Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Pós-Graduação em Ciência da Computação

Combinação de classificadores na categorização de textos

Gustavo Sandini Linden

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação

Orientador: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre, agosto de 2008

Dados Internacionais de Catalogação na Publicação (CIP)

L744c Linden, Gustavo Sandini
Combinação de classificadores na categorização de textos / Gustavo Sandini Linden. – Porto Alegre, 2008.
90 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS
Orientador: Profa. Dra. Vera Lúcia Strube de Lima

1. Informática. 2. Categorização (Lingüística). 3. Lingüística Computacional. 4. Processamento de Textos (Computação). 5. Aprendizagem de Máquina. I. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

PUCRS
Campus Central
Av. Ipiranga, 6681 – prédio 16 – CEP 90619-900
Porto Alegre – RS – Brasil
Fone: +55 (51) 3320-3544 – Fax: +55 (51) 3320-3548
Email: bceadm@pucrs.br
www.pucrs.br/biblioteca

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "***Combinação de Classificadores na Categorização de Textos***", apresentada por Gustavo Sandini Linden, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Artificial, aprovada em 06/12/2007 pela Comissão Examinadora:

Vera Lúcia Strube de Lima

Profa. Dra. Vera Lúcia Strube de Lima – Orientadora — PPGCC/PUCRS

Marcelo Blois Ribeiro

Prof. Dr. Marcelo Blois Ribeiro – PPGCC/PUCRS

Stanley Loh

Prof. Dr. Stanley Loh – UCPel

Homologada em 18/11/08, conforme Ata No. 23/08 pela Comissão Coordenadora.

Fernando Luís Dotti

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

Agradecimentos

À minha família, pelo apoio e amor dedicados em todos estes anos. Em especial, aos meus pais pelas palavras de incentivo. À professora Vera, pessoa especial que encontrei em meu caminho, sempre presente com sua ajuda e disponibilidade. Aos meus amigos e colegas de mestrado, pela amizade e momentos de confraternização.

Resumo

Este trabalho apresenta e avalia uma proposta para Categorização Hierárquica de Textos com uso combinado dos classificadores *k-Nearest Neighbors* (*k*-NN) e *Support Vector Machines* (SVM). O estudo foi embasado numa série de experimentos os quais fizeram uso da coleção Folha-RIcol de textos em língua portuguesa, que se encontram hierarquicamente organizados em categorias. Nos experimentos realizados, os classificadores *k*-NN e SVM tiveram seu desempenho analisado, primeiro individualmente, com uma variante da metodologia de avaliação *hold-out*, e após, de modo combinado. A combinação proposta, denominada *k*-NN+SVM, teve seu desempenho comparado com aquele dos classificadores individuais e com o da combinação por voto. Em síntese, a combinação *k*-NN+SVM não apresentou desempenho superior às demais alternativas, todavia o estudo permitiu a observação do comportamento dos classificadores e seu uso combinado, a identificação de problemas e possíveis soluções, bem como algumas considerações sobre a coleção de documentos utilizada.

Palavras-chave: Categorização Hierárquica de Textos, Aprendizagem de Máquina, classificadores, *k*-NN, SVM.

Abstract

This study presents and evaluates a proposal for Hierarchical Text Categorization combining k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM) classifiers. The study was based on several experiments which made use of Folha-RIcol text collection in Portuguese language. The texts in this collection are hierarchically organized in categories. In the experiments, the performance of k-NN and SVM classifiers was analyzed, individually first, with a variant of hold-out evaluation methodology, and then combined. The proposed combination, referred to as k-NN+SVM, had its performance compared to the individual classifiers and also to the combination by vote. In synthesis, the k-NN+SVM combination did not present better performance to the alternative ones, however the study allowed to observe the classifiers' behavior and its combined use, the identification of problems and possible solutions, as well as taking into consideration the document collection used.

Keywords: Hierarchical Text Categorization, Machine Learning, classifier, k-NN, SVM.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

CT	Categorização automática de Textos	17
AM	Aprendizagem de Máquina	18
RI	Recuperação de Informação	18
SVM	*Support Vector Machines*	19
***k*-NN**	*k-Nearest Neighbors*	19
CHT	Categorização Hierárquica de Textos	19
WEKA	*Waikato System for Knowledge Analysis*	32
STRIVE	*Stacked Reliability Indicator Variable Ensemble*	43
ARFF	*Atribute-Relation File Format*	48

Sumário

Capítulo 1

Introdução

1.1 Motivação

O meio digital é, atualmente, o mais difundido para o armazenamento de informações textuais. Com a adoção em larga escala do armazenamento digital, por empresas, organizações e instituições, surge um problema de organização e gerenciamento dessas informações. Entretanto, devido à quantidade excessiva de dados textuais em formato digital, não é mais possível uma categorização manual.

Uma solução para esse problema é a Categorização automática, ou semi-automática, de Textos (CT). A CT é uma área de pesquisa ligada à Recuperação de Informação (RI), que ultimamente está empregando a Aprendizagem de Máquina [1] na solução desse problema. A categorização se caracteriza por classificar os documentos conforme seu conteúdo [2], ao invés de fazê-lo por autor, título, páginas, relevância ou qualquer outro atributo desejado, como ocorre tradicionalmente em uma classificação. Em ambos os casos, seja no modo manual, seja no modo automático ou semi-automático de categorização, existe a distinção dos documentos por grupos de similaridade, denominados categorias. As duas tarefas, classificação e categorização, são muito similares, tanto que alguns autores referenciam somente a classificação, já que esta agrega, como subárea, o processo de categorização.

Ao classificarmos um documento, este passa a pertencer a um grupo contendo outros documentos que lhe são semelhantes em algum aspecto. Digamos que, em uma biblioteca, queremos classificar livros. Esta classificação pode se dar por autor, data, editor ou assunto, por exemplo. Nesse último caso, ao classificarmos um livro por seu assunto, estamos selecionando como atributo para a classificação o conteúdo do livro, caracterizando assim uma categorização. Esse exemplo serve para mostrar que os documentos são categorizados de acordo com seus assuntos, não que seja obrigatório utilizar apenas o conteúdo para a categorização.

Além do conteúdo, outras informações adicionais quanto aos documentos também podem ser utilizadas. Os melhores exemplos de atributos para a classificação, fora do corpo do documento, são as palavras-chave e títulos. Essas informações são armazenadas de forma explícita

e podem servir como importante fator para a decisão de categorizar ou não um documento em uma categoria.

Um dos aspectos pesquisados na área de CT é a busca por maior acurácia na categorização. Uma das formas de se obter essa maior acurácia é através da combinação de classificadores, ou de seus resultados. Estudos no âmbito da combinação de classificadores baseiam-se na premissa de que: múltiplos classificadores, atuando em conjunto e de maneiras qualitativamente diferente, podem apresentar um ganho com as vantagens de cada classificador. Então, compreender vantagens e desvantagens de cada classificador, ou do uso combinado de classificadores, pode permitir construir um método mais acurado de categorização.

1.2 Uma breve perspectiva sobre a evolução da categorização

A Categorização de Textos, também conhecida como Classificação de Textos, é definida por Sebastiani em [1] como a atividade de rotular textos em linguagem natural com um conjunto pré-definido de categorias. No entanto, a efetiva realização dessa atividade envolve uma evolução da CT com início na década de 60, mas que só tornou-se viável com o barateamento do *hardware* e aperfeiçoamento no *software*.

Na década de 90 a CT começou a ser amplamente explorada com uso das abordagens baseadas em Aprendizagem de Máquina (AM). Antes do uso da AM e antes da aproximação da CT com a Recuperação de Informação, na década de 80, o uso de regras lógicas era a forma de realizar a categorização. As regras eram criadas por especialistas ou engenheiros do conhecimento, pessoas que necessitavam ter um vasto conhecimento do domínio ao qual pertenciam as categorias. Cada classificador possuía regras específicas e, uma vez que o domínio ou as categorias fossem alterados, era refeito todo o trabalho. Muito pouco era reutilizável. Até mesmo mudanças nos documentos significavam alterações nas regras. Com essas desvantagens intensificaram-se as pesquisas voltadas para a automatização do processo de categorização de textos. Atualmente, como menciona Sebastiani [3], em 2006, a CT está situada entre AM e RI.

A automatização do processo de CT utilizando AM foi impulsionada por dois fatores importantes. O primeiro, como ocorre na maioria das automatizações, é o fator tempo. Categorizar manualmente um documento é uma tarefa lenta e onerosa. Para realizar essa tarefa são necessárias a leitura e compreensão de documentos, e isso exige pessoal especializado, o que leva a caracterizar o segundo fator: o custo. Em alguns casos a classificação ocorre em mais de um domínio, ou seja, os documentos podem pertencer a mais de uma área de interesse, implicando a necessidade de um especialista para cada domínio, o que leva ao aumento de pessoal e o conseqüente aumento do custo. Esses motivos ilustram a substituição do simples uso de regras para a AM. A capacidade de aprendizado é a grande vantagem do uso de AM. O aprendizado pode ser refeito a cada alteração de domínio, permitindo a reutilização de trabalhos anteriores. O diferencial desse paradigma é a capacidade de desenvolver classificadores automaticamente por um processo indutivo, onde o aprendizado ocorre em um conjunto de documentos previamente

categorizados sob cada categoria por um especialista no domínio.

Em meados da década de 90 surge uma nova subárea de pesquisa, conhecida como Categorização Hierárquica de Textos (CHT). As categorias passam a ser melhor entendidas como estruturas hierárquicas. Ao processo de CHT é atribuída a tarefa de categorizar documentos em uma ou mais categorias de assuntos, organizadas em uma estrutura hierárquica. Para tanto, igualmente, é utilizada a Aprendizagem de Máquina.

1.3 Objetivo

No contexto deste trabalho, a AM é a alternativa fundamental para a organização e gerenciamento de informações. A grande dificuldade está em determinar qual a melhor forma de utilizar a AM para evitar que intervenções manuais se façam necessárias. Para atender a esse propósito, a AM inclui diversos tipos de algoritmos que atuam como instrumentos para o processo de CT, denominados classificadores.

O objetivo do presente trabalho é analisar o funcionamento dos classificadores denominados *Support Vector Machine* (SVM) e *k-Nearest Neighbors* (*k*-NN) em categorias distribuídas de forma hierárquica. O estudo dos classificadores SVM e *k*-NN não é uma novidade na área de CT, bem como não o é a utilização de CHT [4] [5] [6]. A principal contribuição deste trabalho compreende a análise da aplicação desses algoritmos ao longo das etapas do processo de CHT. Ainda, inclui a discussão quanto à proposta e à análise da integração desses dois classificadores, avaliando o desempenho e eficiência desse processo em *corpora* de textos. Como não foram encontrados na literatura trabalhos de mesma natureza que o aqui proposto, este constitui um estudo exploratório do processo de CHT para a língua portuguesa na combinação de classificadores. Nesse sentido, a combinação por voto é empregada no intuito de melhor avaliar a proposta de integração dos classificadores.

A escolha por um método para a CT depende basicamente da aplicação e da coleção utilizada. Em algumas situações, o desempenho do método pode ser decisivo, enquanto em outras, a eficiência é preferível. Não é uma tarefa simples determinar o melhor classificador para uma aplicação. Um *corpus* estático, ou seja, que não possui alterações em suas categorias, possui um desempenho diferente de um *corpus* dinâmico, onde alterações das categorias e seus domínios são freqüentes. Nesse último caso, é preciso considerar o tempo de pré-processamento, treinamento e testes. Além disso, a estrutura das categorias também pode influenciar na escolha do classificador.

Nesse estudo, os textos são provenientes da coleção Folha-RIcol[1], que contém artigos do ano de 1994 do jornal Folha de São Paulo, manualmente categorizados. A categorização desses textos é feita segundo categorias dispostas em uma hierarquia com dois níveis.

[1]Disponível em: http://www.linguateca.pt/Repositorio/Folha-RIcol/.

1.4 Organização do texto desta dissertação

O restante desta dissertação está dividido em sete capítulos. Os capítulos iniciais tratam da revisão bibliográfica pertinente aos assuntos abordados. Os capítulos intermediários apresentam uma descrição dos experimentos envolvendo o processo de Categorização hierárquica de Textos, realizados durante o desenrolar da dissertação de mestrado.

Sucintamente, no segundo capítulo tem-se o embasamento necessário para a compreensão da Categorização Hierárquica de Textos; no Capítulo 3 são abordados trabalhos relacionados que contribuem para o desenvolvimento desta dissertação; um quarto capítulo apresenta a metodologia dos experimentos, que são descritos em detalhes nos capítulo 5 e 6; o capítulo final contém uma breve conclusão referente aos problemas e peculiaridades encontrados no decorrer deste trabalho.

Capítulo 2

Categorização de Textos

Este capítulo inicia por uma breve revisão teórica da Categorização de Textos para aproximar o leitor das idéias centrais e da atualidade na área. Os conceitos estudados e apresentados no decorrer do capítulo servem como base para compreender o desenvolvimento do trabalho ao longo desta dissertação.

A idéia fundamental da Categorização de Textos é a atribuição de um valor booleano para cada par $(d_i, c_j) \rightarrow D \times C$, onde D é uma coleção de documentos e $C = \{c_j, ..., c_{|C|}\}$ um conjunto de categorias [1]. Partindo desse conceito inicial, a CT é formalmente descrita como a decisão de classificar um documento d_i em uma categoria c_j. No caso de uma decisão negativa, onde $d_i \notin c_j$, o valor-verdade F é associado ao par (d_i, c_j), caso contrário aplica-se o valor-verdade V.

Para que o processo de decisão seja automatizado, é necessário que uma inferência seja realizada. O algoritmo que realiza a conversão das entradas no valor de decisão é comumente denominado classificador. Ele é descrito formalmente por Sebastiani em [1] como uma função-alvo desconhecida:

$$\breve{\Phi} : D \times C \rightarrow \{V, F\}$$

Na realidade, a função $\breve{\Phi}$ é uma aproximação da função Φ. Essa função descreve o modo como os documentos da coleção D devem ser classificados em C. Os dados de entrada dessa função são a coleção de documentos e o conjunto de categorias. A saída é um conjunto de valores {V, F}. O mecanismo de inferência utilizado pela função é comparável a uma "caixa preta". Muitas vezes, não é possível determinar com exatidão como está ocorrendo o processo de classificação. Isso se deve ao fato de a função assumir um comportamento diferente para cada um dos valores de entrada. O resultado final, ou saída, está relacionado ao conjunto de atributos selecionados para a classificação.

O objetivo da Aprendizagem de Máquina é treinar classificadores com exemplos de textos previamente classificados. Com base nesses exemplos, os classificadores são capazes de determinar se um novo documento d_i pertence a uma categoria c_j.

2.1 Tipos de classificadores

Um documento pode pertencer a uma ou várias categorias ou, ainda, a nenhuma das categorias de um conjunto pré-estabelecido. Essa característica diferencia os classificadores em dois tipos: monocategoriais e multicategoriais [1].

A categorização é definida em *n* categorias pertencentes a um conjunto *C* de categorias. Um classificador é monocategorial quando todos os documentos da coleção *D* pertencem a *n* categorias, sendo $0 \leq n \leq 1$; já no caso de $0 \leq n \geq 1$ o classificador é multicategorial.

É possível usar um classificador monocategorial para aplicações multicategoriais, mas o contrário não é válido. Segundo Sebastiani [1], o classificador monocategorial capaz de realizar multicategorização é um classificador *binário*. Ele permite a atribuição de mais de uma categoria a um documento, se tais categorias forem consideradas como estocasticamente independentes umas das outras. Isso significa que o valor de $\breve{\Phi}(d_i, c_j)$ não depende do valor de $\breve{\Phi}(d_i, c'_j)$, considerando que c_j e c'_j são duas categorias distintas. No caso em que c'_j é um subconjunto de c_j, ou seja, um documento pertencente a c'_j também pertence a c_j, as categorias do conjunto $\{c_j, c'_j\}$ não são estocasticamente independentes.

O classificador *binário* determina que um documento d_i deve pertencer à categoria c_j ou então seu complemento, ou seja, pertencer à categoria complementar c'_j. Se o valor da função de uma categoria for independente do valor da função de seu complemento e vice-versa, elas são consideradas independentes; assim o classificador binário pode ser aplicado a cada categoria c independentemente. Isso resulta em um classificador binário para cada categoria, como mostra a função seguinte:

$$\breve{\Phi}_i : D \rightarrow \{V, F\} \tag{2.1}$$

Por essa razão, temos i classificadores $\breve{\Phi}$ atribuindo, cada um deles, um valor V ou F para os documentos da coleção D. Isso implica, para cada par (d_i, c_j), um único valor. A separação das categorias em diferentes classificadores garante que, mesmo uma categoria sendo dependente de outra, os valores de seus documentos podem ser diferentes. Nesse sentido, os classificadores podem julgar um documento d_i como pertencente a uma categoria c_1 mas não à c_2, mesmo sendo c_1 dependente de c_2.

Em aplicações reais é comum encontrar coleções de textos que possuem uma certa organização hierárquica de assuntos, onde existe uma relação de dependência entre assuntos genéricos e específicos. Essa organização hierárquica, se for utilizada no processo de categorização, pode transformar um problema multicategorial em monocategorial. Esse é o assunto tratado na próxima seção.

2.2 Categorização Hierárquica de Textos

Não resta dúvida de que a distribuição de uma coleção de documentos em categorias auxilia na organização, busca e recuperação de informações mas, à medida que o número de documentos e o número de categorias aumentam, essa tarefa torna-se mais complexa para o ser humano. Algumas vezes podemos errar em uma busca, pela impossibilidade de determinar corretamente uma categoria.

Conforme Sun e seus co-autores [7] explicam, uma forma de ajudar a organizar informações no auxílio à compreensão humana é utilizar a organização hierárquica. O uso de categorias hierárquicas estabelece relações entre as categorias e, normalmente, a representação ocorre na forma de uma árvore, onde os nodos filhos são categorias mais específicas de seus pais. Assim, ao percorrer a árvore de hierarquias, é possível encontrar mais facilmente a informação desejada.

Existem três motivações para o uso de hierarquias, enumeradas por Alessio, Murray e Schiaffino em [6]:

- A primeira é que, à medida que o número de categorias cresce, os valores de precisão (*precision*) e abrangência (*recall*) decrescem. Uma das razões para isso é a diversidade de tópicos abordados na categorização. Em muitas situações, áreas de interesse distintas compartilham uma mesma palavra, porém com significados diferentes.

- A segunda motivação leva em consideração a complexidade do processo. É mais difícil resolver um problema grande do que problemas menores, e é exatamente isso que os classificadores hierárquicos fazem. Eles restringem a classificação a um pequeno número de categorias e a um domínio específico.

- A terceira vantagem citada é a possibilidade de resolver a categorização multicategorial com o uso de classificadores monocategoriais. Conceitualmente, à medida que descemos nos níveis da hierarquia, temos a categorização multicategorial apenas quando as categorias pertencerem à mesma ramificação. Isso significa que só irá ocorrer categorização multicategorial quando os nodos forem irmãos. É claro que, ao subirmos em direção ao topo da hierarquia, teremos uma visão multicategorial dos nodos abaixo.

A representação hierárquica de categorias permite utilizar um único classificador, como ocorre na representação plana, onde um único classificador determina a qual ou quais categorias um documento pertence. Também é possível utilizar n classificadores locais, onde cada classificador é responsável por categorizar um subconjunto de categorias na hierarquia.

2.3 Modelo de representação dos documentos

O modelo de representação dos documentos determina a preparação dos documentos para a aplicação dos classificadores. Nesta seção, o modelo de espaço vetorial será apresentado, no

intuito de aprofundar o conhecimento teórico do leitor, proporcionando uma compreensão dos conceitos que se sucedem.

Um modelo de representação consiste em determinar, a partir do conteúdo dos documentos, uma série de atributos para a categorização. Usualmente, os documentos são processados levando em consideração seus termos, ou palavras, individualmente. Nesse caso, os atributos são todas as palavras únicas pertencentes à coleção de documentos. Os atributos também podem ser definidos como parágrafos, sentenças ou um conjunto de palavras em seqüência como, por exemplo, expressões. Apesar dos esforços em determinar novos modelos de representação como, por exemplo, *n*-gramas [1] [2] e *bag-of-concepts* [2] [8], o modelo *bag-of-words* se destaca por sua simplicidade e eficiência. Esse modelo passou a ser amplamente utilizado nas áreas de Recuperação de Informação e Categorização de Textos após o uso por Salton e Buckley na década de 80 [9]. Atualmente, existe uma série de variantes desse modelo que, nesta dissertação, são denominadas genericamente de *bag-of-words*.

Um modelo *bag-of-words* assume que um documento d_i é representado por um vetor de pesos para os seus termos, ou seja, $d_i = \{w_{1i}, ..., w_{xi}\}$, onde x é igual à cardinalidade do vocabulário da coleção D usada pelo classificador.

Os pesos podem ser definidos de forma *booleana*, onde $w_{ki} = 1$ representa a ocorrência e $w_{ki} = 0$ representa a ausência do termo w_k em um documento d_i. Porém, é mais comum utilizar como valor de peso um número real no intervalo [0, 1].

O conceito básico das medidas de peso é fundamentado na Lei empírica de Zipf e no Modelo de Luhn [10]. A Lei de Zipf, no contexto de categorização de textos, determina que, ao se ordenar a freqüência dos termos de uma coleção de documentos em ordem decrescente e aplicar logaritmo, é obtida uma curva decrescente que inicia com uma grande quantidade de termos comuns e termina com um pequeno número de termos raros. Essa curva pode ser vista no gráfico da Figura 2.1, extraído de [10]. A freqüência f dos termos é expressa no eixo y, enquanto os termos são expressos no eixo x.

O modelo de Luhn propõe cortes inferiores e superiores na curva gerada pela Lei de Zipf (gráfico da Figura 2.2, extraído de [10]). O corte superior está próximo ao termo mais comum, enquanto o inferior está próximo ao termo mais raro. Esses cortes pressupõem que os termos acima e abaixo, respectivamente, dos cortes superior e inferior, não possuem tanta importância quanto os termos intermediários, para a representação dos documentos.

Uma característica que surge com o uso do *bag-of-words* é a grande quantidade de atributos. Como cada palavra é um atributo e um documento possui uma certa quantidade de palavras, o vocabulário é superior ao número de palavras de um documento. Muitas dessas palavras ocorrem em um pequeno número de documentos ou então ocorrem com pequenas flexões em suas formas. Isso gera representações, ou vetores de representação, com uma grande quantidade

[1] Modelo no qual os atributos são compostos por palavras que ocorrem em uma seqüência.

[2] Modelo no qual os atributos são compostos por expressões regulares, ou um conjunto de palavras que agregam um único significado.

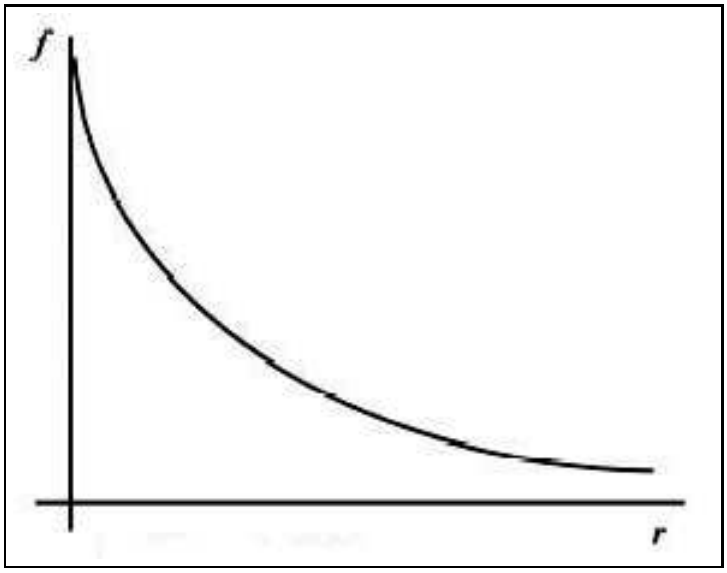

Figura 2.1: Lei de Zipf

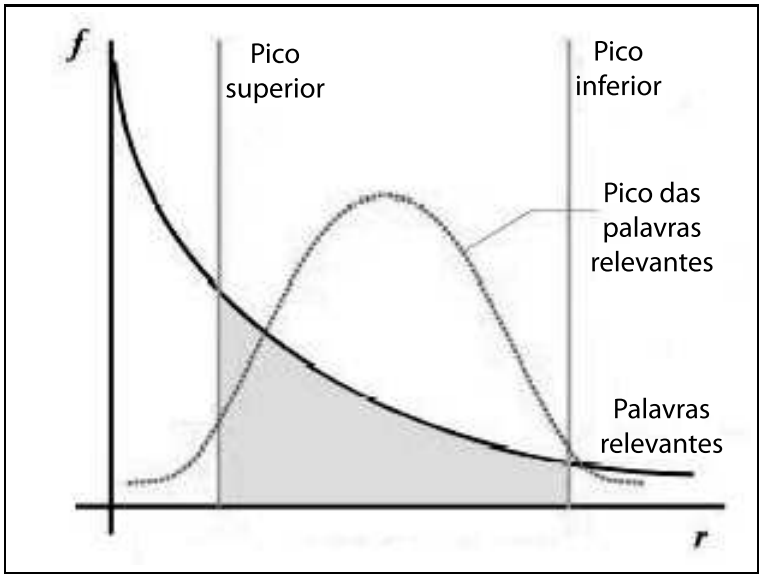

Figura 2.2: Cortes de Luhn

de elementos nulos ($w_{ki} = 0$), denominados vetores esparsos.

Outra característica inerente a essa representação são as palavras similares, seja no sentido semântico, pragmático ou lexical. Sem um devido tratamento, duas palavras similares são consideradas dois atributos independentes, o que pode comprometer, tanto negativa como positivamente, o processo de CT.

Levando em consideração essas duas características, existe a necessidade de expressar os documentos com um número de atributos inferior ao vocabulário, seja pela eliminação, seja pela seleção ou processamento desses atributos. Esse é o tema abordado na subseção que segue.

2.3.1 Seleção de atributos

Os motivos mais comumente citados para justificar a realização de uma seleção de atributos são tempo de processamento, armazenamento e eficiência. Conforme explica Alpaydin em [11], grande parte dos classificadores apresentam a sua complexidade dependente do número de atributos. Quanto maior o número de atributos maior a complexidade. Ainda, determinar atributos irrelevantes evita a extração, o armazenamento e o processamento desnecessários. Classificadores com poucos atributos podem ser mais eficientes em coleções pequenas, uma vez que possuem menos variáveis [11]. Para eliminar o tempo de processamento e diminuir a quantidade de espaço necessário para o armazenamento dos dados, é necessária uma redução na quantidade de atributos, enquanto que, para melhorar a eficiência do processo, a seleção dos melhores atributos é um fator decisivo. Evidentemente, a redução dos atributos também irá melhorar a eficiência, uma vez que os atributos eliminados devem ser os menos significativos possíveis.

Existem dois paradigmas para selecionar os atributos [12]. O primeiro, denominado extração de atributos [11], determina que o processo inicia sem nenhum atributo, sendo os atributos incluídos à medida que os melhores são encontrados. O outro paradigma é o da exclusão, denominado seleção de atributos [11]. Este paradigma determina que, do conjunto de todos os atributos, os irrelevantes devem ser retirados.

Para que a quantidade de termos não seja um fator de depreciação do desempenho, é utilizada uma lista de termos que não contribuem de forma efetiva na categorização. Esses termos são denominados *stoplist*. Termos de um documento que pertençam à lista de *stoplist* são ignorados, seja por serem muito comuns, seja por serem raros. Termos comuns são eliminados por não caracterizarem um bom atributo para categorização. Termos muito raros também não são bons atributos, porque aparecem em uma quantidade limitada de documentos, dificultando a tarefa de aprendizagem.

No processo de seleção de atributos pode-se fazer também o *stemming* das palavras, que é o processo de retirada de sufixos, com a manutenção do radical. Assim, palavras diferentes que possuem um mesmo radical são consideradas como um único atributo. Nesse processo o número de atributos sofre uma redução, mas o peso individual de cada um é aumentado, a cada

ocorrência do radical, independentemente da palavra propriamente dita.

O *stemming* e a utilização da *stoplist* ocorrem em uma etapa anterior à classificação. Muitos estudos já foram realizados tendo comprovado a eficácia dessas técnicas. Tanto o *stemming* quanto a eliminação da *stoplist* trabalham utilizando o conhecimento lexical associado aos termos dos documentos.

A semântica e a pragmática ainda são pouco utilizadas nessa abordagem.

O trabalho de Hidalgo e co-autores [13] utiliza o léxico dos termos para procurar sinônimos na *WordNet*[3]. Essa tentativa é válida, podendo ser, ainda, utilizada para antônimos, ou outros. Mesmo assim, outros tipos de relacionamentos semânticos ainda não têm sido levados em consideração.

2.4 Aprendizagem de Máquina

O presente capítulo, ao enforcar a Categorização de Textos, aborda o processo de Categorização Automática de Textos voltado para a Aprendizagem de Máquina. O processo completo de CT possui quatro etapas. A primeira etapa, o pré-processamento, já foi apresentada na seção 2.3.1 quando foi detalhada a seleção de atributos. A razão para maiores detalhes sobre o pré-processamento não estarem incluídos neste capítulo é que ele não se enquadra exatamente no processo de Aprendizagem de Máquina, mesmo sendo uma etapa indispensável para a eficiência do processo.

O fluxo do processo genérico de Aprendizagem de Máquina pode ser visto na Figura 2.3. As etapas de treinamento, teste e operação serão descritas nas seções que seguem.

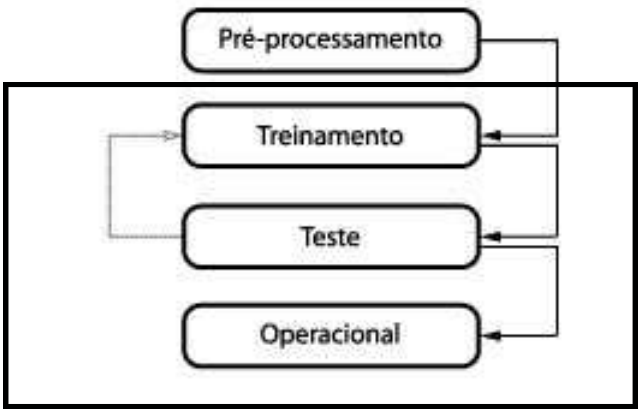

Figura 2.3: Etapas da Aprendizagem de Máquina

[3]Base de dados lexical, que serve como sistema de referência *online* de sinônimos entre palavras da língua inglesa. Disponível em: http://wordnet.princeton.edu/.

2.4.1 Etapas de treinamento e teste

Os componentes básicos da Aprendizagem de Máquina são os dados de entrada, as categorias, o classificador e o resultado da classificação.

Os dados de entrada, previamente categorizados, consistem na entrada para a etapa de treinamento. Isto é, essa etapa utiliza documentos que já foram categorizados manualmente para treinar o classificador. Nessa situação, o aprendizado é "supervisionado", ou seja, o classificador aprende sob supervisão de um agente externo. Nesse caso, o agente externo é a coleção de documentos previamente categorizados.

Fazendo uma analogia, o aprendizado supervisionado é semelhante a um aluno cursando uma disciplina. O aprendizado desse aluno ocorre com a orientação de um professor. O aluno está para o classificador assim como o professor está para a coleção de documentos pré-categorizada.

Essa etapa de treinamento é denominada indutiva [14]. Nela, o aprendizado ocorre através da indução, por enumeração, dos objetos de uma classe. Os objetos são os documentos pertencentes à coleção *D* de treino. Definindo, formalmente, tem-se que o valor final {V, F} da função $\breve{\Phi} : D \times C \rightarrow \{V, F\}$ é conhecido, para todos os pares (d_i, c_j) [1].

Os documentos de entrada são, usualmente, divididos em uma coleção de treino e uma coleção de teste. Nesse sentido, o classificador é treinado com uma coleção (construção indutiva do aprendizado) e validado com outra. A validação é o passo de dedução [14] no qual o aprendizado, realizado no treino, é aplicado à coleção de documentos da etapa de teste.

Em muitos casos, depois de validar o classificador, as duas coleções (treino e teste) são reunidas novamente para realizar o treinamento final (fase operacional). Dessa forma, a validação empírica, inicial, provê uma estimativa pessimista da performance real do classificador [1].

Por utilizar documentos pré-categorizados, pode ocorrer que o classificador apresente *overfitting*, o que significa um treinamento excessivo em uma coleção que não representa todo o domínio. Quando isso ocorre, o classificador torna-se especialista em relação aos dados de treino, classificando corretamente apenas os documentos que pertencem a coleção de treino. No momento em que o classificador é utilizado na etapa de teste, as categorizações, para novos documentos, não serão corretas. Quando um classificador é capaz de categorizar corretamente a quase totalidade dos documentos da coleção de treino, mas não é capaz de identificar corretamente a categoria de novos documentos, diz-se que ocorreu *overfitting* [1]. Caso essa situação ocorra, o processo deve retornar à etapa de treinamento para que um novo aprendizado se estabeleça.

O *underfitting* [14], por sua vez, constitui um aprendizado incompleto por parte do classificador. O aprendizado incompleto resulta em categorizações errôneas de documentos. Esse erro pode ocorrer simplesmente por falta de treinamento, ou porque a coleção de treino não compreende a totalidade do domínio ao qual os documentos pertencem, ou porque o aprendizado não foi completo. Nesse sentido, o processo de treinamento deve ser refeito, ou prolongado,

incluindo novos documentos na coleção de treino.

Tanto em caso de *overfitting* quanto em caso de *underfitting*, novos documentos não serão corretamente categorizados. Seja qual for o motivo do *overfitting* ou *underfitting*, o classificador deve retornar à etapa de treinamento caso uma dessas situações ocorra.

O ideal é um classificador generalista, que seja capaz tanto de identificar corretamente as categorias dos documentos da coleção de treino e teste, quanto de classificar adequadamente novos documentos.

2.5 Classificadores

Nesta seção destacam-se dois modelos de classificadores estatísticos que utilizam a AM para a execução do processo de CT: o *k*-Nearest Neighbors e o *Support Vector Machine*.

As duas subseções seguintes descrevem o funcionamento dos classificadores *k*-NN e SVM, respectivamente. Tanto o classificador *k*-NN como o classificador SVM possuem inúmeros mecanismos internos de inferência (algoritmo). No entanto, as discussões seguintes são relativas aos algoritmos utilizados na implementação desta dissertação.

2.5.1 *k-Nearest Neighbors*

A estratégia desse tipo de classificador é armazenar a coleção de treino em um espaço euclidiano, onde cada documento ocupa um determinado ponto no espaço, de acordo com seus atributos. Nesse sentido, os documentos devem ser representados por uma matriz de documentos versus termos. Os termos, denominados atributos para o propósito da categorização, possuem um peso associado. Os pesos podem ser obtidos através das medidas [9] [1] de freqüência dos termos no documento (TF), inverso da freqüência dos documentos (IDF) e produto da freqüência dos termos pelo inverso da freqüência dos documentos (TFIDF).

No classificador *k*-NN temos os documentos da coleção de treino inseridos em um espaço euclidiano, sendo que novos documentos também pertencerão a esse mesmo espaço. Uma vez que um novo documento é inserido nesse espaço, podemos compará-lo com os *k* documentos mais próximos a ele (vizinhos). A medida de comparação é a distância entre d_i e d_k, onde d_k identifica um número pré-determinado de documentos pertencente ao espaço euclidiano (*corpus* de treino) e d_j simboliza um novo documento a ser categorizado. Essa distância pode ser medida pelo cosseno [9] ou outra medida de proximidade qualquer.

A Figura 2.4 demonstra um exemplo ilustrativo do funcionamento do classificador *k*-NN, onde um novo documento x é comparado com os *k* documentos mais próximos (três nesse exemplo). Como se trata de um exemplo ilustrativo, o espaço euclidiano é descrito por dois vértices, eixos x e y. Em uma situação real, o número de vértices é igual ao número de atributos.

Por último, esse classificador deve fazer uma escolha: dentre as *n* categorias pertencentes aos *k* documentos mais próximos do novo documento, deve escolher qual será a categoria

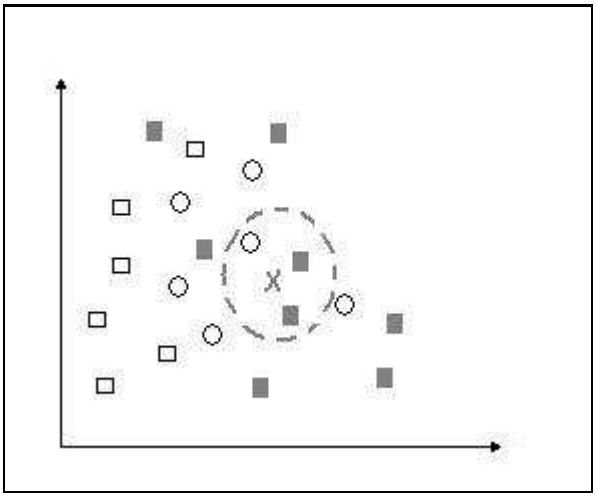

Figura 2.4: Exemplo de um classificador *k*-NN

atribuída a essa nova instância. A escolha mais simples é a escolha da maioria, onde a categoria com maior número de representantes em *k* é a escolhida. Essa escolha e o valor de *k* são os parâmetros mais discutíveis desse tipo de classificador. Formalmente, a categorização do documento d_x pelo classificador *k*-NN é dada por:

$$\sum_{d_k \in Tr_k(d_x)} RSV^4(d_x, d_k) \cdot [\breve{\Phi}(d_k, c_i)]$$

A equação acima determina que o documento d_x deve ser categorizado de acordo com a soma dos *k* documentos mais próximos a ele que maximizam a função $RSV(d_i, d_k)$ [1]. Essa soma retorna o valor final utilizado para a decisão de classificar um documento em c_i. Para isso, $RSV(d_i, d_k)$ deve representar uma métrica entre o novo documento e os *k* documentos mais próximos a ele.

A métrica que representa a função RSV, para os experimentos realizados e descritos nos capítulos 5 e 6 desta dissertação, é a medida de comparação através do uso de cosseno.

Essa escolha toma por base um valor de relevância para os pares (d_i, c_j) tal que possa ser transformado no resultado final (valor binário {V,F}). Essa transformação é obtida através do cálculo de valores de limiar. Esses cálculos realizam "cortes" no conjunto $C_j \in d_k$. Para isso, duas estratégias são utilizadas [15]: *limiar por rank* e *limiar por relevância*.

Basear o limiar em *rank* significa determinar que um número fixo de categorias do topo de uma lista serão atribuídas a d_i. Esse cálculo não garante que todas as categorias sejam corretamente atribuídas. As categorias de um documento podem não ser atribuídas a ele quando estão abaixo do valor de limiar. Ou então, categorias podem ser erroneamente atribuídas, quando o valor de limiar ultrapassar o número de categorias pertencentes ao documento.

Diferente do limiar por *rank*, o limiar baseado em relevância não atribui um número fixo de categorias aos documentos. Nessa estratégia, à atribuição de categorias é aplicada a relevância

[4] A sigla RSV, significaria *Relatedness Status Value*, conforme se entende na leitura de Sebastiani em [1].

estabelecida em cada par (d_i, cj). Esse cálculo estabelece, como o próprio nome o determina, um limiar de relevância, onde as categorias são atribuídas somente quando um determinado valor de relevância mínimo é alcançado.

Uma peculiaridade do *k*-NN é a sua fase de treino diferenciada. O treinamento desse classificador pode ser considerado o tempo de indexação dos documentos que serão comparados [15]. A maior parte do processamento desse classificador é realizada no momento da categorização dos documentos. Considerando essa informação, a complexidade do treinamento passa a ser linear e equivalente ao número de documentos a serem indexados. Por exemplo, em um *corpus* de treino com *n* documentos, o documento d_x é comparado com todos os *n* documentos, resultando em uma complexidade da ordem de O(n).

2.5.2 *Support Vector Machines*

Assim como no *k*-NN, aqui temos um espaço euclidiano onde se encontram os documentos. Porém, ao invés de realizar a busca por documentos semelhantes, esse classificador realiza uma separação de hiperplanos no espaço euclidiano. Um hiperplano pode ser entendido como uma linha que divide o espaço euclidiano em duas regiões, uma contendo os documentos de treino com valores V, e outra contendo os documentos de treino com valores F (exemplos positivos e negativos {-1, +1} de uma categoria, respectivamente). No caso de todos os exemplos possuírem valor-verdade F (-1) em uma região e valor-verdade V (+1) na outra, temos um problema linearmente separável.

Dado um espaço euclidiano contendo os documentos da coleção de treino é possível encontrar infinitos hiperplanos que satisfaçam à divisão das regiões em dois grupos distintos [14]. O melhor hiperplano para a separação é obtido através da seguinte equação:

$$\vec{w} \cdot \vec{x} - b = 0$$

Onde $\vec{w}$ é o vetor de peso perpendicular ao hiperplano, $\vec{x}$ são os atributos de um documento e b é um fator compensador que permite aumentar a margem da separação de hiperplanos. Tanto b como $\vec{w}$ são parâmetros ajustados durante o treinamento [14].

Os documentos mais próximos ao hiperplano são denominados *support vectors* ou vetores de suporte. A idéia por trás do classificador é encontrar o hiperplano que possui a maior distância entre os vetores de suporte das duas regiões. Esse tipo de classificador também é conhecido como classificador da margem máxima, porque a soma das duas distâncias entre os vetores de suporte estabelece a margem do classificador.

A separação dos documentos da coleção de treino, no espaço euclidiano, permite determinar à qual região um novo documento pertence. Ao determinar a região a que um documento pertence, podemos associar ao documento a mesma categoria dos outros documentos da região, contanto que seja satisfeita uma das duas condições [14]:

$$\begin{cases} \vec{w} \cdot \vec{x_i} - b \geq 1, se\ c_j = 1, \\ \vec{w} \cdot \vec{x_i} - b \leq 1, se\ c_j = -1. \end{cases}$$

Essas condições estabelecem que todos os documentos positivos estejam localizados em uma região do hiperplano ($\vec{w} \cdot \vec{x} - b = 1$), assim como os documentos negativos devem estar localizados em outra região do hiperplano ($\vec{w} \cdot \vec{x} - b = -1$). Como é mostrado na Figura 2.5, os documentos positivos (que pertencem à categoria) estão localizados abaixo da margem inferior da figura e os documentos negativos (que não pertencem à categoria) estão localizados acima da margem superior.

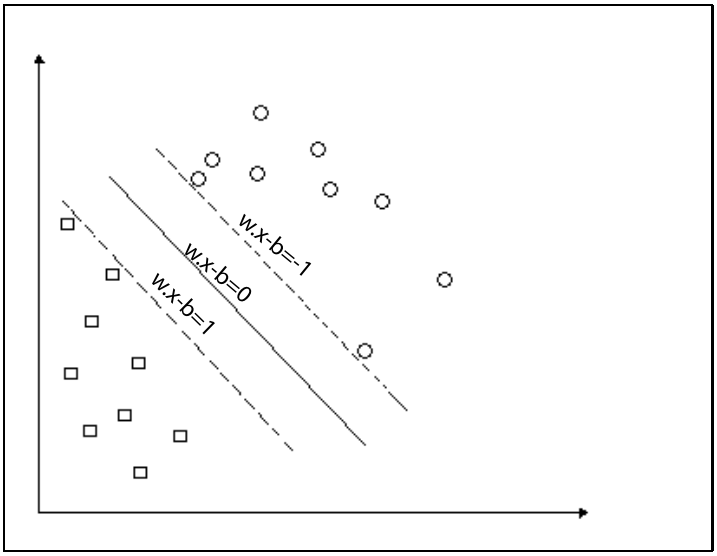

Figura 2.5: Exemplo de um classificador SVM

Formalmente, um classificador SVM possui duas regiões. Isso o torna, conceitualmente, um classificador do tipo monocategorial ou, então, um classificador multicategorial com, no máximo, duas categorias possíveis.

O SVM é um classificador monocategorial quando os documentos em uma de suas duas regiões pertencem a uma categoria e na outra região os documentos não pertencem a essa mesma categoria (podendo pertencer a qualquer outra categoria). Ele é um classificador multicategorial quando os documentos inseridos na segunda região delimitarem, exclusivamente, uma segunda categoria.

O classificador SVM utilizado para este estudo é uma implementação do algoritmo *Sequential Minimal Optimization* (SMO) [16]. A implementação está disponível na ferramenta WEKA, que possui dois tipos de kernel [17] implementados, o kernel polinomial e o kernel gaussiano de base radial (RBF).

2.6 Combinação de classificadores

A combinação de classificadores sugerida aqui como uma alternativa para melhorar o processo de CHT é baseada na seguinte premissa: dada uma tarefa que exige o conhecimento de especialista, uma quantidade *n* de especialistas irá apresentar um resultado de melhor qualidade do que apenas um especialista apresentaria [1]. Na Categorização Automática de Textos, esse esforço consiste em executar um conjunto de *k* classificadores $\{\Phi_1, ..., \Phi_k\}$ para a mesma tarefa de categorização e combiná-los de forma apropriada. Essa combinação de classificadores é caracterizada por uma escolha de *k* classificadores e de uma função combinatória.

A performance de um classificador pode ser ajustada para obter a melhor acurácia possível para uma determinada situação, mas os ajustes são uma tarefa complexa e, ainda, podem existir padrões em que mesmo o melhor classificador apresente falhas na categorização. Segundo Alpaydin em [11], para obter resultados produtivos na combinação de classificadores, as decisões de categorização tomadas pelos classificadores não podem ser as mesmas. Decisões iguais levam a resultados iguais; se for tomada a decisão errada, todos os classificadores erram. Decisões diferentes permitem a um classificador errar enquanto outros classificadores acertam, resultando em uma decisão final correta.

A principal discussão desse método é em relação à função combinatória. Dentre inúmeras funções combinatórias disponíveis, a votação, a seleção dinâmica de classificadores e a combinação adaptativa de classificadores são três funções combinatórias comumente utilizadas [1]. Segue uma breve exposição sobre cada uma, a partir dos pressupostos apresentados por Sebastiani em [1] e Bennett e seus co-autores [18]:

- A função combinatória mais simples é a escolha por voto majoritário. Nela a decisão final é obtida escolhendo-se a categoria com maior número de votos dentre os *k* classificadores. Outra forma de aplicar a votação é atribuir pesos aos classificadores (votação com pesos); nesse caso o voto de cada classificador é relativo ao seu peso, para a decisão final.

- Um exemplo de seleção dinâmica de classificadores é a utilização do classificador mais eficiente para uma determinada categoria, ou categorias. Esta função também pode ser denominada *Best By Class*. Para determinar qual é o classificador mais eficiente, é necessário avaliar todos os classificadores com uma medida de avaliação, em um *corpus* de avaliação.

- A combinação adaptativa de classificadores é descrita como uma função intermediária entre a votação com pesos e a seleção dinâmica de classificadores. Essa função agrega todos os votos dos classificadores, atribuindo pesos para a contribuição da decisão final conforme a eficiência de cada classificador em um *corpus* de avaliação.

Com base nessas funções combinatórias iniciais, diferentes autores descrevem suas próprias propostas de métodos combinatórios que, de uma forma ou outra, fazem uso dos conceitos

expostos por Sebastiani em [1]. A seguir são apresentados alguns exemplos de combinação de classificadores.

O uso de *bagging* em [11] descreve um método de votação onde os classificadores são treinados com pequenas alterações no *corpus*. Nesse caso, os classificadores devem ser treinados usando uma amostra aleatória do *corpus*, diferente a cada iteração. Para garantir tamanhos iguais de amostra para todos os classificadores, os documentos usados em uma iteração são repostos nas próximas iterações. Existe a possibilidade de os documentos serem usados mais de uma vez ou, até mesmo, nenhuma vez. O fator aleatório permite amostras, ao mesmo tempo, semelhantes e diferentes. Semelhantes porque podem compartilhar os mesmos documentos, e diferentes porque documentos diferentes são incorporados à amostra. A principal desvantagem desse método é o fato de que o desempenho dos classificadores é baseado na probabilidade da escolha das amostras.

O método de *boosting* [1] apresenta um conceito intuitivo de aprimoramento, que se encaixa na definição de seleção dinâmica de classificadores. A idéia do *boosting* é aplicar um conjunto de *n* classificadores iterativamente sobre um *corpus* de treino. A cada iteração um novo classificador prioriza a categorização nos documentos onde o classificador anterior obteve a maior taxa de categorizações incorretas. Assim, o treinamento de novos classificadores não é baseado em probabilidade, como ocorre no método *bagging*. A desvantagem desse método é a exigência de um *corpus* suficientemente grande para o treinamento dos classificadores.

O algoritmo AdaBoost [19], freqüentemente citado na literatura, utiliza uma medida de peso para referenciar os documentos incorretamente categorizados, que recebem pesos maiores, enquanto documentos corretamente classificados recebem pesos menores. Utilizando a reposição de documentos, o algoritmo permite a escolha por documentos com pesos menores em *corpora* pequenos.

Alpaydin em [11] exemplifica uma forma de combinação em cascata entre os classificadores. A idéia é ter *k* classificadores utilizados em seqüência, de acordo com sua complexidade ou custo de representação. Assim, os classificadores são aplicados a partir do mais simples ao mais complexo [20]. Cada classificador garante um grau de confiabilidade em sua decisão, para que os classificadores seguintes concentrem o esforço em categorizar os documentos com baixo índice de confiabilidade na categorização. Esse é um método muito semelhante ao *boosting*, a nova característica é o uso de classificadores diferentes no processo de categorização.

2.7 Considerações sobre o capítulo

Esta fundamentação teórica apresentou uma visão geral sobre a Categorização Hierárquica de Textos, destacando conceitos e características da Aprendizagem de Máquina e dos classificadores. Na abordagem utilizada, os classificadores fazem uso de uma coleção de textos previamente categorizados para construir um modelo estatístico de predição, capaz de categorizar novos documentos.

Os diferentes tipos de classificadores utilizam um modelo de representação de documentos, onde os documentos escritos em linguagem natural são mapeados para uma linguagem que permite o tratamento e processamento por estes classificadores.

No processo de classificação, foram identificadas quatro etapas: pré-processamento, treinamento, teste e etapa operacional. No contexto da representação de documentos e da etapa de pré-processamento destaca-se a necessidade de aplicar uma seleção de atributos.

Os conceitos, características e diferenças nos classificadores *k*-NN e SVM foram enfatizados no intuito de demonstrar as distinções existentes nas etapas de treinamento e teste. Também foram apresentados conceitos e exemplos envolvendo métodos de combinação de classificadores.

Capítulo 3

Trabalhos correlatos

Este capítulo descreve os trabalhos relacionados ao tema de estudo desta dissertação e que contribuem para a realização da mesma. Dois dos trabalhos correlatos estão relacionados à Categorizaçao Hierárquica de Textos na língua portuguesa, um dos trabalhos está relacionado à CHT na língua inglesa, e um último trabalho está relacionado à combinação de classificadores que, por sua vez, está relacionado ao objeto de estudo desta dissertação.

A CHT é um assunto amplamente estudado e diferentes tipos de classificadores já foram experimentados em variadas coleções de documentos. Dentre os estudos recentes, é importante citar o uso do classificador *k*-NN em [4,21] na língua portuguesa e, em trabalhos anteriores [15], com a língua inglesa. O classificador SVM é abordado em [17] e [22]. Outras pesquisas [5] em CHT estão voltadas para a categorização em grandes coleções de textos como, por exemplo, textos provenientes da Internet, um trabalho de Dumais e Chen em [23], onde a web pode ser considerada como um grande repositório de textos e permite uma aplicação em larga escala de categorização de textos.

A combinação de classificadores é amplamente pesquisada na área do reconhecimento de textos manuscritos em linguagem natural e reconhecimento de textos em imagens [24]. No entanto, a combinação de classificadores é um assunto que, na CT ou CHT, não vem sendo alvo de uma quantidade significativa de novas propostas. Os trabalhos nesta área são voltados para métodos consolidados na presente literatura, como por exemplo, a votação e o *boosting*. Também existe uma grande quantidade de trabalhos voltados para a combinação de classificadores com o uso de redes neurais. Por esses motivos, neste capítulo é apresentado o trabalho de Bennet, Dumais e Horvitz em [18], uma combinação adptativa de classificadores.

Os resultados dos trabalhos descritos a seguir utilizam as medidas de avaliação[1] de precisão, abrangência, medida F1, macro-média e micro-média, onde: a precisão mede o percentual de documentos corretamente categorizados em uma categoria, dentre todos os documentos da mesma; a abrangência mede o percentual de documentos corretamente categorizados em uma categoria, dentre todos os documentos que deveriam ser categorizados nela; a medida F1 com-

[1] Maiores informações sobre as medidas de avaliação encontram-se no Capítulo 4.

bina os resultados da precisão e da abrangência em um único valor, com mesmo peso para ambas; a macro-média da precisão, abrangência e medida F1 são as média dos resultados de todas as categorias; a micro-média da precisão, abrangência e medida F1 são médias calculadas com o conjunto de todas as categorias. Os valores da macro-média é influenciável pelos resultados das categorias, enquanto que nos valores da micro-média os documentos possuem maior influência.

Dessa forma, os conceitos e idéias dos trabalhos correlatos apresentados no decorrer deste capítulo norteiam o restante desta dissertação.

3.1 O estudo de Langie

O trabalho de Langie [4], que deu origem aos estudos em Categorização de Textos pelo grupo de pesquisas em Processamento da Linguagem Natural da PUCRS, descreve o projeto, a implementação e testes em CHT com o emprego do classificador *k*-NN. O trabalho voltou-se tanto à língua portuguesa como à língua inglesa. No primeiro caso, que será apontado como referência para esse trabalho, foi utilizada a coleção de textos Folha-Hierarq, um subconjunto da coleção Folha-RIcol[2], composta por 2.896 documentos. Para a língua inglesa, Langie utilizou a coleção de textos denominada Reuters-Hierarq que é uma versão da coleção de Sun e Lim em [25], a partir da coleção Reuters-2157 [3].

A coleção de textos Folha-Hierarq foi, previamente, classificada (manualmente) em um conjunto de 28 categorias distintas, resultando em uma árvore de categorias contendo dois níveis. Do total de 28 categorias, dez são situadas no primeiro nível da árvore e o restante no segundo. Como o estudo foi voltado à CHT, não existe um único classificador, mas uma combinação de classificadores, denominados classificadores locais. Todos os nodos da árvore pertencem a uma categoria, com exceção do nodo raiz, sendo que cada nodo não-folha possui um único classificador associado.

A estratégia de testes aplicada foi a de *hold-out*, ou seja, a coleção foi dividida em duas partes, uma para o treino dos classificadores e outra para testes. A coleção de treino contém 1.737 documentos e a coleção de teste, 1.159 documentos. É importante ressaltar que os documentos já se encontravam lematizados, por esse motivo o vocabulário é reduzido e o número de atributos também. Além disso, as palavras constantes na *stoplist*, composta por 365 termos da língua portuguesa, foram retiradas. Essas duas técnicas resultaram na seleção de atributos do processo de categorização.

Para a categorização propriamente dita, foram utilizadas duas estratégias: *limiar baseado em ranking* e *limiar baseado em relevância*.

Na estratégia de *limiar baseado em ranking*, a decisão de classificar um documento é atri-

[2]Coleção derivada do *corpus* da língua portuguesa CETENFolha (*Corpus* do NILC/Folha de São Paulo).

[3]Coleção de textos em língua inglesa composta por reportagens publicadas na agência internacional de notícias Reuters, comumente utilizada na avaliação de CT.

buída à categoria dos documentos vizinhos que possui a maior relevância, não importando o valor de *limiar*, contanto que esse valor seja o mais alto dentre todas as categorias. Essa estratégia limita o processo a uma categorização monocategorial.

A estratégia de *limiar baseado em relevância* assume que os documentos são atribuídos às categorias mais relevantes. Para cada categoria são estabelecidos valores de *limiar*. A categorização em uma certa categoria ocorre quando, para um certo documento, *relevância* $\geq$ *limiar*. O uso de *limiar* permite que mais de uma categoria possa ser atribuída a um documento, contanto que a relevância de uma ou mais categorias seja superior ou igual ao valor de limiar pré-estabelecido. Essa estratégia permite a multicategorização de documentos em uma mesma ramificação.

A Tabela 3.1 apresenta os resultados de Langie em uma estratégia de *limiar baseado em ranking* e com o valor de *k* variável. Os valores de *k* variam entre 13 e 23, dependendo do número de documentos, na coleção de treino, existentes nas categorias. Sendo assim, o valor de *k* aumenta de acordo com o número de documentos. Para cada categoria da tabela são expressas as medidas de precisão (Pr), abrangência (Re) e F1, além das medidas globais de micro-média e macro-média.

Tabela 3.1: Resultado da estratégia de *limiar baseado em ranking* com *k* variável

Categorias	**Pr**	**Re**	**F1**
Agricultura	88%	75%	81%
Área rural	87%	85%	86%
Arte e cultura	91%	88%	90%
Automobilismo	94%	85%	90%
Basquete	100%	96%	98%
carnaval	89%	89%	89%
Ciência	97%	74%	84%
Cinema	65%	79%	71%
Ecologia	100%	57%	73%
Educação	89%	57%	69%
Empregos	64%	81%	71%
Esportes	97%	96%	96%
Finanças	86%	91%	88%
Futebol	90%	96%	93%
Hardware	67%	65%	66%
Imóveis	72%	87%	79%
Informática	94%	96%	95%
Literatura e livros	77%	82%	79%
Medicina e saúde	73%	67%	70%
Moda	100%	46%	63%
Música	86%	79%	82%
Negócios	78%	70%	74%
Pecuária	82%	82%	82%
Política	96%	83%	89%
Software	80%	77%	78%
Turismo	93%	86%	89%
Veículos	91%	86%	89%
Vôlei	100%	83%	91%
Macro-média	87%	80%	82%
Micro-média	88%	85%	86%

Os valores apresentados na Tabela 3.1 demonstram um desempenho acima de 60% para a maioria das categorias, com exceção de poucas categorias que apresentam baixo desempenho em precisão ou abrangência. Dois exemplos ilustrativos são as categorias "Ecologia" e "Moda" que possuem um ótimo desempenho em precisão, porém mediano em abrangência.

A Tabela 3.2 apresenta os resultados de Langie em uma estratégia de *limiar baseada em relevância* e com o valor de *k* variável. Os valores de *k* variam entre 13 e 23, dependendo do número de documentos, na coleção de treino, existentes nas categorias. Para cada categoria da tabela são expressas as medidas de precisão (Pr), abrangência (Re) e F1, além da micro-média e macro-média do total de categorias.

Tabela 3.2: Resultado da estratégia de *limiar baseado em relevância* com *k* variável

Categorias	**Pr**	**Re**	**F1**
Agricultura	85%	85%	85%
Área rural	92%	90%	91%
Arte e cultura	81%	92%	86%
Automobilismo	95%	90%	92%
Basquete	100%	96%	98%
carnaval	53%	89%	67%
Ciência	69%	80%	74%
Cinema	49%	84%	62%
Ecologia	44%	86%	59%
Educação	68%	68%	68%
Empregos	37%	89%	52%
Esportes	92%	98%	95%
Finanças	72%	93%	81%
Futebol	84%	96%	90%
Hardware	53%	84%	65%
Imóveis	52%	85%	64%
Informática	95%	96%	96%
Literatura e livros	58%	80%	67%
Medicina e saúde	39%	73%	51%
Moda	42%	77%	54%
Música	76%	79%	78%
Negócios	35%	80%	49%
Pecuária	74%	91%	82%
Política	86%	93%	90%
Software	67%	85%	75%
Turismo	73%	92%	81%
Veículos	91%	88%	90%
Vôlei	100%	83%	91%
Macro-média	70%	86%	76%
Micro-média	72%	90%	80%

Os documentos são representados segundo o modelo mais comumente conhecido como *bag-of-words*. Os pesos dos atributos dos vetores são definidos através da fórmula *tfidf* [1] seguido de uma normalização, entre 0 e 1. A implementação do classificador adota o algoritmo *k*-NN de Yang [15], que estabelece uma medida de relevância baseada no cosseno.

Através de testes empíricos, Langie constatou que os melhores resultados ocorreram com o uso da estratégia de *limiar baseado em ranking* (Tabela 3.1). A escolha por um valor de *k* variável (entre 13 e 23) também mostrou-se ligeiramente mais eficaz para este classificador e

esta coleção [4].

3.2 O trabalho de Moraes e Lima

O trabalho de Moraes e Lima [21] consiste na CHT para uma coleção de textos em língua portuguesa sem classificação manual prévia (a única informação para a categorização é a seção onde o documento foi publicado, no jornal). Para atingir esse objetivo as autoras utilizaram duas coleções de textos: Folha-Hierarq e PLN-BR CATEG[5]. A primeira é a mesma coleção utilizada por Langie, enquanto a segunda é uma grande coleção contendo trinta mil textos do jornal Folha de São Paulo, nos anos de 1994 a 2005. A coleção PLN-BR CATEG não foi previamente categorizada, portanto os esforços foram voltados à sua categorização propriamente dita, como um passo inicial, no sentido de semi-automatizar o processo. O classificador utilizado no trabalho é o algoritmo *k*-NN utilizado por Langie [4], entretanto implementado em linguagem de programação C.

No processo de CHT foi utilizada a coleção Folha-Hierarq para o treino e a coleção PLN-BR CATEG para os testes dos classificadores. Da última, os textos do ano de 1994 foram retirados por estarem presentes na coleção Folha-Hierarq. A dificuldade em utilizar uma coleção nova está em medir os resultados dos testes. Como não existem categorias manualmente definidas para os documentos, torna-se difícil determinar quando uma categorização é correta ou incorreta. A grande quantidade de documentos na coleção dificulta uma categorização manual. Assim sendo, foi estabelecida uma equivalência de algumas seções do jornal com categorias da hierarquia utilizada por Langie.

Dois experimentos foram realizados, um com a estratégia de *limiar baseado em ranking* e o outro com a estratégia de *limiar baseado em relevância*. Os experimentos, apresentados nas Tabelas 3.3 e 3.4, possuem as mesmas configurações que os experimentos descritos na seção anterior.

A Tabela 3.3 descreve os resultados das medidas de avaliação em precisão (Rr), abrangência (Re) e F1, para as categorias que possuem uma equivalência com as categorias hierárquicas utilizadas por Langie. No final, estão presentes a macro-média e micro-média de todas as categorias.

A Tabela 3.4 descreve os resultados para a precisão (Rr), abrangência (Re) e F1, das categorias que possuem uma equivalência com as categorias hierárquicas utilizadas por Langie. Os resultados desta tabela possuem configuração equivalente aos resultados da Tabela 3.2 apresentada por Langie. No final, estão presentes a macro-média e micro-média de todas as categorias.

A principal contribuição do trabalho de Moraes e Lima está na retomada dos experimentos com o *k*-NN de Langie [4] para uma grande coleção de textos, não categorizados, em língua

[4]Langie realizou outros experimentos, além dos descritos nesta dissertação.

[5]Essa coleção foi obtida através do projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR), apoio CNPq #550388/2005-2.

Tabela 3.3: Resultado da estratégia de *limiar baseado em ranking* com o classificador SVM

Seção	**# documentos**	**Pr**	**Re**	**F1**
Agrofolha	166	27%	83%	41%
Ciência	175	0,8%	59%	13%
Dinheiro	3.790	0,4%	81%	0,80%
Empregos	211	97%	54%	69%
Esportes	4.133	84%	95%	90%
Folha Negócios	36	40%	67%	50%
Imóveis	110	80%	82%	81%
Informática	362	26%	91%	41%
Turismo	429	19%	75%	30%
Veículos	179	30%	84%	44%
Macro-média		41%	77%	47%
Micro-média		32%	89%	47%

Tabela 3.4: Resultado da estratégia de *limiar baseado em relevância* com o classificador SVM

Seção	**# documentos**	**Pr**	**Re**	**F1**
Agrofolha	166	17%	84%	28%
Ciência	175	0,4%	86%	0,7%
Dinheiro	3.790	0,4%	89%	0,7%
Empregos	211	100%	46%	63%
Esportes	4.133	92%	92%	92%
Folha Negócios	36	48%	39%	43%
Imóveis	110	89%	65%	75%
Informática	362	25%	88%	39%
Turismo	429	12%	85%	21%
Veículos	179	17%	88%	29%
Macro-média		40%	76%	40%
Micro-média		24%	88%	37%

portuguesa. A inexistência de trabalhos similiares e de um *corpus* previamente categorizado dificultou a avaliação do classificador. Em geral, os resultados descritos pelas autoras foram muito inferiores aos experimentos de Langie, que utilizou uma coleção menor e previamente categorizada de documentos. As medidas de precisão e abrangência foram afetadas pelas condições do *corpus*, com uma ressalva para três categorias que apresentaram eficiência tanto em precisão como abrangência. No restante das categorias a medida de abrangência mostrou-se melhor que a de precisão. Segundo as autoras, uma das explicações está na evolução do vocabulário ao longo dos anos. Para algumas categorias não existem mudanças significativas de vocabulário, como é o caso das categorias *esportes* e *imóveis*. Outras categorias, como é o caso de *ciência*, por exemplo, possuem uma grande alteração no vocabulário ao longo dos anos.

Esse trabalho demonstra a dificuldade em utilizar a Categorização Hierárquica de Textos em uma grande coleção de textos, onde os mesmos não estão previamente categorizados. As duas maiores dificuldades estão no treinamento e avaliação dos resultados. Mas fica a possibilidade de uma semi-automatização da categorização manual, a partir das respostas obtidas.

3.3 O trabalho de Liu *et al*.

A pesquisa exposta em Liu *et al.* [5] conduziu uma série de experimentos com classificadores SVM, hierárquicos ou não, sobre uma grande coleção de documentos em língua inglesa (o motor de buscas *Yahoo*). O classificador SVM implementado e descrito em Lewis [26] serviu como base para os experimentos.

A disposição usual, em diretórios, das categorias dos documentos, constituiu a representação adotada para o estudo. Classificadores SVM específicos para cada categoria foram treinados com documentos da coleção de treino.

Uma das contribuições dos autores [5] é a análise da complexidade dos algoritmos SVM plano e SVM hierárquico. O custo computacional do classificador hierárquico é 90% inferior ao plano. O motivo para isso é a necessidade do classificador plano de aplicar todos os classificadores para um documento. Já no classificador hieráquico, após a identificação das ramificações não pertencentes ao documento, não existe a necessidade de aplicar o processo de CHT nessas ramificações.

Nessa abordagem, um classificador multicategorial é atribuído a cada nodo não-folha e um classificador monocategorial é atribuído a cada nodo folha. É importante ressaltar que um classificador SVM multicategorial só é capaz de categorizar, no máximo, duas categorias para um documento. Isso representa uma vantagem em nodos que possuem dois nodos filhos, mas quando existem mais de dois filhos a multicategorização pode representar um problema.

Os autores Liu *et al.* [5] encontraram dificuldades em categorias contendo poucos documentos na coleção de treino. Essa situação ocorre com mais freqüência nos níveis mais inferiores da hierarquia. As conclusões ponderadas foram de que o SVM não é adequado quando a quantidade de documentos na coleção de treino é inferior a vinte.

3.4 O trabalho de Bennet, Dumais e Horvitz

A comparação entre diferentes combinações de classificadores é o objetivo do trabalho de Bennet, Dumais e Horvitz em [18]. Além da comparação, o trabalho introduz um método probabilístico para a combinação de classificadores, composto por indicadores de confiabilidade [6]. Esse método acrescenta, à combinação, variáveis que fornecem informações sobre o desempenho dos classificadores em diferentes situações. É aplicado em um algoritmo de categorização, denominado *Stacked Reliability Indicator Variable Ensemble* (STRIVE).

As variáveis ou indicadores de confiabilidade são informações probabilísticas sensíveis ao contexto, fornecidas pelas regras em uma árvore de decisão. Essas informações são referentes aos atributos e à quantidade de categorias às quais eles são forte ou fracamente associados. Em um exemplo citado pelos autores [18], eles consideram três tipos de documentos, onde: (I) as palavras do documento não são relevantes ou estão fortemente associadas a uma categoria; (II)

[6] Do inglês, *reliability indicators*.

as palavras do documento são fracamente associadas a várias categorias; (III) as palavras do documento são fortemente associadas a várias categorias. Os classificadores podem demonstrar padrões de erro diferentes nesses diferentes tipos de documentos e, ao descobrir a qual destes tipos um documento pertence, é possível determinar pesos apropriados para o classificador. A dificuldade está em determinar os diferentes padrões de associações entre as palavras e a estrutura das categorias.

Os experimentos realizados pelos autores descrevem a utilização de classificadores e combinação de classificadores em coleções de textos da língua inglesa, incluíndo o *MSN Web Directory* e Reuters. A avaliação no *corpus* Reuters 21587 envolveu 9.603 documentos para o treino e 3.299 documentos para o teste. De um total de 90 categorias presentes nos documentos, apenas as 10 mais freqüentes foram utilizadas nos experimentos, para evitar que a variação de desempenho resulte em estimativas pouco confiáveis. A seleção de atributos limitou o uso em 300 palavras para cada categoria.

Um dos quatro classificadores utilizados é o classificador polinomial SVM, uma implementação do classificador SMO disponível na ferramenta Smox. Por motivos de simplificação, apenas os experimentos que utilizam o classificador SVM, individualmente ou em um método combinatório, e os experimentos com as coleções *MSN Web Directory* e Reuters 21587 são descritos aqui.

A Tabela 3.5 apresenta uma simplificação da tabela descrita pelos autores. Nela, são mostradas a macro-média e micro-média da medida F1 para a categorização com o classificador SVM (Smox), a combinação *Best By Class* e a combinação STRIVE, utilizada na coleção *MSN Web Directory*.

Tabela 3.5: Desempenho da combinação STRIVE na coleção *MSN Web Directory*

Método	**Macro-média F1**	**Micro-média F1**
Smox	67,0%	70,0%
Best By Class	67,0%	70,0%
STRIVE-Smox	69,4%	72,3%

A Tabela 3.6 apresenta uma simplificação do resultado das categorizações com a coleção Reuters 21587, descritas pelos autores. Nela, são mostrados a macro-média e micro-média da medida F1 para a categorização com o classificador SVM (Smox), a combinação *Best By Class* e a combinação STRIVE.

Tabela 3.6: Desempenho da Combinação STRIVE na coleção Reuters 21587

Método	**Macro-média F1**	**Micro-média F1**
Smox	84,8%	91,0%
Best By Class	85,9%	91,2%
STRIVE-Smox	87,4%	92,3%

Observando-se os resultados das Tabelas 3.5 e 3.6 é possível reparar que a combinação utilizando o algoritmo STRIVE-Smox melhora o desempenho do classificador SVM nas duas

coleções, e que a combinação *Best By Class* apresenta uma melhora no desempenho da coleção Reuters 21587 em relação ao uso individual do classificador SVM.

Nas discussões apresentadas pelos autores, eles descrevem que, com um intervalo de confiança de 95%, não é encontrada uma diferença significativa entre os resultados. Isso não exclui a vialiabilidade das combinações, uma vez que os resultados são consistentes com outros resultados apresentados na literatura.

3.5 Considerações sobre o capítulo

Os trabalhos correlatos que foram apresentados nesse trabalho contribuem para a execução dessa dissertação. A escolha por estes trabalhos se deu por afinidade com os objetivos pretendidos. O trabalho de Langie, é de vital importância para a execução dos experimentos relatados nas próximas seções, uma vez que o protótipo por ele implementado é utilizado como base para os experimentos descritos nessa dissertação.

O objetivo deste capítulo foi destacar os trabalhos correlatos que influenciaram diretamente no processo e metodologia da CHT.

O próximo capítulo descreve a metodologia da pesquisa, empregada para a realização dos experimentos relatados nos Capítulos 5 e 6.

Capítulo 4

Metodologia da pesquisa

Neste capítulo é apresentada a metodologia empregada nos experimentos que fazem parte desta dissertação. Essa metodologia agrega informações e características sobre a coleção de textos utilizada, estrutura hierárquica das categorias, heurística de combinação de classificadores, método de avaliação e medidas de avaliação.

Primeiramente, é apresentada a coleção de textos. A seguir, as categorias e sua estruturação hierárquica são descritas. Para uma compreensão mais minuciosa do processo de CHT, também é mostrado o formato de representação dos documentos, que deve ser interpretado pelos classificadores.

É exposta uma heurística para a combinação dos classificadores *k*-NN e SVM. Essa heurística, denominada *k*-NN+SVM, agrega os dois classificadores em uma única abordagem, onde os classificadores são utilizados de acordo com suas características estruturais.

Para a avaliação dos experimentos realizados e da heurística proposta são apresentados alguns conceitos fundamentais para a metodologia. Dentro dessa metodologia são destacadas algumas medidas de avaliação amplamente empregadas na literatura em estudos semelhantes. Para confirmar os resultados encontrados com a metodologia e as medidas de avaliação, é proposta a aplicação de um teste estatístico Z, no intuito de comprovar se os resultados são estatisticamente significativos.

A metodologia descrita no decorrer deste capítulo é aplicada aos experimentos relatados nos capítulos 5 e 6. Algumas das escolhas, como, por exemplo a heurística proposta e o método de avaliação, estão relacionadas às características da coleção e da estrutura hierárquica de categorias.

O capítulo conclui trazendo algumas considerações pertinentes à escolha da metodologia aqui descrita.

4.1 Coleção de textos

Os experimentos com categorização fazem uso de coleções de documentos que, usualmente, são divididas em duas partes, uma parte para o treino e outra para o teste dos classificadores: *corpus* de treino e *corpus* de teste, respectivamente. Neste trabalho, a coleção de textos utilizada para o desenvolvimento dos experimentos é a Folha-RIcol[1], derivada do *corpus* em língua portuguesa CETENFolha (*Corpus* do NILC/Folha de São Paulo). Essa coleção compreende um conjunto de documentos organizados em 28 categorias hierárquicas.

No total, 2.896 documentos de diferentes seções do jornal constituem a coleção. Os documentos são compostos por notícias do jornal Folha de São Paulo do ano de 1994. De acordo com a vinculação original, cada um dos documentos pertence a uma e apenas uma seção mas, na elaboração manual do *corpus*, um documento pode pertencer a mais de uma categoria e, para fins estatísticos, um mesmo documento que pertença a duas categorias é considerado como dois documentos distintos, no momento da avaliação dos classificadores. O *corpus* é composto por 1.162 documentos associados a uma única categoria e 1.734 documentos com múltiplas categorias.

O conteúdo dos documentos é composto por palavras em sua forma canônica[2]. É preciso levar em consideração que a manipulação das palavras originais acabou por deixar o conteúdo dos documentos bem distinto do seu texto original.

4.2 Organização hierárquica

A estrutura da árvore de categorias está definida no formato representado na Figura 4.1. Ela possui dois níveis na hierarquia, contendo dez categorias no primeiro nível e dezoito no segundo nível. Anterior ao primeiro nível, está o nodo raiz, no entanto ele não é considerado como um nível da hierarquia, pois não possui categoria a ele associada.

O número de documentos por categoria não é uniforme. Algumas categorias possuem poucos documentos, enquanto outras possuem uma quantidade de documentos acima da média. Isso ocorre em função da própria composição do *corpus* jornalístico e da diferente periodicidade de cada seção do jornal. De um modo geral, o número de documentos decresce em níveis inferiores da hierarquia. A Figura 4.1 mostra a organização hierárquica do *corpus*. As categorias alinhadas à esquerda estão no primeiro nível da hierarquia, e as categorias do segundo nível estão indentadas mais à direita.

4.2.1 Representação dos documentos

Para que os classificadores possam realizar as predições, os documentos da coleção Folha-RIcol, tanto os de treino como os de teste, são armazenados em arquivos no formato *Attribute-*

[1] http://www.linguateca.pt/Repositorio/Folha-RIcol/

[2] Verbos no infinitivo e demais palavras na forma masculina singular, se houver.

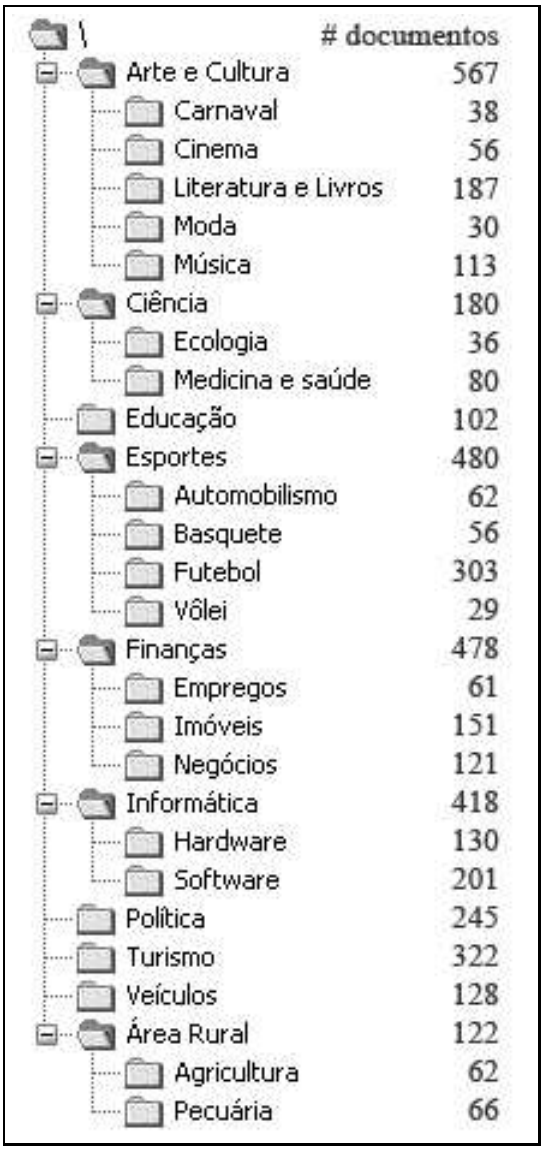

Figura 4.1: Categorias hierárquicas

Relation File Format (ARFF) utilizado pela ferramenta WEKA. A Figura 4.2 mostra um exemplo de arquivo ARFF. Para facilitar a visualização, o número de atributos e documentos foi reduzido e, também, a figura foi dividida em três blocos. O primeiro bloco contém o cabeçalho, o segundo, os atributos e o último, os documentos.

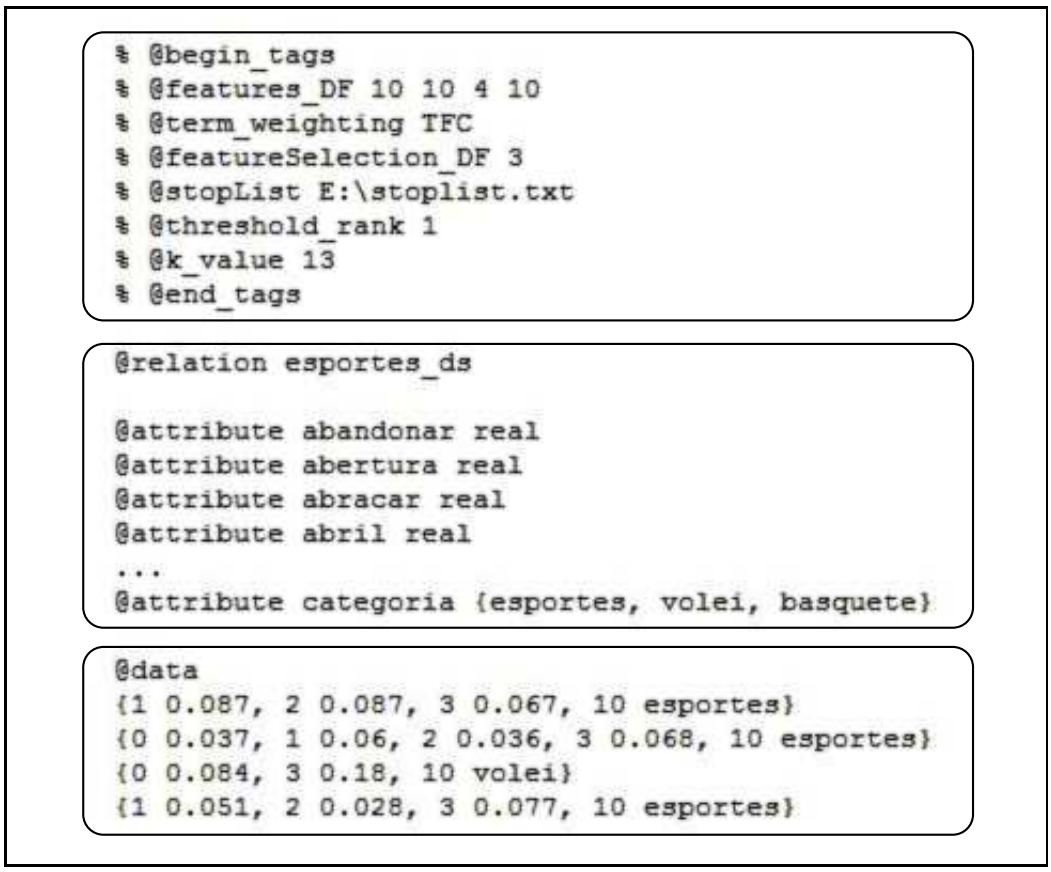

Figura 4.2: Exemplo de um arquivo no formato ARFF

O arquivo ARFF é composto por um cabeçalho, atributos e documentos. O cabeçalho contém informações, nessa ordem, sobre freqüência de atributos, método para calcular o peso dos atributos, seleção de atributos, *stoplist*, estratégia de categorização e valor de *k* (para o classificador *k*-NN[3]).

A parte do arquivo, onde são relacionados os atributos e os documentos, está no padrão do formato estabelecido pela ferramenta WEKA. Após a etiqueta "@relation" vem o nome do arquivo e, nas linhas subseqüentes, estão os atributos. A etiqueta "@data" demarca o início dos documentos.

Ressalta-se aqui a importância de manter a ordem dos atributos nos documentos; do contrário, os classificadores assumem valores de pesos equivocados para os atributos dos documentos. No exemplo da figura 4.2, os documentos são expressos por vetores de atributos separados por vírgula ("‘,"’) e dentro de cada elemento do vetor estão as informações da posição e peso do atributo, respectivamente.

[3]No caso do classificador SVM, os parâmetros do classificador são estabelecidos no próprio algoritmo.

O último atributo é sempre a categoria correta do documento.

4.3 Combinação de classificadores

O principal objeto de estudo desta dissertação é a combinação de classificadores, mais especificamente a heurística proposta nesta seção. No entanto, uma segunda combinação de classificadores também será experimentada, a função combinatória por voto [18] [1]. A maneira mais simples de combinar o resultado de classificadores é pelo voto majoritário, sem a atribuição de peso. Este método é empregado a fim de realizar uma comparação com a heurística proposta.

4.3.1 Votação

Nos experimentos que são relatados no Capítulo 6, a votação é realizada pelo voto majoritário. Na votação empregada nesta dissertação, tanto o classificador *k*-NN como o classificador SVM são aplicados em cada um dos documentos nos *corpora* de treino e teste.

Para atribuir o resultado final, a categorização de um documento, é realizada a média aritmética do valor da predição de cada categoria. Para tanto, os classificadores trabalham com um valor de predição no intervalo entre 0 e 10, onde o valor 0 designa uma categoria com menor probabilidade de categorização e o valor 10 designa uma categoria que deve certamente ser atribuída ao documento.

Em uma categorização monocategorial ou por voto majoritário, a categoria que obtiver o maior valor é atribuída para o documento.

4.3.2 Heurística proposta: *k*-NN+SVM

Nesta seção é descrita uma heurística de combinação dos classificadores *k*-NN e SVM para categorizar a coleção de textos Folha-RIcol.

A idéia é combinar os classificadores *k*-NN e SVM em uma heurística, de forma a tirar proveito das características de cada um deles. Por exemplo, o SVM é essencialmente um classificador monocategorial, enquanto o SVM é multicategorial. É possível aplicar o classificador SVM em problemas multicategoriais mas, no caso de mais de duas categorias, se faz necessário o uso de dois ou mais classificadores SVM em conjunto.

Devido a esses fatores, nesta dissertação é proposta e experimentada uma heurística baseada em complexidade e desempenho.

A heurística proposta é inspirada na proposta de combinação de classificadores descrita por Alpaydin em [11], onde os classificadores são aplicados seqüencialmente, de acordo com o seu custo de complexidade. A principal diferença está na influência dos classificadores para a categorização. Nesta proposta, um único classificador é responsável pela decisão de categorizar ou não um documento.

Estabelecendo uma heurística *k*-NN+SVM de combinação:

- nodos com mais de dois filhos utilizam o classificador *k*-NN;
- nodos não-folhas com dois filhos utilizam o classificador SVM multicategorial;
- nodos com um filho utilizam o classificador monocategorial SVM.

Em nodos com mais de dois filhos a utilização do *k*-NN pode ser mais expressiva, gerando melhores resultados, visto que nesses nodos seria necessário mais de um classificador SVM multicategorial.

No caso do SVM, os classificadores também poderiam ser utilizados em nodos folhas. Para obter uma vantagem com a aplicação de classificadores SVM, seria necessária uma heurística que, ao detectar categorizações incorretas nesses nodos, realizasse algum tipo de correção.

Para a coleção Folha-RIcol são utilizados sete classificadores entre as categorias que constituem nodos não-folha. Cada um destes classificadores é responsável pela classificação de categorias estabelecidas em seus nodos filhos. Por exempo, o classificador do nodo raiz é responsável pela categorização nas categorias presentes no primeiro nível. Nesta heurística este classificador é necessariamente um classificador *k*-NN. Os classificadores do primeiro nível categorizam os documentos em sub-árvores do segundo nível. Esses classificadores podem ser tanto classificadores *k*-NN como classificadores SVM. O segundo nível não possui classificadores, já que não existem outras possibilidades para a categorização.

4.4 Avaliação

Dois aspectos são fundamentais para a execução da tarefa de avaliar o desempenho do processo de CHT: a coleção e as métricas de avaliação. Esses dois aspectos permitem avaliar o desempenho do processo de CHT em relação a outros resultados apresentados na literatura. Muitas vezes não é possível realizar uma comparação direta com resultados anteriores. A coleção de documentos e as métricas de avaliação variam de experimento para experimento, o que dificulta comparações. Existe uma dificuldade em adotar um padrão abrangente o suficiente para todas as variações impostas no desenvolvimento de novas pesquisas.

Um dos principais problemas na avaliação de processos e sistemas de Categorização Automática de Textos é a falta de coleções padronizadas. Como destaca Yang em [27], mesmo a coleção Reuters, destacada por ser usada nas validações e avaliação de sistemas dessa natureza, possui diversas versões; os resultados dependem da divisão do *corpus* de treino e teste, das categorias utilizadas, da representação das categorias e outros fatores. Essas diferenças dificultam a comparação dos resultados obtidos em outros experimentos ou versões. Em CT envolvendo a língua portuguesa, esse problema é agravado, pela escassez, quase ausência, de coleções padronizadas para a avaliação de processos e sistemas.

O ideal para a avaliação seria o uso de uma coleção de avaliação compartilhada por todos os pesquisadores da área de Categorização Automática de Textos, todavia essa não é a situação que se apresenta. As tentativas de criar um padrão constituem soluções viáveis apenas para dois ou três métodos [27]. A solução para esse problema é o uso de medidas globais de avaliação, em conjunto com uma análise crítica dos experimentos.

As medidas de avaliação permitem uma comparação indireta entre diferentes experimentos. As medidas mais comuns são precisão, abrangência, acurácia, erro, *F-measure*, micro-média e macro-média, entre outras. Cada uma delas permite avaliar um aspecto do desempenho do processo de CT [27]. Em conjunto, essas medidas fornecem informações que contribuem para uma comparação quantitativa, indireta, entre diferentes resultados.

Os experimentos desta dissertação fazem uso de medidas locais para cada experimento e medidas globais que sintetizam os resultados de todas as categorias.

4.4.1 Método de avaliação

A avaliação do processo de CHT consiste na divisão do *corpus* inicial em duas partes, resultando em um *corpus* de treino e um *corpus* de teste. O objetivo é construir *corpora* que possibilitem a avaliação de um classificador em uma aproximação do processo real de CHT. Dos métodos de avaliação existentes, dois são destacados neste trabalho: *hold-out* e *cross-validation*.

- O método de avaliação denominado *hold-out* ou *train-and-test* [1] é um método de fácil implementação. Ele consiste em dividir o *corpus* em dois, um para o treino e outro para o teste. O classificador é treinado com o *corpus* de treino e avaliado com o *corpus* de teste. Esse método não garante um resultado realista do processo; dependendo da divisão do *corpus*, o resultado pode ser otimista ou pessimista.

- O método *k-fold cross-validation* [1] consiste em separar o *corpus* inicial em *k* partes e usar *k* classificadores para o treinamento. Esses classificadores são aplicados iterativamente em *corpora* de treino distintos contendo *k-1* partes do *corpus inicial*. Então, para cada iteração existe uma coleção de *k-1* partes para o treino e 1 parte para o teste, sendo que a cada iteração o *corpus* de treino assume uma nova parte *k*. Normalmente, o resultado final é obtido através do cálculo da média de todos os classificadores.

Para os experimentos relatados nos capítulos 5 e 6, é utilizada uma variante do método *hold-out*. A solução adotada é a execução do método *hold-out* em três *corpora* diferentes, com calculo da média e do desvio padrão, e utilização dos mesmos três *corpora* em todos os experimentos. Nesse sentido, os resultados dos experimentos, relatados nos capítulos 5 e 6, permitem realizar uma comparação e análise direta entre os experimentos. Dessa forma, o resultado final é mais expressivo, de um ponto de vista estatístico, do que a execução de apenas um conjunto de *corpus* de treino e teste.

4.4.2 Medidas de avaliação

Para avaliar a CHT, como já foi dito, se faz necessário utilizar métricas de avaliação. Este trabalho utiliza três das medidas [1,7] comumente empregadas: precisão (Pr), abrangência (Re) e a medida F1. Tais medidas são obtidas através de fórmulas envolvendo os valores de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos (tabela de contingência), FP, FN, TP e TN, respectivamente, como exposto a seguir.

A tabela de contigência [1,7], exemplificada na Tabela 4.1, é responsável pela correlação entre as categorias previamente determinadas (categorias corretas) e o resultado da categorização (categorias atribuídas pelo classificador) dos documentos.

Tabela 4.1: Tabela de contingência

		categoria correta de c_j	
		sim	não
categoria atribuída	sim	|TP|	|FP|
	não	|FN|	|TN|

A tabela de contingência compreende o conjunto de todas as categorizações, sejam elas corretas ou não. Na Tabela 4.1 tem-se como:

- Verdadeiros positivos (|TP|) - o conjunto de documentos corretamente categorizados em c_j;
- Falsos positivos (|FP|) - o conjunto de documentos erroneamente categorizados em c_j;
- Verdadeiros negativos (|TN|) - o conjunto de documentos corretamente rejeitados em c_j;
- Falsos negativos (|FN|) - o conjunto de documentos erroneamente rejeitados em c_j.

A precisão de um classificador expressa o percentual de documentos corretamente categorizados em c_j, dentre todos os documentos corretos da categoria c_j.

A abrangência mede o percentual de documentos corretamente categorizados na categoria c_j, dentre todos os documentos que deveriam ser categorizados em c_j.

A medida F combina os resultados da precisão e da abrangência em um único valor. A medida F é denominada medida F1 nesta dissertação, porque a precisão e a abrangência possuem igual valor de peso no cálculo.

Essas medidas de precisão (Pr), abrangência (Re) e F1 são calculadas através das seguintes fórmulas:

$$Pr_i = \frac{|TP_i|}{|TP_i| + |FP_i|} \quad Re_i = \frac{|TP_i|}{|TP_i| + |FN_i|} \quad F1_i = \frac{2Pr_iRe_i}{Pr_i + Re_i} \tag{4.1}$$

Além dessas três medidas, expressas na equação 4.1, que fornecem uma avaliação individual por categorias, é importante utilizar a micro-média e a macro-média. Essas duas medidas [1]

calculam o desempenho global das medidas de precisão, abrangência e F1. A diferença entre elas é que, na macro-média (equação 4.2), as categorias são tratadas com igual importância, enquanto que na micro-média (equação 4.3) são os documentos que possuem igual importância.

$$Pr^M = \frac{\sum_{i=1}^{|C|} Pr_i}{|C|} \quad Re^M = \frac{\sum_{i=1}^{|C|} Re_i}{|C|} \quad F1^M = \frac{\sum_{i=1}^{|C|} F1_i}{|C|} \tag{4.2}$$

A macro-média é uma média dos resultados de todas as categorias para a precisão, abrangência e medida F1.

$$Pr_\mu = \frac{\sum_{i=1}^{|C|} |TP_i|}{\sum_{i=1}^{|C|}(|TP_i| + |FP_i|)} \quad Re_\mu = \frac{\sum_{i=1}^{|C|} |TP_i|}{\sum_{i=1}^{|C|}(|TP_i| + |FN_i|)} \quad F1_\mu = \frac{2Pr_\mu Re_\mu}{Pr_\mu + Re_\mu} \tag{4.3}$$

A micro-média da precisão, abrangência e medida F1 é uma média calculada com o conjunto das tabelas de contingência de todas as categorias.

De acordo com Yang e Liu em [15], a macro-média é influenciada pelo desempenho do classificador em categorias raras e a micro-média é influenciada pelo desempenho do classificador em categorias comuns. Em se tratando de categorias, comuns são aquelas em que os resultados são muito parecidos e raras são aquelas onde a diferença dos resultados é mais visível. Então, a macro-média é mais influenciada por categorias com resultados que divergem da média, sejam eles bons ou ruins. A micro-média não faz distinção entre categorias com resultados bons e ruins, porque não prioriza os documentos ao invés das categorias. Para realizar a análise dos classificadores é importante analisar a micro-média e a macro-média em conjunto. A primeira expressa um desempenho global do processo de categorização, enquanto a segunda expressa o desempenho global das categorias. A análise conjunta das duas medidas ajuda a identificar problemas no processo.

4.4.3 Testes estatísticos

Para comparar os resultados obtidos nos dois próximos capítulos, os testes estatísticos [28] podem ser utilizados no intuito de determinar se os resultados são estatisticamente significativos. Para comparar os resultados são necessários uma hipótese e um intervalo de confiança. A hipótese é comumente a de que não existe relação entre *n* pares de valores observados em uma amostra de tamanho *n*, nesse caso a hipótese bi-caudal, porque é uma distribuição normal bidimensional [28]. O intervalo de confiança representa uma porcentagem de observações onde é garantida a hipótese.

Por exemplo, em uma hipótese que não possui relação estatística e com intervalo de confiança de 95%, presume-se que em 95% dos resultados futuros não deverá existir relação estatística.

Os testes estatísticos servem para validar uma hipótese como, por exemplo, comparar dois

métodos e determinar se existe uma diferença estatística entre eles. No caso de o teste resultar em uma diferença significativa, a hipótese é aceita, caso contrário a hipótese é rejeitada.

Existe também a hipótese uni-caudal onde, dada uma amostra de valores, pretende-se determinar qual a probabilidade da ocorrência para um determinado intervalo. Nesse sentido, a comparação está na amostra e no intervalo de valores. Esse é um teste importante na comprovação de que, em determinado intervalo, o resultado encontrado pode ser repetido em novas ocorrências de um mesmo método. Um exemplo de teste que utiliza hipótese uni-caudal é o teste estatístico de Fischer [28] ou, como é comumente denominado, teste estatístico Z.

Um teste comumente utilizado na comparação de dois métodos distintos é o *chi-square* (qui-quadrado) [11]. Ele visa determinar se existe uma diferença estatisticamente significativa nos resultados de dois métodos. É importante a aplicação desse tipo de teste em pesquisas científicas para verificar se há ou não um embasamento quanto à eficiência de um novo método.

4.5 Considerações sobre o capítulo

Nesta seção foram apresentadas a coleção, a hierarquia das categorias, a representação dos documentos, a heurística proposta e os mecanismos a serem empregados em sua avaliação. Enfim, a metodologia que é utilizada nos experimentos. Essa metodologia é crucial para a realização, descrição e análise dos experimentos descritos nos Capítulos 5 e 6.

A partir das informações apresentadas neste capítulo inicia-se a descrição dos experimentos que serão apresentados e discutidos.

Capítulo 5

Experimentos com classificadores individuais

Os experimentos com o processo de Categorização Hierárquica de Textos, apresentados no decorrer deste capítulo, têm por objetivo comparar qualitativamente os classificadores *k*-NN e SVM entre si. Por questões metodológicas e organizacionais, primeiramente é detalhado o processo de CHT, seguido por uma breve descrição das configurações dos parâmetros dos classificadores e, por fim, é apresentada a análise dos resultados dos experimentos.

Para realizar o processo de CHT, um percentual de 70% do *corpus* é utilizado para treino, e o restante para testes. Este percentual desigual é necessário para priorizar o treinamento do classificador, caso contrário pode ocorrer *underfitting*.

Na primeira etapa, o pré-processamento, são removidas as palavras constantes na *stoplist* e palavras com uma freqüência inferior a três por documento. A *stoplist* é composta pelas palavras mais comumente encontradas e que não agregam um valor significativo para a categorização. As palavras com freqüência inferior a três não são utilizadas por se tratar de palavras raras, que não influenciam no processo de categorização, ou mesmo de erros ortográficos.

Após o pré-processamento, os documentos são armazenados em arquivos no formato ARFF utilizado pela ferramenta WEKA. 70% dos documentos relativos a uma categoria são obtidos aleatoriamente do total de documentos da mesma, para gerar o *corpus* de treino. O restante dos documentos é utilizado para os testes, também armazenado no formato ARFF. Dessa forma, os *corpora* de treino e teste encontram-se separados e estáticos para fins da experimentação.

No total, são utilizados sete classificadores em categorias que constituem nodos não-folha. Cada um destes classificadores é responsável pela categorização em seus nodos filhos. Por exempo, o classificador do nodo raiz é responsável pela categorização nas categorias presentes no primeiro nível, já os classificadores do primeiro nível categorizam os documentos em sub-árvores do segundo nível. O segundo nível não possui classificadores, uma vez que não existem outras possibilidades de categorização.

A avaliação do processo de CHT é realizada por uma variante do método denominado *hold-out*, que é de fácil implementação mas não garante resultados estatisticamente significativos.

Para melhorar o processo, o método *hold-out* é repetido três vezes. A cada iteração o *corpus* inicial é dividido aleatoriamente, gerando *corpora* de treino e teste diferentes. Essa repetição é necessária pela dificuldade em atender uma exigência do método *k-fold cross-validation*. Devido à estrutura hierárquica existe uma dificuldade em manter uma proporção igualitária para cada categoria ao dividir o *corpus* em *k* partes. Como algumas categorias possuem uma quantidade muito pequena de documentos fica difícil manter uma divisão em *k* partes sem diminuir significativamente o número de documentos para o teste dos classificadores.

Os experimentos descritos neste capítulo estão divididos em dois grupos: grupo α e grupo β.

5.1 Grupo α - *k-Nearest Neighbors*

Os experimentos relatados nesta seção são compostos por três execuções do classificador *k*-NN com três *corpora* de treino e teste distintos, execuções estas denominadas grupo α. A Tabela 5.1 descreve a média e o desvio padrão dos resultados de precisão, abrangência e da medida F1 deste grupo.

Tabela 5.1: Média e desvio padrão do grupo α

Categorias	$\overline{Pr}$	$\overline{Re}$	$\overline{F1}$
Carnaval	79,7% ± 0,114	91,1% ± 0,078	84,6% ± 0,059
Cinema	83,3% ± 0,030	84,1% ± 0,071	83,5% ± 0,019
Literatura e livros	76,9% ± 0,037	77,2% ± 0,053	77,0% ± 0,045
Moda	68,2% ± 0,233	65,2% ± 0,197	66,4% ± 0,204
Música	77,6% ± 0,055	85,5% ± 0,028	81,4% ± 0,041
Arte e cultura	90,7% ± 0,027	88,7% ± 0,008	89,7% ± 0,009
Ecologia	72,4% ± 0,240	38,7% ± 0,108	50,3% ± 0,147
Medicina e saúde	79,8% ± 0,088	63,9% ± 0,005	70,9% ± 0,038
Ciência	91,7% ± 0,027	64% ± 0,032	75,4% ± 0,026
Educação	91% ± 0,039	70,4% ± 0,083	79,2% ± 0,065
Automobilismo	95,4% ± 0,040	89,2% ± 0,036	92,1% ± 0,021
Basquete	91,8% ± 0,053	93,2% ± 0,072	92,5% ± 0,062
Futebol	91,1% ± 0,009	95,7% ± 0,012	93,4% ± 0,010
Vôlei	96,7% ± 0,058	92,5% ± 0,066	94,4% ± 0,051
Esportes	97% ± 0,001	95,1% ± 0,009	96% ± 0,005
Empregos	62,2% ± 0,165	51% ± 0,065	55,2% ± 0,067
Imóveis	79,6% ± 0,051	86,7% ± 0,006	82,9% ± 0,026
Negócios	65,7% ± 0,107	69,5% ± 0,098	67,5% ± 0,100
Finanças	87,2% ± 0,027	88,8% ± 0,013	88% ± 0,020
Hardware	79,3% ± 0,066	82,2% ± 0,079	80,7% ± 0,070
Software	75,8% ± 0,036	70,8% ± 0,042	73,2% ± 0,021
Informática	93,9% ± 0,011	94,1% ± 0,017	94,0% ± 0,013
Política	90,5% ± 0,034	93,2% ± 0,024	91,8% ± 0,007
Turismo	89,1% ± 0,036	81,2% ± 0,051	85,1% ± 0,024
Veículos	92,0% ± 0,006	90,4% ± 0,055	91,2% ± 0,030
Agricultura	83,1% ± 0,183	70,2% ± 0,118	75,2% ± 0,160
Pecuária	91,1% ± 0,019	80,5% ± 0,105	85,3% ± 0,062
Área rural	87,4% ± 0,062	76,9% ± 0,044	81,8% ± 0,046
Macro-média	84,3% ± 0,020	79,6% ± 0,009	81,4% ± 0,014
Micro-média	87,3% ± 0,011	84,6% ± 0,004	85,9% ± 0,007

A Tabela 5.1 é o resultado da técnica da avaliação com três repetições do método *hold-out*. Nela, são apresentados os valores para as médias de precisão, abrangência e medida F1 das respectivas categorias, seguidos do desvio padrão. Essa tabela proporciona uma comparação

quantitativa desses resultados com os resultados dos outros classificadores, mas não permite uma análise completa do desempenho do próprio classificador.

O desvio padrão mostra a diferença nos resultados obtidos dos diferentes classificadores e, com exceção de quatro categorias, o desvio padrão obtido para a medida F1 é inferior a 10%. Esse fato caracteriza uma homogeneidade no processo de CHT. Apesar de alguns resultados individuais das categorias divergirem, a micro-média e a macro-média das três execuções são muito semelhantes.

Para realizar uma análise consistente do classificador *k*-NN uma das execuções é apresentada na Tabela 5.2. Nesta tabela, são apresentadas as medidas de precisão, abrangência e F1 para cada uma das categorias, e também são apresentados os totais de verdadeiros positivos (|TP|), falsos positivos (|FP|) e falsos negativos (|FN|) para as respectivas categorias. Ao final das tabelas são apresentadas as medidas de macro-média e micro-média.

Tabela 5.2: Resultado da segunda execução do grupo α

Categorias	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Carnaval	76,9%	100%	87%	10	3	0
Cinema	85,7%	80%	82,8%	12	2	3
Literatura e livros	73,7%	73,7%	73,7%	42	15	15
Moda	72,7%	80%	76,2%	8	3	2
Música	75,8%	83,3%	79,4%	25	8	5
Arte e cultura	90,8%	88,6%	89,7%	148	15	19
Ecologia	100%	50%	66,7%	4	0	4
Medicina e saúde	87%	64,5%	74,1%	20	3	11
Ciência	94,6%	63,6%	76,1%	35	2	20
Educação	88,9%	61,5%	72,7%	24	3	15
Automobilismo	92,9%	86,7%	89,7%	13	1	2
Basquete	95,8%	100%	97,9%	23	1	0
Futebol	92,0%	96,4%	94,2%	81	7	3
Vôlei	100%	87,5%	93,3%	7	0	1
Esportes	97,1%	95,7%	96,4%	132	4	6
Empregos	52,4%	57,9%	55%	11	10	8
Imóveis	80,4%	87,2%	83,7%	41	10	6
Negócios	53,8%	58,3%	56%	21	18	15
Finanças	84,5%	87,3%	85,9%	131	24	19
Hardware	82,5%	89,2%	85,7%	33	7	4
Software	72,5%	69,8%	71,2%	37	14	16
Informática	92,6%	92,6%	92,6%	100	8	8
Política	90,5%	93,8%	92,1%	76	8	5
Turismo	92,5%	81%	86,9%	86	7	19
Veículos	92,5%	92,5%	92,5%	37	3	3
Agricultura	63,6%	58,3%	58,3%	7	4	6
Pecuária	93,3%	82,4%	87,5%	14	1	3
Área rural	80,8%	72,4%	76,4%	21	5	8
Total				1199	186	226
Macro-Média	84,1%	79,7%	81,2%			
Micro-Média	86,5%	84,1%	85,3%			

Nesse grupo, o processo de CHT foi executado em 869 documentos distintos. É possível verificar que o conjunto de verdadeiros positivos, falsos positivos e falsos negativos (tal como exposto na Tabela de Contingência, vide Tabela 4.1) são superiores à quantidade de documentos categorizados. Isso porque, no experimento da Tabela 5.1, o limiar para a escolha das categorias é baseado no *rank* da categoria que possui o maior valor de relevância, ou seja, um classificador atribui apenas uma categoria para cada documento. Isso não significa necessariamente um categorização monocategorial. Como existem diferentes níveis hierárquicos e os classificadores

estão posicionados em dois níveis da hierarquia, mais de um classificador pode fazer parte do processo de CHT. Em uma visão geral do processo, existe uma multicategorização, mas em uma visão específica de cada classificador, o processo é monocategorial.

Essa análise do primeiro grupo de experimentos serve para dar início a uma análise de desempenho dos classificadores, mas não permite ainda um estudo mais detalhado. Isso porque não existe informação sobre a CHT nesse resultado: nele as categorizações são analisadas independentemente e sem informações sobre os níveis hierárquicos. Para aprimorar a análise, é necessário obter informações mais específicas dos classificadores. Para isso, na Tabela 5.3, os resultados estão apresentados em três blocos: primeiro nível, segundo nível e resultado total (correspondente a ambos os níveis da hierarquia). Esse é um fator importante para verificar o efeito da distribuição percentual de documentos na fase de treinamento do processo de CHT, além de permitir verificar qual o efeito da escolha pela categorização da categoria no topo do *ranking*.

Tabela 5.3: Resultado com discernimento por níveis hierárquicos - grupo α

Resultado global do primeiro nível

k-NN	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	90,4%	82,9%	86,1%			
Micro-Média	90,9%	86,6%	88,7%			
Total				790	79	122

Resultado global do segundo nível

k-NN	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	80,6%	78,1%	78,5%			
Micro-Média	79,3%	79,7%	79,5%			
Total				409	107	104

Resultado total

k-NN	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	84,1%	79,7%	81,2%			
Micro-Média	86,5%	84,1%	85,3%			
Total				1199	186	226

A discriminação do resultado por níveis hierárquicos mostra uma queda de desempenho no segundo nível. A depreciação do resultado é esperada, porque as categorizações incorretas no primeiro nível afetam diretamente o desempenho dos classificadores do segundo nível. Um dos possíveis motivos é a escolha do *ranking*, uma vez que o *ranking* de melhor categoria não permite a categorização em duas ramificações da árvore simultaneamente. Outro motivo é a própria representação hieráquica das categorias que agrega categorizações incorretas, de categorizações anteriores, aos classificadores nos níveis inferiores da hierarquia quando um classificador situado diretamente acima não categoriza corretamente um documento.

Na próxima seção prossegue o relato dos experimentos, com o grupo β - *Support Vector Machines*.

5.2 Grupo β - *Support Vector Machines*

Os experimentos do grupo β relatados nesta seção são compostos por três execuções do classificador SVM com três *corpora* de treino e teste distintos. A Tabela 5.4 mostra a média e o desvio padrão dos resultados de precisão, abrangência e da medida F1.

Tabela 5.4: Média e desvio padrão do grupo β

Categorias	$\overline{Pr}$	$\overline{Re}$	$\overline{F1}$
Carnaval	79,6% ± 0,263	60,6% ± 0,215	68,5% ± 0,223
Cinema	97% ± 0,053	62,1% ± 0,251	73,5% ± 0,201
Literatura e livros	63,9% ± 0,041	86,9% ± 0,024	73,5% ± 0,027
Moda	48,8% ± 0,423	25,8% ± 0,250	32,9% ± 0,300
Música	67% ± 0,097	79,0% ± 0,070	72,4% ± 0,081
Arte e cultura	83,1% ± 0,021	94% ± 0,010	88,2% ± 0,013
Ecologia	75,8% ± 0,095	43,5% ± 0,063	54,7% ± 0,041
Medicina e saúde	85,8% ± 0,080	62,4% ± 0,071	71,9% ± 0,044
Ciência	90,6% ± 0,010	68,4% ± 0,050	77,9% ± 0,036
Educação	92,9% ± 0,035	55,6% ± 0,040	69,5% ± 0,024
Automobilismo	100% ± 0,000	78,3% ± 0,029	87,8% ± 0,018
Basquete	97,4% ± 0,044	84% ± 0,083	90% ± 0,050
Futebol	87,6% ± 0,003	96,1% ± 00,24	91,6% ± 0,011
Vôlei	100% ± 0,000	70,8% ± 0,260	81,2% ± 0,171
Esportes	97,9% ± 0,005	94,1% ± 0,015	96% ± 0,009
Empregos	74,7% ± 0,091	44,8% ± 0,045	55,6% ± 0,020
Imóveis	84,0% ± 0,035	79,7% ± 0,030	81,7% ± 0,001
Negócios	57,9% ± 0,068	73,1% ± 0,082	64,6% ± 0,073
Finanças	85,2% ± 0,039	91,7% ± 0,017	88,3% ± 0,029
Hardware	87,4% ± 0,025	65,1% ± 0,063	74,5% ± 0,050
Software	78,6% ± 0,021	72,9% ± 0,028	75,6% ± 0,020
Informática	97,6% ± 0,003	90,9% ± 0,015	94,2% ± 0,009
Política	97,1% ± 0,004	87,9% ± 0,014	92,3% ± 0,007
Turismo	92,3% ± 0,013	82,7% ± 0,062	87,1% ± 0,028
Veículos	93,1% ± 0,042	80,1% ± 0,022	86,1% ± 0,031
Agricultura	88,9% ± 0,111	67,7% ± 0,122	76,8% ± 0,119
Pecuária	100% ± 0,000	71,4% ± 0,120	82,9% ± 0,085
Área rural	98,5% ± 0,026	73,1% ± 0,049	83,9% ± 0,035
Macro-média	85,8% ± 0,030	72,9% ± 0,019	77,6% ± 0,020
Micro-média	86,6% ± 0,008	82,8% ± 0,008	84,7% ± 0,008

O desvio padrão demonstra a diferença nos resultados obtidos dos diferentes classificadores. Novamente, com exceção de cinco categorias, o resultado apresenta um desvio padrão para a medida F1, em três categorias, superior a 10%. Apesar de o classificador SVM apresentar um desvio padrão maior que o do *k*-NN, a macro-média e micro-média dos dois grupos mostram desempenho estatisticamente equivalente no teste do qui-quadrado.

O aumento do desvio padrão em categorias contendo poucos documentos de treino indica um princípio de *underfitting* dos classificadores SVM nessas condições. Um dos motivos para isso pode ser a dificuldade em estabelecer vetores de suporte confiáveis em categorias contendo poucos documentos no *corpus* de treino.

Comparando os resultados da macro-média e da micro-média das tabelas 5.1 e 5.4 (grupo α e grupo β), é possível identificar que a macro-média da abrangência do classificador SVM é 6,7% inferior ao classificador *k*-NN. O mesmo se repete na micro-média, em proporções menores. Isso repercute em uma macro-média da medida F1 inferior para o classificador SVM. Em precisão, o classificador SVM apresenta melhores resultados na macro-média, porém o *k*-NN apresenta melhores resultados na micro-média.

Seguindo com a análise do classificador SVM, a seguir é descrito o experimento realizado com o mesmo *corpus* do grupo α.

O resultado do processo de CHT de uma execução do grupo β é apresentado na Tabela 5.5. Nesta tabela, são apresentadas as medidas de precisão, abrangência e F1 para cada uma das categorias. Também são apresentados os números de verdadeiros positivos (|TP|), falsos positivos (|FP|) e falsos negativos (|FN|) para as respectivas categorias. Ao final das tabelas são apresentadas as medidas de macro-média e micro-média.

Tabela 5.5: Resultados do experimento com o classificador SVM

Categorias	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Carnaval	88,9%	80%	84,2%	8	1	2
Cinema	90,9%	66,7%	76,9%	10	1	5
Literatura e livros	64,6%	89,5%	75%	51	28	6
Moda	71,4%	50%	58,8%	5	2	5
Música	72,2%	86,7%	78,8%	26	10	4
Arte e cultura	82,4%	92,8%	87,3%	155	33	12
Ecologia	75%	37,5%	50%	3	1	5
Medicina e saúde	76,9%	64,5%	70,2%	20	6	11
Ciência	90,9%	72,7%	80,8%	40	4	15
Educação	95,2%	51,3%	66,7%	20	1	19
Automobilismo	100%	80%	88,9%	12	0	3
Basquete	100%	91,3%	95,5%	21	0	2
Futebol	87,9%	95,2%	91,4%	80	11	4
Vôlei	100%	62,5%	76,9%	5	0	3
Esportes	98,5%	94,2%	96,3%	130	2	8
Empregos	72,7%	42,1%	53,3%	8	3	11
Imóveis	87,8%	76,6%	81,8%	36	5	11
Negócios	51,1%	63,9%	56,8%	23	22	13
Finanças	83,4%	90,7%	86,9%	136	27	14
Hardware	89,3%	67,6%	76,9%	25	3	12
Software	77,1%	69,8%	73,3%	37	11	16
Informática	98%	92,6%	95,2%	100	2	8
Política	97,2%	86,4%	91,5%	70	2	11
Turismo	93,3%	80%	86,2%	84	6	21
Veículos	88,6%	77,5%	82,7%	31	4	9
Agricultura	77,8%	53,8%	63,6%	7	2	6
Pecuária	100%	76,5%	86,7%	13	0	4
Área rural	95,5%	72,4%	82,4%	21	1	8
Total				1177	188	248
Macro-Média	86%	73,7%	78,4%			
Micro-Média	86,2%	82,5%	84,3%			

O conjunto de verdadeiros positivos, falsos positivos e falsos negativos deste classificador é superior ao mesmo conjunto apresentado para o classificador *k*-NN. A diferença entre os valores da Tabela 5.2 e Tabela 5.5 mostra que o *k*-NN possui um conjunto maior de verdadeiros positivos, enquanto o SVM possui um conjunto maior de falsos negativos.

A comparação quantitativa da tabela acima com a Tabela 5.2 demonstra que, em se tratando da micro-média e macro-média para este *corpus*, o SVM possui um desempenho superior em precisão, enquanto o *k*-NN possui desempenho superior em abrangência.

A Tabela 5.6 considera, separadamente, as categorias do primeiro e segundo níveis da árvore. Essa tabela permite uma análise do desempenho por níveis distintos da árvore, nela os resultados estão apresentados em três blocos: primeiro nível, segundo nível e resultado total.

A análise individual dos níveis demonstra, novamente, um decréscimo de desempenho do primeiro para o segundo nível. Uma característica interessante é que este decréscimo é pro-

Tabela 5.6: Resultado do grupo β com discernimento por níveis hierárquicos

Resultado global do primeiro nível

SVM	Pr	Re	F1	|TP|	|FP|	|FN|
Macro-Média	92,3%	81,1%	85,6%			
Micro-Média	90,6%	86,3%	84,4%			
Total				787	82	125

Resultado global do segundo nível

SVM	Pr	Re	F1	|TP|	|FP|	|FN|
Macro-Média	82,4%	69,7%	74,4%			
Micro-Média	78,6%	76%	77,3%			
Total				390	106	123

Resultado total

SVM	Pr	Re	F1	|TP|	|FP|	|FN|
Macro-Média	86%	73,7%	78,4%			
Micro-Média	86,2%	82,5%	84,3%			
Total				1177	188	248

porcional para as três medidas de avaliação. Novamente, o *ranking* de melhor categoria e a representação hierárquica podem ser os responsável pela depreciação dos resultados.

5.3 Análise dos grupos α e β

A análise conjunta dos grupos é realizada, nesta seção, com o intuito de permitir observar tendências específicas dos classificadores *k*-NN e SVM. O comportamento é analisado levando em consideração os parâmetros dos classificadores, os três *corpora* de treino e teste e as características da coleção.

A Figura 5.1 mostra um gráfico com resultados obtidos na Tabela 5.3 e Tabela 5.6. Ele descreve a macro-média de precisão e abrangência dos grupos α e β.

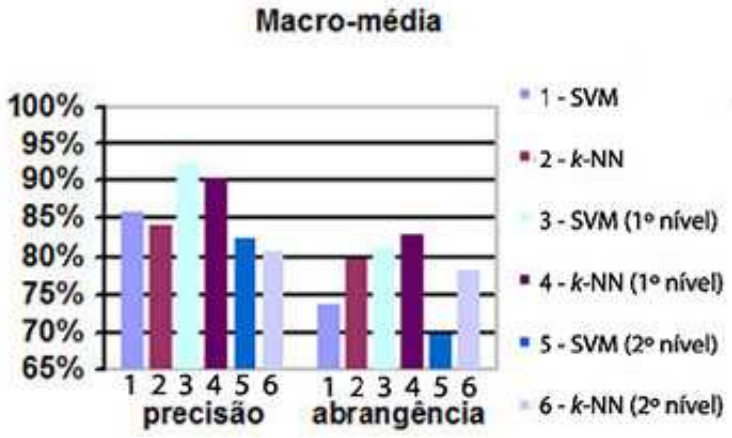

Figura 5.1: Macro-média dos grupos α e β

Com a ajuda da Figura 5.1 é possível identificar que a macro-média da precisão do classificador SVM para o primeiro nível é a mais alta dentre todas e, também, que a macro-média da abrangência do classificador *k*-NN no segundo nível é a mais alta. A macro-média da abrangência do SVM é a mais baixa, especialmente no segundo nível. Em relação à precisão, o

classificador SVM possui desempenho ligeiramente superior ao classificador *k*-NN em todos os níveis. No quesito abrangência, o classificador *k*-NN possui desempenho superior ao classificador SVM. O primeiro nível apresenta os melhores desempenhos, tanto na precisão quanto na abrangência.

A Figura 5.2 mostra um gráfico com resultados obtidos na Tabela 5.3 e Tabela 5.6. Ele descreve a micro-média de precisão e abrangência dos grupos α e β.

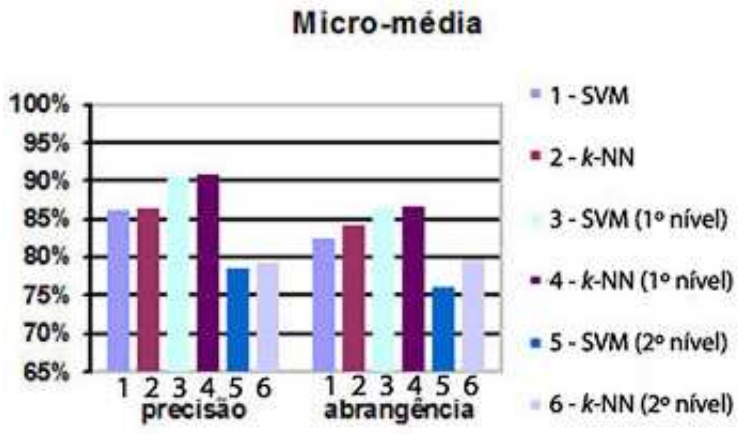

Figura 5.2: Micro-média dos grupos α e β

A micro-média dos classificadores (5.2) é similar em quase todos os aspectos, exceto para a abrangência no segundo nível, onde o classificador SVM apresentou resultado 5% inferior ao *k*-NN. Todavia, o classificador *k*-NN apresenta um desempenho melhor que o classificador SVM, em todos os níveis.

Aparentemente, os classificadores possuem desempenho equivalente de precisão, tanto na macro-média quanto na micro-média. Na abrangência, no entanto, o classificador *k*-NN se sobressai em relação ao classificador SVM, especialmente na macro-média.

O gráfico da Figura 5.3 mostra uma comparação entre os pares de valores de precisão e abrangência; esta comparação permite visualizar uma tendência dos classificadores *k*-NN e SVM para as execuções dos grupos α e β, respectivamente. Este gráfico mostra a tendência de um desempenho melhor em abrangência para o classificador *k*-NN em relação ao classificador SVM. A precisão das execuções dos dois grupos aparenta uma equivalência entre os dois classificadores, com uma pequena vantagem em desempenho para o classificador SVM.

Um agravante no desempenho, para a macro-média da abrangência, que se mostra inferior do grupo β, é a diferença dos resultados obtidos em categorias em níveis inferiores da árvore. Como dito anteriormente, a dificuldade em determinar os melhores vetores de suporte, em categorias com poucos documentos, pode ser o motivo desse impacto negativo. Na macro-média o impacto é significativo, na micro-média nem tanto. Isso indica que o classificador SVM é sensível à quantidade de documentos na etapa de treinamento, sem afetar significativamente o processo de CHT. Em um *corpus* balanceado, com número suficiente de documentos de treino, essa dificuldade não deve ser relevante.

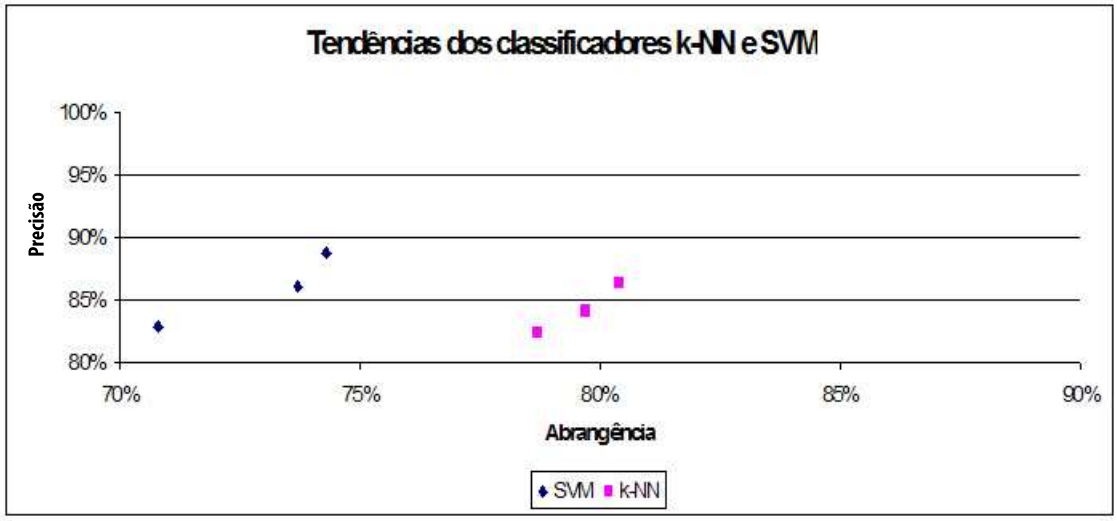

Figura 5.3: Tendência da macro-média dos grupos α e β

A análise das tabelas de contingência dos grupos α e β, mostra que o número de falsos negativos é superior ao número de falsos positivos. Em relação aos níveis da árvore, o comportamento é o mesmo. Isso é resultado da combinação das características da coleção, com a escolha por monocategorização do melhor documento do *ranking*. A coleção é composta por documentos que pertencem a mais de uma categoria. Os classificadores locais realizam monocategorização. Então, com essa estratégia de categorização monocategorial baseada em *ranking*, alguns documentos serão categorizados em menos categorias que o esperado. Essas características afetam principalmente as categorias do segundo nível, uma vez que o classificador do nodo raiz é monocategorial e os documentos podem ter mais de uma categoria. Por isso, os classificadores do primeiro nível são comprometidos pelo resultado do classificador no nodo raiz.

5.3.1 Análise da coleção

Nesta subseção são apresentadas características da coleção Folha-RIcol que contribuem, aparentemente, para o desempenho do processo de CHT. A Tabela 5.7 apresenta as categorias que possuem os melhores resultados individuais de precisão, abrangência e F1.

Uma das primeiras conclusões observadas a partir da Tabela 5.7 é a categoria "esportes" como sendo aquela que possui os melhores resultados, aparecendo em seis ocasiões com uma das três melhores classificações, tanto na precisão e abrangência como na medida F1. Em segundo lugar estão as categorias "futebol" e "vôlei" , aparecendo três vezes cada.

De uma análise das categorias em que as notícias são mais corretamente categorizadas, principalmente "Esportes", "Futebol" e "Vôlei", obtém-se as seguintes constatações:

1. Todas elas possuem um vocabulário próprio e distinto do restante das notícias;

2. No vocabulário aparecem constantemente nomes de times (no mínimo uma vez para cada

Tabela 5.7: Colocação das categorias com melhor desempenho

Desempenho em categorias do grupo α			
	Pr	**Re**	**F1**
1o	Esportes	Futebol	Esportes
2o	Vôlei	Esportes	Vôlei
3o	Automobilismo	Informática	Informática
Desempenho em categorias do grupo β			
	Pr	**Re**	**F1**
1o	Automobilismo	Futebol	Esportes
2o	Pecuária	Esportes	Informática
3o	Vôlei	Arte e cultura	Futebol

time e mais de uma vez para um dos times se ele possui destaque na notícia); no vocabulário aparecem constantemente nomes de jogadores; tanto os nomes de times como os de jogadores parecem ser interessantes na discriminação das categorias;

3. Essas categorias possuem uma grande quantidade de documentos, o que resulta em um treinamento adequado do classificador.

Dessas observações se deduz que a caracterização do conteúdo é mais fácil nessas categorias. Em outras, onde os valores não foram os melhores, não é tão fácil identificar termos comuns.

Nos casos em que as categorias apresentam um desempenho inferior à media das outras categorias pode-se observar que o erro está associado a outras categorias. Por exemplo, a categoria "software" apresenta uma quantidade significativa de falsos negativos devido a, pelo menos, dois fatores:

- Quantidade de atributos - os documentos dessa categoria estão constantemente associados à categoria "hardware", principalmente, em descrições de produtos ou novidades. Como essas descrições relatam uma série de periféricos (*hardware*) e apenas um Sistema Operacional (*software*), os classificadores tendem a categorizar os documentos na categoria "hardware" porque a freqüência dos atributos associados à categoria "hardware" é superior à freqüência dos atributos associados à categoria "software";

- Tecnologia - diversas outras áreas além da informática mencionam tecnologias ligadas ao *software*. Isso ocorre, por exemplo, na ciência e educação. Por isso, em alguns casos, a terminologia usada nessas áreas pertence ao domínio da área do *software*. Muitos dos avanços na ciência envolvem o uso de *software*. Na educação, existem muitos projetos governamentais com apoio de *software* e também existe uma crescente aplicação de *software* ligado a instituições de ensino.

Intuitivamente, esses fatores levam a considerar a categoria *software* como secundária, em relação a outras categorias. Por exemplo, na educação um projeto ou aplicação é a notícia principal, e a tecnologia pode ser vista como um fator secundário. Por isso, a ligação da categoria "software" com outras áreas prejudica uma categorização correta, na mesma, quando a estratégia de categorização é baseada no melhor *ranking* (monocategorial).

A combinação da estratégia de categorização de limiar por *ranking* e os fatores descritos acima inviabilizam que, para esta coleção, um ajuste de parâmetros, ou outra forma de melhorar o desempenho, venham a influenciar significativamente para a redução no conjunto de falsos negativos da categoria "software".

5.4 Considerações sobre o capítulo

Este capítulo apresentou um conjunto de experimentos realizados com a coleção de textos Folha-RIcol. Os experimentos foram executados em dois grupos distintos, α e β, com a distinção por classificadores *k*-NN e SVM, respectivamente. Além dos resultados, os pontos fortes e fracos de cada um dos classificadores, encontrados durante a execução dos experimentos, foram ressaltados e analisados, buscando identificar possíveis causas e discutir possíveis soluções.

Os experimentos relatados neste capítulo servem como uma base para a comparação dos experimentos do Capítulo 6. Além destes experimentos, outros experimentos foram realizados. Um outro experimento realizado foi a execução dos classificadores com uma proporção, no *corpus* de treino e no *corpus* de teste, de 60% por 40%, respectivamente. Este experimento não teve prosseguimento e não chegou a ser relatado nesta dissertação porque o comportamento dos classificadores não foi alterado, com exceção de uma pequena depreciação nos resultados.

O próximo capítulo é dedicado à combinação de classificadores com ênfase na heurística proposta no Capítulo 4.

Capítulo 6

Experimentos combinando classificadores

Este capítulo é voltado aos experimentos que adotam a combinação de classificadores no processo de CHT. Os classificadores usados nas duas combinações são o *k*-NN e o SVM. A metodologia dos experimentos é a descrita no Capítulo 4. A avaliação dos experimentos consiste em uma análise e comparação dos resultados entre si e, também, com os resultados do capítulo anterior (grupos α e β). Ao final são apresentadas algumas considerações sobre os resultados obtidos, identificando problemas e considerando possíveis soluções.

Este capítulo tem por objetivo principal analisar a eficiência da heurística proposta no Capítulo 4. Para tanto, os experimentos utilizam dois métodos combinatórios de classificadores: a votação e a heurística *k*-NN+SVM. É importante ressaltar que qualquer combinação de classificadores agrega em si um aumento no processamento, seja ele significativo ou não. Não é o objetivo deste trabalho analisar a otimização do processo ou dos algoritmos.

Para realizar o processo de CHT, um percentual de 70% do *corpus* é utilizado para treino e o restante para testes. Este percentual desigual é necessário para priorizar o treinamento do classificador, caso contrário pode ocorrer *underfitting*.

Na primeira etapa do processo, o pré-processamento, são removidas as palavras constantes na *stoplist* e palavras com uma freqüência inferior a três por documento. A *stoplist* é composta pelas palavras mais comumente encontradas e que não agregam um valor significativo para a categorização.

Após o pré-processamento, os documentos são armazenados em arquivos no formato ARFF utilizado pela ferramenta WEKA. 70% dos documentos relativos a uma categoria são obtidos aleatoriamente do total de documentos da mesma, para gerar o *corpus* de treino. O restante dos documentos é utilizado para os testes, também armazenado no formato ARFF.

Para evitar que o desempenho dos classificadores seja prejudicado por motivos alheios à combinação, tanto os *corpora* de treino e teste como os parâmetros dos classificadores são mantidos constantes, em relação aos experimentos do capítulo anterior, nos experimentos detalhados neste capítulo. Uma modificação nos documentos, atributos ou parâmetros dos classificadores pode resultar em uma diferença no mecanismo de inferência dos classificadores. Essa alteração modificaria o desempenho de maneira imprópria para a avaliação desejada e, deve ser evitada.

Assim como nos experimentos relatados no Capítulo 5, são utilizados sete classificadores entre as categorias que constituem nodos não-folha.

A avaliação do processo de CHT consiste na repetição do método *hold-out* com os mesmos *corpora* de treino e teste obtidos nos experimentos do Capítulo 5.

Os experimentos deste capítulo estão divididos em dois grupos: grupo χ e grupo δ.

6.1 Grupo χ - Combinação por voto

A maneira mais simples de combinar o resultado de classificadores é pelo voto majoritário, sem a atribuição de peso. Este método é empregado a fim de realizar uma comparação com a heurística proposta nesta dissertação, além da comparação com os classificadores *k*-NN e SVM.

As três execuções do grupo χ, descritas nesta seção, possuem os mesmos *corpora* de treino e teste, citados anteriormente. Da mesma forma que nos classificadores anteriores, a Tabela 6.1 descreve a média e o desvio padrão dos resultados de precisão, abrangência e medida F1 dessas três execuções.

Tabela 6.1: Média e desvio padrão dos classificadores - grupo χ

Categoria	$\overline{Pr}$	$\overline{Re}$	$\overline{F1}$
Carnaval	88,6% ± 0,025	67,9% ± 0,062	76,7% ± 0,029
Cinema	83,9% ± 0,068	54,9% ± 0,257	64,9% ± 0,200
Literatura e livros	72,8% ± 0,141	83,7% ± 0,023	77,5% ± 0,088
Moda	81,9% ± 0,033	52,8% ± 0,285	61,9% ± 0,206
Música	66,5% ± 0,154	79,2% ± 0,106	72,2% ± 0,135
Arte e cultura	92,4% ± 0,031	94% ± 0,016	93,2% ± 0,024
Ecologia	75,4% ± 0,069	44,6% ± 0,155	55,5% ± 0,137
Medicina e saúde	92,1% ± 0,068	72,7% ± 0,079	81,3% ± 0,076
Ciência	94% ± 0,030	72,7% ± 0,105	81,8% ± 0,077
Educação	96,8% ± 0,028	78,6% ± 0,077	86,6% ± 0,058
Automobilismo	100% ± 0,000	83% ± 0,032	90,8% ± 0,018
Basquete	93,6% ± 0,003	95,9% ± 0,036	94,7% ± 0,016
Futebol	94,1% ± 0,051	97,3% ± 0,006	95,6% ± 0,023
Vôlei	100% ± 0,000	100% ± 0,000	100% ± 0,000
Esportes	98,5% ± 0,013	94,5% ± 0,009	96,4% ± 0,011
Empregos	85,5% ± 0,074	60% ± 0,173	70,1% ± 0,142
Imóveis	85,8% ± 0,086	87,9% ± 0,068	86,8% ± 0,077
Negócios	74,5% ± 0,081	81,2% ± 0,074	77,7% ± 0,078
Finanças	92,2% ± 0,048	92,6% ± 0,015	92,4% ± 0,031
Hardware	88,4% ± 0,052	84,1% ± 0,047	86,2% ± 0,050
Software	85,2% ± 0,072	74,5% ± 0,127	79,4% ± 0,103
Informática	98,6% ± 0,005	94,3% ± 0,022	96,4% ± 0,014
Política	96,9% ± 0,027	96,4% ± 0,009	96,6% ± 0,009
Turismo	92,7% ± 0,035	89,4% ± 0,014	91,1% ± 0,024
Veículos	96,3% ± 0,032	94,6% ± 0,047	95,5% ± 0,039
Agricultura	94,9% ± 0,044	74,3% ± 0,065	83,3% ± 0,058
Pecuária	96,3% ± 0,032	83,6% ± 0,024	89,5% ± 0,001
Área rural	95,7% ± 0,038	80,8% ± 0,042	87,6% ± 0,040
Macro-média	89,8% ± 0,046	80,9% ± 0,068	84,3% ± 0,063
Micro-média	91,3% ± 0,046	87,3% ± 0,043	89,3% ± 0,044

O desempenho na macro-média e micro-média das medidas de avaliação, na combinação por voto dos classificadores, é superior tanto no caso do classificador *k*-NN como no caso do classificador SVM. Esse resultado por si só não é significativo: o desvio padrão deste classificador, nas mesmas circunstâncias, é muito superior ao desvio padrão dos outros classificadores. Isso significa uma variação dos resultados obtidos entre as execuções, que é demonstrada pela

grande quantidade de categorias com um desvio padrão superior a 10%. Analisar o resultado obtido pela melhor execução, nesse caso, não permite uma compreensão correta do desempenho desse classificador. No entanto, uma análise dessa discrepância permite identificar os motivos que levaram ao resultado obtido, apresentado na Tabela 6.1.

Nas medidas de macro-média e micro-média o desvio padrão encontrado é o mais alto dentre os observados nesta dissertação. Levando em consideração esses valores, a variação possível dos resultados ocorre entre 4,3% e 6,8%, dependendo da medida de avaliação. Por exemplo, a macro-média da medida F1 é de 84,3% com desvio padrão de 6,3% o que leva a considerar o real valor dessa medida como entre 78% e 90,6%. Essa variação de 12,6% do valor verdadeiro é muito elevada e leva a uma reflexão.

Um dos motivos que pode levar à diferença dos resultados entre as execuções é que, na segunda execução, os resultados são muito otimistas. O desempenho esperado pela combinação dos classificadores é positivo para a segunda execução, enquanto nas duas outras execuções não há uma diferença significativa. Deduz-se que, nessa situação, a escolha dos documentos de treino influenciaria diretamente no resultado, como mostra a Tabela 6.2. A análise da tabela de contingência dessas execuções mostra que a eficiência da categorização melhora nas categorias "literatura e livros", "música", "arte e cultura", "empregos", "finanças" e "negócios" onde se concentra a maioria dos falsos positivos. No entanto, essas categorias apresentam o maior desvio padrão.

Outro ponto crucial é o conjunto de execuções realizadas. Três execuções podem não determinar o desempenho real dessa combinação de classificadores. Entretanto, devido ao tempo necessário para a preparação e execução do processo de CHT envolvendo a combinação de classificadores, esse é o número comumente adotado para os experimentos. É preciso considerar que um número maior de execuções tende a diminuir a variação dos resultados e produzir, assim, um resultado mais condizente com a realidade. Como esse não é o foco dessa dissertação e sim um ponto de comparação com a combinação de classificadores expressa pelo grupo δ, os valores encontrados são considerados suficientes para uma análise.

Passando à análise detalhada de uma execução do grupo χ, tem-se os resultados exibidos na Tabela 6.2.

Nos dados mostrados na Tabela 6.2, o número de falsos positivos sofre um decréscimo substancial. Já a redução no número de falsos negativos não é tão significativa. O fator que diferencia os dois conjuntos da tabela de contingência é a escolha da categorização pelo melhor escore. Considerando essa estratégia, o conjunto de falsos positivos pode ser minimizado com alterações nas condições e heurísticas dos experimentos. Já o conjunto de falsos negativos não sofre variações devido às razões anteriormente apresentadas (Seção 5.3) e à estratégia de melhor escore.

A seguir, o Teste estatítisco Z é utilizado para verificar a probabilidade de uma nova ocorrência de resultados semelhantes aos da Tabela 6.2.

Com base na macro-média da medida F1 (84,3%) e em seu desvio padrão (0,063) (Ta-

Tabela 6.2: Resultado da votação dos classificadores

Categorias	Pr	Re	F1	|TP|	|FP|	|FN|
Carnaval	85,7%	75,0%	80,0%	6	1	2
Cinema	91,7%	84,6%	88,0%	11	1	2
Literatura e livros	89,1%	86,4%	87,7%	57	7	9
Moda	85,7%	85,7%	85,7%	6	1	1
Música	84,2%	91,4%	87,7%	32	6	3
Arte e cultura	95,9%	95,9%	95,9%	165	7	7
Ecologia	83,3%	62,5%	71,4%	5	1	3
Medicina e saúde	100,0%	81,8%	90,0%	18	0	4
Ciência	97,5%	84,8%	90,7%	39	1	7
Educação	100,0%	87,5%	93,3%	28	0	4
Automobilismo	100,0%	86,7%	92,9%	13	0	2
Basquete	93,3%	100,0%	96,6%	14	1	0
Futebol	100,0%	96,6%	98,3%	85	0	3
Vôlei	100,0%	100,0%	100,0%	10	0	0
Esportes	100,0%	95,5%	97,7%	128	0	6
Empregos	94,1%	80,0%	86,5%	16	1	4
Imóveis	95,7%	95,7%	95,7%	44	2	2
Negócios	83,9%	89,7%	86,7%	26	5	3
Finanças	97,8%	94,3%	96,0%	132	3	8
Hardware	94,4%	89,5%	91,9%	34	2	4
Software	93,5%	89,2%	91,3%	58	4	7
Informática	99,2%	96,8%	98,0%	121	1	4
Política	100,0%	95,3%	97,6%	81	0	4
Turismo	96,8%	91,1%	93,9%	92	3	9
Veículos	100,0%	100,0%	100,0%	38	0	0
Agricultura	100,0%	81,8%	90,0%	9	0	2
Pecuária	100,0%	80,8%	89,4%	21	0	5
Área rural	100,0%	85,7%	92,3%	30	0	5
Total				**1319**	**47**	**110**
Macro-média	95,1%	88,7%	91,6%			
Micro-média	96,6%	92,3%	94,4%			

bela 6.1) é possível verificar a probabilidade de uma nova ocorrência (x) com uma macro-média da medida F1 igual ou superior a 91,6%, realizando o cálculo abaixo:

$$prob(x \geq 91,6\%) =$$

$$prob(z \geq (0,916 - 0,843)/0,063) =$$

$$prob(z \geq 1,158) =$$

$$1 - prob(z \leq 1,158) =$$

$$1 - 0,874 =$$

$$0,126$$

Obtém-se um valor de Z $\geq$ 1,158, que, convertendo na tabela Z corresponde a 0,874, resultando em uma probabilidade de 12,6% para $x \geq 91,6\%$. Esta é uma probabilidade muito baixa; mesmo levando em consideração a quantidade de repetições e um intervalo de confiança de 95%, o valor real da macro-média da medida F1 está entre 77,6% e 91,4%, demonstrando que o valor 91,6% está acima do limiar real dessa medida. Dito isto, os resultados da Tabela 6.2 não são estatisticamente significativos para a avaliação desta combinação de classificadores. Dificilmente irá repetir-se essa situação. Portanto, a hipótese de uma nova ocorrência com resultados semelhantes aos da Tabela 6.2 deve ser rejeitada.

A Tabela 6.3 considera, separadamente, as categorias do primeiro e segundo nível da árvore. Como descrito anteriormente, os resultados expressos por esta execução devem ser rejeitados por não representarem um valor estatisticamente significativo. Entretanto, a análise destes resultados pode encaminhar possíveis identificações de problemas e soluções.

Tabela 6.3: Resultado do grupo χ com discernimento por níveis hierárquicos

Resultado global do primeiro nível

votação	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	98,7%	92,6%	95,5%			
Micro-Média	98,2%	94,0%	96,1%			
Total				854	15	54

Resultado global do segundo nível

votação	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	93%	86,5%	89,4%			
Micro-Média	93,5%	89,2%	91,3%			
Total				465	32	56

Resultado total

votação	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	95,1%	88,7%	91,6%			
Micro-Média	96,6%	92,3%	94,4%			
Total				1319	47	110

Os dados resultantes da análise por níveis hierárquicos apresentam uma pequena diferença no conjunto de falsos negativos e uma grande diferença no conjunto de verdadeiros positivos e falsos positivos.

O decréscimo no desempenho do primeiro nível para o segundo nível existe mas, aparentemente, a diferença é menor do que o decréscimo expresso nos classificadores *k*-NN e SVM, grupo α e grupo β. O comportamento observado em experimentos anteriores mantém-se, ou seja, em níveis superiores da hierarquia o desempenho dos classificadores é melhor que em níveis inferiores.

Cabe ressaltar um aspecto interessante: nos classificadores dos grupos α e β estes mesmos *corpus* de treino e *corpus* de teste não apresentam o melhor desempenho do classificador, enquanto que a combinação por voto apresenta os melhores resultados para estes mesmos *corpora*. Analisando a tabela de contingência, é possível verificar que o conjunto de falsos positivos diminui, na votação, nas categorias *arte e cultura*, *finanças* e *turismo* e em categorias em nodos filhos da categoria *arte e cultura*.

6.2 Grupo δ - Heurística *k*-NN+SVM

Esta seção descreve os experimentos compostos por três execuções da heurística *k*-NN+SVM, grupo δ, com os mesmos três diferentes conjuntos de *corpora* de treino e teste utilizados nos experimentos anteriores.

A heurística *k*-NN+SVM de combinação prevê que:

- nodos com mais de dois filhos utilizem o classificador *k*-NN;

- nodos não-folhas com dois ou menos filhos utilizem, respectivamente, classificadores multicategoriais e monocategoriais SVM.

A Tabela 6.4 exibe a média e o desvio padrão dos resultados de precisão, abrangência e da medida F1, individual para cada categoria do grupo δ. Ao final são mostradas a macro-média e micro-média global, com as mesmas medidas de avaliação.

Tabela 6.4: Média e desvio padrão do grupo δ com a heurística *k*-NN+SVM

Categorias	$\overline{Pr}$	$\overline{Re}$	$\overline{F1}$
Carnaval	79,7% ± 0,114	91,1% ± 0,078	84,6% ± 0,059
Cinema	83,3% ± 0,030	84,1% ± 0,071	83,5% ± 0,019
Literatura e livros	76,9% ± 0,037	77,2% ± 0,053	77% ± 0,045
Moda	68,2% ± 0,233	65,2% ± 0,197	66,4% ± 0,204
Música	77,6% ± 0,055	85,5% ± 0,028	81,4% ± 0,041
Arte e cultura	90,7% ± 0,027	88,7% ± 0,008	89,7% ± 0,009
Ecologia	68,3% ± 0,161	34,5% ± 0,051	45% ± 0,043
Medicina e saúde	83,4% ± 0,049	60,3% ± 0,029	69,8% ± 0,006
Ciência	91,7% ± 0,027	64% ± 0,032	75,4% ± 0,026
Educação	91% ± 0,039	70,4% ± 0,083	79,2% ± 0,065
Automobilismo	95,4% ± 0,040	89,2% ± 0,036	92,1% ± 0,021
Basquete	91,8% ± 0,053	93,2% ± 0,072	92,5% ± 0,062
Futebol	91,1% ± 0,009	95,7% ± 0,012	93,4% ± 0,010
Vôlei	96,7% ± 0,058	92,5% ± 0,066	94,4% ± 0,051
Esportes	97% ± 0,001	95,1% ± 0,009	96% ± 0,005
Empregos	62,2% ± 0,165	51% ± 0,065	55,2% ± 0,067
Imóveis	79,6% ± 0,051	86,7% ± 0,006	82,9% ± 0,026
Negócios	65,7% ± 0,107	69,5% ± 0,098	67,5% ± 0,100
Finanças	87,2% ± 0,027	88,8% ± 0,013	88% ± 0,020
Hardware	86,8% ± 0,011	65,9% ± 0,073	74,7% ± 0,045
Software	75,9% ± 0,008	74,3% ± 0,057	75,1% ± 0,033
Informática	93,9% ± 0,011	94,1% ± 0,017	94% ± 0,013
Política	90,5% ± 0,034	93,2% ± 0,024	91,8% ± 0,007
Turismo	89,1% ± 0,036	81,5% ± 0,051	85,1% ± 0,024
Veículos	92% ± 0,006	90,4% ± 0,055	91,2% ± 0,030
Agricultura	79,8% ± 0,142	68,7% ± 0,141	73,8% ± 0,141
Pecuária	90,5% ± 0,040	78,6% ± 0,106	83,7% ± 0,055
Área rural	87,4% ± 0,062	76,9% ± 0,044	81,8% ± 0,046
Macro-média	84,4% ± 0,026	78,8% ± 0,013	80,9% ± 0,018
Micro-média	87,5% ± 0,011	84,2% ± 0,009	85,9% ± 0,010

As medidas de avaliação para a macro-média e a micro-média dos resultados da Tabela 6.4 possuem valores quase idênticos aos resultados obtidos na tabela equivalente para o grupo α (classificador *k*-NN).

Um dos motivos que pode ter influenciado nesse resultado é o maior número de classificadores *k*-NN do que classificadores SVM. No total são quatro classsificadores *k*-NN e três classificadores SVM. Outro motivo é que os classificadores *k*-NN estão posicionados nas categorias onde existe o maior número de documentos. Esses valores representam em um maior volume de categorizações de documentos nos classificadores *k*-NN, sejam eles para os conjuntos de verdadeiros positivos, falsos positivos ou falsos negativos.

A análise individual de uma execução do grupo δ exibe as medidas de precisão, abrangência e medida F1, em cada uma das categorias, expostas na Tabela 6.5. Também são apresentados os números de verdadeiros positivos (|TP|), falsos positivos (|FP|) e falsos negativos (|FN|) para as respectivas categorias. Ao final da tabela são apresentadas as medidas de macro-média e micro-média.

Tabela 6.5: Resultado da heurística *k*-NN+SVM

Categorias	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Carnaval	76,9%	100%	87%	10	3	0
Cinema	85,7%	80%	82,8%	12	2	3
Literatura e livros	73,7%	73,7%	73,7%	42	15	15
Moda	72,7%	80%	76,2%	8	3	2
Música	75,8%	83,3%	79,4%	25	8	5
Arte e cultura	90,8%	88,6%	89,7%	148	15	19
Ecologia	75%	37,5%	50%	3	1	5
Medicina e saúde	85,7%	58,1%	69,2%	18	3	13
Ciência	94,6%	63,6%	76,1%	35	2	20
Educação	88,9%	61,5%	72,7%	24	3	15
Automobilismo	92,9%	86,7%	89,7%	13	1	2
Basquete	95,8%	100%	97,9%	23	1	0
Futebol	92%	96,4%	94,2%	81	7	3
Vôlei	100%	87,5%	93,3%	7	0	1
Esportes	97,1%	95,7%	96,4%	132	4	6
Empregos	52,4%	57,9%	55%	11	10	8
Imóveis	80,4%	87,2%	83,7%	41	10	6
Negócios	53,8%	58,3%	56%	21	18	15
Finanças	84,5%	87,3%	85,9%	131	24	19
Hardware	86,2%	67,6%	75,8%	25	4	12
Software	75%	67,9%	71,3%	36	12	17
Informática	92,6%	92,6%	92,6%	100	8	8
Política	90,5%	93,8%	92,1%	76	8	5
Turismo	92,5%	81,9%	86,9%	86	7	19
Veículos	92,5%	92,5%	92,5%	37	3	3
Agricultura	63,6%	53,8%	58,3%	7	4	6
Pecuária	86,7%	76,5%	81,2%	13	2	4
Área rural	80,8%	72,4%	76,4%	21	5	8
Total				1186	183	239
Macro-Média	83,2%	77,9%	79,8%			
Micro-Média	86,6%	83,2%	84,9%			

É nítido que os resultados mostrados na Tabela 6.5 são condizentes com os resultados da Tabela 6.4, que também faz parte das execuções do grupo δ. Diferente dos resultados do grupo χ, os valores encontrados para esta tabela possuem uma variação inferior a 1% com relação ao esperado, com base no desvio padrão da Tabela 6.4. Os resultados da precisão e abrangência para a macro-média demonstram um desempenho intermediário em relação ao desempenho dos classificadores *k*-NN e SVM. A medida de abrangência, nessa combinação, é superior aos resultados do grupo β e, a precisão possui desempenho próximo à melhor execução do classificador *k*-NN. Cabe ressaltar que esta execução não consitui o melhor desempenho encontrado para esta combinação.

Um aspecto interessante nos resultados da Tabela 6.5 é que, na maioria das categorias em que é aplicado o classificador SVM, o desempenho da medida de precisão dessas categorias decai, em relação ao classificador SVM do grupo β. Um bom desempenho para a medida de precisão é uma das características apresentadas nos experimentos do grupo β, entretanto nos resultados do grupo δ apenas a categoria *ciência* apresenta uma melhora no desempenho.

A Tabela 6.6 considera, separadamente, as categorias do primeiro e segundo nível da árvore. Lembrando, são quatro classificadores *k*-NN e três classificadores SVM. O esperado é um desempenho equivalente ao classificador *k*-NN no primeiro nível, com um bom desempenho em abrangência e, uma melhora na precisão nas categorias em que se encontram o classificador SVM, uma vez que a abrangência no primeiro nível deve ser melhor que no grupo β.

A heurística da abordagem combinatória não prevê nenhum classificador SVM para o pri-

Tabela 6.6: Resultado do grupo δ com discernimento por níveis hierárquicos

Resultado global do primeiro nível

k-NN+SVM	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	90,4%	82,9%	86,1%			
Micro-Média	90,9%	86,6%	88,7%			
Total				790	79	122

Resultado global do segundo nível

k-NN+SVM	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	79,1%	75,1%	76,4%			
Micro-Média	79,2%	77,2%	78,2%			
Total				396	104	117

Resultado total

k-NN+SVM	**Pr**	**Re**	**F1**	**|TP|**	**|FP|**	**|FN|**
Macro-Média	83,2%	77,9%	79,8%			
Micro-Média	86,6%	83,2%	84,9%			
Total				1186	183	239

meiro nível, por isso os valores do classificador *k*-NN e da heurística *k*-NN+SVM são idênticos (tabelas 5.1 e 6.4). Já no segundo nível, a heurística apresenta um desempenho, aparentemente, pior que o desempenho do classificador *k*-NN, e melhor que o desempenho do classificador SVM. Esse desempenho é o esperado, uma vez que os classificadores *k*-NN são predominantes em número. Entretanto, não é possível observar nenhuma diferença significativa no desempenho da medida de precisão, o que indica uma depreciação no resultado da heurística em relação ao resultado encontrado individualmente no classificador SVM. Conclui-se que os resultados pouco atrativos encontrados aqui estão próximos aos resultados individuais do classificador *k*-NN.

Novamente, o desempenho do classificador é melhor em níveis superiores da árvore. No entanto, a depreciação dos resultados por níveis hierárquicos continua, uma vez que as duas características, relatadas anterioremente, são constantes nos quatro experimentos.

6.3 Análise dos grupos χ e δ

A análise conjunta dos grupos é realizada, nesta seção, com o intuito de permitir observar tendências específicas do comportamento da combinação dos classificadores *k*-NN e SVM. O comportamento é analisado levando em consideração os parâmetros dos classificadores, os três *corpora* de treino e teste e as características da coleção.

O gráfico da Figura 6.1 mostra uma comparação entre os pares de valores de precisão e abrangência que descreve uma tendência da combinação dos classificadores *k*-NN e SVM, por voto e para a heurística *k*-NN+SVM, para as execuções dos grupos χ e δ, respectivamente.

Este gráfico mostra a tendência de um desempenho melhor em abrangência no grupo δ em relação ao grupoχ. A precisão das execuções dos dois grupos aparenta uma vantagem no desempenho para a combinação por voto.

Um agravante no desempenho, para a macro-média da abrangência, inferior do grupo χ,

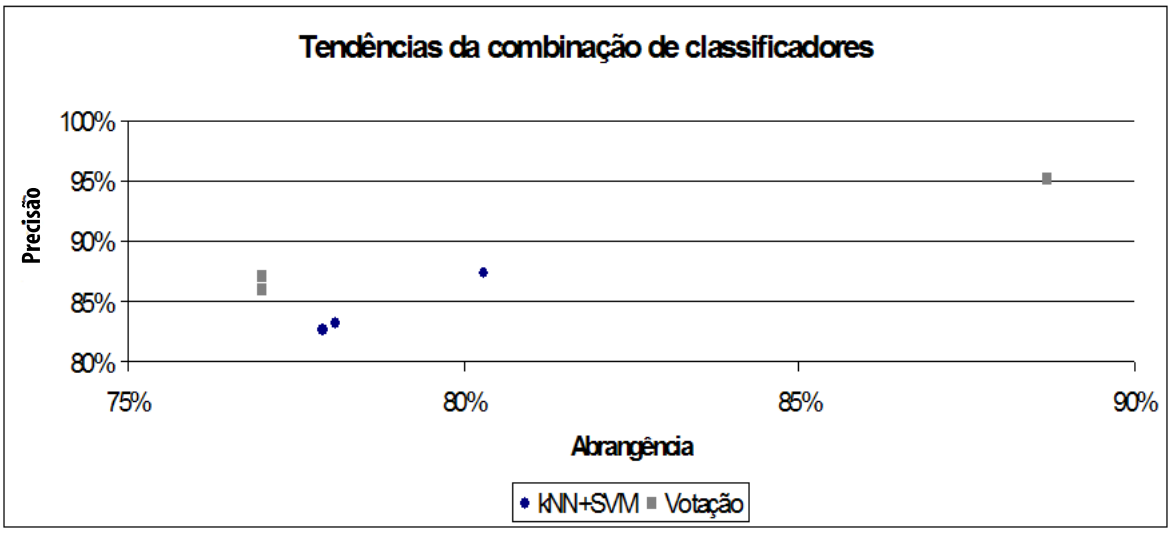

Figura 6.1: Tendência da macro-média dos grupos χ e δ

pode ser a influência que o voto do classificador SVM agrega consigo os problemas inerentes a esses classificadores. Como já foi explicado anteriormente, a dificuldade em determinar os melhores vetores de suporte, em categorias com poucos documentos, pode ser o motivo desse impacto negativo. Na macro-média o impacto é significativo, na micro-média nem tanto. Isso indica que o classificador SVM é sensível à quantidade de documentos na etapa de treinamento, sem afetar significativamente o processo de CHT. Em um *corpus* balanceado, com número suficiente de documentos de treino, essa dificuldade não deve ser relevante. No grupo δ esse problema não é um agravante ao resultado final porque ora a decisão é do classificador SVM, ora a decisão é do classificador *k*-NN.

A análise das tabelas de contingência dos grupos χ e δ, mostra que o conjunto de falsos negativos é superior ao conjunto de falsos positivos. Em relação aos níveis da árvore o comportamento é o mesmo. Isso é resultado da combinação das características da coleção com a escolha por monocategorização do melhor documento do *ranking*. Como os documentos pertencem, em média, a 1,5 categorias, ampliar a estratégia de categorização para categorizar o documento na primeira e segunda categoria do *ranking* deve aumentar significativamente o número de falsos negativos, sem garantir um decréscimo no conjunto de falsos positivos. Essas características afetam principalmente as categorias do segundo nível, uma vez que o classificador do nodo raiz é monocategorial e os documentos podem ter mais de uma categoria. Por isso, os classificadores do primeiro nível são comprometidos pelo resultado do classificador no nodo raiz.

6.4 Considerações sobre o capítulo

Este capítulo apresentou um conjunto de experimentos realizados com a combinação de classificadores. Os experimentos foram executados em dois grupos distintos, χ e δ, com a distinção de combinação por voto e a heurística *k*-NN+SVM, respectivamente. Além dos resultados encontrados durante a execução dos experimentos, os mesmos foram analisados e comparados com os experimentos do capítulo anterior, buscando identificar possíveis problemas e discutir possíveis soluções. Em uma comparação, o Teste Z foi aplicado em uma das execuções do grupo χ, que apresentou valores fora do padrão das outras execuções do mesmo grupo.

Os experimentos relatados no Capítulo 5 serviram para embasar a comparação dos experimentos deste capítulo. A compreensão das características e dificuldades encontradas nos classificadores e na coleção, adquiradas no capítulo anterior, ajudaram na identificação e explicação das dificuldades encontradas neste capítulo.

Para sumarizar os resultados obtidos com os quatro grupos de experimentos é apresentada a Tabela 6.7.

Tabela 6.7: Comparação dos quatro grupos

	Macro-média			**Micro-média**		
	Pr	Re	F1	Pr	Re	F1
***k*-NN**	84,1%	79,7%	81,2%	79,3%	79,7%	79,5%
SVM	86%	73,7%	78,4%	86,2%	82,5%	84,3%
votação	95,1%	88,7%	91,6%	96,6%	92,3%	94,4%
***k*-NN+SVM**	83,2%	77,9%	79,8%	86,6%	83,2%	84,9%

Em uma comparação direta, a heurística *k*-NN+SVM não apresentou um ganho de desempenho em relação à votação. Os resultados encontrados para a heurística estão muito próximos aos resultados dos classificadores *k*-NN e SVM, aplicados individualmente. Exceto para a votação, não houve um ganho no desempenho. Alguns dos motivos que podem ter ocasionado esta situação são:

- as características da coleção. No Capítulo 5 foi apresentada uma análise da coleção que revelou a diferença na quantidade de documentos por categoria e o emprego de nomes próprios, específicos para algumas categorias, como duas características que modificam o desempenho dos classificadores.

- as características inerentes aos classificadores. As características do classificador *k*-NN, em obter melhor desempenho para a abrangência, e do classificador SVM, em obter melhor desempenho na precisão, não foram observadas na heurística *k*-NN+SVM. Na realidade, os resultados foram contrários ao esperado.

- o comportamento hierárquico dos classificadores. A dificuldade em analisar individualmente um classificador, sem levar em consideração o ciclo completo da Categorização Hierárquica de Textos, não permite que os resultados das combinações de classificadores

sejam completamente compreendidos. Isto, porque na CHT o resultado de um classificador está associado ao resultado obtido por classificadores nos níveis anteriores da hierarquia.

Capítulo 7

Conclusão

Este trabalho apresentou uma proposta de combinação de classificadores e sua avaliação para a Categorização Hierárquica de Textos em uma coleção da língua portuguesa, a coleção Folha-RIcol. Nesse tipo de categorização, um conjunto de categorias são hierarquicamente organizadas em uma estrutura de árvore. Nessa estrutura, os documentos podem ser categorizados em qualquer categoria. A proposta consistiu em uma combinação dos classificadores *k*-NN e SVM em uma heurística, denominada *k*-NN+SVM. Para avaliar essa proposta, foram realizados experimentos com os classificadores *k*-NN e SVM, em combinações (vide Capítulo 6) e separadamente (vide Capítulo 5).

Para o desenvolvimento deste trabalho foi buscada uma fundamentação téorica sobre a CHT, enfatizando os procedimentos e características comuns aos classificadores e métodos de categorização utilizados nesta dissertação. Esta fundamentação teórica apresentou uma visão geral sobre a Categorização Hierárquica de Textos, destacando conceitos e características da Aprendizagem de Máquina e dos classificadores. Na abordagem adotada, os classificadores fazem uso de uma coleção de textos previamente categorizados para construir um modelo estatístico de predição capaz de categorizar novos documentos.

No processo de categorização, foram identificadas quatro etapas: pré-processamento, treinamento, teste e etapa operacional. Os conceitos, características e diferenças dos classificadores *k*-NN e SVM foram enfatizados no intuito de demonstrar as distinções existentes nas etapas de treinamento e teste. No contexto da representação de documentos e da etapa de pré-processamento, destaca-se a necessidade de aplicar uma seleção de atributos.

Os trabalhos correlatos também contribuíram para a execução dessa dissertação. A escolha dos trabalhos teve um impacto na aplicação de idéias e conceitos. O trabalho de Langie, cujo protótipo foi utilizado como base estrutural para os experimentos realizados nesta dissertação, é de vital importância. O trabalho de Moraes e Lima, demonstrou a dificuldade na realização da CHT em uma coleção de textos que não foi previamente categorizada. O trabalho de Liu *et al.* identificou uma fragilidade presente nos classificadores SVM com poucos documentos para treino. O trabalho de Bennet, Dumais e Horvitz, apresentou uma melhora no desemempenho da categorização com a utilização da combinação STRIVE.

Na elaboração da metodologia, foram apresentadas a coleção, a hierarquia das categorias, a representação dos documentos, a heurística proposta e a avaliação; enfim, a metodologia que foi utilizada nos experimentos. Essa metodologia é crucial para a realização, descrição e análise dos experimentos, permitindo identificar e discutir os problemas encontrados.

A análise dos resultados obtidos através de experimentos permitiu a observação de características, a identificação de problemas e possíveis soluções, bem como tecer algumas considerações sobre os classificadores, sobre as combinações e mesmo sobre a coleção de textos empregada.

7.1 Contribuições

Nesta seção são destacadas algumas contribuições resultantes do desenvolvimento desta dissertação:

- **Processo de CHT** - o processo descrito nesta dissertação não é inovador, todavia a sintetização de idéias e conceitos de diferentes sub-áreas de pesquisa constitui uma contribuição para a realização de novos estudos na área.
- **A proposta da heurística *k*-NN+SVM** - embora a heurística proposta não tenha apresentado resultados superiores aos encontrados com o uso do classificadores *k*-NN e SVM individualmente, essa proposta constitui uma contribuição ao utilizar idéias e conceitos na tentativa de aperfeiçoar o desempenho do processo de CHT. A contribuição pode ser vislumbrada com a identificação e análise dos problemas encontrados.
- **Avaliação** - a avaliação dos experimentos realizados contribuiu para a utilização de classificadores, combinados ou independentes, em trabalhos futuros, de acordo com suas características e a coleção de textos utilizada.

O processo, a metodologia e a avaliação, por si só, não são uma inovação em relação aos estudos existentes na literatura, mas o seu conjunto agrega conhecimentos importantes que viabilizam a realização de novos estudos na área.

A heurística proposta permitiu detalhar as características inerentes aos classificadores e à coleção. Apesar de essa proposta não demonstrar um ganho que aponte para viabilidade de uso, permite avaliar o rumo a ser tomado em novas pesquisas, com esta coleção ou em coleções com características similares.

7.2 Trabalhos futuros

Durante o desenvolvimento desta dissertação foram destacados diversos procedimentos, características, dificuldades e possíveis soluções envolvendo o processo de CHT. Este estudo permitiu uma compreensão dos esforços e avanços presentes nesta área de pesquisa. No entanto,

foram o teste, avaliação e análise dos classificadores e da coleção de textos Folha-RIcol que trouxeram as maiores contribuições aqui evidenciadas. Ao longo do processo de criação, desenvolvimento e teste da heurística *k*-NN+SVM foram observadas especifidades dos classificadores e da coleção de textos. Estas observações, permitiram vislumbrar dificuldades e carências que podem dar continuidade ao trabalho desenvolvido.

Um próximo trabalho interessante seria estudar o comportamento dos classificadores ou combinação de classificadores em uma fase operacional. Para tanto, existe a coleção de textos PLN-BR CATEG que já foi objeto de estudo no trabalho de Moraes e Lima. Até o presente momento, esta coleção não está categorizada. Portanto, existe a necessidade de um apoio a sua categorização com o uso de ferramentas computacionais, uma vez que o trabalho de Moraes e Lima demonstrou as dificuldades em categorizar automaticamente esta coleção. Uma possível solução seria o emprego da combinação de classificadores, na realização de um processo semi-automático, onde a decisão final seria realizada por pessoas.

Outras perspectivas de trabalhos futuros são alterações na heurística, como por exemplo, inverter a ordem dos classificadores. Nesse caso, onde é utilizado o classificador *k*-NN passam a ser utilizado o classificador SVM e vice-versa, no intuito de analisar e compreender mais especificamente o comportamento dos dois classificadores.

Referências

[1] SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys*, ACM Press, New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. ISSN 0360-0300.

[2] DUMAIS, S. *et al*. Inductive learning algorithms and representations for text categorization. In: *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*. New York, NY, USA: ACM Press, 1998. p. 148–155. ISBN 1-58113-061-9.

[3] SEBASTIANI, F. Classification of text, automatic. In: BROWN, K. (Ed.). *The Encyclopedia of Language and Linguistics*. Second. Amsterdam, NL: Elsevier Science Publishers, 2006. v. 2, p. 457–463.

[4] LANGIE, L. C. *Um estudo sobre a aplicação do algoritmo kNN à categorização hierárquica de textos.* 110 p. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2004.

[5] LIU, T.-Y. *et al*. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explorations Newsletter*, ACM Press, New York, NY, USA, v. 7, n. 1, p. 36–43, 2005. ISSN 1931-0145.

[6] D'ALESSIO, S.; MURRAY, K.; SCHIAFFINO, R. The effect of using hierarchical classifiers in text categorization. In: *Proceedings of RIAO-00: 6th International Conference Recherche d'Information Assistée par Ordinateur*. 2000. p. 302–313.

[7] SUN, A.; LIM, E.-P.; WEE-KEONG, N. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology*, Wiley, New York, NY, v. 54, n. 11, p. 1014–1028, 2003. ISSN 1532-2882.

[8] SAHLGREN, M.; CöSTER, R. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In: *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004. p. 487.

[9] SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Process Management*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, 1988. ISSN 0306-4573.

[10] LOSEE, R. M. Term dependence: A basis for luhn and zipf models. *Journal of the American Society of Information Science*, v. 52, n. 12, p. 1019–1025, 2001.

[11] ALPAYDIN, E. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge: The MIT Press, 2004. Hardcover. ISBN 0262012111.

[12] CHAKRABARTI, S. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002. ISBN ISBN 1-55860-754-4.

[13] HIDALGO, J. M. G. *et al*. Integrating lexical knowledge in learning-based text categorization. In: *Proceedings of JADT-02, 6th International Conference on the Statistical Analysis of Textual Data*. St-Malo, FR. 2002.

[14] TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

[15] YANG, Y.; LIU, X. A re-examination of text categorization methods. In: *22nd Annual International SIGIR*. Berkley: ACM Press, 1999. p. 42–49.

[16] PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. In: ____. *Advances in kernel methods: support vector learning*. Cambridge, MA, USA: MIT Press, 1999. p. 185–208. ISBN 0-262-19416-3.

[17] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: NéDELLEC, C.; ROUVEIROL, C. (Ed.). *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998. p. 137–142.

[18] BENNETT, P. N.; DUMAIS, S. T.; HORVITZ, E. The combination of text classifiers using reliability indicators. *Information Retrieval*, Kluwer Academic Publishers, Hingham, MA, USA, v. 8, n. 1, p. 67–100, 2005. ISSN 1386-4564.

[19] FREUND, Y.; SCHAPIRE, R. E. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, v. 5, n. 14, p. 771–780, 1999.

[20] ALPAYDIN, E. Techniques for combining multiple learners. In: *Proceedings of Engineering of Intelligent Systems*. 1998. v. 2, p. 6–12.

[21] MORAES, S. M. W.; LIMA, V. L. S. de. Um estudo sobre categorização hierárquica de uma grande coleção de textos em língua portuguesa. p. 1–10, 2007.

[22] CESA-BIANCHI, N.; GENTILE, C.; ZANIBONI, L. Hierarchical classification: combining bayes with svm. In: *ICML '06: Proceedings of the 23rd international conference on Machine learning*. New York, NY, USA: ACM Press, 2006. p. 177–184. ISBN 1-59593-383-2.

[23] DUMAIS, S.; CHEN, H. Hierarchical classification of web content. In: *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2000. p. 256–263. ISBN 1-58113-226-3.

[24] FAN, J.; GAO, Y.; LUO, H. Hierarchical classification for automatic image annotation. In: *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2007. p. 111–118. ISBN 978-1-59593-597-7.

[25] SUN, A.; LIM, E.-P. Hierarquical text classification and evaluation. In: *Proceedings of ICDM-01, IEEE International Conference on Data Mining*. 2001. p. 521–528.

[26] LEWIS, D. D. *et al*. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, MIT Press, Cambridge, MA, USA, v. 5, p. 361–397, 2004. ISSN 1533-7928.

[27] YANG, Y. An evaluation of statistical approaches to text categorization. *Information Retrieval*, Kluwer Academic Publishers, v. 1, n. 1/2, p. 69–90, 1999.

[28] FONSECA, J. S.; MARTINS, G. A.; TOLEDO, G. L. *Estatística Aplicada*. São Paulo: Atlas, 1985. ISBN 85-224-1901-9.

Apêndice A

A.1 Algoritmos

Neste apêndice estão reunidos dois dos principais algoritmos usados no desenvolvimento desta dissertação. Os algoritmos estão descritos em "alto nível"com uma sintaxe semelhante à utilizada na linguagem Java.

O primeiro algoritmo é uma contribuição do estudo de Langie, extraída de [4].

Algorithm 1 Algoritmo da estratégia de *limiar baseado em rank*

```
1:  chosenCategory = table.getCategory(0);
2:  // Encontra a categoria com maior valor de relevância
3:  for ( do int i = 1; i < table.size(); i++; )
4:      category = table.getCategory(i);
5:      if ( category.getRelevance() > chosenCategory.getRelevance() ) then
6:          chosenCategory = category;
7:      end if
8:  end for
9:  // Retorna null se a categoria com maior valor de relevância
10: // for a categoria na qual o classificador está sendo executado
11: if ( table.isActiveCategory(chosenCategory) ) then
12:     return null;
13: end if
14: return chosenCategory;
```

O segundo algoritmo é uma contribuição desta dissertação.

Algorithm 2 Algoritmo da heurística *k*-NN+SVM

```
1: chosenNode = table.getCategoryNode();
2: if ( chosenNode.getChildCount() > 2 ) then
3:     classifier = k-NN;
4: else
5:     classifier = SVM;
6: end if
7: return classifier;
```

A.2 Tabela Z

Tabela da Distribuição Normal

Área entre 0 e Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990