadamente apresentaram melhora em suas taxas de erro à medida que os intervalos de geração dos conjuntos sintéticos cresce, pois tal procedimento dificulta a etapa de classificação, visto que aproxima os elementos de todas as classes, tornando as fronteiras entre essas classes sobrepostas.

Além disso, realizar experimentos com o modelo linear em que toda a variabilidade é considerada. Nesse modelo, os coeficientes das regressões para as classes são estimados conjuntamente.

Referências Bibliográficas

[1] BOCK, H.; DIDAY, E. ***Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data***. Berlin, Heidelberg: Springer, 2000. 1, 2.2

[2] BILLARD, L.; DIDAY, E. ***Symbolic Data Analysis: Conceptual Statistics and Data Mining***. Wiley Series In Computational Statistics. John Wiley and Sons, 2006. 1

[3] HAYKIN, S. *Neural networks: A comprehensive foundation*. 3rd. ed. Prentice Hall, 2008. 2.1.2

[4] MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, p. 115–133, 1943. 2.1.2

[5] BILLARD, L.; DIDAY, E. **From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis**. *Journal of the American Statistical Association*, v. 98, n. 462, 2003. 2.2

[6] ICHINO, M.; YAGUCHI, H.; DIDAY, E. ***A fuzzy symbolic pattern classifier***. Berlin: Springer, 1996. p. 92–102. 2.2.3

[7] DE SOUZA, R. M. C. R. ***Symbolic approach to SAR image classification***. 1999. Dissertação (Mestrado em Computação) - Federal Univesity of Pernambuco, 1999. 2.2.3

[8] D'OLIVEIRA, S.; CARVALHO, F. D.; DE SOUZA, R. **Classification of sar images through a convex hull region oriented approach**. *11th International Conference on Neural Information Processing (ICONIP-2004), Lectures Notes in Computer Science - LNCS 3316*, p. 769–774, 2004. 2.2.3

[9] CIAMPI, A.; DIDAY, E.; LEBBE, J.; PERINEL, E.; VIGNES, R. **Growing a tree classifier with imprecise data**. *Pattern Recognition Leters, 21*, p. 787–803, 2000. 2.2.3

[10] ROSSI, F.; CONAN-GUEZ, B. **Multi-layer perceptron interval data**. *Classification, Clustering and Data Analysis (IFCS2002)*, p. 427–434, 2002. 2.2.3

[11] MALI, K.; MITRA, S. **Symbolic classification, clustering and fuzzy radial basis function network**. *Fuzzy sets and systems, 152*, p. 553–564, 2005. 2.2.3

[12] APPICE, A.; D'AMATO, C.; ESPOSITO, F.; MALERBA, D. **Classification of symbolic objects: A lazy learning approach**. *Intelligent Data Analysis. Special issue on Symbolic and Spatial Data Analysis: Mining Complex Data Structures*, 2005. 2.2.3

[13] SILVA, A. P. D.; BRITO, P. **Linear Discriminant Analysis for Interval Data**. *Computational Statistics, 21*, p. 289–308, 2006. 2.2.3, 3.6.3, 4.3, 4.3

[14] HOSMER, D.; LEMESHOW, S. ***Applied Logistic Regression***. 2nd. ed. John Wiley and Sons, 2000. 3.2

[15] FAHRMEIR, L.; TUTZ, G. ***Multivariate Statistical Modelling Based on Generalized Linear Models***. 2nd. ed. Springer, 2001. 3.2

[16] DE SOUZA, R. M.; CYSNEIROS, F. J. A.; DE QUEIROZ, D. C.; DE A. FAGUNDES, R. A. A symbolic pattern classifier for interval data based on binary probit analysis. *KI 2008*, p. 340–347, 2008. 3.3

[17] DE SOUZA, R. M.; DE QUEIROZ, D. C.; DE A. CYSNEIROS, F. J.; FAGUNDES, R. A. wo pattern classifiers for interval data based on binary regression models. *ICDIM 2008*, 2008. 3.3

[18] DE SOUZA, R. M.; DE QUEIROZ, D. C.; DE A. CYSNEIROS, F. J.; DE A. FAGUNDES, R. A. A multi-class logistic regression model for interval data. *SMC 2008*, 2008. 3.3

[19] BILLARD, L.; E.DIDAY. **Regression analysis for interval-valued data**. *Data Analysis , Classification and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS00)*, p. 369–374, 2000. 3.4

[20] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. ***The Elements of Statistical Learning, Data Mining, Inference and Prediction***. Springer, 2001. 3.4

[21] DUDA, R.; HART, P.; D.G.STORK. ***Pattern Classification***. 2nd. ed. John Wiley and Sons, 2001. 3.5.2

[22] ALEXANDRE, L.; CAMPILHO, A.; KAMEL, M. **On combining classifiers using product and sum rules**. *Pattern recognition Letters, 22*, p. 1283–1289, 2001. 3.5.2, 3.6.4

[23] CHOUAKRIA, A.; CAZES, P.; DIDAY, E. ***Symbolic Principal Component Analysis***. Berlin, Heidelberg: Springer, 2000. p. 200–212. 3.6.3

[24] CARVALHO, F. D. **Fuzzy c-means Clustering Methods for Symbolic Interval Data**. *Pattern recognition Letters, 28*, p. 423–437, 2007. 4.3

[25] CARVALHO, F. D.; BRITO, P.; BOCK, H. H. **Dynamic Clustering for Interval Data Based on L2 Distance**. *Computational Statistics (Zeitschrift)*, p. 231–250, 2006. 4.3

[26] CARVALHO, F. D.; PIMENTEL, J.; BEZERRA, L.; DE SOUZA, R. **Clustering symbolic interval data based on a single adaptive Hausdorff distance**. *Proceedings of the 2007 IEEE International Conference on Systems, Man and Cybernetics (IEEE SMC 2007)*, p. 451–455, 2007. 4.3

[27] DE SOUZA, R.; CARVALHO, F. D.; PIZZATO, D. **A Partitioning Method for Mixed Feature-Type Symbolic Data using a Squared Euclidean Distance**. *29th Annual German Conference on Artificial Intelligence (KI2006)*, p. 260–273, 2006. 4.3

[28] CARVALHO, F. D.; DE SOUZA, R.; BEZERRA, L. **A dynamical clustering method for symbolic interval data based on a single adaptive Euclidean distance**. *Proceedings of the Brazilian Symposium on Artificial Neural Networks (SBRN 2006), 8*, 2006. 4.3

[29] ALLWEIN, E.; SHAPIRE, R.; SINGER, Y. **Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers**. *Journal of Machine Learning Research*, p. 113–141, 2000.

APÊNDICE A

Assinaturas

Renata Maria Cardoso Rodrigues de Souza
Orientadora

Diego Cesar Florencio de Queiroz
Aluno