EFEITOS TEMPORAIS EM CLASSIFICAÇÃO DE TEXTOS: CARACTERIZAÇÃO E ENGENHARIA DE DADOS

RENATA BRAGA ARAÚJO
Orientador: Wagner Meira Junior

EFEITOS TEMPORAIS EM CLASSIFICAÇÃO DE TEXTOS: CARACTERIZAÇÃO E ENGENHARIA DE DADOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Belo Horizonte
Novembro de 2008

Araújo, Renata Braga

A663e Efeitos temporais em classificação de textos: caracterização e engenharia de dados. / Renata Braga Araújo. — Belo Horizonte, 2008
xx, 102 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Prof. Wagner Meira Junior.

1. Recuperação de Informação - Tese. 2. Mineração de Dados - Tese. 3. Banco de Dados - Tese. I. Orientador. II. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Efeitos Temporais em Classificação de Textos: Caracterização e Engenharia de Dados

RENATA BRAGA ARAUJO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. MARCOS ANDRÉ GONÇALVES - Co-orientador
Departamento de Ciência da Computação - UFMG

DRA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - ICEx - UFMG

PROF. ALBERTO FERREIRA DE SOUZA
Departamento de Informática - UFES

Belo Horizonte, 04 de dezembro de 2008.

Aos meus pais e irmãos, pelo amor incondicional, e ao Guilherme pelo apoio durante toda essa trajetória.

Agradecimentos

Agradeço primeiramente aos meus pais, Edson Rotéia Araújo e Telma Braga Araújo, pela criação que me deram, por estarem sempre presentes em minha vida me dando apoio em todos os momentos, e por me darem uma base familiar sólida que possibilitou a conquista de mais um dos meus sonhos. Agradeço também o amor, paciência e compreensão que sempre tiveram quando em alguns momentos foi preciso me dedicar aos estudos, me ausentando de momentos familiares.

Agradeço também aos meus irmãos, Robson Braga Araújo e Edson Rotéia Araújo Júnior, pelo companheirismo e apoio. Ao Júnior pelos momentos de distrações e divertimento, e ao Robson, por ser o meu exemplo a ser seguido e uma pessoa com a qual sempre posso e poderei contar. A todos os meu familiares e amigos, agradeço por estarem sempre ao meu lado, me proporcionando momentos de muita alegria e descontração. À Natália Guimarães por ser minha amiga de todos momentos e à Flávia Fradico que mesmo distante sempre me deu muito apoio e foi muito importante durante essa caminhada.

Agradeço aos professores e colegas de mestrado, que me ajudaram a crescer profissionalmente, me passaram muitos ensinamentos, sendo fundamentais nessa conquista. Ao Wagner Meira por ter sido meu orientador, e proporcionar a realização deste trabalho e ao Fernando Mourão que teve uma participação fundamental no desenlvolver do mesmo.

Enfim, gostaria de agradecer especialmente ao Guilherme Henrique Trielli Ferreira, por desempenhar tantos papéis diferentes em minha vida ao mesmo tempo. Por ter sido meu amigo quando essa trajetória começou, por ter sido meu colega de mestrado, me ajudando inúmeras vezes com os trabalhos que desempenhei, e, principalmente, por ter sido e ser meu companheiro de todas as horas, com seu apoio e amor, sem os quais nada disso seria possível.

Resumo

A Classificação Automática de Documentos (CAD) tem se tornado um tópico de pesquisa importante devido à crescente quantidade de informação disponível na Internet. A CAD normalmente segue uma estratégia de aprendizado supervisionada, em que, primeiramente, um modelo de classificação é construído utilizando documentos pré-classificados e, em seguida, esse modelo é utilizado para classificar novos documentos. Um grande desafio para a CAD, em diversos cenários, é que as características dos documentos e das classes às quais eles pertencem podem mudar ao longo do tempo. Entretanto, a maioria das técnicas recentes para a CAD são aplicadas sem considerar a evolução temporal da coleção de documentos.

Neste trabalho, caracterizamos detalhadamente a evolução temporal na CAD, com base em uma metodologia de análise dos efeitos temporais, e propomos estratégias de engenharia de dados para tratar esses efeitos. Na metodologia de análise, foi mostrado que a evolução temporal pode ser explicada por três fatores: distribuição de classes, distribuição de termos e similaridade de classes. Aplicamos também metodologias experimentais e métricas capazes de isolar cada um desses fatores para que eles sejam analisados separadamente. Além disso, apresentamos estratégias de engenharia de dados que incorporam os aspectos temporais nas bases, através dos processos de filtragem e transformação dos dados. Enquanto a filtragem de dados envolve apenas uma seleção dos documentos que devem compor o conjunto de treino, a transformação dos dados envolve um processo de modificação dos termos dos documentos da base de dados, atribuindo a eles um novo rótulo de forma a incorporar os aspectos temporais.

Através da estratégia exaustiva de filtragem, mostramos que, utilizando apenas 69% da base de dados da ACM, foi possível obter uma acurácia de 89,76%, e com apenas 25% da coleção MedLine, uma acurácia de 87,57%. Isso significa um ganho de até 20% na eficácia do classificador, com conjuntos de treino muito menores do que a base de dados inteira. Entretanto, sabemos que utilizar tal estratégia em cenários reais é inviável. Por outro lado, com nossas estratégias de transformação de dados, obtivemos um ganho de até 6,5% na acurácia do processo de classificação, sendo essas estratégias aplicáveis em cenários reais e extensíveis à utilização de outros algoritmos.

Abstract

The Automatic Document Classification (ADC) has become an important research topic due to the increasing amount of information available on the Internet. ADC usually follows a standard supervised learning strategy, in which we first build a classification model using pre-classified documents and then this model is used to classify new documents. One major challenge for ADC in many scenarios is that the characteristics of the documents and the classes to which they belong may change over time. However, most of the current techniques for ADC are applied without taking into account the temporal evolution of the collection of documents.

In this work, we characterize the temporal evolution in ADC in details, based on an analysis methodology for the temporal effects, and we propose data engineering strategies to deal with these effects. In the analysis methodology, we show that the temporal evolution may be explained by three factors: class distribution, term distribution and class similarity. We employ experimental methodologies and metrics capable of isolating each of these factors in order to analyze them separately. Moreover, we present some data engineering strategies that incorporate the temporal aspects in the databases, through processes of data filtering and transformation. While data filtering consists of selecting the documents that will be part of the training set, data transformation is a process in which the terms of the documents in the database are changed, assigning them a new label that will somehow incorporate the temporal aspects.

Using an exhaustive filtering strategy, we showed that, with only 69% of the ACM database, we are able to have an accuracy of 89.76%, and with only 25% of the MedLine, an accuracy of 87.57%, which means gains of up to 20% in the accuracy with much smaller training sets than the entire database. However, we know that this strategy is not feasible in real scenarios. On the other hand, with our data transformation strategies, we obtained a gain of up to 6.5% in the accuracy, and these strategies may me applied in real scenarios and even extended to the use of other algorithms.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Devido ao crescente uso da Internet, a quantidade de informação sendo armazenada e acessada através da Web se torna cada vez maior. Essa informação é freqüentemente organizada como documentos de texto e tem sido alvo principal de máquinas de busca e outras ferramentas de recuperação de informação, que realizam tarefas como buscas e filtragens. Uma estratégia comum para lidar com essa quantidade de informação é associá-la a categorias semanticamente significantes, sendo tal processo conhecido como Classificação Automática de Documentos (CAD) Borko e Bernick [1963].

Historicamente, em função da introdução de documentos digitais, a automação dessa tarefa tem sido tratada como um ramo importante de pesquisa. A partir dos anos 60, uma lista de sucessos e de problemas foi reportada na literatura sobre o tema e, muito embora o entusiasmo sobre classificação tenha enfraquecido em determinada época, o rápido crescimento da Web fez reviver o interesse pela CAD. Dessa forma, a associação de documentos a classes é utilizada e contribui atualmente em várias tarefas, tais como: assinalamento automático de rótulos a documentos, construção de diretórios de tópicos, identificação do estilo de escrita de documentos, criação de bibliotecas digitais, melhoria da precisão de buscas na Web e, até mesmo, ajudando usuários a interagir com máquinas de busca. Outras aplicações relevantes envolvem também a filtragem de SPAMs, a detecção de conteúdos adultos e a detecção de plágio.

A CAD normalmente segue uma estratégia de aprendizado supervisionada, em que, primeiramente, um modelo de classificação é construído utilizando documentos pré-classificados e, em seguida, esse modelo é utilizado para classificar novos documentos. Construir um modelo de classificação de textos significa achar um conjunto de características que melhor definem classes de documentos. Um grande desafio em diversos cenários é que as características dos documentos e das classes às quais eles pertencem podem mudar com o tempo, já que novas informações são criadas, novos termos são introduzidos, novas áreas surgem e algumas áreas são divididas em subáreas mais espe-

cíficas. Conseqüentemente, os padrões existentes no vocabulário que caracterizam uma determinada classe evoluem com o tempo, fazendo com que muitas diferenças entre classes, que já foram úteis para distingüi-las entre si, não existam mais.

Recentemente, foi possível observar um exemplo interessante do impacto da evolução das classes. Em 2005, foi lançado um carro com o nome *Tucson*. Até essa data, entretanto, o termo *Tucson* se referia principalmente à cidade localizada no estado do Arizona, nos Estados Unidos. Assim, documentos que continham esse termo, pertencentes a anos anteriores a 2005, tinham grande probabilidade de serem associados a temas relativos à cidade. Porém, a partir dessa data, houve um aumento significativo no número de documentos que se referem a *Tucson* como um carro. Modelos de classificação criados depois dessa mudança devem possuir, conseqüentemente, um aumento na probabilidade de que um documento que contém esse termo seja associado ao assunto "carros". Entretanto, tais modelos só são possíveis quando os aspectos temporais dos documentos são levados em consideração, o que não é o caso da maioria das estratégias para a construção de classificadores.

Apesar da potencial redução na qualidade dos modelos de classificação devido à evolução do vocabulário, a maioria das técnicas recentes para CAD, tais como, *nearest-neighbor* Lam e Ho [1998], métodos de seleção de atributos Mladenic e Grobelnik [1998], classificação Bayesiana McCallum e Nigam [1998], *support vector machines* Joachims [1999], e classificação baseada em associação Friedman et al. [1996]; Veloso et al. [2006], são aplicadas sem levar em consideração os efeitos dos aspectos temporais na coleção de documentos. Embora seja reconhecido que a evolução temporal é um fator relevante no contexto da CAD, ainda não é muito claro como essa evolução influencia no desempenho dos classificadores, quais são as características desse aspecto temporal que afetam a qualidade do classificador e como explorar essas propriedades para melhorar a acurácia do classificador. Responder tais perguntas é um dos objetivos deste trabalho.

Há alguns trabalhos em CAD que consideram a evolução temporal das coleções, mas eles focam apenas no desafio de tratar o desequilíbrio da freqüência das classes e como essa freqüência muda com o tempo, fator esse que será tratado aqui como "distribuição de classes". Neste trabalho, entretanto, iremos além, considerando outros dois fatores relacionados com a evolução temporal. Um desses fatores é como a relação entre termos e classes muda ao passar do tempo, já que termos aparecem, desaparecem e variam seu poder discriminativo entre as classes. O terceiro fator é como a similaridade entre classes, representada em função dos termos que ocorrem em seus documentos, varia com o tempo. Por exemplo, duas classes podem ser similares em um determinado momento e não mais em um momento posterior.

Para entender e caracterizar melhor esses três fatores, serão avaliados neste trabalho a evolução temporal e seus efeitos em duas bases de dados: a biblioteca digital da ACM, que contém documentos da área de ciência da computação, e a MedLine, uma biblioteca digital relacionada com a área de medicina. Assim, será proposta e aplicada uma nova metodologia para avaliar o impacto da evolução temporal no desempenho de classificadores. Essa metodologia permitirá analisar separadamente cada um dos fatores, destacando as causas que resultam na degradação do desempenho do classificador. Serão também apresentados alguns resultados de uma análise empírica exaustiva, que demonstra a potencial melhora ao se considerar os aspectos temporais. Finalmente, serão feitas adaptações nas bases de dados, de forma que esses dados agreguem informações temporais dos documentos. A partir dessas adaptações, mesmo não alterando algoritmos de classificação propriamente ditos, espera-se que o classificador melhore sua acurácia uma vez que estará considerando os aspectos temporais (já que informações sobre esses aspectos foram inseridas nas bases de dados).

1.1 Objetivo

Este trabalho tem como um de seus objetivos comprovar e mensurar o impacto dos aspectos temporais no comportamento dos classificadores automáticos. Além disso, a partir desses estudos, pretende-se tratar bases de dados de forma que elas incorporem características temporais em seus dados, através de processos de engenharia de dados. Com isso, espera-se obter uma melhora na acurácia do processo de classificação.

Portanto, será avaliado como o surgimento de novos contextos e a evolução do vocabulário (com o conseqüente surgimento ou desaparecimento de classes) influenciam na qualidade das técnicas de classificação atuais, atribuindo, assim, uma perspectiva evolutiva para essa classe de algoritmos. Com base nas respostas obtidas e nos estudos realizados, são propostas estratégias engenharia de dados, utilizando-se processos de filtragem, transformação e adaptação dos mesmos, de forma que estejam inseridos neles os aspectos temporais dos documentos. Assim, espera-se que, mesmo sem alterar códigos de algoritmos de classificação, obtenha-se uma melhora da acurácia desses algoritmos, uma vez que estarão considerando implicitamente os aspectos temporais analisados.

1.2 Motivação

As principais motivações deste trabalho surgem da necessidade de definir de forma clara e detalhada como os aspectos temporais influenciam a acurácia do classificador

ao longo do tempo, já que, apesar de conhecido na literatura, esse fator nunca foi amplamente estudado e compreendido.

Além disso, espera-se que, a partir dessa compreensão em detalhes do processo de evolução do vocabulário e dos seus efeitos no desempenho dos classificadores, seja possível criar um modelo em que os fatores temporais são levados em consideração e, dessa forma, proporcionar uma melhora na acurácia dos classificadores, que são atualmente de suma importância em diversas áreas da ciência da computação, como mencionado anteriormente.

1.3 Contribuições

As principais contribuições deste trabalho são:

- Definição e aplicação de uma nova metodologia para avaliar o impacto da evolução temporal na acurácia de classificadores;
- Comprovação e quantificação do impacto dos aspectos temporais no comportamento dos classificadores automáticos;
- Análise de três fatores distintos (distribuição de classes, distribuição de termos e similaridade de classes) separadamente, destacando as causas que resultam na degradação do desempenho do classificador;
- Análise empírica exaustiva que demonstra a potencial melhora ao se considerar os aspectos temporais, utilizando o processo de filtragem de dados;
- Definição e aplicação de estratégias de engenharia de dados (filtragem e transformação), de forma que agreguem informações temporais, objetivando-se melhorar a acurácia de classificadores.

1.4 Organização do texto

O restante do texto desta dissertação está organizado da seguinte forma: no Capítulo 2 estão apresentados os trabalhos encontrados na literatura que abordam de alguma forma o tema deste trabalho. Assim, fazemos uma breve comparação da diferença existente entre eles e o que foi desenvolvido e objetivado com o tema aqui proposto.

Alguns conceitos e termos utilizados, assim como alguns aspectos que envolvem a execução dos experimentos realizados ao longo deste trabalho, estão apresentados no Capítulo 3, para que os mesmos possam ser completamente compreendidos, possibilitando um bom entendimento do que foi desenvolvido.

O Capítulo 4 é responsável por apresentar a metodologia para avaliar o impacto da evolução temporal na acurácia de classificadores, comprovando e mensurando este impacto no comportamento dos classificadores automáticos. É neste capítulo que é feita uma análise individual de cada um dos fatores: distribuição de classes, distribuição de termos e similaridade de classes, destacando as causas que resultam na degradação do desempenho do classificador.

Após serem estudados os efeitos temporais, propomos estratégias de transformação da base de dados de forma a incorporar esses efeitos. Assim, no Capítulo 5, apresentamos os experimentos que envolvem essas transformações, assim como o impacto delas na acurácia dos classificadores.

Por fim, no Capítulo 6 apresentamos as conclusões e os trabalhos futuros a serem realizados, seguido pelas Referências Bibliográficas deste trabalho.

Capítulo 2

Trabalhos Relacionados

Apesar de a classificação de documentos ser um assunto amplamente estudado, a análise dos aspectos temporais nessa classe de algoritmos é relativamente recente - tendo sido estudada apenas na última década. Pode-se apontar três amplas áreas em que existiram esforços significativos nessa direção: a classificação adaptativa de documentos, a filtragem adaptativa de informação e o *concept drift*. Cada uma delas será discutida a seguir neste capítulo.

Uma das áreas na qual os aspectos temporais têm sido estudados em mais detalhes é a Classificação Adaptativa de Documentos Cohen e Singer [1999], que abrange um conjunto de técnicas relacionadas a aspectos temporais, entre outros, com o objetivo de melhorar a eficiência e a acurácia de classificadores de texto, através de uma adaptação incremental e eficiente Liu e Lu [2002]. A Classificação Adaptativa de Documentos traz uma variedade de desafios para a mineração de texto, discutidos nos parágrafos a seguir.

O primeiro desafio é a noção de contexto e como ele pode ser explorado para se obter modelos de classificação melhores. Um contexto é uma partição semanticamente significativa do espaço de dados, sendo utilizado como uma estratégia de reduzir a complexidade associada à geração de um modelo de classificação. Pesquisas anteriores em classificação de documentos identificaram duas formas essenciais de contextos: termos vizinhos que estão próximos a uma determinada palavra-chave Lawrence e Giles [1998] e termos que indicam o escopo e a semântica de um documento Caldwell et al. [2000]. A maioria dos estudos se concentram no primeiro tipo de contexto. Entretanto, determinar os termos semanticamente significativos para cada classe tem uma grande aplicabilidade. Esses termos podem ser utilizados para definir contextos de categorias. Ao contrário dos termos vizinhos, que são extraídos baseados na proximidade a um único termo, termos de contexto podem ser determinados através da análise de múltiplos documentos de múltiplas categorias.

O segundo desafio é criar os modelos de classificação incrementalmente, devido ao fato de que tanto o conteúdo quanto o vocabulário da coleção podem evoluir à medida que novos documentos são adicionados. Como a coleção evolui, o contexto que caracteriza cada classe também evolui. Obviamente, já que o vocabulário pode evoluir, nenhum contexto pode ser pressuposto. O contexto pode até evoluir para cobrir todos os termos que aparecem nos documentos, enquanto termos inapropriados podem introduzir ineficiência e erros no classificador de documentos. Portanto, a construção de um contexto "ótimo" consiste de uma série de processos de regulagem, que talvez precisem ser refeitos devido à adição de um novo documento. Refazer todo o processo de construção de um classificador para cada documento, entretanto, é claramente impraticável computacionalmente.

O terceiro desafio é a eficiência dos classificadores de documentos. O uso de um classificador é, na maioria das vezes, mais freqüente que sua construção. A eficiência computacional do seu uso é, portanto, essencial para o processo de gerenciamento de informação e conhecimento. Assim, adaptar classificadores de forma incremental e eficiente é um fator chave para aprimorar o desempenho dos classificadores.

Apesar de haver diversos esforços para identificar novos contextos e lidar com os desafios mencionados anteriormente, nenhum desses esforços está preocupado em realizar a classificação de documentos em seu contexto temporal original, isto é, categorizar o documento de acordo com o período de tempo a que ele pertence. Além disso, nenhum trabalho explorou a relação entre essa evolução de contexto e o tempo propriamente dito. Os estudos encontrados na literatura focam apenas em identificar o surgimento de um novo contexto, sem relacionar esses contextos ao seu tempo cronológico.

Outro campo em que existe uma preocupação sobre os aspectos temporais é a Filtragem Adaptativa de Informação Haykin [1999]. A Filtragem de Informação Belkin e Croft [1992] é descrita como um problema de classificação binária, em que documentos são classificados como relevantes ou não de acordo com o interesse do usuário. Segundo Dumais et al. [1998], a modelagem de classificadores em domínios estáticos é suficientemente bem controlada. Várias aplicações assumem que a distribuição do conjunto de treino e a distribuição dos dados novos são similares. Isso pode ser verdade para aplicações que duram um pequeno período de tempo, mas se torna inválido para aplicações de longos períodos. Nesse caso, é inevitável adaptar os classificadores a novas situações, para manter a qualidade da classificação.

Nesse contexto, Klinkenberg desenvolveu trabalhos interessantes sobre Filtragem Adaptativa Klinkenberg e Renz [1998], em que ele explora métodos para reconhecer mudanças de conteúdo e determinar janelas de interesse no conjunto de treino, cujo tamanho é fixo ou adaptável automaticamente ao tamanho corrente da mudança de

contexto. Entretanto, a eficácia da abordagem varia de acordo com a mudança do interesse do usuário ao longo do tempo. Outro trabalho relevante é apresentado por J. Allan Allan [1996]. Nesse trabalho, os autores verificaram a relevância incremental das respostas dos usuários no processo de filtragem de informação, também considerando mudanças no interesse dos usuários. J. Allan também avaliou os aspectos dinâmicos na detecção de tópicos e rastreamentos Allan et al. [1998]. Em Yang e Kisiel [2003], os autores desenvolveram uma nova técnica, chamada *Margin-based Local Regression*, que automaticamente otimiza a acurácia do classificador ao longo do tempo, baseado em quão importante é uma resposta parcial no sistema de recuperação de documentos em ordem cronológica.

Em Jones e Diaz [2007], os autores mostram que é importante considerar os aspectos temporais ao gerar resultados para uma consulta, uma vez que isso pode melhorar a qualidade dos resultados obtidos. Analisando a seqüência de tempo de um conjunto de resultados para uma consulta, é possível caracterizar o quão temporalmente dependente é o tópico, e o quão relevante serão os resultados. Os autores concluem que os meta-dados associados a uma consulta podem ser combinados com técnicas de recuperação de texto para melhorar a compreensão e o tratamento de buscas de textos em documentos que possuem uma data associados a eles. Esse trabalho é, de alguma forma, similar apresentado aqui, no sentido em que investiga como os aspectos temporais podem ser utilizados para melhorar a qualidade de certos processos. Entretanto, enquanto nos concentramos no processo de classificação, eles exploram esse assunto para a recuperação de informação.

Considerando ainda estudos que endereçam mudanças de textos, o trabalho de C. Lanquillon e I. Renz Lanquillon e Renz [1999] baseia-se na idéia de assegurar a acurácia da classificação. Basicamente, existem duas formas de adaptar um classificador. O classificador pode ser completamente retreinado, de acordo com o conjunto de treinamento atual, ou um classificador existente pode ser atualizado com base em alguns exemplos apenas. No artigo, são discutidas as dificuldades associadas a cada técnica e, devido a esse fato, o trabalho restringiu-se apenas a detectar mudanças de textos, através de métricas de acurácia, não tratando a fase de adaptação. Lewis Lewis [1995], por sua vez, realizou alguns trabalhos sobre sistemas autônomos de classificação de textos, em que ele sugere o monitoramento do comportamento a fim de adaptá-los às mudanças, se necessário.

Por fim, *Concept Drift* abrange um outro conjunto relevante de esforços Forman [2006]; Scholz e Klinkenberg [2007]. Em Forman [2006], é apresentada a Classificação Diária de Tarefas (CDT), um arcabouço que lida com classificação de documentos tentando considerar os aspectos temporais. Nesse trabalho, o tempo é discretizado

em períodos, por exemplo, dias. Para cada período de tempo, uma amostra aleatória de tamanho limitado é provida como um conjunto de treinamento rotulado. Uma métrica de desempenho, como, por exemplo, a acurácia da classificação, é calculada para cada dia da base de dados, sendo a média reportada sobre todos os dias. Em Scholz e Klinkenberg [2007], os autores propõem um método para treinar classificadores baseados em entradas de fluxos de dados. Essa técnica é naturalmente adaptada ao *concept drift* e permite quantificar as mudanças em relação ao aprendizado de base. É empiricamente provado que tal método é mais eficiente que algoritmos de aprendizado que ignoram o conceito de *drift*. Em ambos os trabalhos, os autores consideram a existência de efeitos temporais sem entender exatamente o que são esses efeitos temporais. Eles argumentam que, apesar de as classes apresentarem uma explosão de elementos em alguns períodos, a sua identidade real não se modifica. Neste trabalho, por outro lado, apresentamos uma explicação detalhada dos aspectos temporais, baseada em resultados empíricos que demonstram a existência de outros efeitos além da explosão de elementos de algumas classes. Além disso, apresentamos uma análise empírica exaustiva que demonstra como podemos melhorar os classificadores usando efeitos temporais.

Capítulo 3

Fundamentos Teóricos e Metodologia Experimental

Neste capítulo serão apresentados conceitos importantes utilizados ao longo deste trabalho, cuja compreensão é fundamental para um bom entendimento do mesmo. Dessa forma, objetiva-se relembrar conceitos já conhecidos na literatura, assim como definir a utilização de alguns termos, para que o seu significado e a maneira em que foram aplicados seja completamente compreendida.

Além disso, ao longo dos próximos capítulos, serão apresentados diversos experimentos realizados que contribuem para a análise e discussão do problema proposto nesta dissertação. Tanto a metodologia de análise criada para caracterizar os efeitos temporais no processo de classificação, quanto as estratégias de transformação de dados propostas para que esses efeitos sejam considerados em busca de um aprimoramento dos classificadores, foram embasadas em dados experimentais que contribuíram significativamente para um entendimento mais profundo das questões levantadas e discutidas.

Dessa forma, antes de apresentarmos essas análises, é importante ressaltar ainda, neste capítulo, alguns aspectos que envolvem a execução desses experimentos, tais como a descrição das bases de dados utilizadas, as métricas de avaliação adotadas, e alguns detalhes de como o algoritmo de classificação (no caso, o SVM Joachims [1999, 2006]) foi utilizado neste processo.

3.1 Processo de Classificação

O trabalho aqui apresentado consiste na avaliação do processo de classificação de documentos sob o âmbito dos efeitos temporais, e como esse processo pode ser aprimorado, dada uma compreensão mais profunda do problema. Assim, é de fundamental importância o entendimento deste conceito de Classificação Automática de Documentos.

Portanto, primeiramente, será descrito como o processo de classificação acontece, assim como serão destacadas as suas principais fases.

A Classificação Automática de Documentos geralmente utiliza uma estratégia de aprendizado supervisionado. Na primeira fase do processo, é construído um modelo de classificação baseado em documentos já conhecidos e previamente classificados. Na segunda fase, esse modelo construído é utilizado para classificar novos documentos. Assim, a construção de modelos de classificação equivale a encontrar um conjunto de termos com o poder discriminativo para melhor identificar as classes dos documentos.

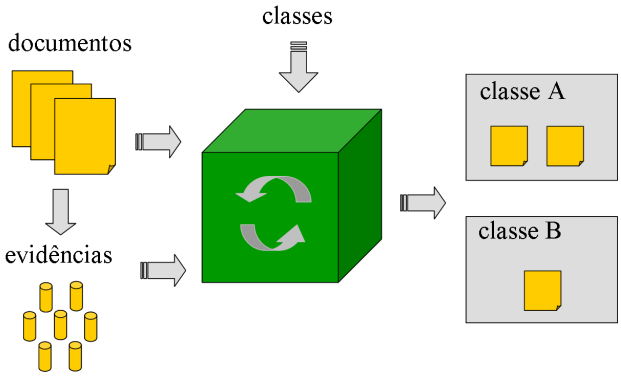

Figura 3.1. Processo de Classificação

A Figura 3.1 ilustra esse processo, em que documentos são analisados e, a partir deles, são encontradas evidências (termos discriminativos) para a construção do modelo de classificação. Os documento utilizados nessa fase são chamados de base de treino e essa primeira fase de construção do modelo é chamada de fase de treinamento. A seguir, documentos novos são classificados utilizando o modelo de classificação construído. Esses documentos novos são denominados base de teste e o processo de assinalamento

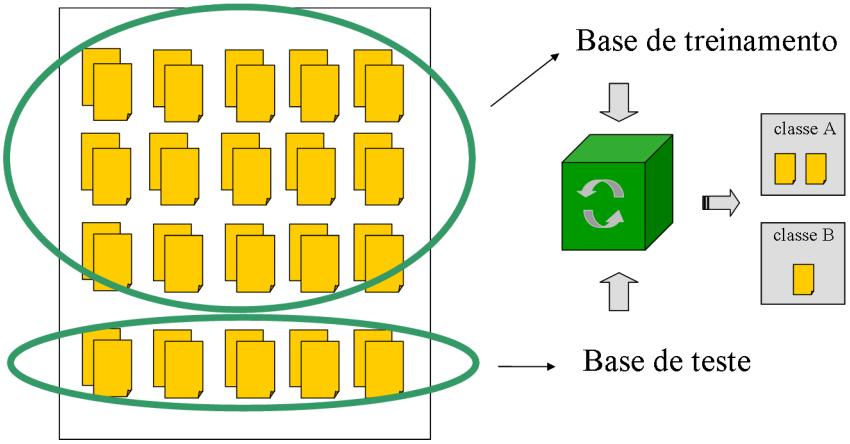

Figura 3.2. Base de Treinamento e Base de Teste

de classes a tais documentos é chamado de fase de teste. A Figura 3.2 ilustra a base de treino (ou treinamento) e a base de teste, ambas utilizadas no processo de classificação.

3.2 Descrição das Bases de Dados

Para a realização dos experimentos desta dissertação, foram utilizadas duas bases de dados distintas: a coleção de documentos da ACM e a coleção de documentos da MedLine. A seguir, será apresentada uma descrição de cada uma delas.

3.2.1 Coleção da ACM

A primeira coleção de documentos utilizada em nossos experimentos consiste em um conjunto de 30.000 documentos da biblioteca digital da ACM, contendo artigos relacionados a tópicos da área de computação. Os artigos considerados pertencem aos anos entre o intervalo de 1980 e 2002.

Nessa coleção, os documentos são associados às classes do esquema de classificação da ACM pelos próprios autores. Apesar de que nesse esquema originalmente existem classes e subclasses, utilizamos apenas o primeiro nível da taxonomia composto por 11 categorias. A Tabela 3.1 mostra o nome das 11 classes do primeiro nível da ACM, assim como a abreviação desses nomes que foi utilizada ao longo do trabalho, como, por exemplo, em alguns gráficos.

Categoria	**Abreviação**
General Literature	GLit
Hardware	HW
Computer Systems Organization	CSO
Software	SW
Data	Data
Theory of Computation	TheoryC
Mathematics of Computing	MathC
Information Systems	InfoS
Computing Methodologies	CMethodo
Computer Applications	CAppl
Computing Milieux	CMilieux

Tabela 3.1. Classes da coleção ACM

É importante ressaltar que o primeiro nível dessa taxonomia não sofreu mudanças durante o período considerado, possibilitando, dessa forma, uma execução apropriada dos experimentos.

Além disso, na coleção considerada, temos que todos os documentos são associados a uma única classe e, para a realização dos experimentos, são avaliados apenas o título e o *abstract* desses artigos.

3.2.2 Coleção da MedLine

Além do conjunto de artigos pertencentes à coleção da ACM, avaliamos também uma segunda coleção de documentos composta por artigos da MedLine. A MedLine é uma base de dados da literatura internacional da área médica e biomédica, que foi produzida pela NLM (*National Library of Medicine*, USA).

Para a realização dos experimentos, consideramos um conjunto de 4.112.069 documentos, com artigos datando de 1970 a 2006. Nessa base de dados os artigos encontram-se classificados em em 7 classes distintas, que correspondem às áreas definidas pela própria PubMed para acesso aos documentos da área de medicina. Na Tabela 3.2, estão apresentados os nomes dessas 7 categorias, assim como a abreviação utilizada para uma delas ao longo deste trabalho.

Categoria	**Abreviação**
Aids	Aids
Bioethics	Bioethics
Cancer	Cancer
Complementary Medicine	CMedicine
Toxicology	Toxicology

Tabela 3.2. Classes da coleção MedLine

É importante ressaltar que originalmente nessa base de dados, um mesmo documento pode pertencer a mais de uma classe. Entretanto, para a realização dos experimentos, isso foi tratado removendo os documentos em que isso acontecia. Novamente, foram considerados apenas o título e *abstract* dos artigos para realizar o processo de classificação.

3.3 Métricas de Avaliação

Nesta seção, serão descritas as métricas de avaliação utilizadas nos diversos experimentos realizados ao longo deste trabalho.

3.3.1 Acurácia de um Classificador

Neste trabalho, objetivamos melhorar a acurácia de um classificador à medida que consideramos os aspectos temporais para realizar a classificação. Dessa forma, é importante que o conceito de acurácia utilizado aqui esteja bem definido, assim como o conceito de ganho dessa acurácia.

A acurácia de um classificador depende do número de documentos que foram corretamente classificados (DCC - Documentos Corretamente Classificados) em relação ao número total de documentos classificados (TDC - Total de Documentos Classificados), podendo ser avaliada, em porcentagem, através da seguinte fórmula:

$$\texttt{Acurácia} = \frac{DCC}{TDC} \times 100$$

Assim, quando mencionamos que a acurácia de um classificador é igual a 75%, por exemplo, isso significa que 75% dos documentos analisados foram classificados corretamente.

Além disso, quando mencionamos que houve um ganho de, por exemplo, 20% na acurácia do classificador, estamos nos referindo a um ganho absoluto. Ou seja, se a acurácia do processo de classificação era 65% ao se classificar um conjunto de documentos e, com as melhoras realizadas no processo de classificação adotado, houve um ganho de 20%, isto significa que a acurácia do classificador para aquele mesmo conjunto de documentos de teste passou a ser 85%.

3.3.2 Métricas de Precisão, Revocação e $\mathbf{F}_1$

Após aplicarmos as estratégias de transformação dos dados, é interessante avaliar, para cada uma das classes, como essa transformação influencia no processo de classificação dos documentos. Com esse objetivo, utilizamos as métricas de precisão, revocação e F_1. A seguir, apresentaremos a definição de cada uma delas.

A precisão é definida como a relação entre o número de documentos classificados corretamente em uma dada classe (DCA - Documentos Corretamente Associados à classe) e o número de documentos classificados (corretamente ou não) naquela classe (TDA - Total de Documentos Associados à classe). Pela fórmula, temos:

$$\texttt{precisão} = \frac{DCA}{TDA}$$

A revocação, por sua vez, consiste na relação entre o número de documentos classificados corretamente em uma dada classe (DCA) e o número de documentos, entre todos documentos que foram classificados, que realmente pertencem àquela classe (DPC

- Documentos Pertencentes à Classe). Assim, pela fórmula, temos:

$$\texttt{revocação} = \frac{DCA}{DPC}$$

Por fim, a métrica F_1 faz uma ponderação da precisão e da revocação, podendo ser calculada como:

$$F_1 = \frac{2 \times \texttt{precisão} \times \texttt{revocação}}{\texttt{precisão} + \texttt{revocação}}$$

É importante observar que as três métricas apresentadas são aplicadas por classe. Assim, quando trabalhamos com diversas classes, calculamos as métricas para cada classe separadamente. Pode ser interessante, entretanto, sumarizar esses valores através de métricas globais. Portanto, definimos aqui também a MacroF$_1$, que representa a média da métrica F$_1$ obtida para todas as classes. Considerando $|C|$ o número de classes presente na base de dados, podemos definir a MacroF$_1$ através da fórmula apresentada a seguir:

$$\texttt{MacroF}_1 = \frac{\sum_i^{|C|} \texttt{F}_{1i}}{|C|}$$

3.3.3 Similaridade de cosseno

O conceito de similaridade de cosseno será utilizado na execução do experimento da Subseção 4.2.3, em que objetiva-se verificar o quão semelhante dois vocabulários são entre si. Dessa forma, apresentamos a definição desse conceito a seguir.

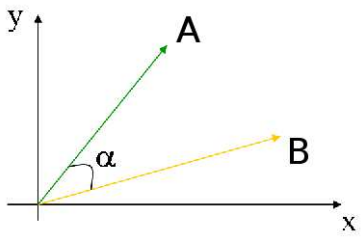

Figura 3.3. Similaridade de Cosseno

Similaridade de cosseno é uma medida de similaridade entre dois vetores de n dimensões, encontrada através da determinação do cosseno do ângulo entre eles (Figura 3.3). Dados dois vetores de atributos, A e B, a similaridade de cosseno é representada usando o produto escalar, de magnitude igual a:

$$\texttt{similaridade} = cos(\alpha) = \frac{A.B}{||A|| \times ||B||}$$

O resultado da similaridade varia de -1 (significando que os vetores são totalmente diferentes) a 1 (significando que são exatamente os mesmos), sendo que valores entre eles indicam uma similaridade ou dissimilaridade intermediária. Entretanto, para os experimentos realizados neste trabalho, não faz sentido valores negativos para a similaridade de cossenos e, portanto, os limites desse intervalo são os valores 0 e 1.

3.3.4 Validação Cruzada

Em vários experimentos realizados neste trabalho, utilizamos o conceito de validação cruzada (ou, como mais conhecido, *cross-validation*) na execução dos testes. A seguir, é apresentada uma breve descrição do mesmo.

A técnica de *cross-validation* é um método estatístico em que uma amostra dos dados é particionada em subconjuntos, de forma que a análise é inicialmente realizada em um único subconjunto, enquanto os outros subconjuntos são guardados para um uso subseqüente, no qual é feita a confirmação e a validação da análise inicial. O subconjunto inicial dos dados é chamado conjunto de treino, enquanto os outros subconjuntos são chamados de conjuntos de teste ou validação.

No *K-fold cross-validation*, a amostra original é particionada entre K subamostras. Das K subamostras, apenas uma subamostra é usada como os dados de validação para testar o modelo, e as $K - 1$ subamostras restantes são usadas como dados de treino. O processo de *cross validation* é, então, repetido K vezes, com cada uma das K subamostras sendo usadas exatamente uma vez como os dados de validação. Pode-se então fazer a média dos K resultados das execuções para produzir uma única estimativa. A vantagem desse método se repetir sobre subamostras randômicas é que todas as observações são usadas tanto no treino quanto na validação, e cada observação é usada para validação apenas uma vez.

3.3.5 Janela Temporal

Nesta dissertação, são realizados vários experimentos em que o conceito de janela temporal é utilizado. Nestes experimentos, a janela temporal funciona como um filtro para selecionar os documentos utilizados na fase de treinamento em que o modelo de classificação é construído.

Assim, quando temos uma janela temporal de tamanho 10, por exemplo, para classificar um documento do ano de 1991, isso significa que utilizaremos os documentos entre os anos de 1986 e 1996 como documentos de treino para construir o modelo de classificação. Ou seja, a janela é aplicada tomando-se o ano do documento de teste como o ponto central, de forma que ela cresce simetricamente para o passado e futuro

até atingir o tamanho estabelecido. Portanto, no exemplo dado, são considerados os 5 anos posteriores a 1991 e os 5 anos anteriores a 1991.

É importante ressaltar que para os anos das coleções que se encontram nas extremidades do conjunto de documentos considerados para formar as bases de dados, foi necessário adotar um procedimento diferentes. A coleção da ACM considerada neste trabalho possui documentos entre os anos de 1980 a 2002. Assim, se considerarmos uma janela temporal de tamanho 8 para um documento de teste que pertence ao ano 1981, por exemplo, não há como considerar os 4 anos anteriores a 1981, uma vez que nossa base de dados não possui documentos de 1987, 1988 e 1989. Portanto, quando isso acontece, uma janela de 8 será considerada, mas a partir do ano de 1980, ou seja, temos uma janela de 1980 a 1988. Esse mesmo procedimento também é adotado, para as duas coleções, quando não há como considerar os anos posteriores ao ano do documento de teste.

3.4 Algoritmo SVM

O SVM (*Support Vector Machine*) é uma uma técnica de classificação textual que representa o estado da arte em algoritmos de classificação. Esse algoritmo foi utilizado na execução de todos os experimentos realizados neste trabalho. Portanto, a seguir, é apresentada uma breve explicação de como ele funciona.

O algoritmo SVM é baseado na idéia de que, dados dois conjuntos de pontos linearmente separáveis no espaço, ao se encontrar um hiperplano que separe esses dois conjuntos, a classificação de um novo ponto torna-se trivial, pois basta verificar o sinal do resultado ao se inserir esse ponto na equação do hiperplano encontrado. No entanto, é imediata a percepção de que podem existir infinitos hiperplanos que separam dois conjuntos de pontos linearmente separáveis no espaço.

Demonstra-se que, para minimizar o risco de se classificar erroneamente um novo ponto, deve-se encontrar o hiperplano cuja distância para os pontos mais próximos (vetores suporte) é a maior possível. O desafio do algoritmo SVM é, portanto, encontrar esse hiperplano. A Figura 3.4 ilustra de forma simplificada o processo descrito, em que o algoritmo busca encontrar o hiperplano que maximize essa distância (representado no quadro D).

Entretanto, como as coleções de documentos aqui consideradas possuem mais de duas classes, nos experimentos realizados neste trabalho, foi utilizada a abordagem um-para-todos Vapnik [1998], a fim de adaptar o classificador binário SVM para a classificação multi-classe.

Após ter sido realizada a definição de todos esses conceitos, vamos abordar, no

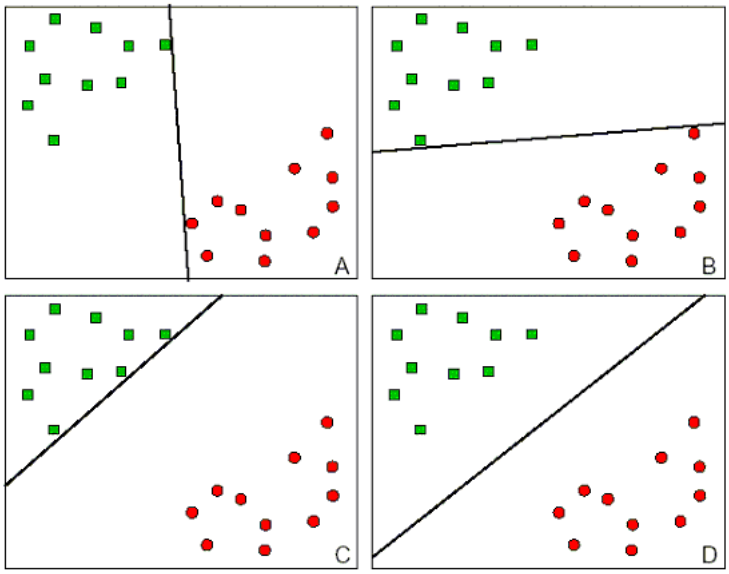

Figura 3.4. Algoritmo SVM

próximo capítulo, os efeitos temporais no processo de classificação. Dessa forma, tentaremos caracterizar a evolução temporal, através de uma metodologia de análise, para avaliar o impacto desses efeitos em classificadores automáticos.

Capítulo 4

Efeitos Temporais na Classificação

Durante o processo de construção do modelo de classificação e da classificação de documentos propriamente dita, vários desafios, que não são necessariamente relacionados a aspectos temporais, devem ser considerados. A seguir, serão discutidos alguns deles.

O primeiro desafio é lidar com a distribuição de classes. Algumas classes são muito mais freqüentes que outras, fazendo com que seja difícil gerar modelos que não sejam tendenciosos. O segundo desafio é relacionado ao poder discriminativo dos termos em uma classe específica, o que denominamos de distribuição dos termos. Nesse caso, pode-se notar que alguns termos são mais discriminativos em algumas classes do que em outras. O terceiro desafio é a sobreposição das classes entre si, já que termos podem estar associados a mais de uma classe, o que chamamos de similaridade de classes. Quanto maior for a sobreposição, mais difícil é gerar um bom modelo, que seja capaz de distingüí-las. É importante perceber que a intensidade de cada desafio é inversamente proporcional à acurácia e à generalidade do modelo.

Conforme será discutido em seções posteriores, a evolução temporal pode agravar os desafios previamente citados, uma vez que modelos baseados na coleção inteira de documentos podem não funcionar bem. Assim, considerar a evolução da coleção é essencial para uma classificação eficaz. Mostramos ainda que a evolução temporal pode ser uma tendência clara e bem definida ou uma variação (aparentemente) randômica, dificultando a construção de classificadores precisos e afetando significativamente a tarefa de classificação. A seguir, cada desafio será novamente discutido, mas agora considerando a evolução temporal juntamente a cada um deles.

A evolução temporal faz com que a distribuição de classes varie ao longo do tempo, o que deve ser quantificado para evitar a criação de um modelo tendencioso. Classes podem aparecer e desaparecer, o que pode ocorrer devido a separações e junções, respectivamente, de classes já existentes. Por exemplo, as subclasses *Information Retrieval* e *Artificial Intelligence*, no esquema de classificação da ACM em 1964, pertenciam

à mesma classe: *Applications*. No novo esquema de classificação da ACM, cada uma delas pertence a uma classe diferente, *Information Systems* e *Computing Methodologies*, respectivamente. Dessa forma, o efeito da evolução temporal na distribuição de classes dever ser caracterizado e medido, para que sejamos capazes de introduzí-lo no modelo de classificação.

Sob a perspectiva da evolução temporal, o segundo desafio abrange a evolução da distribuição dos termos. Nós mensuramos e caracterizamos quando termos aparecem, desaparecem, migram entre classes, ou simplesmente sofrem uma variação de seu poder discriminativo. Para ilustrar isso, é possível usar o mesmo exemplo dado no Capítulo 1 sobre *Tucson*. Nesse caso, o termo *Tucson* perdeu poder discriminativo na classe Geografia e ganhou mais poder discriminativo na classe Automobilismo. Intuitivamente, pode-se notar que, na maioria dos casos, essa evolução acontece de forma gradativa, fazendo com que períodos de tempo que estão temporalmente próximos também sejam mais similares.

O terceiro desafio trata da evolução da similaridade entre as classes, sem considerar as razões por trás da intensidade do grau de similaridade. Por exemplo, há algum tempo atrás, as classes Crime e Biologia não eram muito semelhantes. Entretanto, recentemente, elas se tornaram mais parecidas, já que a análise de DNA começou a ser usada nas investigações criminais. A similaridade de classes quantifica o quão diferentes ou semelhantes são duas classes. Pode-se considerar dois tipos de similaridade de classes que aparecem ao se considerar a evolução temporal: global e local. A similaridade global considera toda a coleção, enquanto a similaridade local considera apenas uma parte dela. Percebe-se que lidar com a similaridade local pode ser difícil, não só devido à sua variação, mas também devido à granularidade variável que ela pode assumir. Por exemplo, duas classes podem ser similares considerando um período de 10 anos, mas podem se tornar diferentes nos últimos dois anos desse intervalo.

É importante notar que os últimos dois desafios lidam com a ocorrência de termos nas classes, porém de forma diferente. Ou seja, a distribuição de termos considera essa ocorrência dentro de uma mesma classe (características intraclasse) enquanto a similaridade de classes considera a ocorrência de termos entre classes distintas (características interclasse).

Um claro compromisso que aparece ao se considerar a evolução temporal é encontrar a granularidade do intervalo de tempo que devemos utilizar ao se construir os modelos. A primeira questão associada a esse compromisso é o efeito amostral no contexto da evolução temporal da coleção. O efeito amostral surge como conseqüência da estratégia popular de construir o modelo baseado em uma amostragem da coleção de documentos e do fato de que isso pode degenerar características importantes da base

(ao conter termos discriminativos insuficientes, causar a distorção de classes ou obter uma distribuição incorreta dos termos). Por exemplo, se for considerada a distribuição das classes de forma geral, elas podem parecer balanceadas, mas ao verificar essa distribuição em períodos de tempo específicos, o efeito amostral pode ser ressaltado. A existência desse efeito amostral sugere que se deve aumentar o tamanho da amostra, enquanto a evolução temporal induz o contrário, já que usar períodos de tempo muito longos pode resultar em evidências conflitantes. Determinar a granularidade apropriada, ou seja, o período de tempo em que os conceitos de classificação estão estáveis e a amostra seja grande o suficiente, é uma tarefa difícil e depende das características da coleção. A abordagem aqui tomada consiste em, primeiramente, entender esse compromisso estudando cada um desses fatores que fazem parte dele e, então, tentar analisá-los, para possibilitar a construção de um modelo de classificação mais preciso. Nós não conhecemos nenhum outro trabalho que faz uma investigação detalhada desse compromisso, conforme relatado no Capítulo 2.

A seguir, serão analisados os dois fatores que contribuem para o compromisso discutido anteriormente: o efeito amostral e a evolução temporal. Primeiramente, o efeito amostral será estudado, para comprovar sua existência e influência na Classificação Automática de Documentos. Além disso, esta análise é relevante uma vez que mostra ser necessário isolar o efeito amostral nos experimentos ao se estudar os efeitos temporais, de forma a remover sua influência e proporcionar uma análise apenas da evolução temporal. Portanto, após estudar o efeito amostral, será apresentada e caracterizada uma metodologia para qualificar e quantificar a evolução temporal, que pode influenciar a acurácia de classificadores automáticos. Para tal, serão analisados cada um dos desafios apresentados anteriormente sob a perspectiva da evolução temporal, o que chamamos de efeitos temporais.

Para demonstrar a existência dos efeitos temporais, que podem influenciar a acurácia de classificadores automáticos, foi executada uma série de experimentos usando o algoritmo SVM (Support Vector Machine), um classificador do estado da arte, e duas coleções de documentos. A primeira coleção consiste em um conjunto de 30.000 documentos da biblioteca digital da ACM, contendo artigos dos anos entre 1980 e 2002, e a segunda coleção de documentos é a MedLine, que contém 4.112.069 documentos, com artigos datando de 1970 a 2006, como descrito no Capítulo 3.

4.1 Efeito Amostral

Apesar de o efeito amostral ser um fenômeno bem conhecido na literatura, por razões de completude, foi conduzido um conjunto de experimentos para ilustrar o impacto

desse efeito nos classificadores. No nosso caso específico, é necessário isolar o efeito temporal do efeito amostral. Para tal, foi construído um conjunto de documentos D com documentos pertencentes a apenas um ano, mais especificamente ao ano de 1999, para a coleção da ACM, e ao ano de 1970, para a coleção da MedLine. Dessa forma, garantimos que o efeito temporal estará muito reduzido ou quase ausente nesse conjunto.

A partir desse conjunto D, criamos vários subconjuntos S_X, cujo índice X representa a porcentagem de documentos retirados de D para constituir esse subconjunto, representando, portanto, o seu tamanho. Cada S_X será submetido ao algoritmo de classificação SVM, realizando-se um processo de *10-fold cross-validation* Brieman e Spector [1992] para cada um e calculando a média dos seus resultados. Variamos o valor de X de 20% a 100%. Já que todos os documentos pertencem ao mesmo ano, foi possível analisar como a acurácia do processo de classificação é afetada pelo tamanho da amostra. A Figura 4.1 e a Figura 4.2 mostram os resultados desses experimentos utilizando as bases da ACM e da MedLine, respectivamente. O eixo x desses dois gráficos representa o tamanho da amostra (ou seja, o tamanho do subconjunto S_X, construído a partir do conjunto D), e o eixo y representa a média da acurácia obtida pelo SVM ao classificar os documentos de cada subconjunto S_X, utilizando-se o processo de *10-fold cross-validation*. Pode-se perceber que, nas duas coleções (ACM e MedLine), à medida que aumentamos o tamanho da amostra, a acurácia do classificador também melhora, conforme esperado.

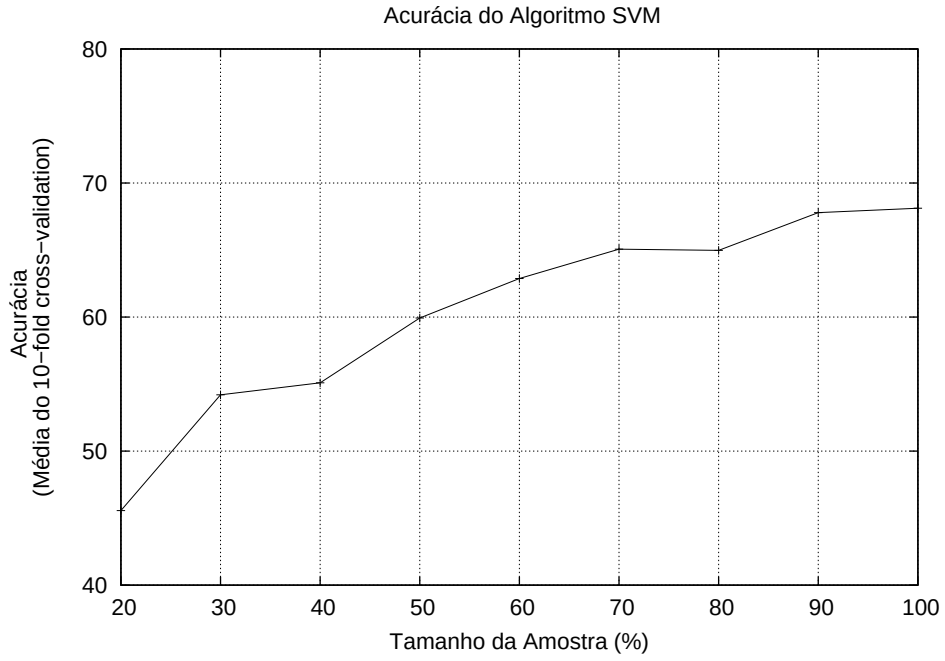

Figura 4.1. Efeito Amostral - ACM

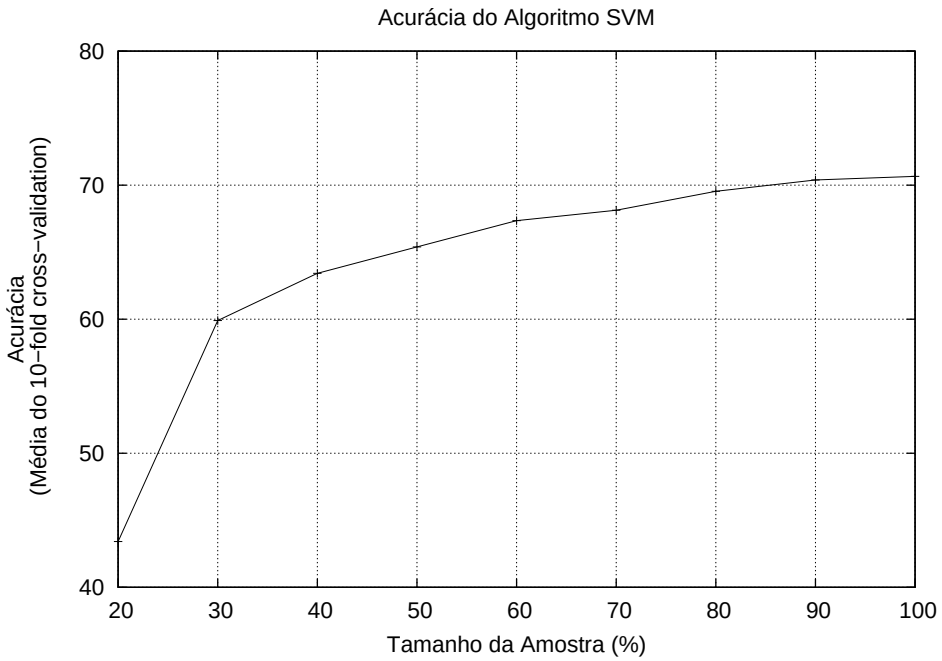

Figura 4.2. Efeito Amostral - MedLine

Além disso, temos que em ambas as coleções, o efeito amostral é também influenciado pela variação da quantidade de documentos ao longo dos anos. Na coleção da ACM, a média do número de documentos por ano, no intervalo de 1980 a 2002, é 1.086,5, com um desvio padrão de 550,8 documentos. O ano com o menor número de documentos é 1980, com 441 documentos, enquanto o ano de 1999 é o que contém o maior número de documentos, com um total de 2.728 documentos. Na coleção da MedLine, o número médio de documentos por ano é 100.789, com um desvio padrão de 51.649,3 documentos. O número mínimo de documentos por ano é 34.755 e ele ocorre no ano de 1970, enquanto o número máximo de documentos é 208.878 e ele ocorre no ano de 2005.

Esses resultados mostram que é impossível analisar o efeito temporal se não o isolarmos do efeito amostral, que desempenha um papel importante na Classificação Automática de Documentos. Assim, o efeito amostral deve ser considerado ao se construir um modelo de classificação.

4.2 Efeito Temporal

Esta seção está dividida em duas partes. Na primeira, será apresentado um conjunto de experimentos que foram realizados com o intuito de demonstrar a existência do efeito temporal e sua influência no processo de classificação. A seguir, será analisado cada um dos desafios mencionados anteriormente neste capítulo, sob a perspectiva da evolução

temporal. É importante notar que, para cada um dos desafios, isolamos os efeitos de outros fatores que também podem influenciar a tarefa de classificação, garantindo um ambiente sem ruídos para analisarmos o fenômeno em questão em cada caso.

4.2.1 Quantificação do Efeito Temporal

Para demonstrar e quantificar o efeito temporal, foram conduzidos dois conjuntos de experimentos. No primeiro conjunto, dividimos a coleção original por ano e, para cada um dos anos, foi considerado o mesmo número de documentos, sendo esses documentos selecionados aleatoriamente. Para determinar qual seria o número de documentos considerados por ano, utilizamos o número mínimo de documentos encontrado em um dado ano, ou seja, 441 para a coleção da ACM e 34.755 para a coleção da MedLine. Ao utilizar o mesmo número de documentos para cada um dos anos, nossa intenção foi separar o efeito temporal do efeito amostral, para que fosse possível analisar o efeito temporal de forma isolada. Em seguida, realizamos o processo de classificação para os conjuntos de documentos de cada um dos anos, aplicando o processo de *3-fold cross-validation* (uma vez que o tempo de execução para se realizar o *10-fold cross-validation* seria muito grande) e obtendo a média desses resultados. Ou seja, nesse experimento, utilizamos como conjunto de treino, documentos que pertencem ao mesmo ano dos documentos do conjunto de teste.

Para contrastar com os resultados obtidos a partir dessa primeira configuração, selecionamos aleatoriamente o mesmo número de documentos (441 para a base da ACM e 34.755 para a base da MedLine) considerando documentos da base toda para formar um novo conjunto de documentos (e não somente documentos de um mesmo ano, como feito anteriormente). A seguir, esse conjunto foi aleatoriamente dividido em três partes iguais, para ser possível, posteriormente, realizar o processo de *3-fold cross-validation* (utilizando cada combinação de duas partes desse conjunto para criar modelos de classificação). Assim, testamos cada um desses modelos, criados a partir das diferentes combinações, com um conjunto de documentos pertencente a um ano específico A_i, criado no primeiro experimento. Nós variamos o i de forma a cobrir todos os anos da coleção e calculamos o desempenho médio dos três modelo gerados pelo *cross-validation* para cada ano. É importante notar que, no primeiro experimento, o conjunto de treino é constituído de documentos que pertencem ao mesmo ano dos documentos de teste. Já no segundo experimento, existem documentos no conjunto de treino de todos os anos da base de dados.

A Figura 4.3 e a Figura 4.4 mostram a comparação dos resultados obtidos a partir desses experimentos para a coleção ACM e MedLine, respectivamente. O eixo x dos gráficos apresentados nessas figuras representa o ano a que os documentos a serem

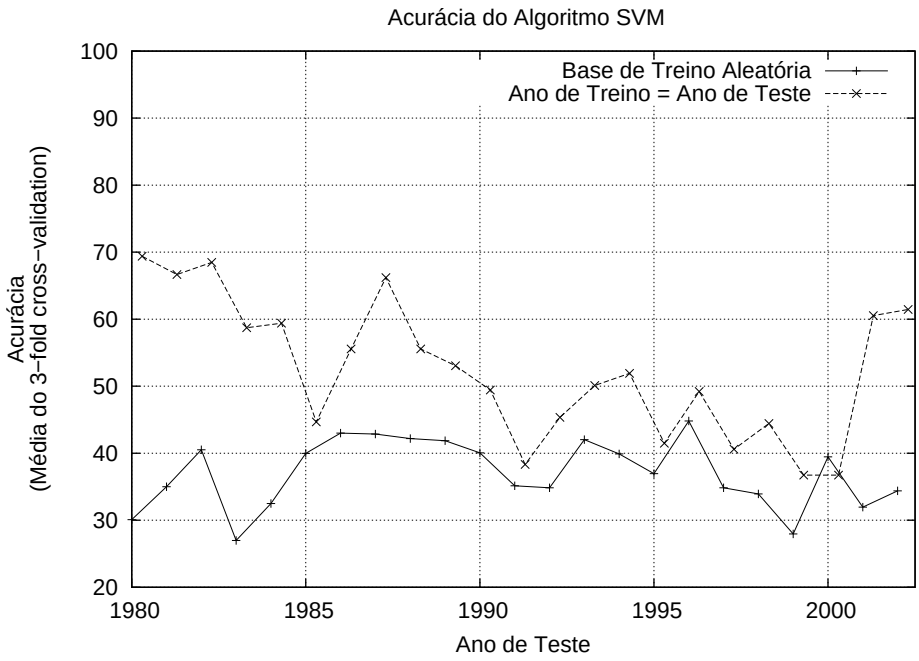

Figura 4.3. Média do Efeito Temporal - ACM

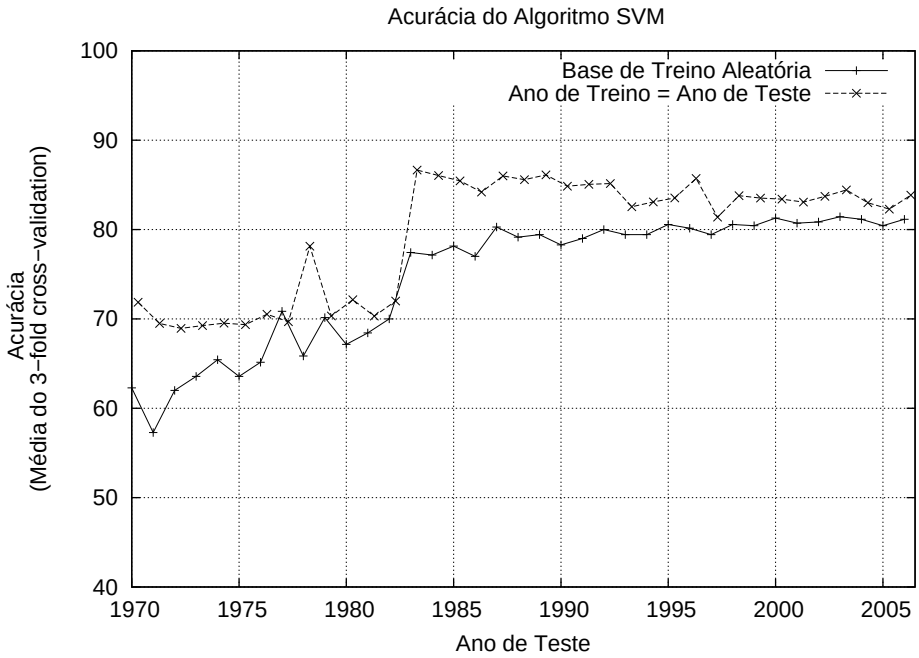

Figura 4.4. Média do Efeito Temporal - MedLine

testados pertencem, e o eixo y representa a média da acurácia obtida pelo SVM ao classificar o conjunto de documentos pertencente àquele ano específico (utilizando-se o processo de *3-fold cross-validation*), tanto para uma base de treino composta por documentos que pertencem a diferentes anos: "Base de Treino Aleatória" (curva mais

abaixo nos gráficos), quanto para uma base de treino utilizando somente documentos de treinos que pertencem ao mesmo ano que os documentos de teste: "Ano de Treino = Ano de Teste" (curva mais acima nos gráficos).

Pode-se perceber que, para a maioria dos anos, em ambas as coleções, a acurácia da tarefa de classificação quando selecionamos como conjunto de treino somente documentos que pertencem ao mesmo ano que os documentos do conjunto de teste é melhor do que a acurácia da tarefa de classificação quando consideramos documentos aleatórios para formar o conjunto de treino. É importante ressaltar ainda que a formação do conjunto de treino com documentos aleatórios é o que é normalmente usado em experimentos de classificação.

Nós caracterizamos ainda o efeito temporal utilizando uma estratégia diferente. Nesse segundo conjunto de experimentos, assim como foi feito anteriormente, dividimos a coleção original em anos e usamos o mesmo número de documentos para cada ano. Cada ano A_i foi randomicamente dividido em três partes de mesmo tamanho. Nós usamos duas das três partes como o conjunto de treino para gerar o modelo de classificação. Esse modelo é avaliado para todos os documentos de cada ano A_j das coleções, em que $j \neq i$. Além disso, também testamos o modelo com a parte de A_i que não foi utilizada como conjunto de treino. Nós repetimos esse processo para cada uma das três combinações geradas usando o processo de *3-fold cross-validation*. Após a execução de cada um dos três modelos gerados a partir de A_i, calculamos a média da acurácia dos três modelos usados em cada ano analisado. Em seguida, normalizamos os valores da acurácia das classificações, de forma que o maior valor obtido se torne 100 e que os demais valores sejam calculados em relação a ele.

A Figura 4.5 e a Figura 4.6 mostram os resultados quando usamos 1988 e 1987 como os anos de treino, na coleção da ACM e na coleção da MedLine, respectivamente. Assim como no conjunto de gráficos anterior, o eixo x representa o ano a que os documentos a serem testados pertencem. Já o eixo y dos gráficos representa a média da acurácia relativa (através da normalização explicada anteriormente) obtida pelo SVM ao classificar o conjunto de documentos pertencente a cada ano específico (utilizando-se o processo de *3-fold cross-validation*), a partir de uma base de treino composta por documentos que pertencem ao ano de 1988 (para a coleção da ACM) e a partir de uma base de treino composta por documentos que pertencem ao ano de 1987, para a coleção da MedLine.

É possível ver que há um pico da acurácia do classificador quando os documentos do conjunto de teste pertencem ao mesmo ano que os documentos do conjunto de treino, ou a anos próximos a ele. Há também uma degradação visível da acurácia do classificador à medida que o ano de teste se distancia do ano de treino.

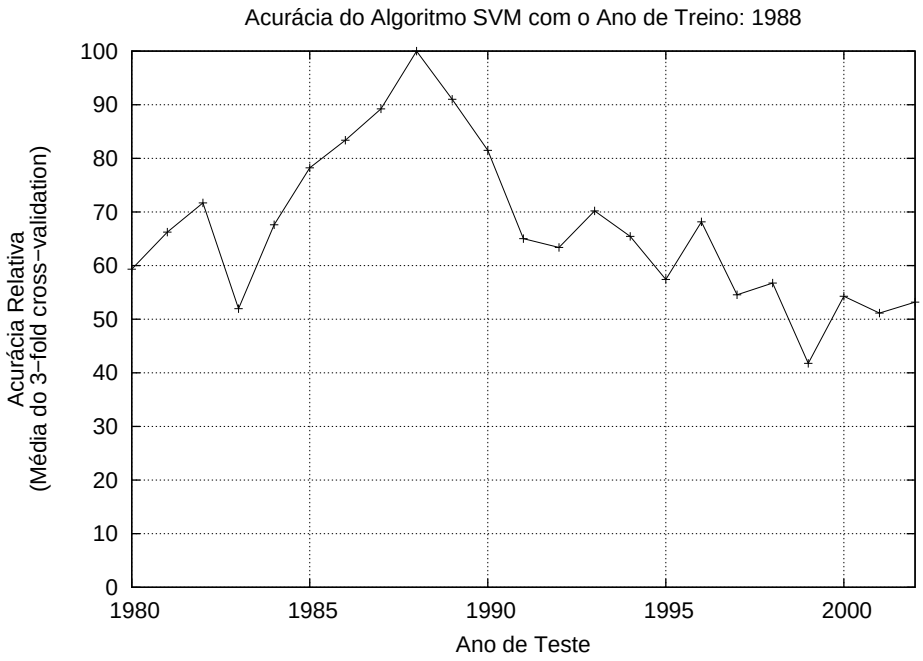

Figura 4.5. Localidade Temporal - ACM

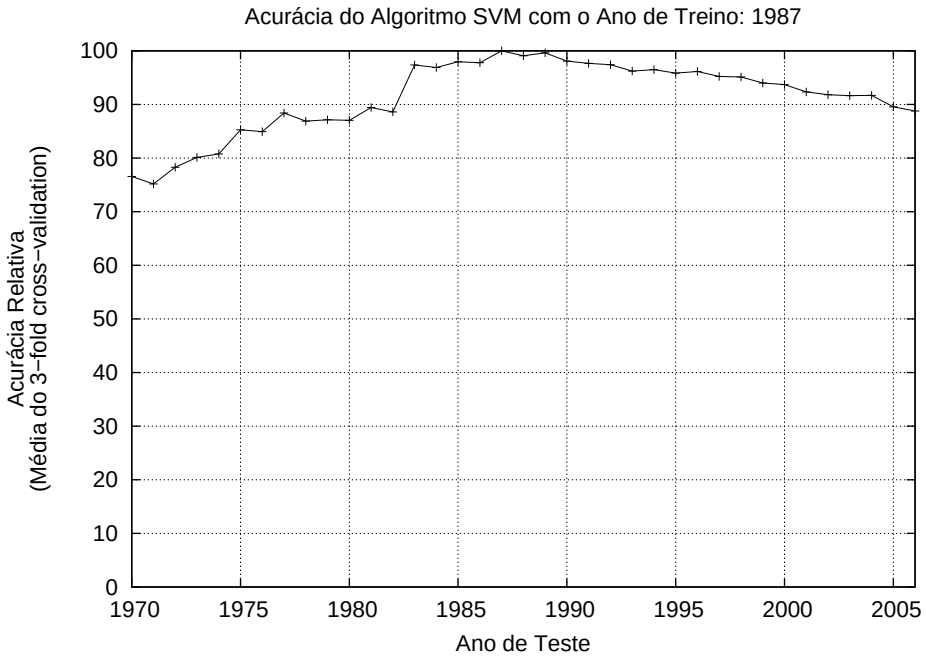

Figura 4.6. Localidade Temporal - MedLine

A Figura 4.7 e a Figura 4.8 agregam os experimentos descritos para ambas as coleções ACM e MedLine, respectivamente. Para cada ano de treino, obtivemos os resultados dos experimentos após o processo de *cross-validation* e os adaptamos da mesma forma que anteriormente. Em seguida, calculamos o desempenho médio (que

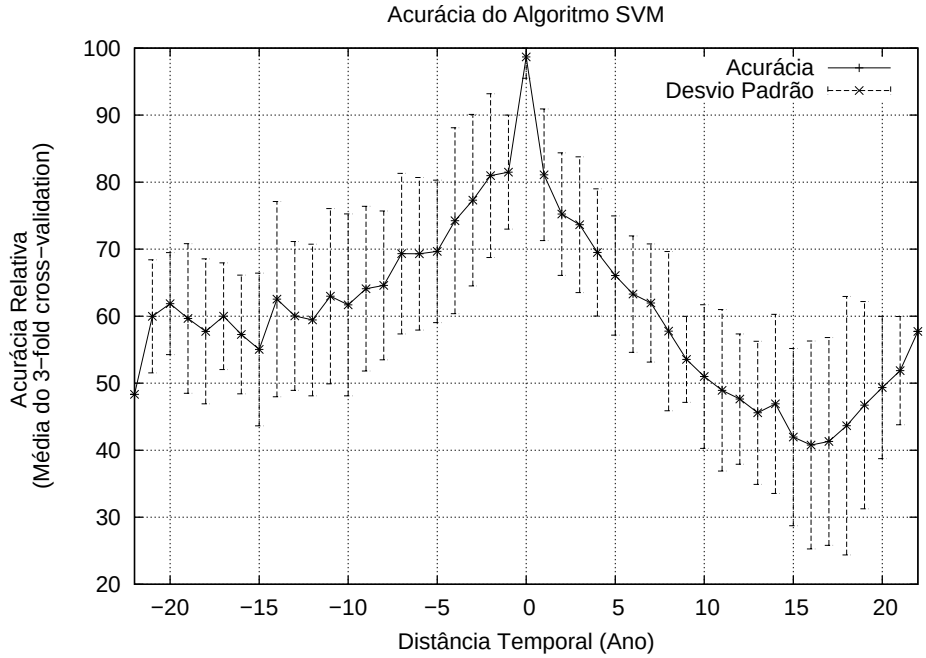

Figura 4.7. Variação da Localidade Temporal - ACM

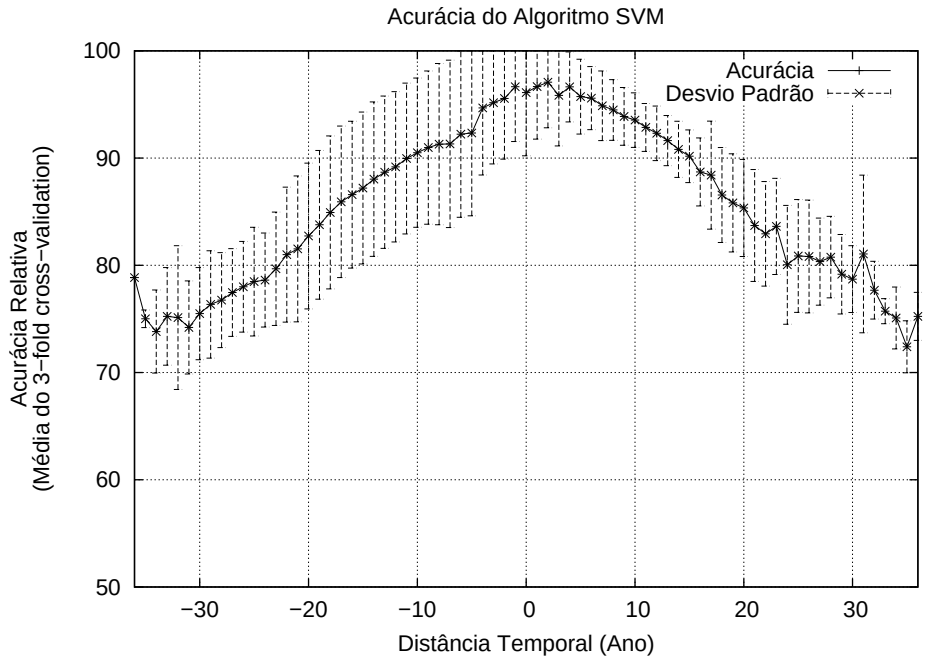

Figura 4.8. Variação da Localidade Temporal - MedLine

correspondem aos pontos escuros no gráfico) que obtivemos para cada distância no tempo dos documentos de teste em relação ao ano do treino. Ou seja, calculamos o desempenho médio quando os documentos de teste pertencem a um ano a frente ao ano dos documentos de treino (índice 1 no eixo x do gráfico), quando os documentos

de teste pertencem a um ano anterior ao ano dos documentos de treino (índice -1 no eixo x do gráfico), e assim por diante. Dessa forma, uma distância positiva no tempo, digamos 10, no eixo x, significa que os documentos de teste utilizados são dez anos mais novos que o ano dos documentos de treino e uma distância negativa no tempo, digamos 5, significa que os documentos de teste utilizados são cinco anos mais velhos que o ano dos documentos de treino. Para cada média, mostramos também o intervalo determinado pelo desvio padrão.

Como pode ser visto, na maioria dos casos, o melhor cenário é encontrado quando os documentos de teste e treino pertencem ao mesmo ano (intervalo zero). Somente em alguns poucos casos, um desempenho um pouco maior foi alcançado usando documentos de teste e treino pertencentes a anos distintos. Entretanto, mesmo nesses casos, o melhor desempenho foi alcançado usando documentos de teste e treino que estão perto temporalmente.

Os experimentos descritos mostram evidências fortes da existência do efeito temporal como um fator que contribui para a degradação da acurácia de classificadores automáticos. A seguir, apresentamos e analisamos os três desafios mencionados anteriormente que contribuem para essa degradação, sendo eles: distribuição de classes, distribuição de termos e similaridade de classes.

4.2.2 Distribuição de Classes

Nesta seção, analisamos o efeito temporal na distribuição de classes mais detalhadamente. A Figura 4.9 e a Figura 4.10, onde plotamos, para cada ano, a distribuição de classes, ilustram a variação em termos da representatividade das classes ao longo do tempo para as coleções da ACM e da MedLine, respectivamente. Dessa forma, temos que o eixo x dos gráficos representa cada ano existente na coleção de documentos, e o eixo y representa a porcentagem de ocorrência de documentos de cada uma das classes naquele ano específico.

Como podemos perceber, a oscilação na ocorrência das classes ao longo do tempo é um fenômeno freqüente. Por exemplo, existem documentos na classe *Aids* datados de 1973, mas a classe só se torna significante depois do ano 1985. Outras classes, como a classe *MathC* da ACM, se tornaram menos freqüentes ao passar do tempo.

Para mostrar o efeito dessa oscilação da distribuição de classes isoladamente, conduzimos o seguinte experimento. Para cada coleção, foram selecionados aleatoriamente o mesmo número de documentos de duas classes diferentes (C_k e C_l), em um dado ano, para criar outras sub-coleções. Usando essas novas sub-coleções, geramos subconjuntos T_X de tamanhos iguais. Esses subconjuntos foram criados da seguinte maneira: $X\%$ dos documentos de T_X são aleatoriamente selecionados da classe C_k e os documentos

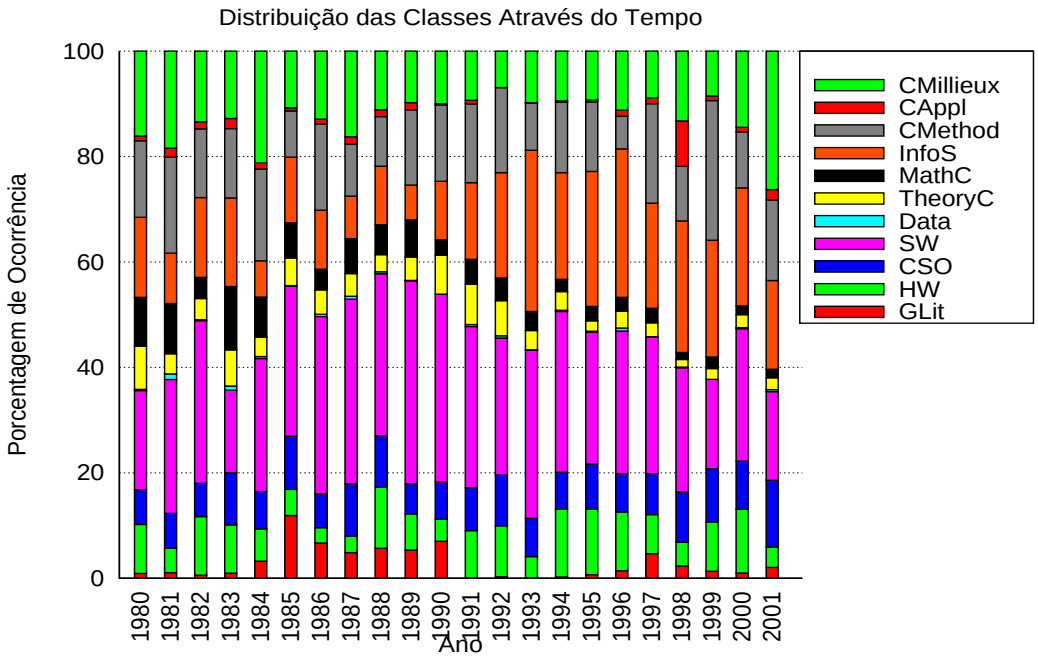

Figura 4.9. Variação da Ocorrência das Classes ao Longo do Tempo - ACM

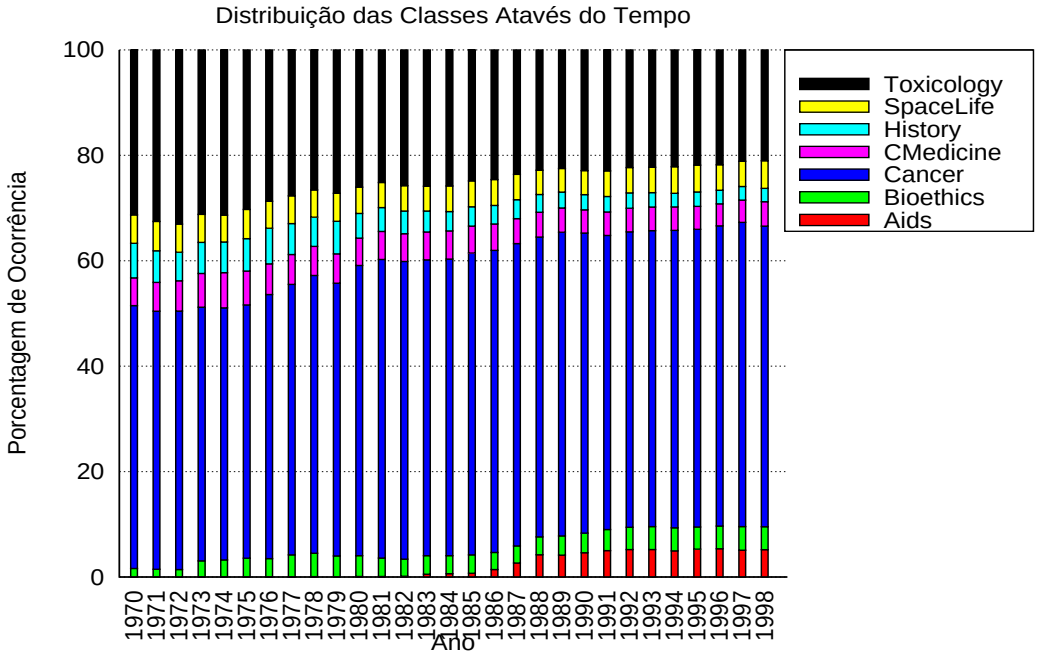

Figura 4.10. Variação da Ocorrência das Classes ao Longo do Tempo - MedLine

restantes de T_X são aleatoriamente selecionados da outra classe C_l. Consideramos os valores possíveis de X iguais a 25, 50 e 75. A seguir, cada subconjunto T_X foi aleatoriamente dividido em três partes. Para cada combinação possível de duas das três partes, geramos um modelo de classificação. Cada modelo criado foi testado com os outros subconjuntos T_X (ou seja, se $X = 25$, por exemplo, esse modelo é testado para

o subconjunto T_X em que $X = 50$ e o subconjunto T_X em que $X = 75$), além de ser testado com a parte restante do mesmo subconjunto T_X ($X = 25$) a partir do qual foi criado.

Nós conduzimos esses experimentos em ambas coleções e apresentamos seus resultados nos gráficos da Figura 4.11 e da Figura 4.12. O eixo x desses gráficos representa a porcentagem de ocorrência de documentos na base de teste que pertencem a uma determinada classe a ser analisada (classe "InfoS" para a coleção da ACM, e classe "Bioethics" para a coleção da MedLine), e o eixo y representa a a média da acurácia obtida pelo SVM ao classificar esses documentos (utilizando-se o processo de *3-fold cross-validation*), a partir de uma base de treino que possui 75% de seus documentos pertencentes à classe sendo analisada.

Através desses gráficos, podemos perceber como a acurácia do processo de classificação aumenta à medida que a distribuição dos documentos de uma dada classe na fase de teste é próxima à distribuição dos documentos do conjunto de treino. Dessa forma, fica claro que a distribuição dos documentos entre as classes influencia o desempenho dos classificadores.

Esses resultados mostram que devemos ser muito cuidadosos ao criarmos modelos de classificação, uma vez que podemos gerar modelos tendenciosos que podem não ser apropriados para os dados sendo testados. Quando consideramos que a freqüência das classes está constantemente mudando ao longo do tempo, isso se torna um problema

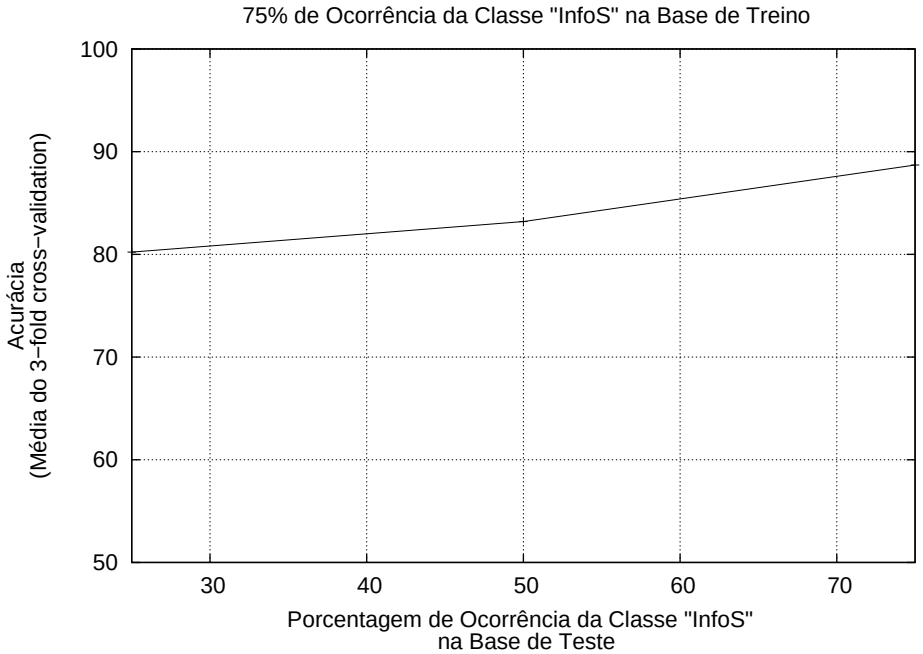

Figura 4.11. Distribuição de Classes - ACM

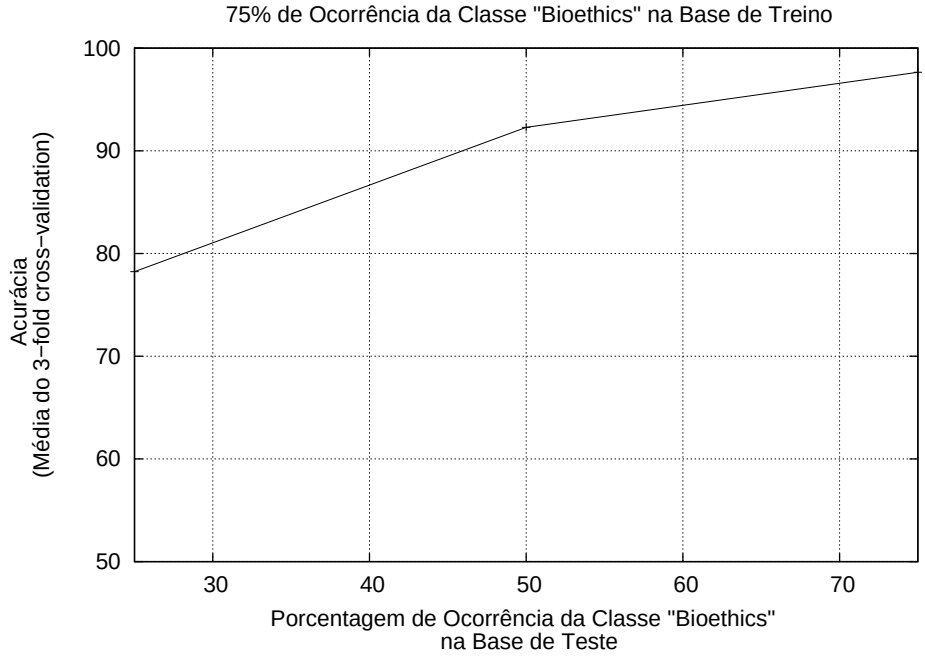

Figura 4.12. Distribuição de Classes - MedLine

ainda maior, que deve ser levado em consideração.

4.2.3 Distribuição de Termos

Nesta seção, analisamos o segundo desafio: a evolução da distribuição dos termos. Esse desafio ocorre devido ao movimento de termos em uma dada classe ao longo do tempo e está associado a características intraclasse. Os termos aparecem, desaparecem, migram entre classes e se tornam menos ou mais discriminativos para uma classe em diferentes períodos, o que deve ser medido e caracterizado para que sejamos capazes de considerar esse aspecto no modelo de classificação.

Para mensurar esse efeito, criamos um vocabulário $V_{k,i}$ contendo as t palavras de maior *info-gain* Forman [2003] na classe C_k em um dado ano A_i. Elas podem ser consideradas os termos mais discriminantes dessa classe em um dado ano. Nós representamos esse vocabulário em um espaço vetorial em que cada coordenada do vetor representa um termo e contém o peso desse termo baseado em sua freqüência na classe C_k no ano A_i. Em seguida, calculamos a união de todos os vetores da mesma classe C_k para todos os anos. Dessa forma, calculamos o quão constante é o vocabulário de cada classe ao longo do tempo. Em teoria, para uma dada classe C_k, se todos os seus vetores $V_{k,i}$ do vocabulário forem constantes (ou seja, eles não mudaram ao passar do tempo), a união dos vetores de C_k conteria exatamente t palavras. Por outro lado, se todos os vetores de C_k forem completamente diferentes uns dos outros, então o tamanho da

união dos vetores seria t vezes o total do número de anos da coleção. Para a coleção da ACM, usamos $t = 50$ e para a coleção da MedLine, utilizamos $t = 100$. É importante lembrar que os documentos da coleção da ACM abrangem um período de tempo igual a 23 anos e os documentos da MedLine abrangem um período de 37 anos. Assim, o número total de termos distintos, para a coleção ACM, pode atingir o valor de 1.150 termos, enquanto que, para a coleção MedLine, esse valor é de 3.700 termos.

A união dos vetores para cada classe é mostrada na Figura 4.13 e na Figura 4.14, para a coleção da ACM e para a coleção da MedLine, respectivamente. Nesses gráficos, o eixo x representa as classes existentes nas coleções de documentos, e o eixo y representa o número de termos distintos encontrados ao realizarmos a união de todos os vetores com as t palavras de maior *info-gain* de cada classe para todos os anos.

É importante observar que representamos o gráfico para a coleção ACM em uma escala que varia de 0 a 1.150 termos, uma vez que esse é o valor máximo de termos distintos que se pode obter para essa coleção através do processo descrito, como explicado anteriormente. Já para a coleção da MedLine, o gráfico foi apresentado em uma escala que varia de 0 a 3.700 termos, uma vez que o número máximo de termos distintos para essa coleção é 3.700 termos, como também explicado anteriormente. Isso foi feito para que fosse possível ter uma percepção mais precisa da diferença de variação desses termos para cada uma das classes, em relação ao intervalo máximo possível dessa variação.

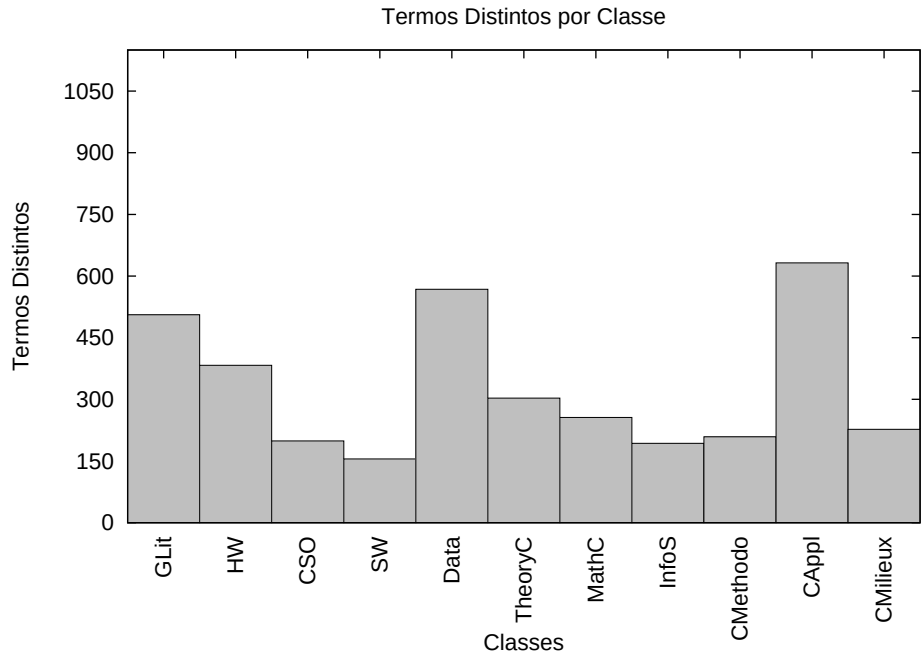

Figura 4.13. Análise do Movimento de Termos em Cada Classe - ACM

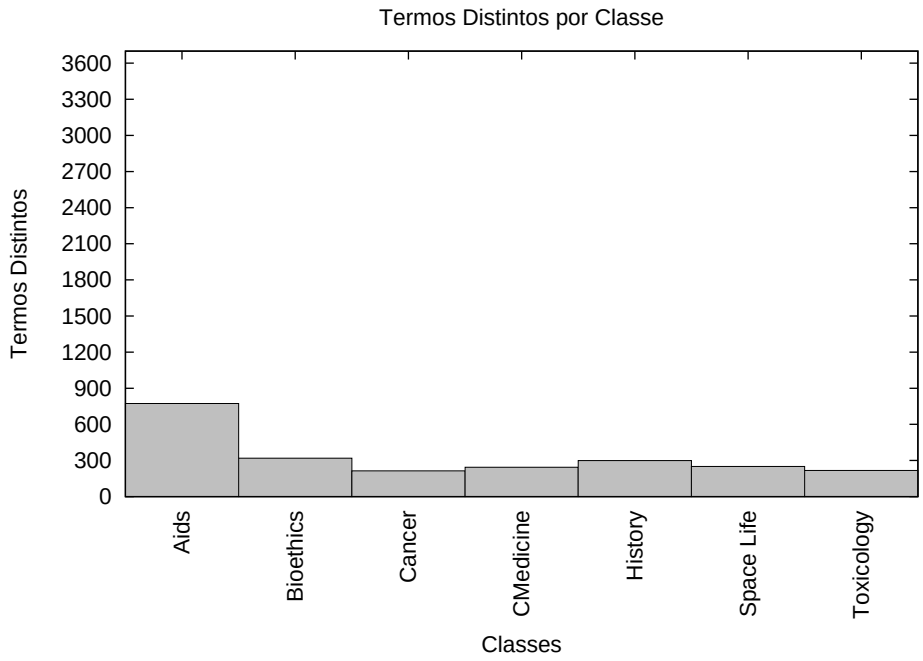

Figura 4.14. Análise do Movimento de Termos em Cada Classe - MedLine

Como se pode notar, o movimento dos termos é diferente para cada classe, o que significa que existem classes que são mais dinâmicas do que outras em ambas as coleções. Na coleção da ACM, por exemplo, temos que a classe *Software* foi a que se manteve mais estável, enquanto a classe *Applications* foi a mais dinâmica. Já na coleção MedLine, temos que a classe *Aids* foi a mais dinâmica enquanto a classe *Cancer* foi a que se manteve mais estável.

Para caracterizar melhor o fenômeno da distribuição de termos, realizamos ainda um outro experimento: dividimos a base de dados por anos e, para cada ano, separamos os documentos de acordo com suas respectivas classes. Para cada classe C_k de um determinado ano A_i, criamos novamente um vocabulário $V_{k,i}$, contendo os t termos de maior *info-gain* nessa dada classe desse ano. Em seguida, representamos esse vocabulário como um vetor e calculamos a similaridade de cosseno Salton e McGill [1986] entre os vetores do vocabulário $V_{k,i}$ e $V_{k,j}$ de uma mesma classe, em que $i \neq j$, ao longo de todos os anos da coleção.

Os resultados desse experimento estão apresentados na Figura 4.15 e na Figura 4.16. O eixo x dos gráficos representa cada ano existente na coleção de documentos. O eixo y, para a coleção ACM, representa a similaridade de cosseno entre o vetor com os t termos de maior *info-gain* da Classe "HW" naquele ano específico, e o vetor com os t termos de maior *info-gain* da Classe "HW" no ano de 1999. Já para a coleção MedLine, o eixo y representa a similaridade de cosseno entre o vetor com os t termos de maior

info-gain da Classe "Aids" naquele ano específico, e o vetor com os t termos de maior *info-gain* da Classe "Aids" no ano de 1985.

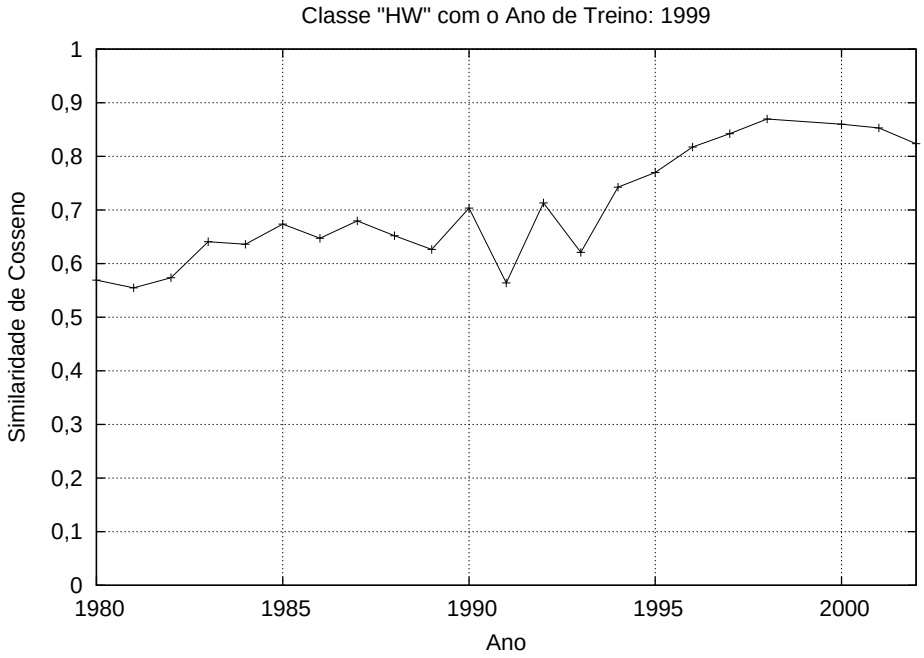

Figura 4.15. Distribuição dos Termos - ACM

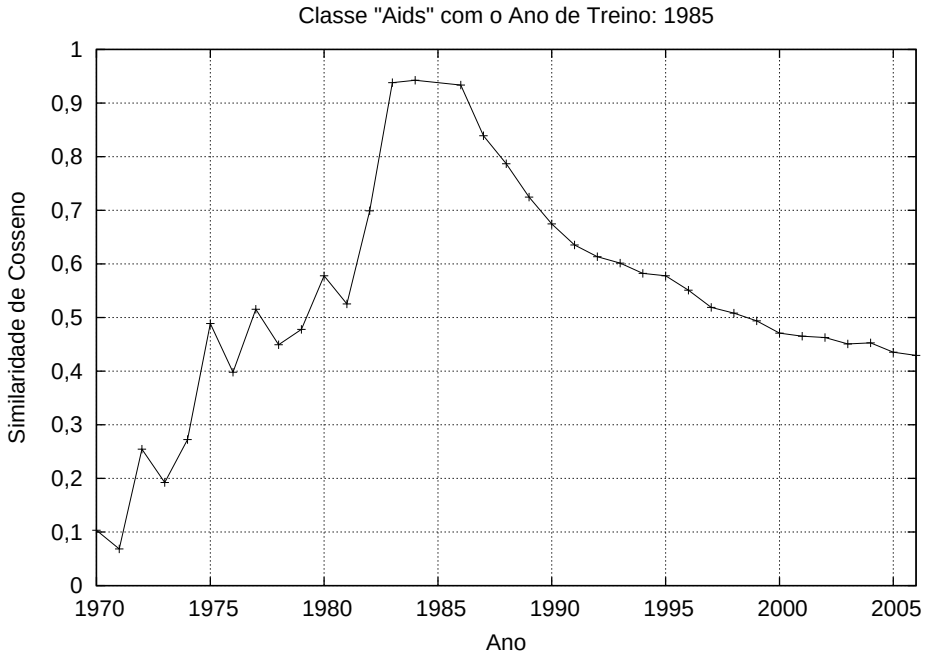

Figura 4.16. Distribuição dos Termos - MedLine

Pode-se perceber que os vocabulários de anos que estão próximos ao ano considerado tendem a ser mais similares em ambas as coleções. Assim, Quanto mais distantes no

tempo são dois documentos de uma dada classe, menor é a probabilidade de existirem termos em comum entre eles. É importante notar que os gráficos dessas figuras não contêm o valor da similaridade de cosseno para o ano de treino, isto é, em que $i = j$, já que esse valor é sempre igual a 1.

A partir desses experimentos realizados, é possível ver que, mesmo quando uma dada classe continua existindo (ou seja, a classe não desaparece, não é dividida e nem agrega outras classes), o seu assunto ou suas características principais podem mudar ao longo do tempo. Por exemplo, apesar de a classe *Artificial Intelligence* estar presente no esquema de classificação da ACM desde sua inserção, há um certo período de tempo um dos assuntos mais estudados nessa área tem sido redes neurais, enquanto anteriormente o assunto mais estudado era lógica de primeira ordem. É também interessante notar que as classes *Hardware*, na coleção da ACM, e a classe, *Aids*, na coleção MedLine, têm comportamentos diferentes. Enquanto a curva da similaridade de cosseno da classe *Hardware* ("HW") apresenta apenas um leve declínio, essa mesma curva para a classe *Aids* apresenta um forte declínio. Ou seja, o vocabulário da classe *Hardware* muda lentamente ao longo do tempo enquanto o vocabulário da classe *Aids* é muito mais dinâmico. Conseqüentemente, o tempo tem um impacto maior na classificação de documentos da classe *Aids* do que na classificação de documentos da classe *Hardware*.

Por fim, a Figura 4.17 e a Figura 4.18 mostram a média da similaridade de cosseno quando variamos a distância no tempo entre os diferentes vetores do vocabulário das classes em ambas as coleções. A distância zero significa que estamos comparando um vocabulário com ele mesmo, o que obviamente corresponde ao máximo da similaridade. Assim, o eixo x desses gráficos representa a distância temporal em anos entre os vetores dos vocabulários de uma determinada classe, e o eixo y representa a média da similaridade de cossenos desses vetores de cada classes, para uma determinada distância temporal do eixo x.

Pode-se observar que, quanto maior é a distância no tempo entre os vocabulários, menos similares eles são entre si, o que demonstra uma evolução dos vocabulários das classes ao longo dos anos. Um exemplo interessante é a classe *Aids* da coleção MedLine. Diferentemente de outras classes, a similaridade entre os seus vetores de vocabulário diminui rapidamente, mostrando novamente que a classe *Aids* é muito dinâmica.

Como demonstramos que o vocabulário evolui, é evidente que um modelo de classificação gerado considerando documentos de um certo período de tempo pode ser menos eficaz quando testado com documentos pertencentes a um outro período de tempo, uma vez que o vocabulário pode ter evoluído de forma que as premissas construídas não são mais verdadeiras, ou seja, os termos discriminantes podem não ser mais os mesmos. Isso torna esse desafio muito interessante.

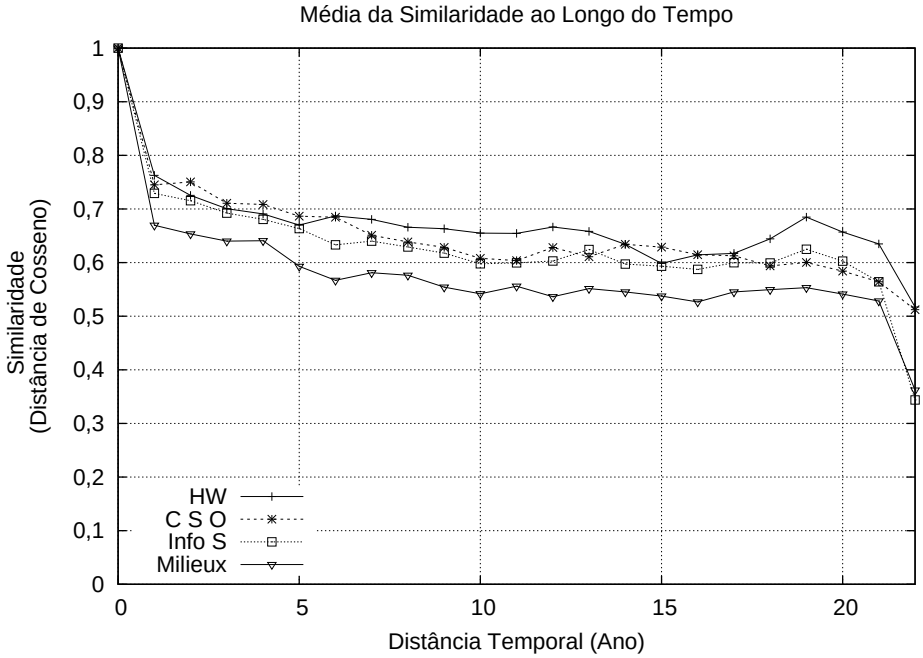

Figura 4.17. Média da Distribuição dos Termos - ACM

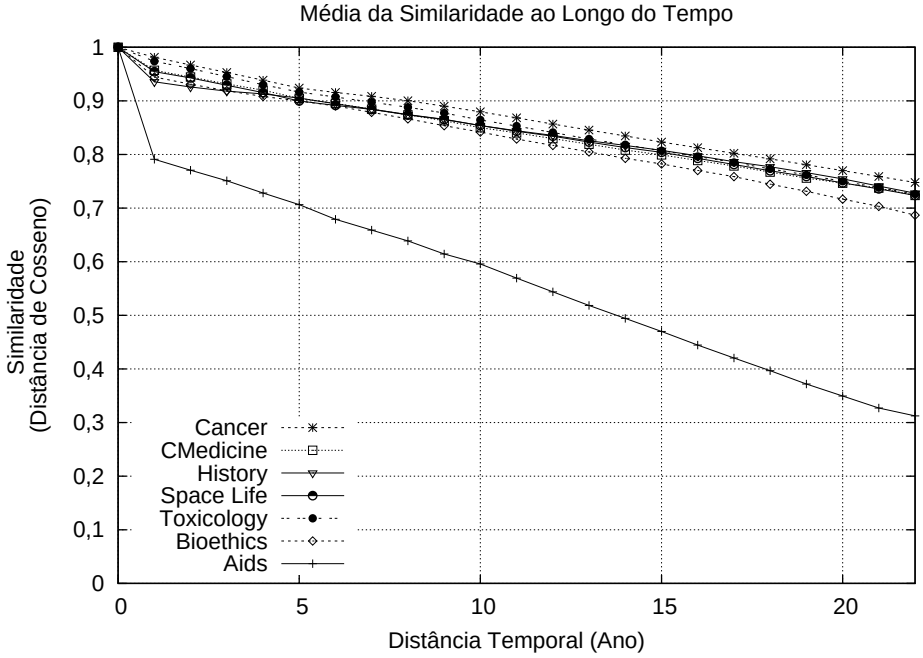

Figura 4.18. Média da Distribuição dos Termos - MedLine

4.2.4 Similaridade de Classes

Por fim, nesta seção analisamos o último desafio: a evolução da similaridade de classes. Esse desafio é interessante dada a migração e a variação da freqüência dos termos nos vocabulários das classes ao longo do tempo. É importante perceber que isso está

associado às características interclasses.

Para analisar esse fenômeno mais detalhadamente, conduzimos o experimento descrito a seguir. Novamente a coleção foi dividida em anos e selecionamos aleatoriamente o mesmo número de documentos por classe em cada ano. Além disso, como também feito anteriormente, criamos um vocabulário $V_{k,i}$, que é formado pelas t palavras que contêm os maiores valores de *info-gain* na classe C_k no ano A_i. Para simplificar a análise do experimento, consideramos apenas duas classes e calculamos a similaridade de cosseno entre os vetores de vocabulário dessas duas classes em um dado ano A_i, variando A_i para todos os anos das coleções. Assim, a Figura 4.19 mostra um dos gráficos obtidos para a coleção da ACM enquanto a Figura 4.20 mostra um exemplo dos gráficos obtidos para a coleção da MedLine. Nesses gráficos apresentados, o eixo x representa então cada ano existente na coleção de documentos. O eixo y, para a coleção da ACM, representa a similaridade de cosseno entre o vetor de vocabulário com os t termos de maior *info-gain* da Classe "Data" e o vetor de vocabulário com os t termos de maior *info-gain* da Classe "CMilieux", para cada um dos anos. Já para a coleção da MedLine, o eixo y representa a similaridade de cosseno entre o vetor de vocabulário com os t termos de maior *info-gain* da Classe "Aids" e o vetor de vocabulário com os t termos de maior *info-gain* da Classe "Toxicology", para cada um dos anos.

Como se pode observar, os vetores de vocabulário de duas classes são menos ou mais semelhantes em diferentes períodos de tempo. Por exemplo, as classes *Data* e *Mileux*

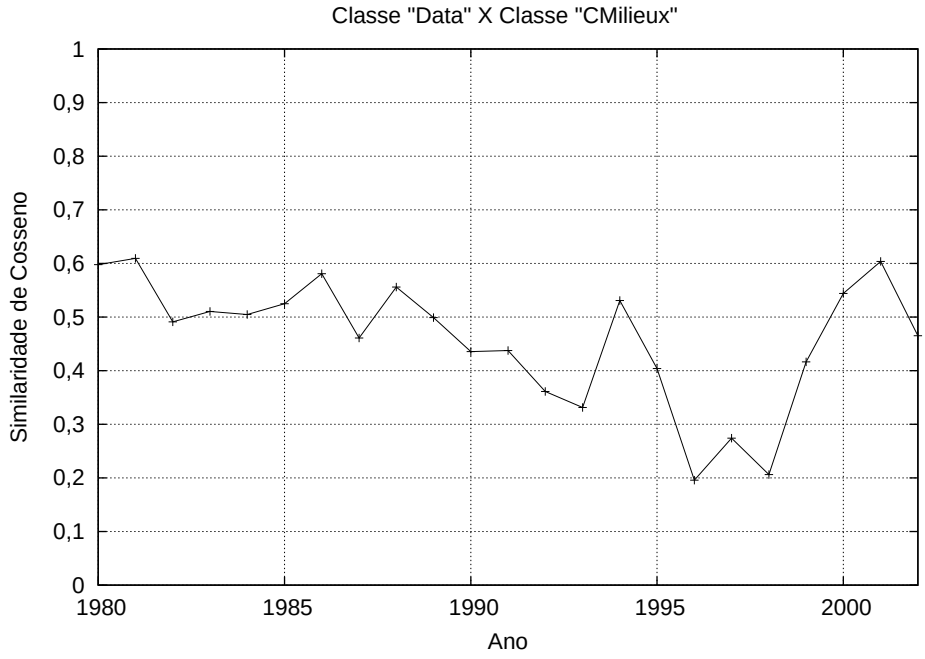

Figura 4.19. Similaridade de Cossenos ao Longo do Tempo - ACM

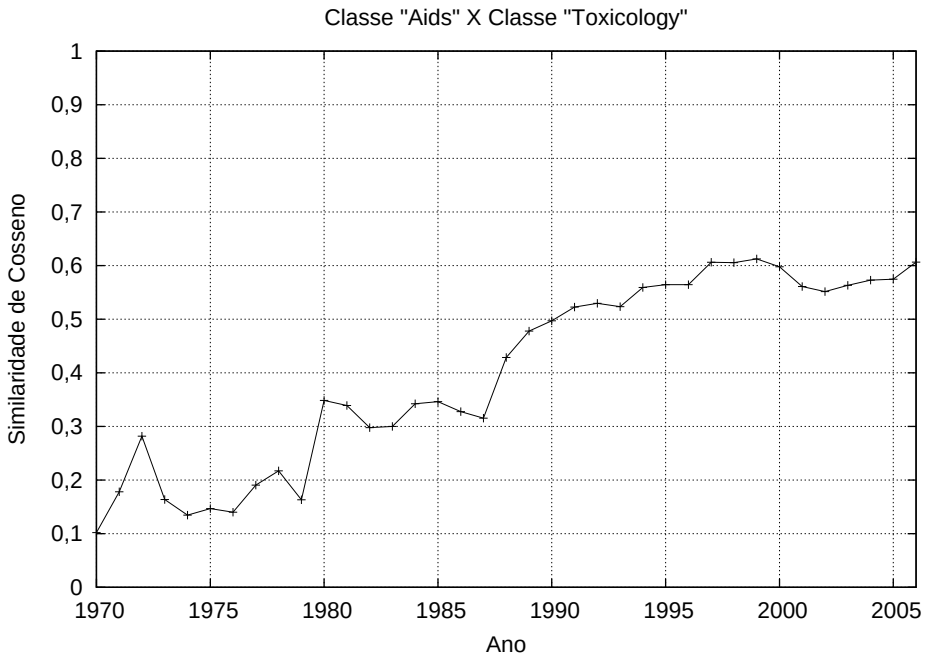

Figura 4.20. Similaridade de Cossenos ao Longo do Tempo - MedLine

da coleção da ACM são menos similares entre si no período entre 1996 a 1998. Isso indica que se formos classificar documentos que se encontram nesse intervalo de tempo utilizando como base de treino apenas documentos desse período específico, a acurácia do classificador pode ser melhor do que quando utilizamos documentos provenientes da base toda como conjunto de treino. Isso se deve ao fato de que, se os vocabulários usados para treinar o classificador para duas classes forem diferentes dos vocabulários dessas classes no período em que elas realmente apresentam mais diferenças entre si, as duas classes tendem a ser consideradas mais similares do que realmente são.

Observações similares são válidas para as classes *Aids* e *Toxicology* da coleção da MedLine. A Figura 4.20 mostra que essas classes estão se tornando mais similares entre si desde a década de 70. Conseqüentemente, a classificação de documentos de ambas as classes é mais fácil de ser realizada para documentos das décadas de 70 e 80, do que para documentos do ano de 2000.

As Tabelas 4.1 e 4.2 mostram a variação ao longo do tempo na similaridade entre cada par de classes para a coleção da ACM e da MedLine, respectivamente. Um valor na tabela representa o desvio padrão em relação à média das similaridades ao longo de todos os anos entre o par de classes relativo à linha e à coluna da tabela. Como é possível observar, a variação da similaridade é muito alta para alguns pares de classes. Por exemplo, na coleção da MedLine, o desvio padrão em relação à média de similaridade entre a classe *Complementary Medicine* (CM) e a classe *History* (Hist) é

21%, sendo esse valor muito maior que os demais. Isso significa que essas duas classes podem ter sido muito similares em alguns períodos de tempo e não muito similares em outros. Conseqüentemente, a dificuldade em diferenciar as duas classes também varia consideravelmente ao longo do tempo. Outra observação importante é que a classe *Cancer* (da coleção da MedLine) difere das outras classes, já que mantém uma baixa variação de sua similaridade em relação a todas as outras classes ao longo dos anos. Isso indica que o efeito temporal afeta menos essa classe do que outras classes da MedLine.

	Lit	HW	C S O	SW	Data	Theory	Math	Info S	C M	C App	Milieux
Lit	0	0.14	0.12	0.12	0.12	0.29	0.14	0.13	0.14	0.12	0.29
HW	-	0	0.08	0.13	0.11	0.12	0.11	0.12	0.10	0.10	0.13
C S O	-	-	0	0.10	0.09	0.10	0.08	0.07	0.08	0.10	0.13
SW	-	-	-	0	0.09	0.06	0.09	0.10	0.11	0.12	0.13
Data	-	-	-	-	0	0.05	0.08	0.09	0.10	0.13	0.13
Theory	-	-	-	-	-	0	0.14	0.13	0.07	0.06	0.29
Math	-	-	-	-	-	-	0	0.13	0.10	0.09	0.15
Info S	-	-	-	-	-	-	-	0	0.10	0.08	0.15
C M	-	-	-	-	-	-	-	-	0	0.11	0.13
C App	-	-	-	-	-	-	-	-	-	0	0.12
Milieux	-	-	-	-	-	-	-	-	-	-	0

Tabela 4.1. Matriz do Desvio Padrão da Similaridade - ACM

	Aids	Bio	Cancer	C M	Hist	Space L	Tox
Aids	0	0.19	0.16	0.18	0.19	0.18	0.19
Bio	-	0	0.04	0.20	0.17	0.19	0.12
Cancer	-	-	0	0.04	0.03	0.04	0.05
C M	-	-	-	0	0.21	0.08	0.05
Hist	-	-	-	-	0	0.20	0.11
SpaceL	-	-	-	-	-	0	0.05
Tox	-	-	-	-	-	-	0

Tabela 4.2. Matriz do Desvio Padrão da Similaridade - MedLine

Em resumo, baseado nas caracterizações dos desafios apresentadas anteriormente, foi possível mostrar evidências da variação de classes e termos ao longo do tempo, assim como evidências de que a similaridade entre essas classes também varia. Além disso, confirmamos que esses fenômenos têm grande influência na acurácia dos classificadores. Esses resultados estão também apresentados em Mourão et al. [2008].

Uma vez confirmada a existência e a influência dos efeitos temporais no processo de Classificação Automática de Documentos, o próximo passo é, então, tratar esses efeitos para tornar a tarefa de classificação mais eficaz. Dessa forma, no próximo capítulo, apresentaremos várias estratégias com tal objetivo.

Capítulo 5

Estratégias Temporais de Engenharia de Dados

Neste capítulo, mostraremos algumas estratégias que foram implementadas e avaliadas, com o objetivo de melhorar o processo de classificação. Essas estratégias consistem em realizar um processo de engenharia de dados, através da filtragem, adaptação e transformação dos mesmos, de forma que seja possível tratar os efeitos temporais previamente analisados. Como foi demonstrado que esses efeitos têm uma influência significativa nos classificadores automáticos de documentos, esperamos tornar a tarefa de classificação mais eficaz através do uso das estratégias criadas.

Antes de optarmos pelo processo de engenharia de dados, consideramos também uma outra vertente possível para lidar com os efeitos temporais no processo de classificação: a adaptação de algoritmos de classificação. Entretanto, apesar de interessante, a adaptação de um algoritmo seria uma alternativa um pouco mais restrita em relação à extensibilidade da solução, já que se restringiria a apenas um algoritmo específico, não sendo possível aplicá-la a outros que utilizassem métodos diferentes para realizar a tarefa de classificação. A engenharia de dados, por sua vez, não se concentra em nenhuma característica específica de algum algoritmo, sendo possível aplicar essas estratégias a vários algoritmos de classificação. Assim, essa última opção se torna mais atraente do que a adaptação de um algoritmo e será, portanto, utilizada neste trabalho.

Surge, então, o desafio de como alterar a base de dados para tratar os efeitos temporais, objetivando-se que, de alguma forma, a transformação dos dados contribua para o classificador atuar de forma mais eficaz, dado o contexto temporal. A seguir, apresentaremos algumas estratégias que foram criadas para explorar esse contexto.

5.1 Metodologia

Como mencionado anteriormente, utilizaremos um processo de engenharia de dados com o intuito de tratar os aspectos temporais na tarefa de classificação. Podemos dividir esse processo em dois subprocessos principais: a filtragem dos dados e a transformação dos dados.

O processo de filtragem dos dados, ao contrário do processo de transformação, não realiza uma modificação no conteúdo da base de dados. Ele consiste em selecionar documentos, através de algum critério específico, para constituírem o conjunto de treino que irá gerar o modelo de classificação. O critério escolhido neste trabalho foi a proximidade temporal desses documentos aos documentos que serão testados. Assim, podemos construir uma janela temporal ao redor dos documentos de teste, que determinará quais os documentos devem ser considerados para o conjunto de treino. É interessante ainda variar o tamanho dessa janela temporal, de forma a descobrirmos qual é o seu tamanho ótimo, ou seja, aquela através da qual iremos obter a melhor acurácia do algoritmo de classificação. Para tal, propomos uma estratégia exaustiva de filtragem, que será apresentada na Seção 5.2.

O processo de transformação de dados, por sua vez, realmente modifica os dados presentes na base de dados e não apenas faz uma filtragem dos documentos do conjunto de treino. Lembramos aqui que as bases de dados são constituídas pelos documentos presentes nela, em que cada linha representa um documento e, nessa linha, temos os termos que aparecem nesse documento específico. Assim, uma forma que encontramos de transformar esses dados consiste em gerar novos rótulos para esses termos, de forma que eles passem a incorporar alguma informação temporal. Ao fazermos isso, estamos modificando o espaço em que o algoritmo SVM atua, uma vez que estamos gerando "novos termos" (novos rótulos), ao mesmo tempo que alguns termos podem desaparecer, caso todas as suas ocorrências tenham ganhado novos rótulos. Portanto, podemos acabar alterando a dimensionalidade do espaço do SVM, ao mudarmos os rótulos dos termos. Ao mesmo tempo, podemos aproximar dois documentos que estavam distantes nesse espaço, caso a transformação faça com que tenham mais termos em comum, e distanciar dois documentos que estavam originalmente próximos nesse espaço, caso a transformação faça com que os termos que eles tinham em comum sejam modificados.

Um exemplo simples da modificação dos termos seria a concatenação de cada termo do documento com o ano a que ele pertence. Por exemplo, um documento do ano de 1980 que contenha o termo "Tucson" teria esse termo modificado para "Tucson1980" e um documento de 2006 que tenha esse mesmo termo, passaria a ter o termo "Tucson2006". Portanto, esses dois documentos, que antes da transformação continham um termo em comum, após a transformação não mais o possuem. Isso pode ser interessante

uma vez que, devido à evolução temporal, realmente faça sentido que esses termos sejam considerados diferentes. No caso desse exemplo, em 2005 foi lançado um carro que tem o nome "Tucson". Antes dessa data, a palavra se referia principalmente à cidade localizada no estado do Arizona, nos Estados Unidos. Portanto, existe a possibilidade que o documento de 2006 esteja se referindo ao carro e não à cidade e, dessa forma, não o consideraremos da mesma maneira. Assim, estamos inserindo informações temporais na base de dados, para que o modelo de classificação não utilize padrões que não são mais verdadeiros, devido à evolução temporal.

Percebe-se, então, que a transformação dos termos pela atribuição de novos rótulos é uma maneira através da qual consegue-se influenciar o processo de classificação, tratando-se os aspectos temporais. Assim, analisaremos agora a intuição utilizada em cada estratégia de transformação criada, uma vez que elas foram elaboradas de forma incremental. Nesta seção, isso será feito de forma superficial, apenas para mostrar como elas foram desenvolvidas e como as idéias evoluíram. Posteriormente, entretanto, será apresentada cada uma das estratégias de forma detalhada, assim como os resultados obtidos através da aplicação das mesmas.

A primeira estratégia de transformação dos dados, proposta na Seção 5.3, é uma forma simples de incorporar informação temporal na base de dados e consiste exatamente em concatenar o ano de cada documento aos seus termos, conforme exemplificado anteriormente. Ela é uma tentativa de abordar problemas como o descrito para o caso do termo "Tucson", em que devido à evolução temporal, ele pode ter um significado diferente em períodos de tempo diferentes da base de dados.

Em nossa segunda estratégia, apresentada na Seção 5.4, incorporamos a idéia de termos predominantes em um dado ano. Chamamos de termo predominante aquele que julgamos ter um poder discriminativo maior que outros termos, dado o conjunto analisado. Assim, se um termo é considerado predominante nos anos de 1982, 1990 e 2001, por exemplo, ele receberá um novo rótulo (igual para todos os anos) nos documentos desses anos, enquanto nos outros documentos ele não terá o seu rótulo modificado. Dessa forma, estamos tentando separar as épocas em que a ocorrência daquele termo possui um poder discriminativo maior (que pode ajudar a determinar a qual classe o documento deve ser associado), das épocas em que aquele termo não possui esse poder discriminativo no conjunto de documentos.

Na Seção 5.5, apresentamos a terceira estratégia, que mantém o conceito de termos predominantes, porém, ao invés de considerar um termo predominante em um dado ano, ela analisa se um termo é predominante em uma dada classe em um dado ano. Ou seja, estamos aumentando a granulação da análise ao introduzirmos uma outra dimensão: a de classes. Como nosso objetivo é melhorar a acurácia do processo de assinalamento

de documentos às classes que eles pertencem, faz sentido fazermos essa análise sob essa condição. Assim, se um termo for predominante na classe A nos anos 1985, 1987 e 1999, por exemplo, ele terá o seu rótulo modificado somente em documentos que pertençam à classe A naqueles anos. Estamos, portanto, separando termos com poder discriminativo em cada classe nos períodos de tempo em que eles possuem esse poder. Como, em cada classe, a freqüência mínima para que um termo seja considerado discriminante pode variar, aplicamos duas estratégias diferentes baseadas no conceito explicado: uma que contém um limiar mínimo fixo para determinar se o termo é predominante ou não; e outra que contém esse limiar mínimo flexível, que pode variar de classe para classe.

Por fim, na Seção 5.6, propomos uma quarta estratégia, que tenta associar o conceito de localidade temporal à ultima estratégia criada. Assim, para determinar se um termo é predominante em uma dada classe em um dado ano, não consideramos apenas as informações do termo no ano em questão, mas também verificamos as informações desse termo naquela classe nos anos próximos ao ano do documento sendo analisado. Assim, mesmo se o termo não for considerado predominante naquela classe e ano específicos, se ele o for nos anos próximos temporalmente a ele, ele pode passar a ser considerado predominante no ano em questão. Com isso, determinamos uma janela temporal dentro da qual consideraremos os documentos que estão temporalmente próximos ao ano do documento sendo analisado, e variamos o tamanho dessa janela para avaliar o impacto que ela tem na acurácia do classificador. Novamente, essa estratégia foi aplicada utilizando-se um limiar de predominância mínimo fixo e um limiar de predominância mínimo variável.

Assim, nas próximas seções, apresentaremos primeiramente nossa estratégia em que os dados são filtrados, objetivando-se ponderar o compromisso entre o problema de amostragem e a evolução temporal, já mencionado anteriormente. E, em seguida, apresentamos algumas estratégias de transformação utilizadas para a alteração das bases de dados e o impacto delas no classificador SVM.

5.2 Estratégia Exaustiva de Filtragem de Dados

Nesta seção, apresentaremos uma estratégia exaustiva de filtragem de dados. Através da utilização dessa estratégia, objetivamos aprimorar o processo de classificação, de forma que ele se torne mais eficaz.

Melhoras podem ser obtidas considerando tanto o efeito amostral quanto o efeito temporal. O desafio, entretanto, é lidar com esses dois efeitos de forma simultânea, uma vez que eles exigem estratégias opostas. Apenas aumentar descriteriosamente o tamanho do conjunto de treino, por exemplo, pode não produzir uma melhora na efi-

cácia. Isso se deve ao fato de que esse aumento implica em uma maior quantidade de dados provenientes de anos que não estão temporalmente próximos aos documentos de teste e, conforme já discutido anteriormente, isso pode introduzir ruídos na base de dados. Por outro lado, reduzir o conjunto de treino a apenas documentos temporalmente próximos aos documentos de teste certamente reduzirá o tamanho da amostra de treino. Portanto, existe um compromisso entre esses dois efeitos.

Nossa estratégia é otimizar esse compromisso e, para tal, o classificador deve utilizar um processo para selecionar documentos para a base de treino que considere os efeitos temporais. Ou seja, para cada documento sendo testado, selecionamos para constituir o conjunto de treino apenas os documentos que estão temporalmente próximos ao documento de teste. A proximidade é definida por uma janela que pode crescer simetricamente nas duas direções, passado e futuro, partindo do ano ao qual o documento de teste pertence. Para considerar também o efeito amostral, variamos o tamanho da janela de 0 (ou seja, o próprio ano A_i ao qual o documento de teste pertence) a N, em que N representa o número de anos antes e depois de A_i. O valor de N é definido como o valor que maximiza a acurácia do classificador para os documentos que pertencem ao ano A_i.

A Figura 5.1 e a Figura 5.2 mostram a acurácia do classificador à medida que variamos o tamanho da janela para os documentos de teste dos anos (aleatoriamente escolhidos) 1990 e 1995, respectivamente, na coleção da ACM. A Figura 5.3 e a Figura 5.4 mostram a acurácia do classificador ao variarmos o tamanho da janela para documentos de teste dos anos (também escolhidos aleatoriamente) 1971 e 1974, respectivamente, na coleção da MedLine. O eixo x desses gráficos representa o tamanho da janela temporal em anos utilizada, e o eixo y representa a acurácia obtida pelo SVM ao classificar os documentos de um determinado ano (no caso, 1990, 1995, 1971 e 1974), a partir de uma base de treino composta por documentos que pertencem a cada janela temporal considerada.

A partir dos exemplos dados, podemos ver que, em ambas as coleções, não existe apenas um tamanho ótimo único da janela para a base de dados inteira, ou seja, um tamanho ótimo da janela pode ser encontrado para cada ano específico. Por exemplo, na coleção da ACM, para documentos de 1990, o tamanho ótimo da janela é 10 anos. Para os documentos de 1995, entretanto, o tamanho ótimo é 19 anos. Na coleção da MedLine, por sua vez, o tamanho ótimo para os documentos de 1971 é 0 (ou seja, deve-se considerar como treino apenas documentos do próprio ano de 1971), enquanto para os documentos de 1974, esse tamanho ótimo é 7 anos.

Apesar de o mecanismo utilizado para determinar a janela temporal de tamanho ótimo para cada ano ser apropriado para uma análise da evolução temporal, ele é inviá-

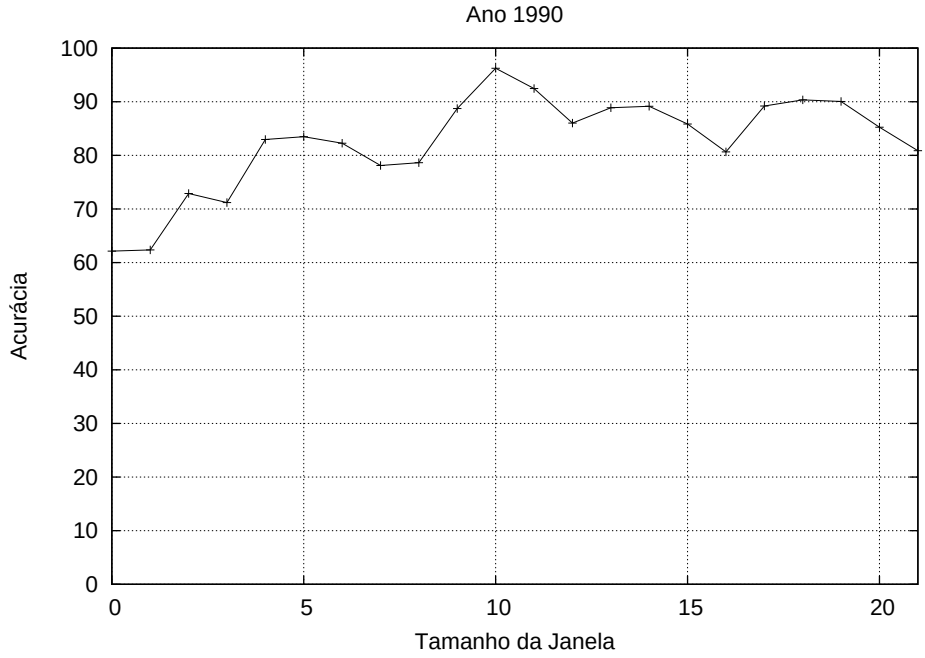

Figura 5.1. Classificação com Janela Deslizante - 1990 - ACM

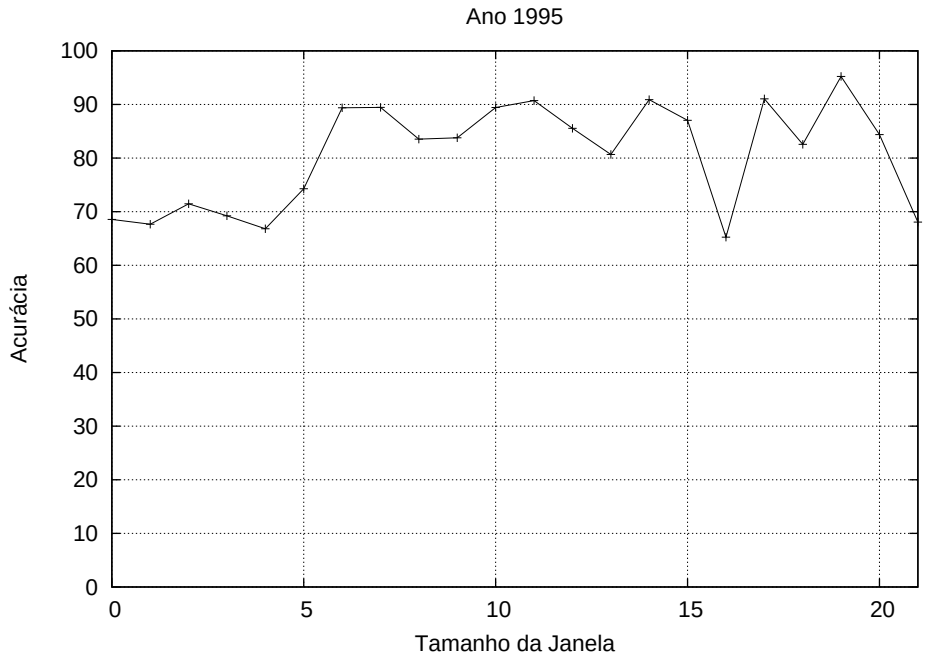

Figura 5.2. Classificação com Janela Deslizante - 1995 - ACM

vel em cenários reais, uma vez que requer uma avaliação exaustiva para cada um desses anos. Entretanto, como será explicado a seguir, a acurácia do processo de classificação quando consideramos uma janela temporal, mesmo que não seja ótima para cada um dos anos, é maior do que quando a base de treino inteira é utilizada. Temos ainda que

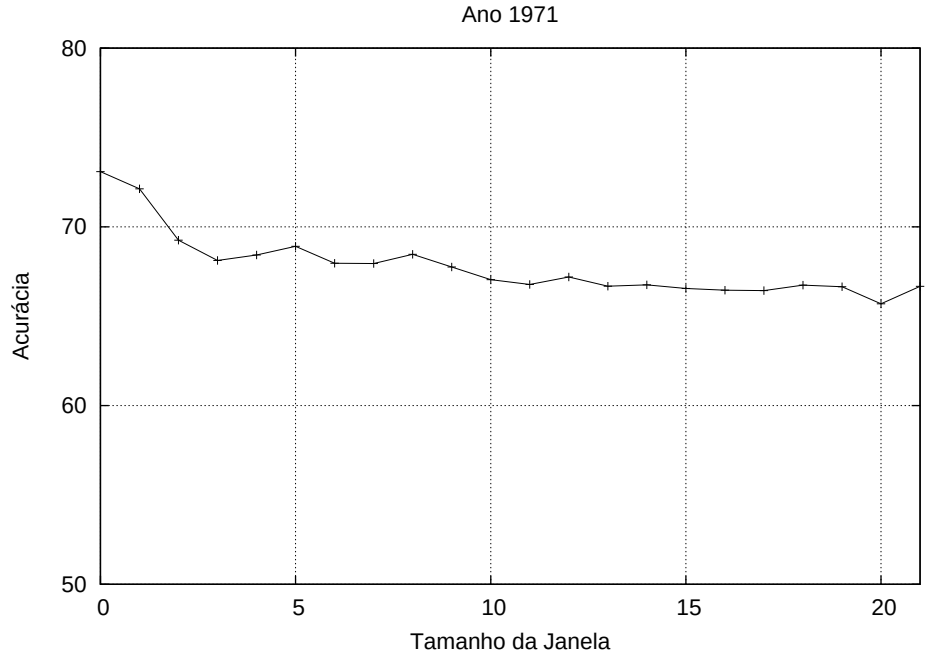

Figura 5.3. Classificação com Janela Deslizante - 1971 - MedLine

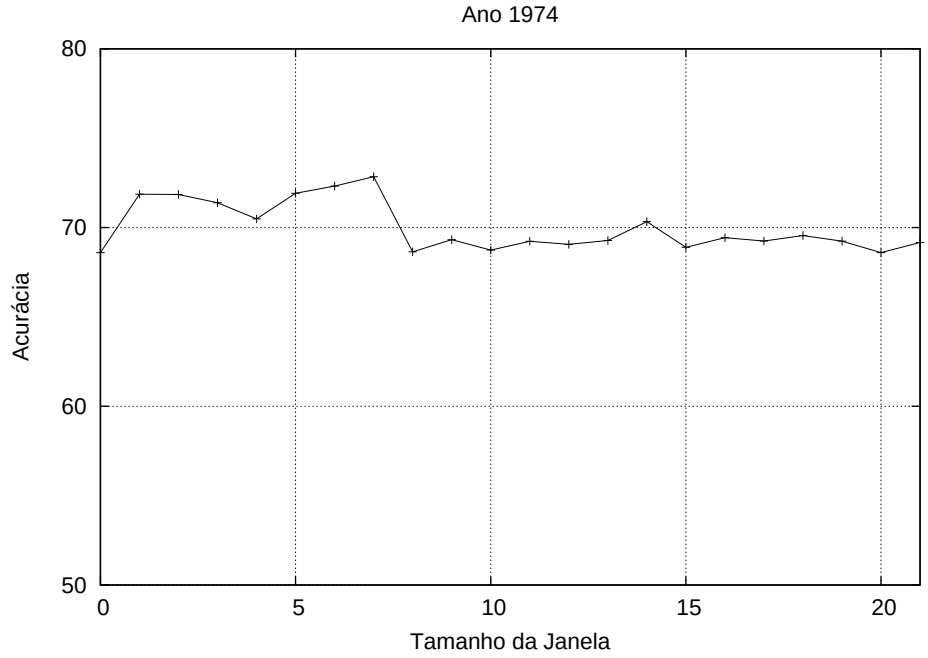

Figura 5.4. Classificação com Janela Deslizante - 1974 - MedLine

um classificador que considera os efeitos temporais, além de melhorar a acurácia, pode melhorar também o seu desempenho em relação à eficiência computacional, já que ele pode ser utilizado com um conjunto de treino muito menor.

Assim, primeiramente, mostraremos que melhoras podem ser obtidas com o uso de

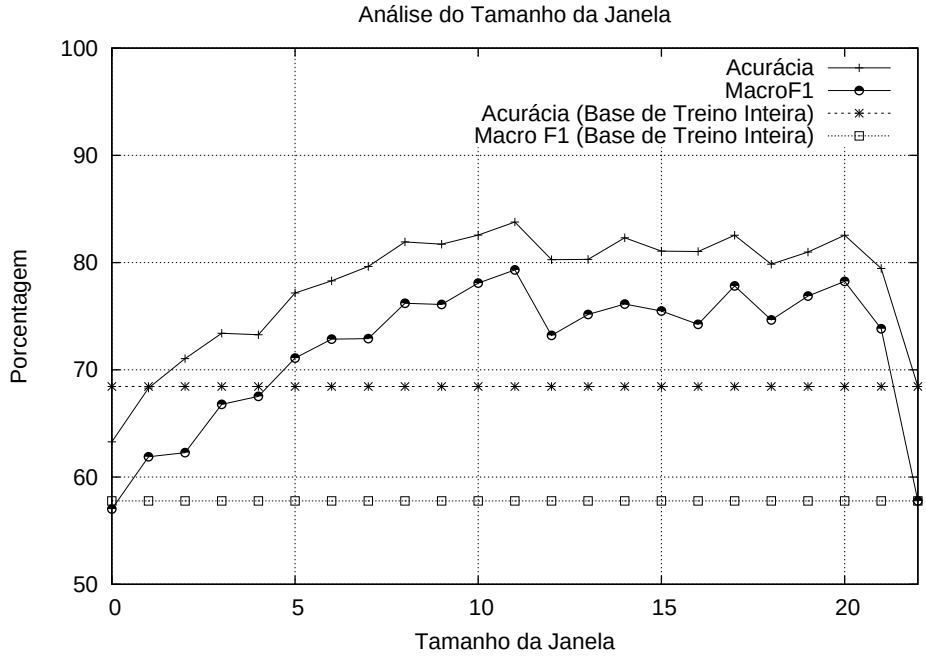

Figura 5.5. Acurácia Variando-se o Tamanho da Janela - ACM

uma janela de tamanho médio, isto é, a mesma janela para todos os anos. A Figura 5.5 e a Figura 5.6 mostram, para a coleção da ACM e para a coleção da MedLine, respectivamente, a acurácia do algoritmo SVM quando utilizamos a base de treino inteira e

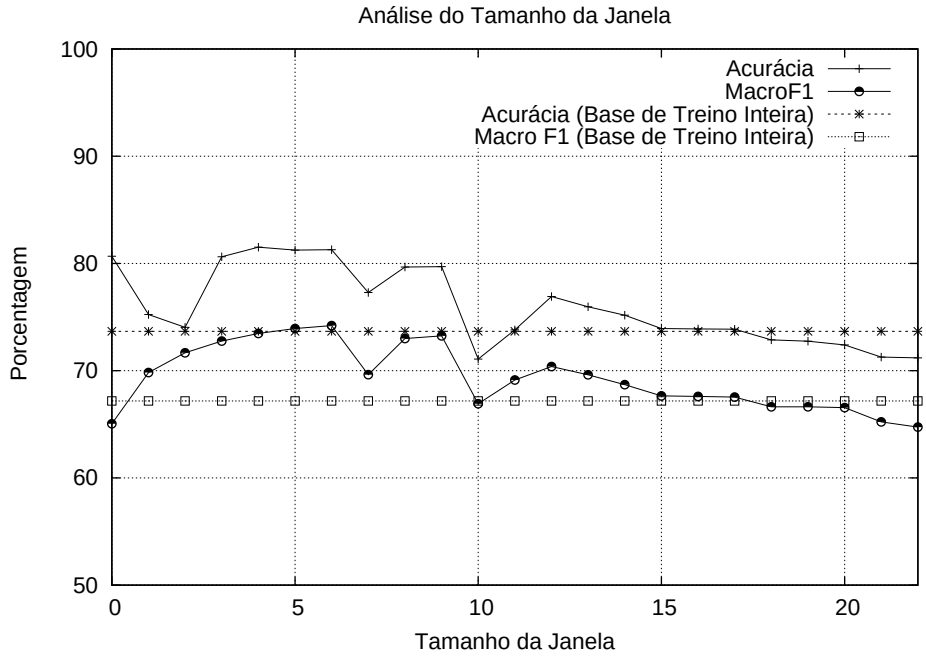

Figura 5.6. Acurácia Variando-se o Tamanho da Janela - MedLine

quando utilizamos uma janela temporal de tamanho fixo para todos os anos. Além da acurácia, utilizamos também a métrica MacroF$_1$ (definida na Seção 3.3.2), para avaliar o processo de classificação, sendo essa métrica também apresentada nos gráficos. O eixo x desses gráficos representa novamente o tamanho da janela temporal em anos. Já o eixo y representa a porcentagem da acurácia e a porcentagem da MacroF$_1$, obtidas pelo SVM, ao classificar os documentos a partir de uma base de treino composta por todos os documentos da base (obtendo-se duas curvas constantes, uma vez que, nesse caso, não houve variação do tamanho da janela temporal), e representa também a porcentagem da acurácia e a porcentagem da MacroF$_1$, obtidas pelo SVM, ao classificar os documentos a partir de uma base de treino composta por documentos pertencentes a cada janela temporal considerada.

Para ambas as coleções, pode-se perceber que, mesmo quando não consideramos uma janela ótima para cada ano, utilizar uma janela ótima global é ainda melhor do que utilizar todo o conjunto de treino.

Os gráficos da Figura 5.7 e da Figura 5.8, por sua vez, mostram a relação do tamanho do conjunto de treino e do tamanho da janela temporal, para as coleções ACM e MedLine, respectivamente. O eixo x desses gráficos representa o tamanho em anos das janelas temporais consideradas, e o eixo y representa o tamanho da base de treino gerada por cada janela temporal em relação ao tamanho da base de treino original. Assim, essa informação foi obtida através do cálculo da média do número

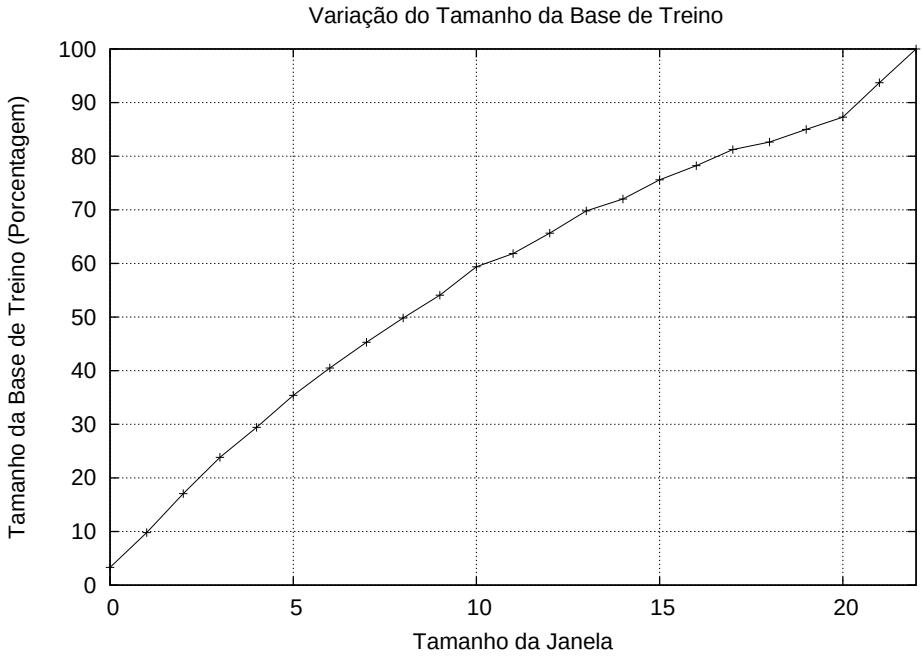

Figura 5.7. Tamanho da Coleção - ACM

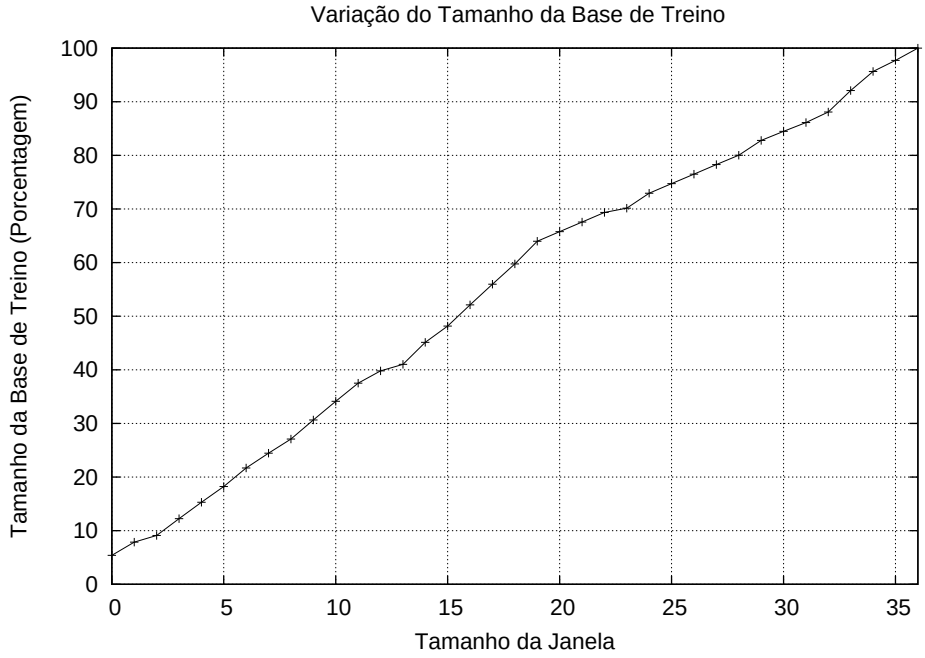

Figura 5.8. Tamanho das Coleções - MedLine

de documentos presentes dentro da janela temporal de um dado tamanho, quando a deslizamos sobre todos os anos, em relação ao número total de documentos presentes na base de treino original.

Analisando a Figura 5.5, pode-se perceber que, com o tamanho da janela temporal igual a 5, as duas métricas (acurácia e MacroF$_1$) apresentam-se melhores do que ao se utilizar todo o conjunto de treino. Relacionando esse dado com a informação obtida na Figura 5.7, temos que isso é alcançado utilizando-se apenas 33% de todo o conjunto de treino da ACM. Para essa coleção, pode-se notar ainda que o tamanho ótimo da janela global é 11 anos, o que corresponde a 61% de todo o conjunto de treino. Já para a base de dados da MedLine, podemos perceber que o uso de apenas 13% (equivalente ao tamanho de janela igual a 3) é suficiente para atingir uma acurácia consideravelmente maior e uma MacroF$_1$ também superior à MacroF$_1$ atingida ao se utilizar toda a coleção. Para essa coleção, o melhor resultado foi atingido usando um tamanho de janela igual a 4, o que representa 16% de todo o conjunto de treino.

Após ter sido feita essa análise da janela global ótima, vamos estudar agora a estratégia de explorar uma janela de tamanho ótimo para cada ano. Na Figura 5.9 e na Figura 5.10, mostramos o tamanho ótimo da janela temporal para cada ano, para a coleção ACM e MedLine, respectivamente. Dessa forma, o eixo x desses gráficos representa cada ano existente na coleção de documentos, e o eixo y representa o tamanho em anos da janela temporal ótima para cada ano específico. Através dos gráficos pode-se

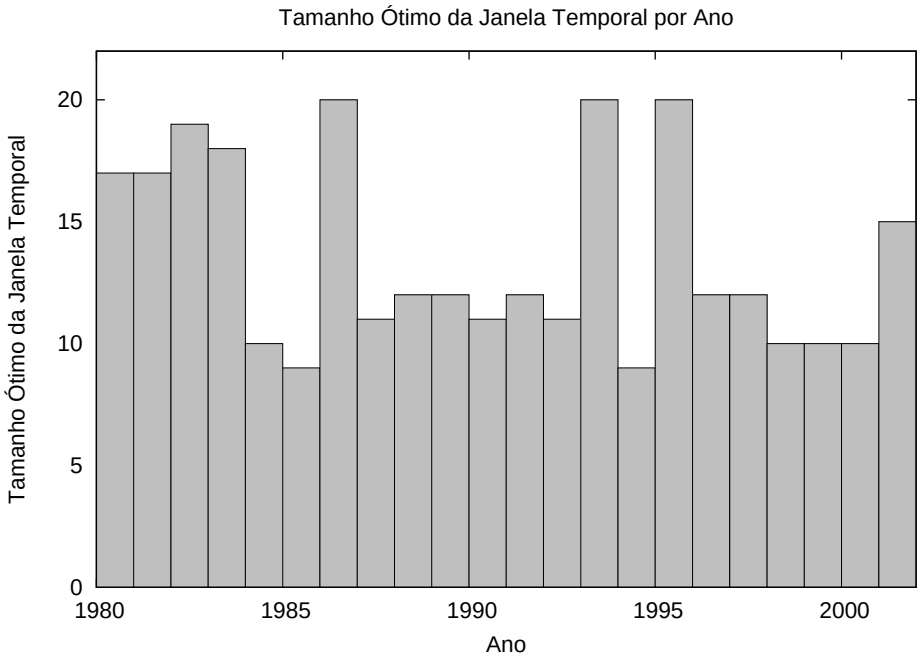

Figura 5.9. Análise Da Janela Temporal Ótima Por Ano - ACM

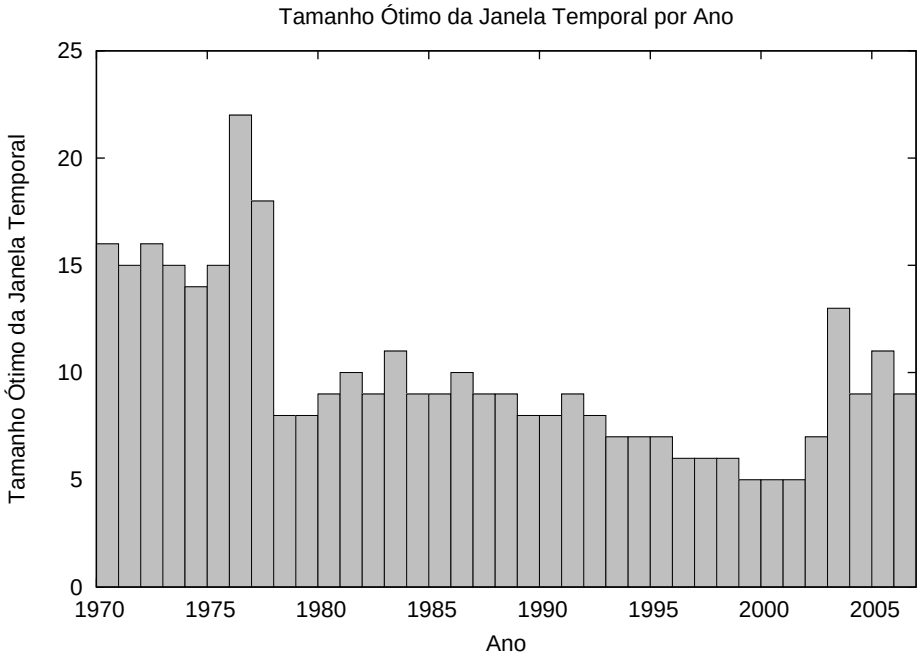

Figura 5.10. Análise Da Janela Temporal Ótima Por Ano - MedLine

perceber que esse valor varia consideravelmente.

Já a Figura 5.11 e a Figura 5.12 mostram os maiores valores da acurácia encontrados (quando utilizamos uma janela temporal ótima por ano) para cada ano nas duas coleções. Assim, o eixo x desses gráficos representa novamente cada ano existente na

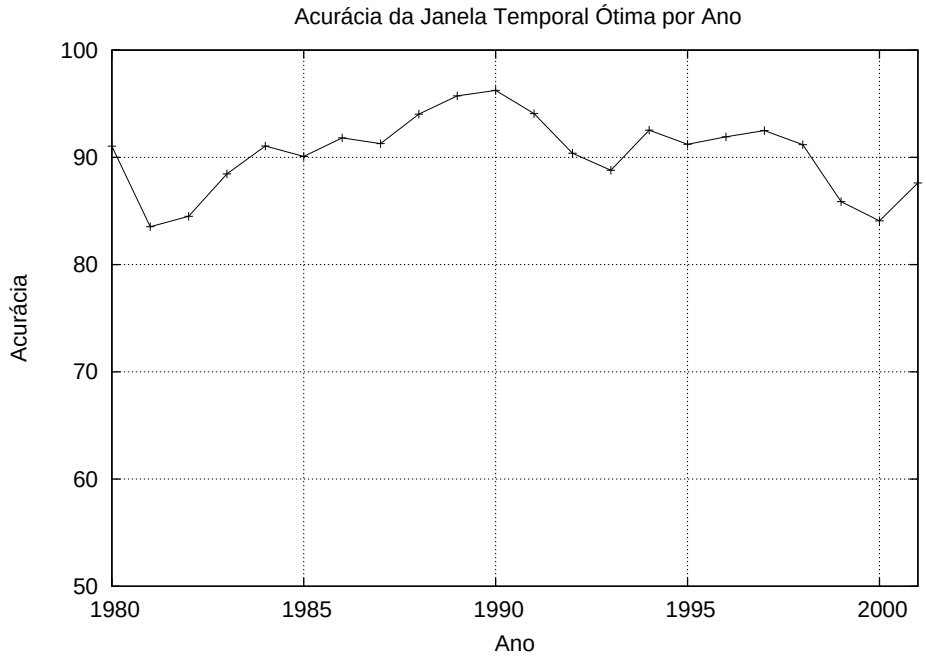

Figura 5.11. Acurácia Utilizando a Janela Temporal Ótima Por Ano - ACM

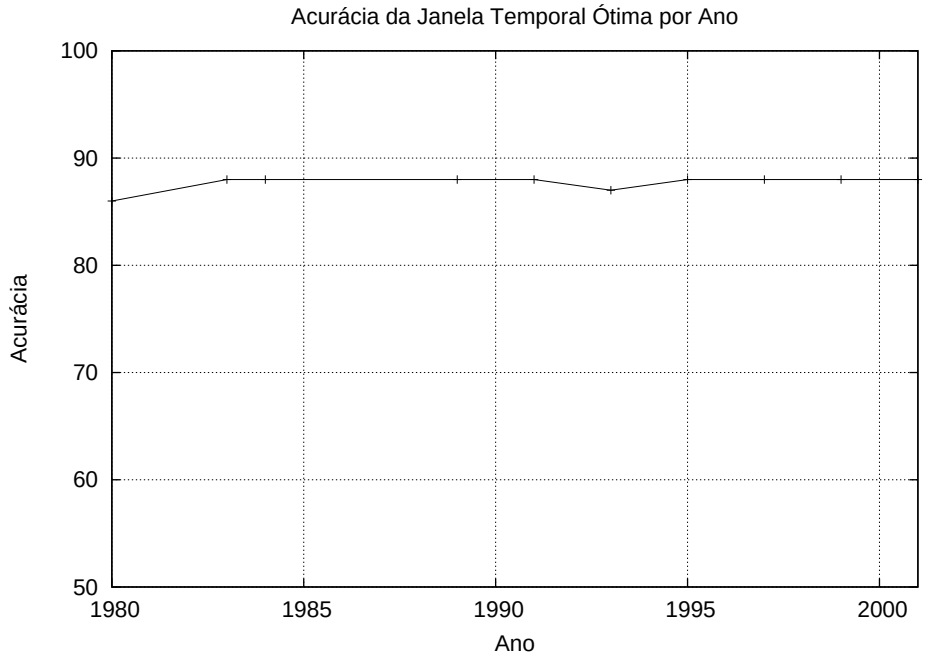

Figura 5.12. Acurácia Utilizando a Janela Temporal Ótima Por Ano - MedLine

coleção de documentos, e o eixo y representa a acurácia obtida pelo SVM ao classificar os documentos de cada ano, utilizando como base de treino os documentos que estão dentro da janela ótima encontrada para aquele ano. Analisando esses gráficos, podemos perceber que a média da acurácia do classificador é maior do que aquela encontrada

quando utilizamos apenas um tamanho de janela global para todos os anos (em que obtemos uma acurácia máxima de 83,7% e 81,5% para as coleções da ACM e MedLine, respectivamente, como mostrado na Figura 5.5 e na Figura 5.6). Além disso, quando calculamos todos os documentos classificados corretamente em todo o conjunto de teste (quando utilizamos uma janela ótima para cada ano), obtemos uma acurácia de 89,76% para a ACM e 87,57% para a MedLine. Assim, variando o tamanho da janela temporal para cada ano, obtemos valores consideravelmente maiores para ambas as coleções.

Esses resultados são ainda melhores quando comparados aos resultados encontrados quando usamos todo o conjunto de treino para o processo de classificação, em que não são considerados os aspectos temporais. Nesse último, obtivemos uma acurácia de apenas $68,44\%$ para a coleção da ACM e $73,67\%$ para a coleção da MedLine. Dessa forma, para o método que considera uma janela temporal ótima para cada ano, temos um ganho de 21% na acurácia, na coleção da ACM, e um ganho de aproximadamente 14% para a coleção da MedLine. Ganhos no tempo de execução também são obtidos quando aplicamos uma janela temporal ótima diferente para cada ano. A janela ótima gera um conjunto de treino de aproximadamente 69% do conjunto de treino da coleção original da ACM e 25% para a coleção da MedLine. Uma vez que o tempo de execução do SVM está diretamente relacionado ao tamanho do conjunto de treino, sua redução faz com que o tempo de execução também seja menor.

	Acurácia	Ganho	% da Base de Treino
Base original	68,44%	-	100
Janela Global Ótima	83,70%	15,26%	61
Janela Ótima por Ano	89,76%	21,32%	69

Tabela 5.1. Resultados da Filtragem - ACM

	Acurácia	Ganho	% da Base de Treino
Base original	73,67%	-	100
Janela Global Ótima	81,50%	7,83%	22
Janela Ótima por Ano	87,57%	13,90%	25

Tabela 5.2. Resultados da Filtragem - MedLine

Os resultados obtidos com esse estudo estão sumarizados na Tabela 5.1 e na Tabela 5.2. Nessas tabelas, apresentamos, para as duas coleções, as acurácias obtidas para a janela temporal ótima global, para a janela temporal ótima por ano e para a base original sem nenhum processo de filtragem. Além disso, apresentamos o ganho dessas acurácias em relação àquela obtida utilizando-se a base original sem a aplicação

de nenhuma janela temporal. Por fim, a informação do tamanho das bases de treino utilizadas em cada um desses processos também está contida nessas tabelas.

Em resumo, nossa análise empírica mostra que, ao explorar os aspectos temporais de forma apropriada, podemos obter melhoras relevantes no processo de classificação. Apesar de a estratégia exaustiva ser adequada para uma avaliação analítica, como aplicada no contexto deste trabalho, certamente é inviável em cenários reais. Desse ponto de vista, surge, como um desafio maior, a necessidade de novas estratégias para melhorar a qualidade da Classificação Automática de Documentos, considerando os efeitos temporais.

Assim, apresentaremos, a seguir, estratégias que envolvem a transformação dos dados. É importante ressaltar que, nas próximas estratégias propostas, mostraremos os resultados dos experimentos apenas para a base da ACM, com o intuito de aumentar a clareza e facilitar o entendimento dos mesmos. Optamos pela coleção da ACM, e não da MedLine, uma vez que ela foi a que apresentou um maior ganho ao se considerar os aspectos temporais, sendo, portanto, mais afetada por eles.

5.3 Agregação do Ano a Termos

Após ter sido realizada uma estratégia exaustiva de filtragem dos documentos que irão compor o conjunto de treino (utilizado para a construção do modelo de classificação), será apresentada a primeira das estratégias que transformam os dados da base de documentos.

A abordagem utilizada aqui é uma forma simples de incorporar informação temporal na base de dados. Assim como as outras estratégias de transformação de dados a serem apresentadas nas próximas seções, ela consiste em gerar novos rótulos para os termos dos documentos. Como mencionado anteriormente, a base de dados é composta pelos termos que cada documento possui. Dessa forma, cada linha representa um documento, contendo o seu identificador e a classe a que pertence, e nela estão os termos pertencentes àquele documento, como mostra a Figura 5.13. Portanto, nossa primeira transformação será agregar a informação do ano do documento a cada um de seus termos.

Para tal, extraímos de cada documento a informação do ano a que ele pertence, armazenando essa informação em uma tabela de indexação documento-ano. É importante observar que essa informação não era utilizada pelo algoritmo no processo de classificação anteriormente. Assim, concatenamos a cada termo desse documento o ano. Por exemplo, suponha que os termos t_1, t_2, t_3 e t_4 pertençam a um documento do ano de 1980. Após modificarmos a base com essa transformação, que chamaremos de

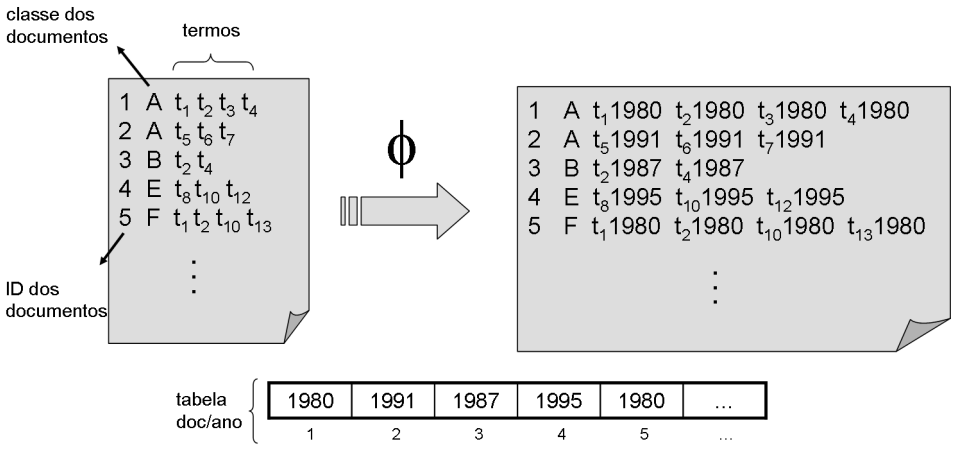

Figura 5.13. Transformação ϕ da Base de Dados

ϕ, esses termos passarão a ser $t_1$1980, $t_2$1980, $t_3$1980 e $t_4$1980, como ilustrado na Figura 5.13. Estamos, portanto, diferenciando o termo t_2 que aparece em um documento de 1980, por exemplo, do termo t_2 que aparece em um documento de 1987.

Isso pode ser interessante, uma vez que o mesmo termo pode ter significados e/ou poder discriminativo diferentes em períodos de tempos diferentes. Recordando o exemplo do termo "Tucson", discutido anteriormente, temos que, até 2005, ele se referia a uma cidade do estado do Arizona, nos Estados Unidos. A partir de 2005, entretanto, um modelo de carro foi criado com esse mesmo nome. Assim, se usarmos modelos de classificação gerados com documentos anteriores a 2005 para classificar um documento que contém o termo "Tucson" pertencente a 2006, por exemplo, eles tenderão a classificar esse documento como relacionado à cidade de "Tucson", o que pode não refletir a realidade. Portanto, pode ser interessante diferenciarmos o termo "Tucson" de um documento de 1990 do termo "Tucson" que aparece em um documento de 2006.

Essa estratégia explora o efeito temporal com uma granularidade bastante fina, já que, ao concatenar o ano aos termos do documento, ela faz com que documentos pertencentes ao mesmo ano possuam mais termos em comum do que documentos pertencentes a anos diferentes, que na verdade não possuirão nenhum termo em comum. Dessa forma, ao se realizar o processo de classificação, esses documentos que estão próximos temporalmente (pertencentes ao mesmo ano) ao documento que está sendo testado serão mais utilizados para derivar o modelo classificação.

A base de dados original da ACM sujeita à transformação continha 1.378.634 termos, sendo 56.449 termos diferentes entre si, ou seja, uma média de aproximadamente 27,5 ocorrências para cada termo. Após a transformação ela continuou contendo 1.378.634 termos, entretanto com 164.833 termos diferentes entre si, obtendo uma média de 8,4 ocorrências para cada termo. Aumentamos, portanto, a dimensionalidade do

espaço no qual o SVM vai atuar em 108.384 dimensões, ou seja, cerca de três vezes mais dimensões do que o espaço original. Esse aumento se deve ao fato de que multiplicamos cada termo diferente pelo número de anos em que ele ocorre.

Para descobrir o impacto dessa transformação na acurácia do classificador, realizamos a tarefa de classificação utilizando a nova base de dados. O processo de *3-fold cross-validation* foi aplicado e calculamos a média das acurácias. Obtivemos um valor igual a $66,29\%$, ou seja, uma acurácia para os documentos da coleção da ACM um pouco menor do que aquela encontrada sem alteração da base de dados (igual a $68,44\%$, como citado na seção anterior).

Além disso, para analisar melhor o impacto na classificação dos documentos em cada uma das classes, usamos as métricas de precisão, revocação e F_1. A Figura 5.14, a Figura 5.15 e a Figura 5.16 mostram os resultados dessas métricas, respectivamente. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada).

Através dos gráficos, pode-se perceber que, apesar de a precisão do processo de classificação ter melhorado em três classes (como a classe *MathC*, por exemplo), ela piorou para a maioria delas. Ou seja, houve um número menor de documentos corretamente associados a essas classes, em relação a todos os documentos que foram classificados

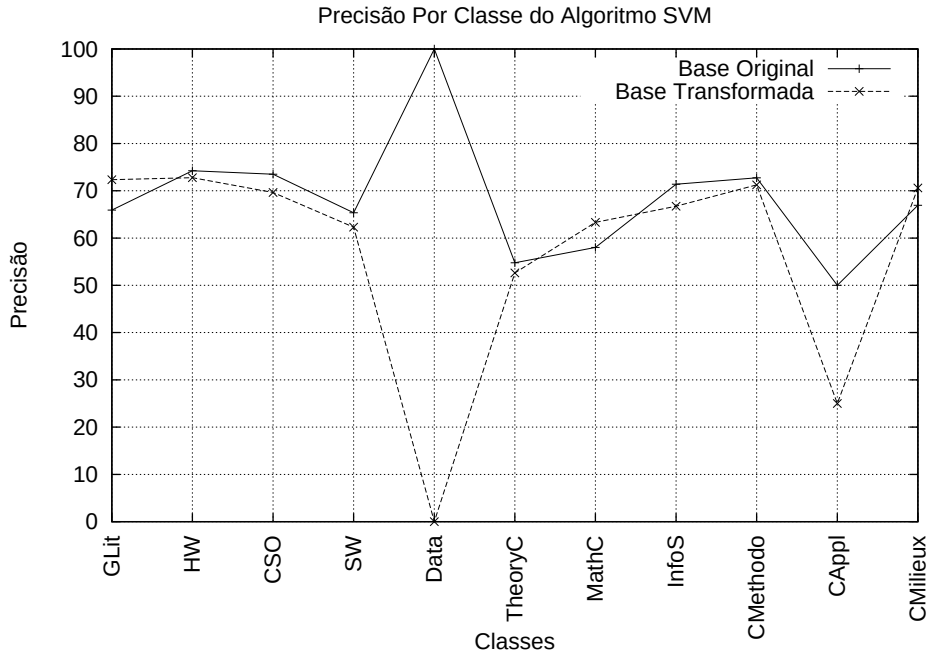

Figura 5.14. Análise da Precisão por Classe

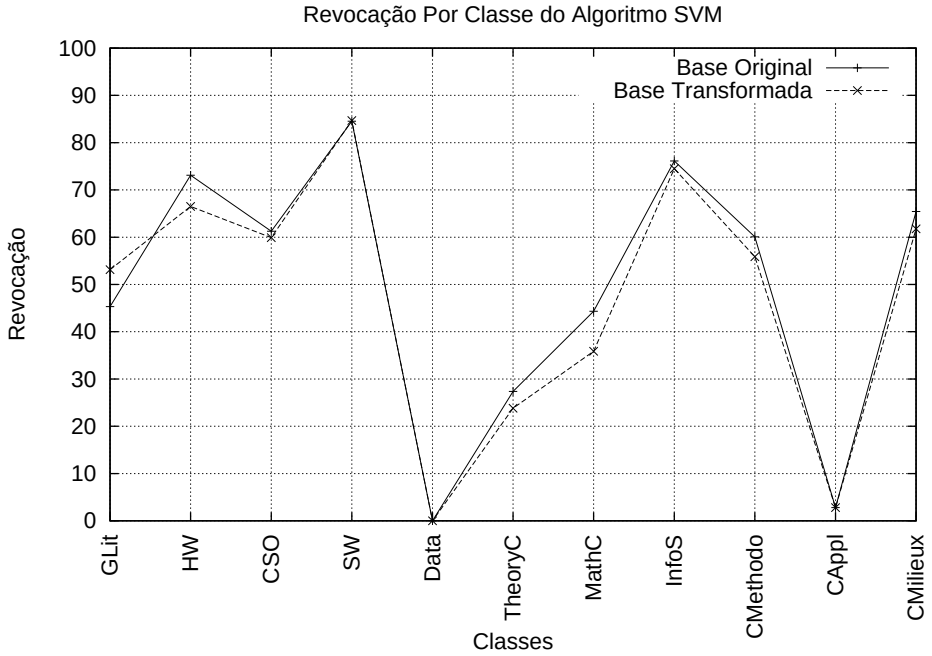

Figura 5.15. Análise da Revocação por Classe

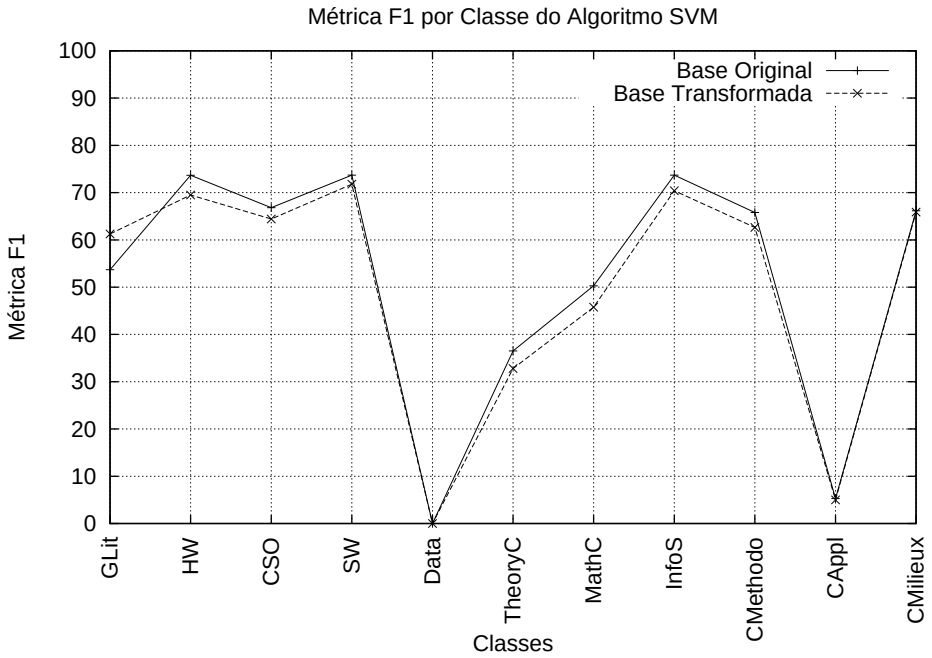

Figura 5.16. Análise da F1 por Classe

nelas. Analisando a métrica revocação, por sua vez, temos que ela foi menor ou igual à revocação obtida sem a transformação da base de dados, para praticamente todas as classes. Isso significa que, entre os documentos que deveriam ter sido associados a uma dada classe, tivemos um número menor de documentos realmente associados

a ela. Por fim, a métrica F_1, que pondera essas duas medidas, apresentou-se menor que a F_1 obtida para a base não transformada para todas as classes, exceto a *GLit.* É importante observar também que essas métricas tiveram um valor baixo para a classe *Data* (exceto o valor da precisão encontrado com base de dados não transformada) e para a classe *CAppl.* Isso se deve ao fato de que, para a base de dados transformada, apenas um documento foi associado à classe *Data* e esse documento foi associado a ela erroneamente. Na base de dados original, entretanto, nenhum documento foi assinalado à *Data*, o que ocasionou que a métrica precisão assumisse o valor de 100%. Já para a classe *CAppl*, somente quatro documentos foram associados a ela, sendo que apenas um foi classificado corretamente nela.

Os valores obtidos da acurácia e dessas outras métricas podem ser explicados devido ao fato de que, apesar de a estratégia explorar informações temporais, esse processo é feito de uma forma muito rígida, com uma granularidade muito fina. Isso faz com que mesmo os termos com o mesmo significado, que pertencem a documentos de anos próximos (mas não idênticos), sejam considerados diferentes. Assim, ao introduzir a informação temporal dessa maneira, há também uma perda de informação em relação às características de uma dada classe em anos próximos uns aos outros, já que seus termos são considerados totalmente diferentes. Isso pode gerar uma perda de informação maior do que o ganho obtido devido à informação temporal agregada. Portanto, temos que encontrar uma forma de realizar essa agregação da temporalidade dos documentos, sem perder outras informações importantes já inerentes à base de dados.

5.4 Agregação do Ano a Termos Predominantes

A segunda estratégia de transformação dos dados, para que eles agreguem informações que ajudem no processo de classificação, consiste também em atribuir novos rótulos aos termos dos documentos. Entretanto, esse processo é feito agora utilizando o conceito de termos predominantes, que será definido mais adiante nesta seção.

Assim, ao invés de estarmos distinguindo os termos por ano (em que, após a transformação, um termo t_1 em 1983 se tornará diferente desse mesmo termo t_1 em 1991), estamos fazendo essa distinção entre os anos em que o termo é predominante e os anos em que esse mesmo termo não é. Dessa forma, o termo t_1 receberá um novo rótulo único para todos os anos em que ele for predominante e permanecerá com o mesmo rótulo nos anos em que não for. É importante notar que, já que o novo rótulo associado ao termo t_1 nos anos predominantes é único, não temos mais uma separação dos termos para cada ano, como acontecia na estratégia anterior, e sim por conjuntos (grupos) de anos.

Desenvolvemos, então, essa abordagem da seguinte forma. O primeiro passo consiste em estabelecer todos os anos em que cada termo é predominante. Para tal, definimos, a predominância P de um termo t em um ano A através da fórmula a seguir:

$$P_t = \frac{T_f(t, A)}{\sum_{i=1}^{k} T_f(t, A_i)}$$

em que $T_f(t, A)$ é o número de vezes que o termo t aparece no ano A e k é o número máximo de anos existentes na base de dados. Essa é uma medida de relevância (importância) de um termo para um dado ano, o que implicará diretamente no fato de o mesmo ser ou não um discriminante da classe. Note ainda que o valor da predominância de um termo será um número entre 0 e 1. Temos que contabilizar então, para cada ano, o número de vezes que o termo aparece nos documentos que pertencem àquele ano, independentemente de qual classe o documento pertence (corresponde ao numerador da fórmula), assim como o número de vezes que o mesmo termo aparece em documentos ao longo de todos os anos (corresponde ao denominador da fórmula). Estabelecemos também um limite de predominância mínimo γ, passado como parâmetro, que fará a distinção entre o termo ser considerado predominante ou não em um determinado ano. Um termo t é considerado predominante se $P_t > \gamma$, e não predominante, caso contrário. Após uma primeira leitura da base de dados, é possível então estabelecer todos os anos nos quais um dado termo t é predominante.

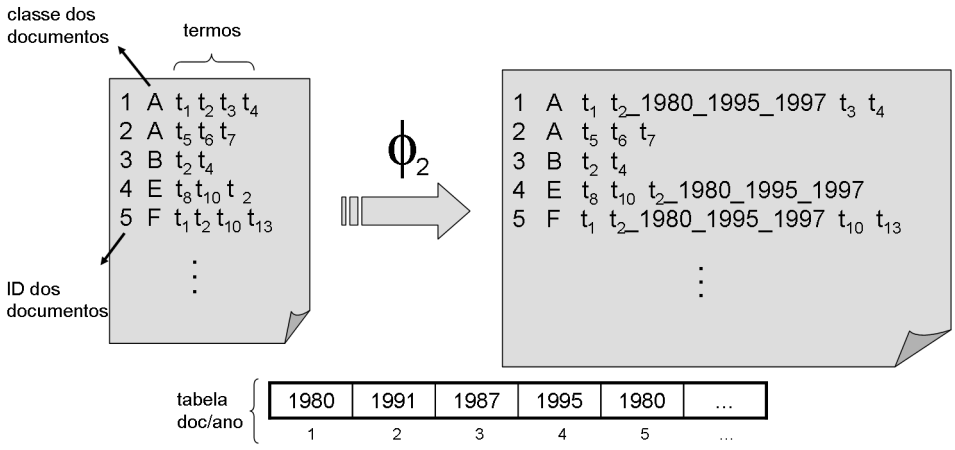

Figura 5.17. Transformação ϕ_2 da Base de Dados

O segundo passo consiste em fazer a substituição dos termos predominantes por um novo rótulo. Essa substituição é feita da seguinte forma: se o termo t_2, por exemplo, for predominante nos anos 1980, 1995 e 1997, todo documento que possui o termo t_2 e pertença a um desses anos, terá seu termo t_2 substituído por um novo rótulo da forma t_2_1980_1995_1997, como mostra a Figura 5.17. Ou seja, a partir dessa

transformação, denominada ϕ_2, estamos unificando em um mesmo rótulo o termo t_2 dos anos em que ele é predominante, e separando o termo t_2 dos anos em que ele não é predominante (por exemplo, o termo t_2 do documento 3 que pertence ao ano de 1987), uma vez que não receberá um novo rótulo e se tornará diferente.

Ao utilizar essa estratégia, portanto, temos o intuito de melhorar a qualidade dos dados (eliminando-se ruídos), uma vez que não é interessante utilizar padrões, para gerar modelos de classificação, a partir de informações de um dado termo que não é considerado predominante naquele período de tempo. Assim, essa é uma forma de considerar a evolução temporal do termo, distinguindo os períodos de tempo em que ele se apresenta discriminante, dos períodos de tempo em que ele não é. Além disso, não estamos mais utilizando uma granularidade tão fina ao se considerar o aspecto temporal, em que o termo era considerado diferente em cada ano, e isso pode contribuir para que essa estratégia seja mais bem sucedida.

Para realizar os experimentos e avaliar o impacto da transformação ϕ_2 no processo de classificação, variamos o parâmetro γ entre $0,1$ e $0,9$ (de $0,1$ em $0,1$), e geramos uma nova base de dados para cada uma dessas configurações. Antes de analisarmos o impacto na acurácia do processo de classificação, utilizando cada uma dessas bases, apresentaremos algumas características das mesmas. Lembramos aqui que a base original, isto é, antes do processo de transformação, possui 1.378.634 termos, sendo 56.449 termos diferentes entre si.

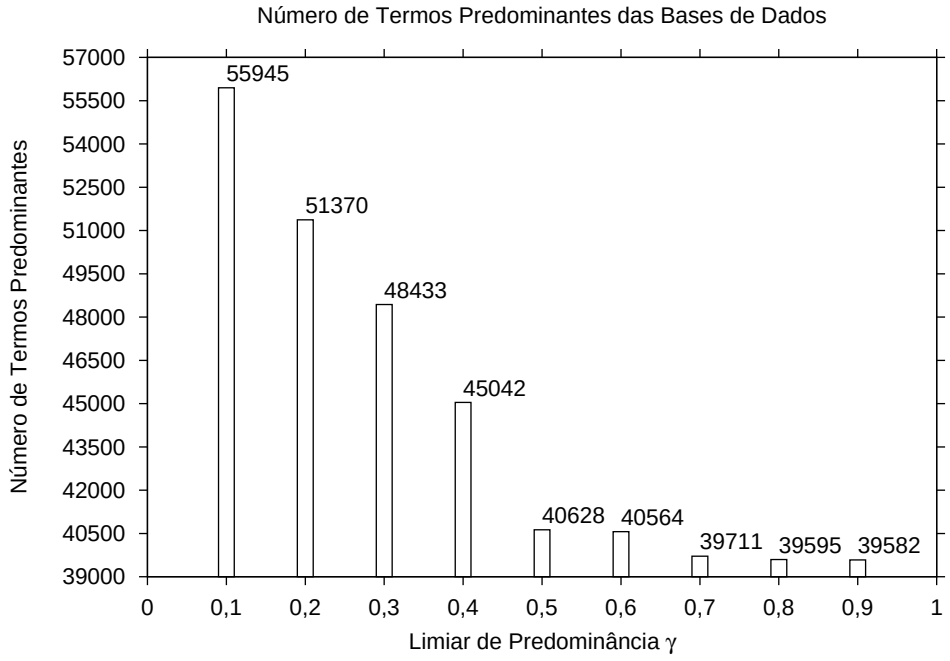

Figura 5.18. Número de Termos Predominantes

A Figura 5.18 mostra o número de termos que foram considerados predominantes em pelo menos um ano, quando utilizamos diferentes limiares mínimos de predominância para gerar as bases. Dessa forma, o eixo x representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número de termos predominantes, em pelo menos um ano, encontrados quando cada um dos limiares mínimos é utilizado.

A partir do gráfico, podemos perceber que, à medida que aumentamos o limiar de predominância mínimo γ, menor é o número de termos considerados predominantes em pelo menos um ano, o que era esperado pela definição. É interessante notar que, mesmo para o limiar de predominância mínimo mais alto considerado ($\gamma = 0,9$), cerca de 70% dos termos foram considerados predominantes em pelo menos um ano, enquanto que para $\gamma = 0,1$, esse valor chega a 99%.

Para analisar melhor as características das bases geradas, contabilizamos também, para cada termo que foi considerado predominante, o número de anos em que ele é predominante. Assim, para cada base de dados gerada, calculamos o número médio da quantidade de anos em que esses termos foram predominantes. O resultado dessa análise se encontra na Figura 5.19, em que o eixo x representa novamente o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número médio de anos em que os termos foram considerados predominantes.

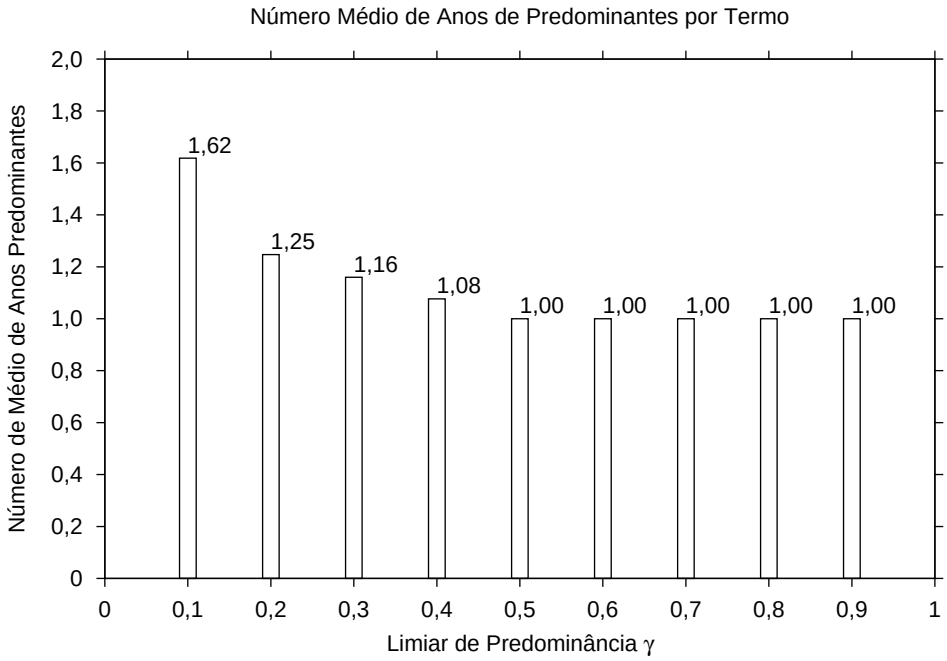

Figura 5.19. Número Médio de Anos Predominantes

Pode-se perceber que, para as bases de dados geradas com um limiar de predomi-

nância mínimo maior que $0,5$, caso um termo seja considerado predominante, o número máximo de anos em que ele pode ser predominante é 1. Isso se deve ao fato de que não é possível que um termo possua uma freqüência de ocorrência maior do que 50% em dois anos diferentes. Além disso, temos que, mesmo para baixos limiares mínimos de predominância, a maior parte dos termos predominantes o são em um número pequeno de anos, sendo que essa média chega ao máximo a $1,6$, quando $\gamma = 0,1$.

Após analisarmos algumas características das bases de dados geradas através do processo de transformação ϕ_2, realizamos a tarefa de classificação, utilizando essas bases, com o intuito de verificar como essas transformações impactam na eficácia do algoritmo SVM. Para tal, para cada uma das bases de dados, aplicamos o processo de *3-fold cross-validation* e calculamos a média das acurácias obtidas. O resultado desse experimento está apresentado na Figura 5.20, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas.

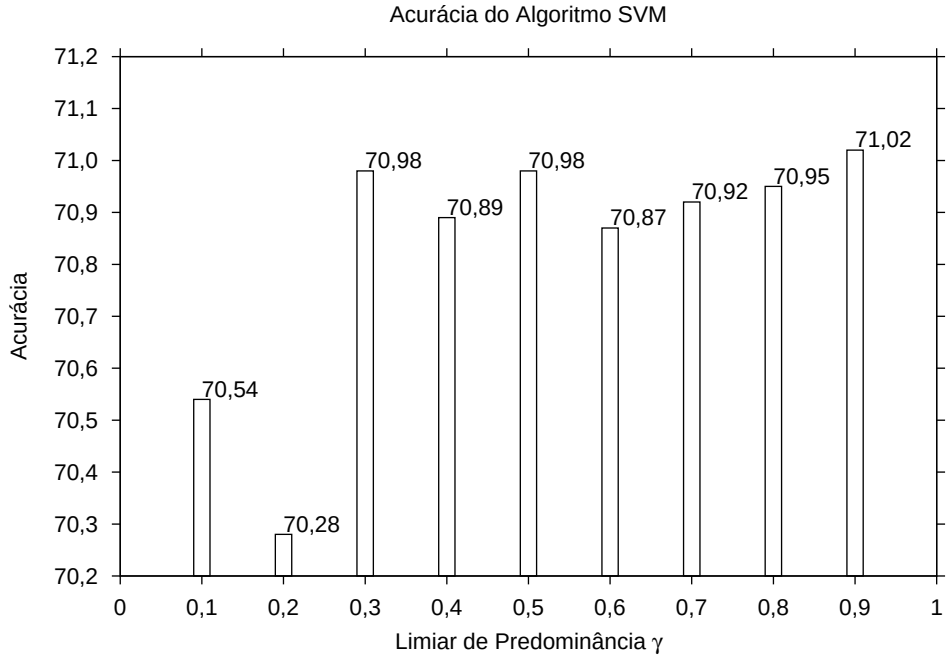

Figura 5.20. Análise da Acurácia - Termos Predominantes

Através do gráfico, pode-se perceber que, apesar de a variação do limiar de predominância mínimo γ não apresentar grandes impactos na acurácia do algoritmo de classificação, existe uma pequena melhora a partir da predominância mínima igual a $0,3$. Isso é intuitivo uma vez que valores de predominância muito baixos não são su-

ficientes para distinguir os termos que são realmente discriminantes em um dado ano e podem permitir a inserção de ruídos. De uma forma geral, entretanto, obtemos uma acurácia em torno de 71%, cerca de $2,5\%$ a mais do que aquela obtida sem alteração da base de dados.

Por fim, verificamos também o impacto dessa transformação em cada classe. Para ilustrar esse impacto, vamos analisar a base de dados que proporcionou o maior ganho de acurácia (ou seja, a base gerada com $\gamma = 0,9$). A Figura 5.21, a Figura 5.22 e a Figura 5.23 mostram, respectivamente, as métricas de precisão, revocação e F_1 obtidas para essa base de dados transformada, em comparação a essas mesmas métricas para a base de dados original. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se um limiar de predominância mínimo igual a $0,9$).

Através dos gráficos, pode-se perceber que a precisão aumentou em todas as classes, exceto nas classes *GLit* e *CAppl*. É interessante observar que essas duas classes estão exatamente entre as que mais sofrem com a evolução temporal, tanto na evolução de seus termos (como mostrado na Figura 4.13 - Seção 4.2.3), quanto na sua distribuição ao longo dos anos (como apresentado na Figura 4.9 - Seção 4.2.2). Portanto, essas classes exigem estratégias mais elaboradas para que realmente sejamos

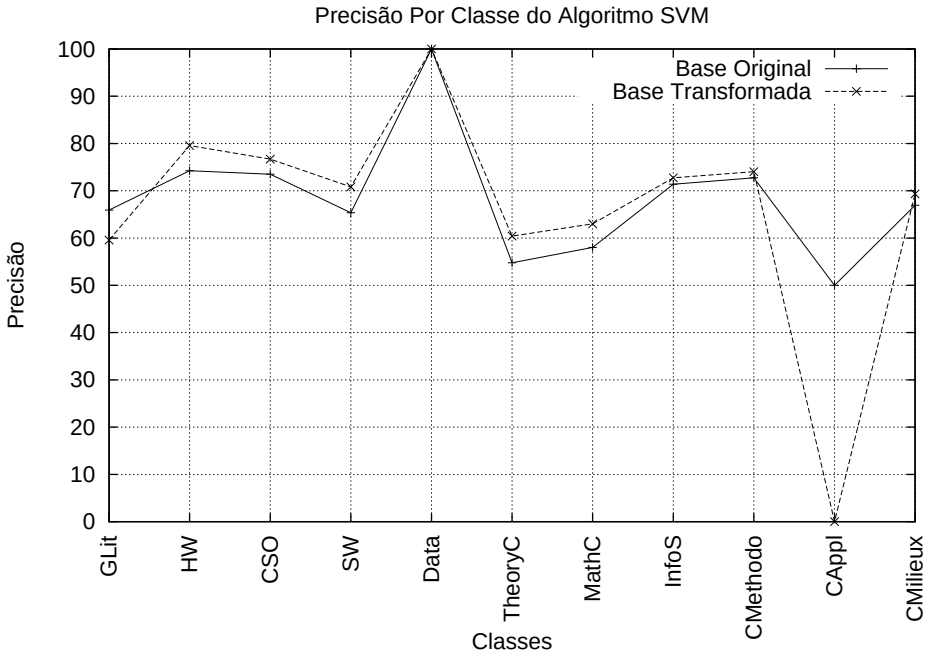

Figura 5.21. Análise da Precisão por Classe

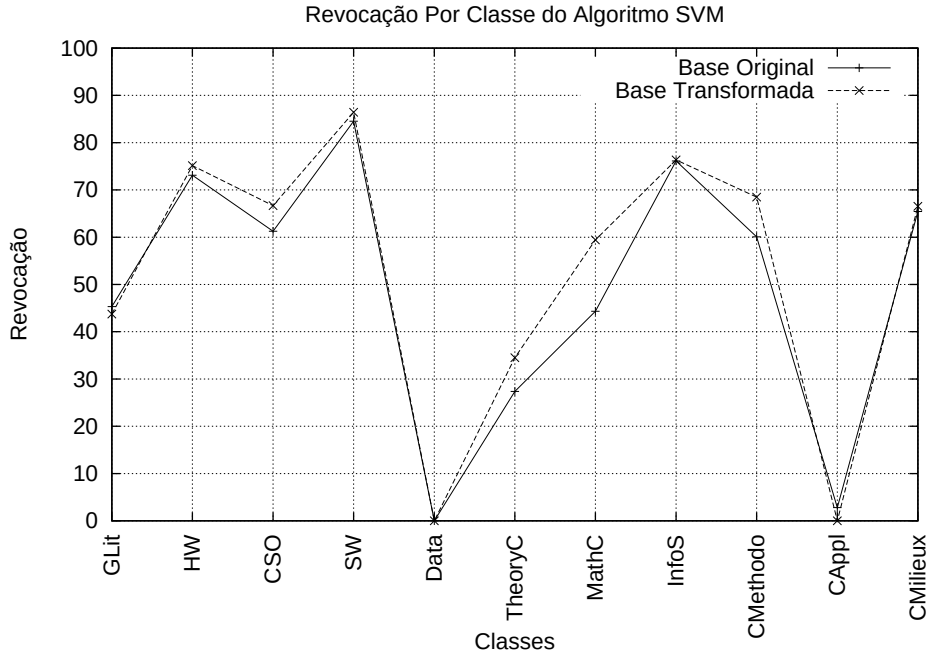

Figura 5.22. Análise da Revocação por Classe

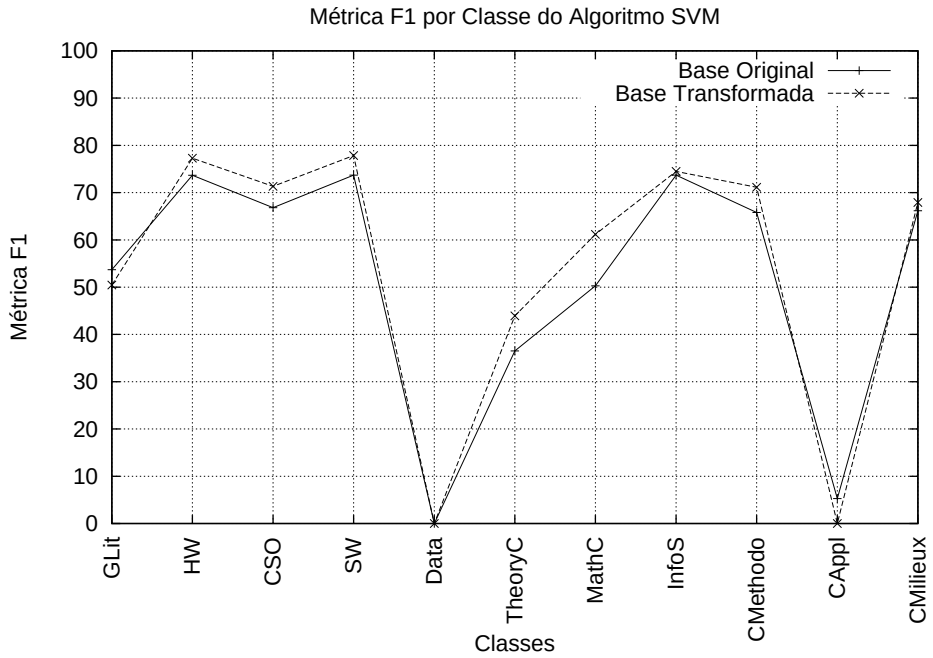

Figura 5.23. Análise da F1 por Classe

capazes de tratar a evolução temporal que elas apresentam. A métrica de revocação, por sua vez, apresentou-se melhor ou praticamente igual para todas as classes quando comparada à base original, não havendo nenhuma que tenha se destacado positiva ou negativamente após a transformação. Já a métrica F_1, que pondera essas duas outras

métricas, também se apresenta melhor em todas as classes, exceto nas classes *GLit* e *CAppl*. É importante ainda notar que, novamente, os baixos valores dessas métricas para as classes *Data* e *CAppl* são devidos à pequena quantidade de documentos que foram associados a elas (nenhum documento foi associado à classe *Data* e apenas quatro documentos foram classificados na classe *CAppl*).

5.5 Agregação do Ano e da Classe a Termos Predominantes

A estratégia proposta nesta seção adiciona uma outra dimensão ao conceito de termos predominantes abordado na seção anterior. Ao invés de considerarmos se um termo é predominante em um determinado ano, em relação à sua ocorrência na base de dados ao longo de todos os anos, propomos aqui analisar a predominância do termo em uma dada classe em um dado ano, em relação à sua ocorrência em todas as outras classes naquele mesmo ano.

Estabelecemos, portanto, um novo conceito de predominância, agora por classe, que foi definido da forma a seguir. A predominância P de um termo t em uma classe C, em um dado ano, é definida por:

$$P_{t_C} = \frac{T_f(t, C)}{\sum_{i=1}^{m} T_f(t, C_i)}$$

em que $T_f(t, C)$ é o número de vezes que o termo t aparece na classe C no ano analisado, e m é o número máximo de classes existentes na base de dados. É interessante observar que, enquanto na estratégia anterior foi estabelecida uma predominância global (calculada relativamente à ocorrência do termo na base de dados toda), a predominância definida aqui é uma predominância local, que é calculada relativamente à ocorrência do termo em um determinado ano, para se determinar em quais classes ele é predominante naquele ano. Ou seja, enquanto na estratégia anterior o conceito determinado de predominância relacionava o quanto um termo é discriminante para um determinado ano em comparação a todos os anos da base de dados, essa estratégia relaciona, dado um determinado ano, o quanto um termo é discriminante para uma classe em relação a todas as outras classes naquele ano. Novamente, foi estabelecido um limite mínimo de predominância γ que fará a distinção entre o termo ser considerado predominante ou não, em uma dada classe em um determinado ano.

Assim, desenvolvemos essa abordagem da seguinte forma. O primeiro passo consiste em achar, para cada ano, as classes nas quais o termo é predominante naquele ano. Para tal, foi contabilizado, para cada ano e para cada uma das classes, o número de

vezes que o termo aparece naquela classe naquele ano (que corresponde ao numerador da fórmula), assim como foi contabilizado o número de vezes que o termo aparece naquele ano em todas as classes (que corresponde ao denominador da fórmula). Após uma primeira leitura da base de dados, torna-se possível então identificar se o termo t é predominante ou não em cada uma das classes em cada um dos anos, de acordo com o limite de predominância mínimo γ, passado como parâmetro. Dessa forma, um termo t é considerado predominante em um dado ano se $P_{t_C} > \gamma$, e não predominante, caso contrário. Modificando o ponto de vista da análise, percebe-se ainda que podemos determinar todos os anos nos quais um dado termo t é predominante em uma dada classe C.

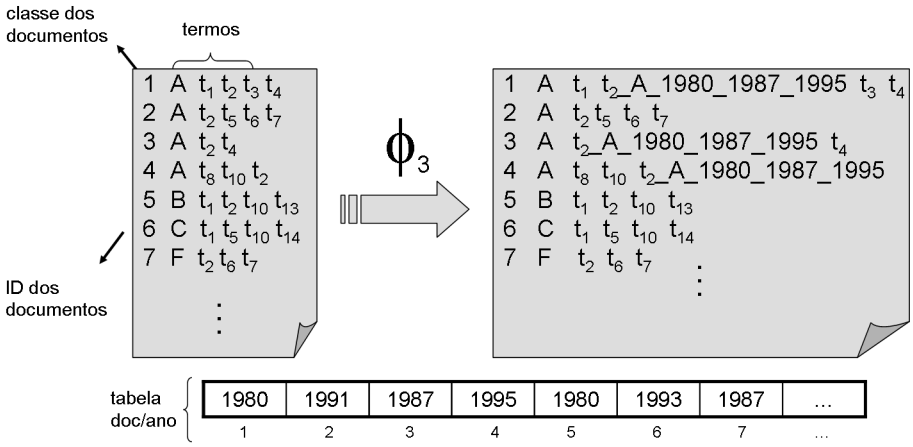

Figura 5.24. Transformação ϕ_3 da Base de Dados

O segundo passo consiste em fazer a substituição dos termos predominantes por um novo rótulo. Isso é feito da seguinte forma: se o termo t_2 for predominante na classe A nos anos 1980, 1987 e 1995, por exemplo, todo documento que possui o termo t_2 e pertença à classe A em um desses anos, terá seu termo t_2 substituído por um novo rótulo da forma $t_2\_A\_1980\_1987\_1995$, como mostra a Figura 5.24. Ou seja, a partir dessa transformação, estamos unificando em um mesmo rótulo o termo t em uma dada classe C nos anos em que ele é predominante, e separando o termo t dos anos em que ele não é predominante naquela classe, uma vez que não receberá um novo rótulo e se tornará diferente.

Ao utilizar essa estratégia, portanto, estamos de certa forma eliminando ruídos da base de dados. Ela é muito similar à estratégia anterior, no sentido que aborda a questão temporal fazendo a distinção entre os períodos de tempo em que os termos possuem um maior poder discriminativo e os períodos em que esses mesmos termos não apresentam esse poder. Entretanto, nessa estratégia, isso é feito considerando também as classes a que esses termos estão associados. Essa distinção é importante, uma vez

que não é interessante utilizar padrões para gerar modelos de classificação a partir de informações de um dado termo que não seja considerado predominante em uma dada classe naquele período de tempo.

A seguir, apresentaremos os experimentos realizados utilizando os conceitos aqui definidos. Primeiramente, iremos mostrar um experimento conduzido que objetiva apenas avaliar a capacidade de aprendizado do algoritmo SVM, ao transformarmos a base de dados sobre a qual ela atua. Após feita essa análise, descreveremos duas abordagens propostas para a estratégia apresentada nesta seção.

5.5.1 Avaliação da Capacidade do Classificador SVM

Antes de apresentarmos as abordagens desenvolvidas para a estratégia descrita anteriormente, iremos descrever um teste que realizamos, cujo intuito é derivar, experimentalmente, o limite teórico do classificador SVM. Ou seja, iremos criar um cenário fictício em que proporcionaremos ao algoritmo, através da transformação da base de dados, toda a informação necessária para classificar os documentos que ela contém e analisaremos como o classificador se comporta com essas informações, avaliando sua capacidade de aprendizado para essa base.

Assim, como a estratégia proposta nesta seção considera os termos predominantes separadamente para cada classe, e não somente por anos, surge a questão de como analisar os termos pertencentes a documentos do conjunto de teste. Isso porque, na prática, não temos a informação de qual classe esses documentos pertencem, sendo essa informação exatamente o que se deseja descobrir com o processo de classificação. Com o intuito de realizar essa avaliação da capacidade de aprendizado do classificador, entretanto, optamos por considerar a informação das classes às quais os documentos do conjunto de teste pertencem. Dessa forma, no processo de contabilização de quantas vezes o termo aparece em uma determinada classe em um determinado ano, os termos dos documentos de teste também serão contados, uma vez que sabemos à qual classe o documento deve ser associado. Porém, isso seria inviável na prática, uma vez que, obviamente, não possuímos essa informação em cenários reais de classificação.

Para realizar os experimentos e verificar o impacto da transformação no processo de classificação, variamos o parâmetro γ entre $0,0$ e $0,9$ (de $0,1$ em $0,1$). Nesta análise, achamos interessante também considerar um limite mínimo de predominância 0, para avaliar a acurácia do algoritmo quando todos os termos são considerados predominantes e recebem um novo rótulo. Como feito anteriormente, geramos uma nova base para cada uma dessas configurações. Entretanto, antes de analisarmos o impacto na acurácia do processo de classificação utilizando cada uma dessas bases, apresentaremos algumas características das mesmas.

Primeiramente, para cada uma das bases, analisamos o número de termos que foram considerados predominantes em pelo menos uma classe, considerando todos os anos. A Figura 5.25 mostra esse resultado, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número de termos predominantes, em pelo menos uma classe (considerando todos os anos), encontrados quando cada um dos limiares mínimos é utilizado.

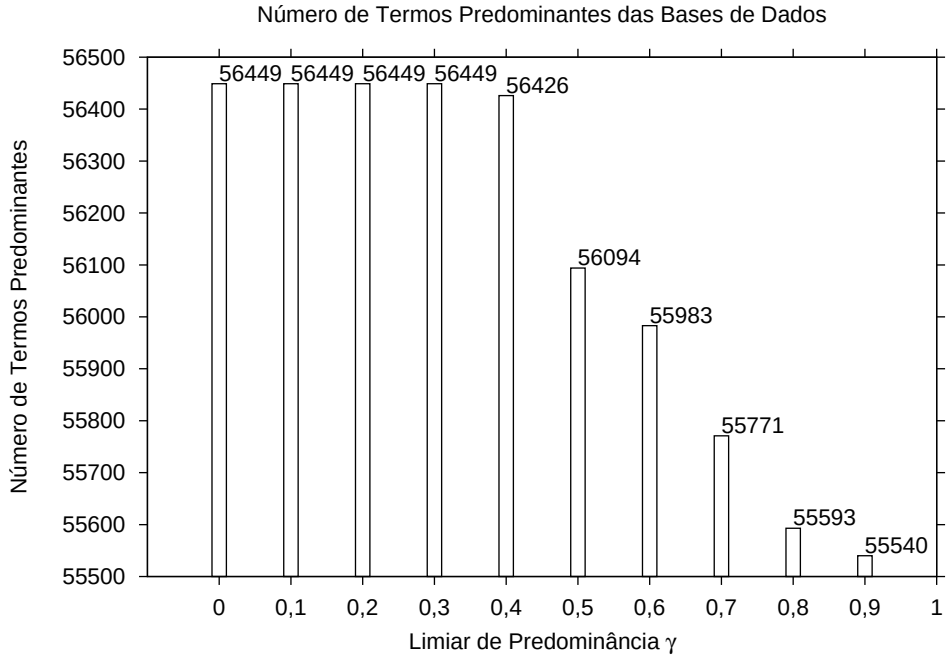

Figura 5.25. Número de Termos Predominantes

A partir do gráfico, podemos perceber que, para um limiar mínimo de predominância menor do que $0,4$, todos os termos são considerados predominantes. Ou seja, ele foi predominante em pelo menos uma classe em um dado ano. Além disso, mesmo para um limiar de predominância igual a $0,9$, 98% dos termos foram considerados predominantes. Isso significa que praticamente todos os termos geram, no mínimo, um novo outro rótulo, sendo duplicada a dimensionalidade do espaço de dados.

Para continuar a análise da transformação do espaço de dados, verificamos também o número de classes em que cada termo é considerado predominante, ao longo de todos os anos. Assim, fizemos uma média desse número de classes por termo, para cada base de dados, e apresentamos os resultados na Figura 5.26. Esse número corresponde também ao número médio de novos rótulos que cada termo gera. Dessa forma, o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número médio de classes em que cada termo foi considerado predominante.

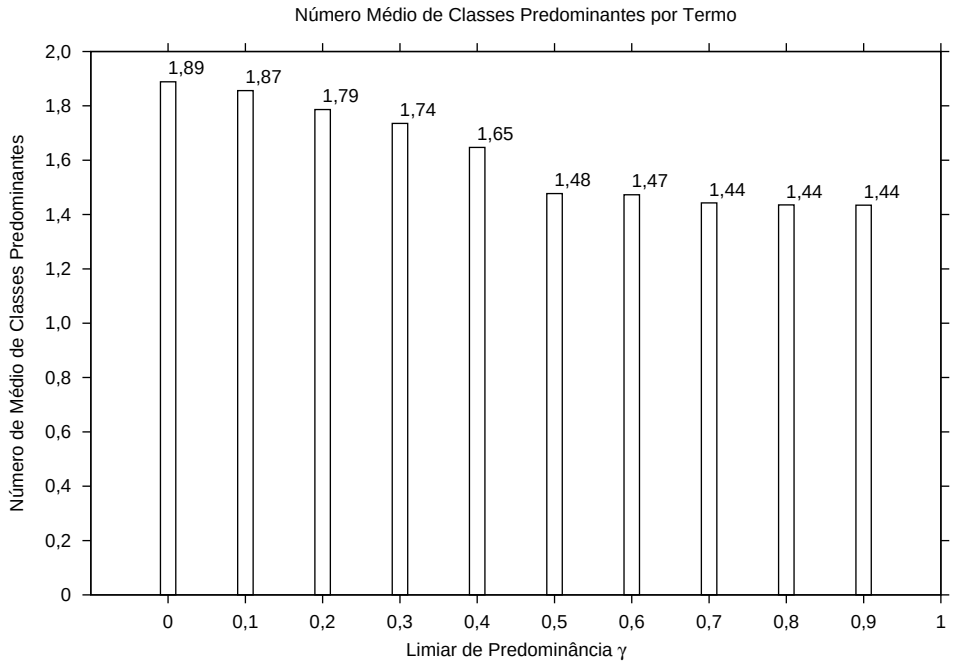

Figura 5.26. Número Médio de Classes Predominantes

Através do gráfico, pode-se perceber que, mesmo quando diminuímos o valor do limite mínimo de predominância, o maior número médio de classes em que um termo é predominante é aproximadamente $1,9$ classes. Isso significa que, em média, cada termo proporciona no máximo a geração de dois novos rótulos, triplicando a dimensionalidade do espaço de dados, já que o próprio termo também é mantido para os anos e classes em que ele não é predominante.

Após as análises feitas, realizamos agora o processo de classificação utilizando essas bases de dados geradas, com o intuito de verificar como essas transformações impactam na eficácia do algoritmo SVM. Assim, para cada uma das bases de dados, aplicamos o processo de *3-fold cross-validation* e calculamos a média das acurácias obtidas. O resultado desse experimento está apresentado na Figura 5.27, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas.

O gráfico nos mostra que, quando aproximamos o γ do valor zero, a acurácia do algoritmo sobe para aproximadamente $98,6\%$. Isso acontece uma vez que realizamos a transformação dos dados, agregando a eles informações que são as mais precisas possíveis, realmente refletindo a freqüência dos termos na base de dados. É por isso que, nesse caso, quanto mais termos são substituídos com a nova informação (ou seja, quanto

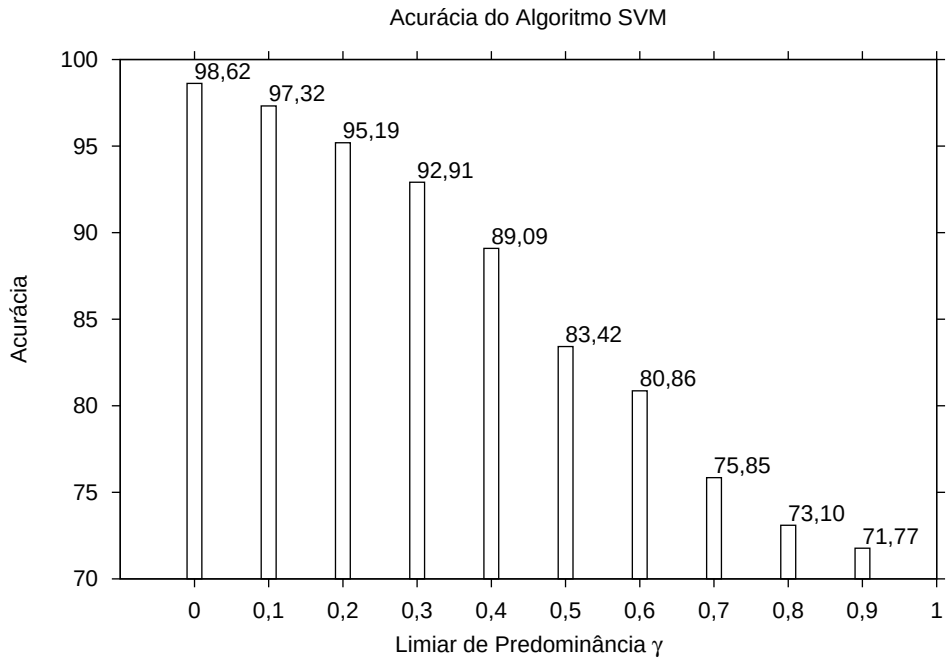

Figura 5.27. Análise da Acurácia - Avaliação do SVM

menor for o limite inferior de predominância γ), melhor é a acurácia do classificador.

Analisamos, ainda, o impacto da transformação por classes. Novamente, para ilustrar esse impacto, consideramos a base de dados transformada que gerou o maior ganho na acurácia da tarefa de classificação. A Figura 5.28, a Figura 5.29 e a Figura 5.30 mostram, respectivamente, as métricas de precisão, revocação e F_1 obtidas para essa base de dados, em comparação a essas mesmas métricas para a base de dados original. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se um limiar de predominância mínimo igual a 0).

Através dos gráficos, pode-se perceber que a precisão é de 100% em praticamente todas as classes da base transformada, enquanto a revocação obtida para essa base também foi muito alta para a maioria das classes, se apresentando um pouco mais baixa (cerca de 70%) para as classes *Data* e *CAppl*. Esses baixos valores para essas duas classes ocorrem devido ao fato de que o conjunto de teste apresenta poucos documentos pertencentes a elas (principalmente para a classe *Data*, em que esse número é igual a 16). Assim, a classificação de um documento errado dessa classe tem um impacto grande nessa métrica. Por fim, a métrica F_1, que pondera as duas outras métricas, também apresenta valores muito próximos a 100% para praticamente todas as classes.

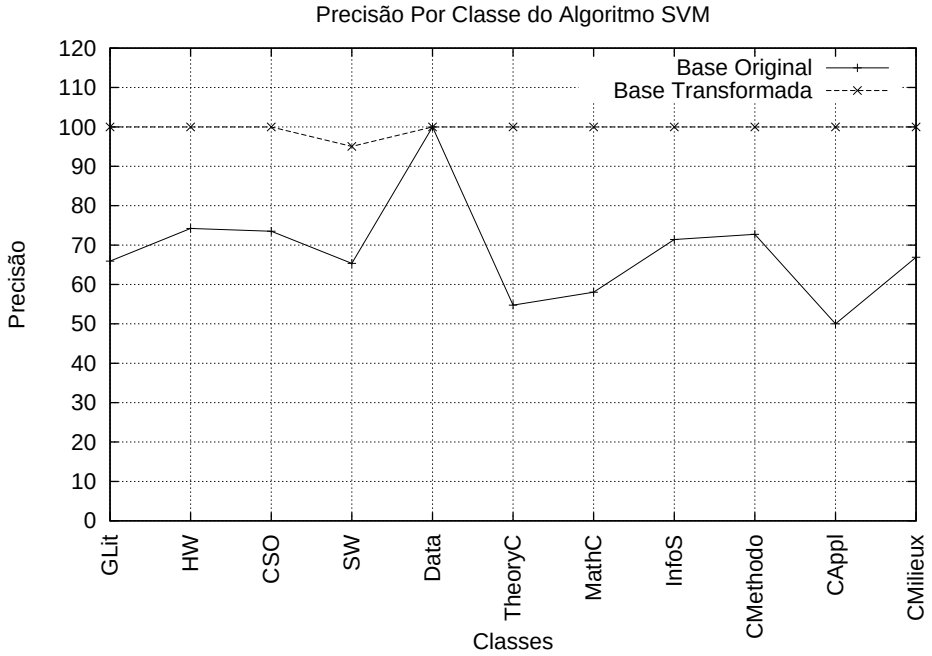

Figura 5.28. Análise da Precisão por Classe

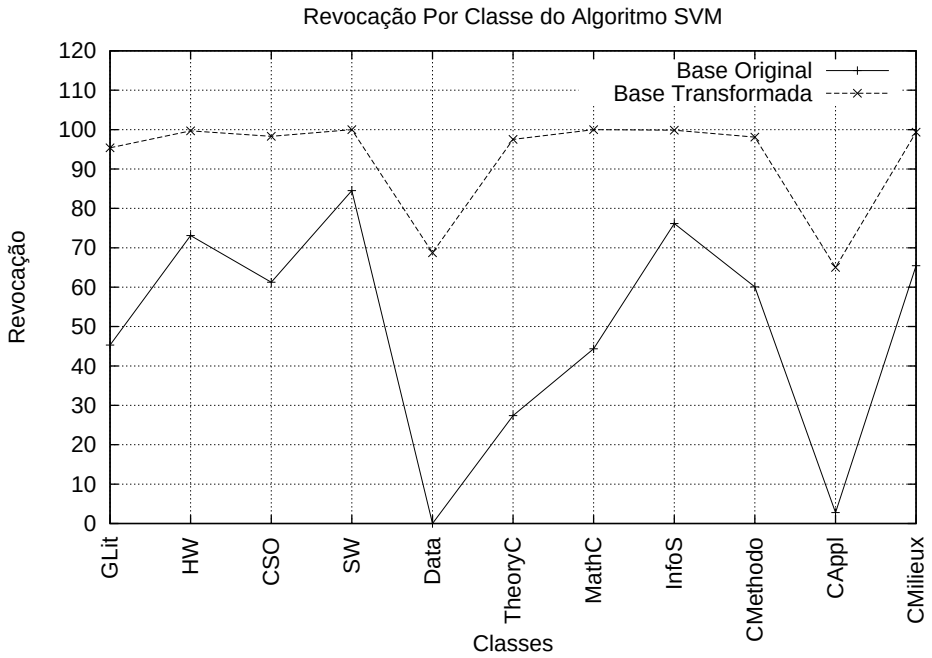

Figura 5.29. Análise da Revocação por Classe

Através desse experimento, comprovamos então a capacidade de aprendizado do algoritmo SVM ao lidar com uma base de documentos que sofreu uma transformação em seus dados. Ou seja, a abordagem de incorporar informações na base de dados realmente pode contribuir significativamente para a melhora da acurácia do processo

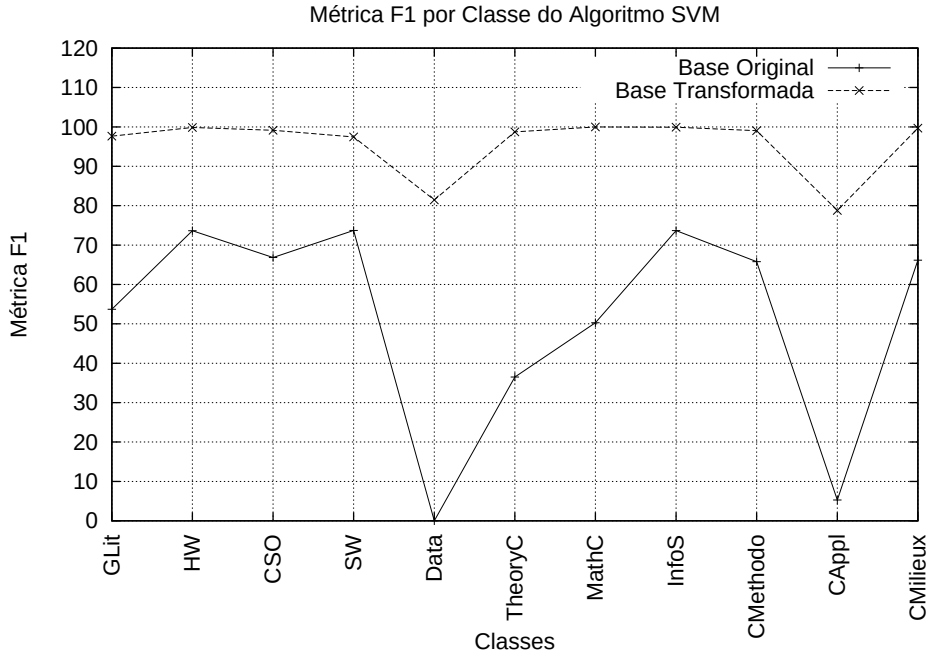

Figura 5.30. Análise da F1 por Classe

de classificação. Temos, então, que desenvolver estratégias inteligentes, capazes de incorporar informações relevantes, para que essa melhora ocorra.

5.5.2 Limiar de Predominância Fixo

No experimento anterior, utilizamos a informação das classes às quais os documentos do conjunto de teste pertencem, o que torna inviável que essa transformação da base de dados seja feita em um cenário real, em que não possuímos tal informação. Por outro lado, podemos encontrar uma forma de induzir informações relevantes para os documentos do conjunto de teste, tendo como base os dados que existem nos documentos do conjunto de treino. Uma estratégia simples encontrada para fazer isso é explicada a seguir.

Ao contabilizarmos os termos que pertencem a uma dada classe em um dado ano, temos que ignorar aqueles termos que pertencem ao documento de teste, uma vez que não possuímos a informação de sua classe. Após feita essa primeira leitura da base de dados, contabilizando os termos e ignorando aqueles pertencentes a documentos de teste, é estabelecido, de acordo com um limite mínimo de predominância γ (definido anteriormente), quais termos são predominantes em cada classe em cada ano.

A seguir, é feita a substituição dos termos predominantes por novos rótulos que incorporam a informação da classe e dos anos em que eles são predominantes, como também explicado anteriormente. É no momento de fazer essa substituição que nos

deparamos com o problema dos termos dos documentos do conjunto de teste. Assim, a medida adotada foi a seguinte: ao lermos um termo do documento de teste, temos a informação de qual ano ele pertence. Dessa forma, procuramos naquele ano, em qual classe esse termo ocorreu mais vezes. Se, a partir desse valor da freqüência do termo, esse termo é considerado predominante naquela classe então ele também receberá um novo rótulo. Se, entretanto, mesmo na classe em que ele ocorreu mais vezes naquele ano, ele não supera o limite mínimo de predominância estabelecido, isso significa que ele não é predominante em nenhuma classe naquele ano e, portanto, não deve ganhar um novo rótulo. Estamos, então, utilizando uma estratégia gulosa para tentar agregar informações também para os documentos de teste, em que o termo de um documento de teste é considerado da classe em que ele é "mais predominante".

Utilizamos, novamente, a mesma configuração para realizar os experimentos, ou seja, variamos o parâmetro γ entre $0,1$ e $0,9$ (de $0,1$ em $0,1$), e geramos uma nova base para cada uma dessas configurações. A seguir, apresentaremos as características das bases de dados geradas.

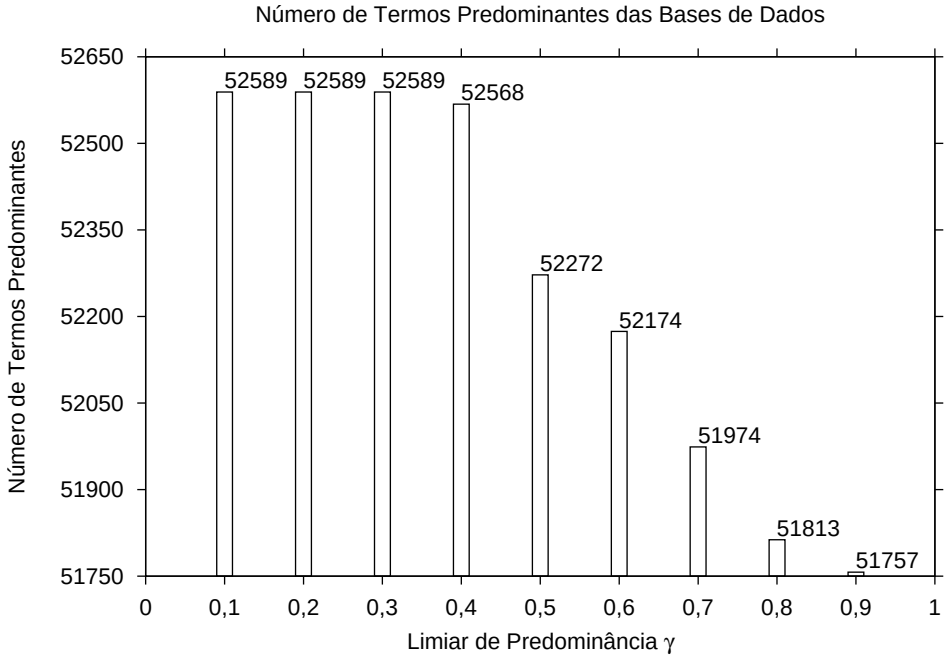

Figura 5.31. Número de Termos Predominantes

Assim como feito na subseção anterior, para cada uma das bases, analisamos o número de termos que foram considerados predominantes em pelo menos uma classe, considerando todos os anos. Apresentamos esses resultados na Figura 5.31, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número de termos predominantes, em

pelo menos uma classe (considerando todos os anos), encontrados quando cada um dos limiares mínimos é utilizado.

Através do gráfico, pode-se perceber que ele possui um comportamento muito semelhante, ao se variar o limiar mínimo de predominância, àquele constatado no experimento para avaliar a capacidade de aprendizado do SVM. Entretanto, neste experimento, foram encontrados cerca de 4.000 termos predominantes a menos do que os encontrados no da subseção anterior, para todas as configurações. Isso se deve ao fato de que, nesta estratégia, não contabilizamos os termos dos documentos de teste para determinar a freqüência dos termos na base de dados, influenciando assim na determinação de um termo ser predominante ou não.

Novamente, verificamos também o número de classes nas quais cada termo é considerado predominante ao longo de todos os anos, calculando em seguida a média desse número de classes por termo para cada base de dados. A Figura 5.32 apresenta esses resultados. Dessa forma, o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número médio de classes em que cada termo foi considerado predominante.

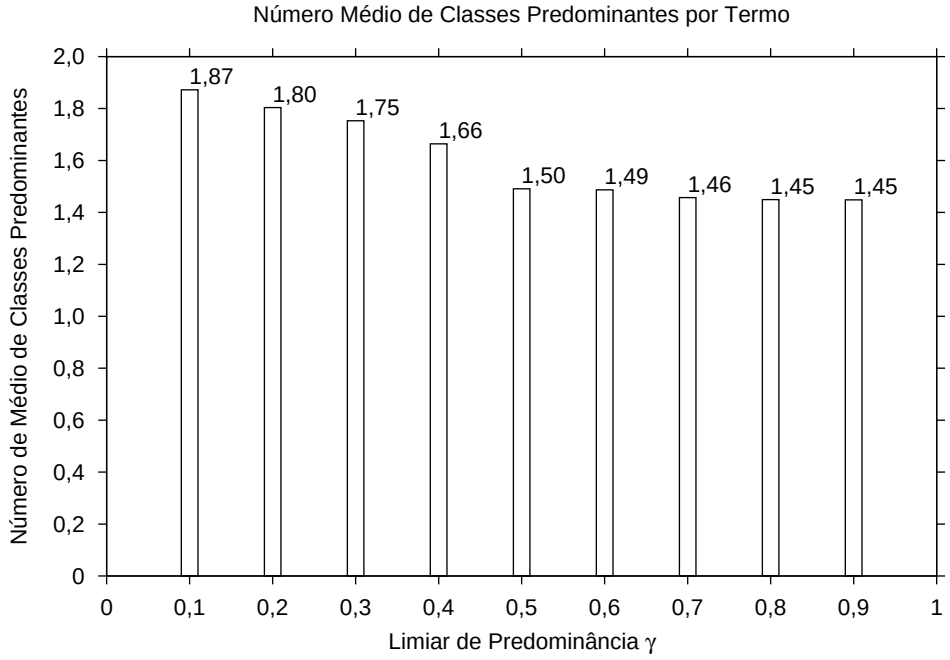

Figura 5.32. Número Médio de Classes Predominantes

Fazendo uma comparação com os resultados obtidos nesse mesmo experimento na subseção em que avaliamos o classificador SVM, temos que eles se mostram bastante semelhantes. Apesar de o gráfico anterior ter nos mostrado que menos termos são considerados predominantes com esta estratégia, para aqueles que o são, o número

médio de classes em que essa predominância ocorre para cada termo é praticamente o mesmo.

Por fim, fizemos também uma análise dos termos pertencentes aos documentos do conjunto de teste, uma vez que eles foram considerados de forma especial no tratamento de substituição por novos rótulos. Assim, podemos entender melhor o funcionamento da heurística gulosa ao tratar esses termos. Para tal, contabilizamos, entre os termos pertencentes aos documentos do conjunto de teste, aqueles que foram considerados predominantes e aqueles que não foram considerados predominantes. Isso nos proporciona uma noção de o quanto a heurística gulosa transformou os dados das bases de dados. Além disso, entre os termos de teste considerados predominantes em uma dada classe, contabilizamos também quantos foram considerados predominantes na classe certa e quantos foram considerados predominantes, mas o associando à classe errada. Através desses dados, podemos analisar o quão a heurística utilizada é eficaz. A Figura 5.33 mostra o gráfico contendo essas informações, em que o eixo x representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa o número de termos.

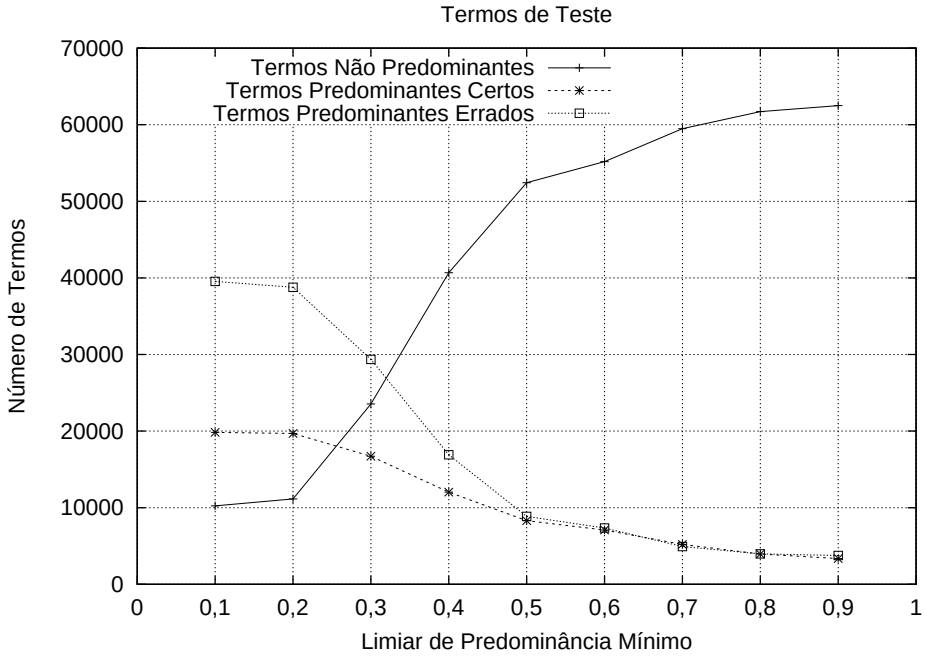

Figura 5.33. Análise do Número de Termos de Teste

Podemos notar que há cerca de 70.000 ocorrências [1] de termos nos documentos de teste (equivalente a soma dos termos das três curvas do gráfico). À medida que

[1] É importante notar que, enquanto nos outros gráficos quando nos referimos ao número de termos estamos considerando o número de termos diferentes, aqui estamos contabilizando o número total de *ocorrências* dos termos.

aumentamos o limiar mínimo de predominância, temos uma redução significativa do número de ocorrências de termos que são considerados predominantes. Para $\gamma = 0,1$ esse número chega a 60.000, enquanto que para $\gamma = 0,9$, esse número é apenas 10.000. Assim, ressaltamos que a heurística atua com uma intensidade considerável para baixos limiares mínimos de predominância, enquanto não se apresenta tão atuante quando esses limiares são mais altos. Por outro lado, quando $\gamma = 0,9$, temos que a eficácia da heurística se encontra em torno de 50% (entre os termos considerados predominantes, cerca de 4.000 termos foram associados à classe certa e 4.000 termos foram associados a classe errada). Isso é também verdade para os limiares mínimos de predominância a partir de 0,5. Porém, para limiares menores, essa eficácia chega a 33% ($\gamma = 0,1$).

Após ter sido feita essa análise das bases de dados resultantes da transformação, aplicamos o algoritmo de classificação SVM a cada uma delas e verificamos o impacto dessas transformações na eficácia da tarefa de classificação. Assim, realizamos o processo de *3-fold cross-validation* para cada uma das bases e calculamos a média das acurácias obtidas. O resultado desse experimento está apresentado na Figura 5.34, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas.

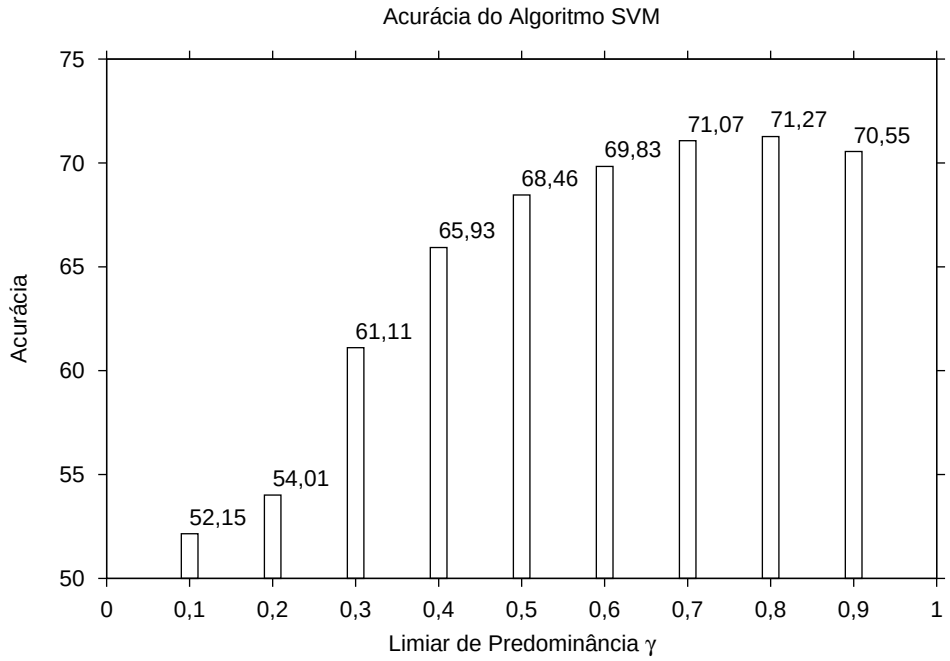

Figura 5.34. Análise da Acurácia - Limiar de Predominância Fixo

Através do gráfico, podemos perceber que à medida em que aumentamos o limiar

de predominância mínimo, a acurácia do algoritmo também melhora. Além disso, obtivemos acurácias ruins para baixos limiares de predominância. Isso pode ser justificado pelo fato de que, como mencionado anteriormente, uma predominância mínima baixa não é capaz de separar os termos mais discriminantes para uma dada classe, permitindo a introdução de ruídos. Além disso, neste caso, temos a aplicação de uma heurística gulosa para termos dos documentos de teste e essa heurística pode causar a introdução de uma informação não verdadeira (mais ruídos) na base de dados. Assim, quanto menor é o limiar mínimo de predominância, maior o número de termos que devem ser substituídos e maior é o uso da heurística. Se a heurística não for muito inteligente, ela pode gerar mais erros do que acertos nas informações introduzidas por ela, como foi verificado anteriormente, piorando a acurácia do algoritmo. Obtivemos, portanto, a melhor acurácia do algoritmo quando $\gamma = 0,8$, uma vez que um valor de γ maior restringe muito os termos que serão considerados predominantes. Dessa forma, conseguimos obter uma acurácia de 71,27%, o que representa um ganho de quase 3% em relação à acurácia obtida quando não há transformação da base de dados.

Para entendermos melhor a transformação realizada na base de dados, e como ela contribuiu para uma melhora da eficácia do classificador, vamos também analisar o impacto em cada uma das classes, utilizando a base de dados que proporcionou o maior ganho da acurácia ($\gamma = 0,8$). A Figura 5.35, a Figura 5.36 e a Figura 5.37 mostram, respectivamente, as métricas de precisão, revocação e F_1 obtidas para essa

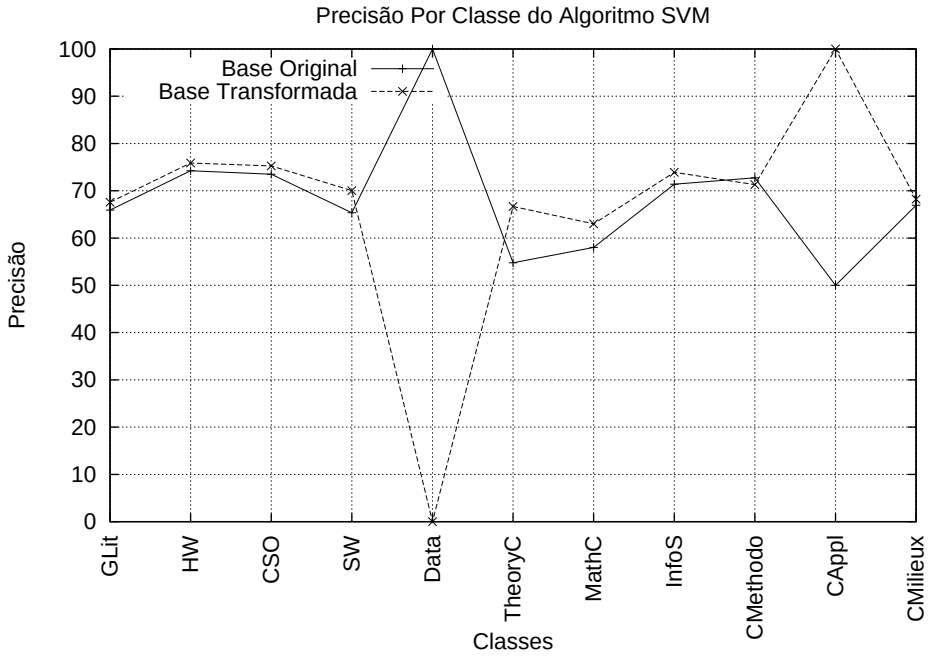

Figura 5.35. Análise da Precisão por Classe

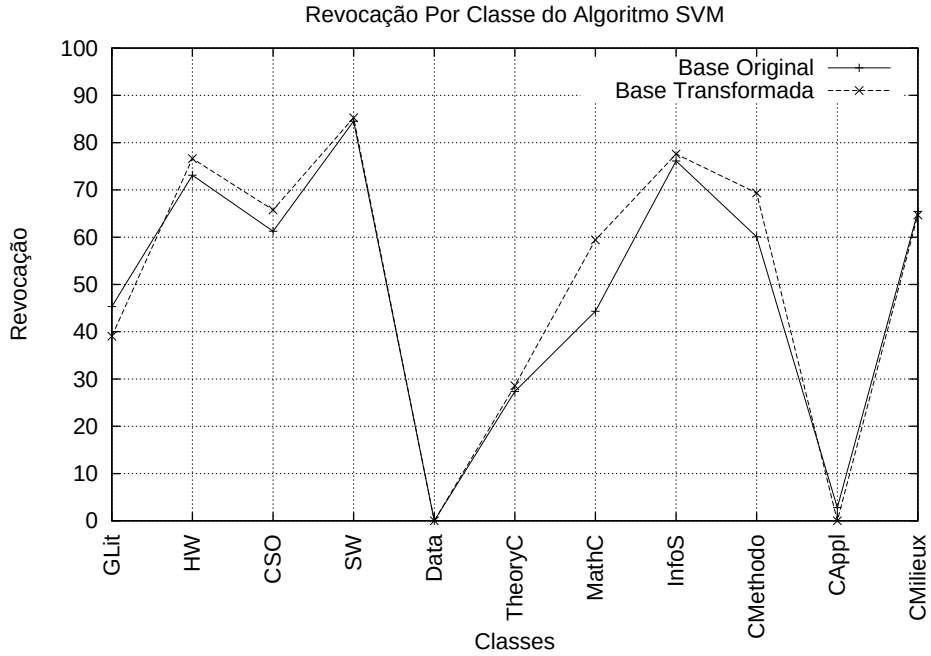

Figura 5.36. Análise da Revocação por Classe

base de dados, em comparação a essas mesmas métricas para a base de dados original. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo

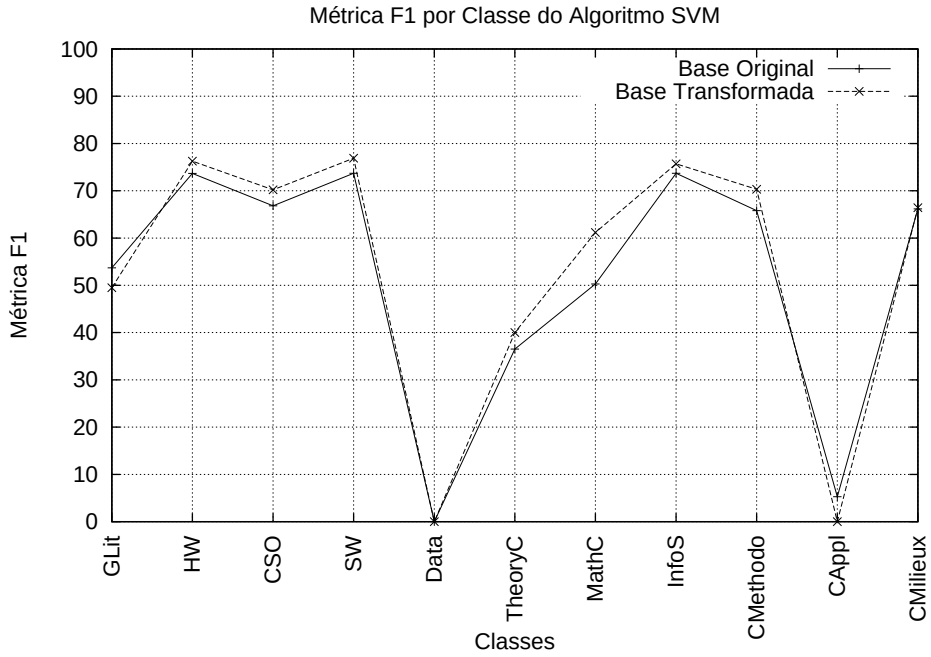

Figura 5.37. Análise da F1 por Classe

SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se um limiar de predominância mínimo igual a $0,8$).

Podemos observar que, para todas as classes, com exceção da classe *Data* e da classe *CMethodo* (em que os valores para as duas bases de dados foram praticamente iguais), a precisão apresentou uma melhora em relação à base de dados original. A métrica de revocação também se apresentou muito perto ou melhor para todas as classes na base de dados transformada, quando comparamos com a base original. Por fim, a F_1, que retrata a ponderação dessas duas métricas, também apresentou resultados próximos ou melhores para a base transformada. É importante observar que a precisão para a classe *CAppl* atingiu um valor de 100%, enquanto que, para a classe *Data*, esse valor foi 0%, para essa configuração da base de dados. Isso se deve ao fato de que nenhum documento foi associado à classe *CAppl* e apenas um documento foi associado à classe *Data*, sendo esse documento associado a ela erroneamente. Isso explica também os baixos valores das métricas revocação e F_1 para essas duas classes.

5.5.3 Limiar de Predominância Variável

Através dos experimentos realizados anteriormente, pode-se perceber que o parâmetro γ exerce uma influência fundamental na transformação da base de dados, uma vez que ele é o responsável por filtrar a quantidade de informação que será agregada ou não à base. Entretanto, há um compromisso em relação à variação desse parâmetro. Por um lado, ao se diminuir o valor de predominância mínima, temos que mais informações serão agregadas à base de dados, o que pode tanto contribuir com o processo de classificação quanto prejudicá-lo, ao inserir ruídos. Por outro lado, ao se aumentar o valor de γ, menos informações são agregadas à base de dados, mas a qualidade dessas informações é melhor. Assim, é difícil estabelecer um valor de γ que consiga ponderar essas duas características. Além disso, o valor ideal desse parâmetro também pode variar entre classes e ao longo dos anos. Ou seja, enquanto para uma classe, em um determinado período de tempo, o limite de predominância mínimo para encontrar os termos discriminantes possa ser alto (em torno de $0,8$, por exemplo), para uma outra classe (ou em outro período de tempo), esse limite pode ser muito pequeno.

A partir dessa hipótese, consideramos então uma nova estratégia, em que o limite mínimo de predominância é variável. Primeiramente, é feita uma leitura da base de dados em que os termos são contabilizados para cada classe em cada ano. É importante lembrar que não contabilizamos os termos pertencentes a documentos do conjunto de teste. Em seguida, deve-se determinar quais termos são predominantes em cada classe em cada ano. Nesse momento, ao invés de termos um limite de predominância mínimo

fixo, ele será flexível e poderá variar de acordo com a classe e ao longo dos anos. Para cada ano, verificamos, então, em qual classe o termo é mais freqüente, e consideramos que o termo é predominante nessa classe. Dessa forma, para cada ano, o termo será considerado predominante em apenas uma classe, o que não acontecia nas estratégias anteriores nas quais, caso o termo ultrapassasse o limite mínimo de predominância em mais de uma classe, ele era considerado predominante em todas elas. Assim, utilizando essa estratégia, ao mesmo tempo que, em determinados anos, um termo poderá ser considerado predominante com o valor de $0,9$ em uma determinada classe, em outros anos, um termo poderá ser considerado predominante com um valor de $0,2$ para uma dada classe, dado que esse termo não possui um valor maior de predominância em nenhuma outra classe.

É importante observar que, mesmo que um documento do conjunto de treino pertença à classe D, por exemplo, se for averiguado que um dos seus termos é mais predominante na classe F, esse termo receberá um novo rótulo que o associará à classe F e não à classe D. Isso é coerente uma vez que espera-se que um documento da classe D, mesmo contendo termos que são mais predominantes em outras classes, contenha um número considerável de termos que são mais predominantes na classe à qual ele pertence. Portanto, não estamos introduzindo ruídos na base de dados. Nas estratégias propostas anteriormente nesta seção, entretanto, considerávamos a classe do documento de treino para verificar se os termos eram predominantes nessa classe específica, sendo que eles só recebiam um novo rótulo caso fossem predominantes na mesma classe à qual o documento pertencia. Isso apresenta uma diferença considerável na transformação dos dados.

Apesar de o limite mínimo de predominância ser variável nesse caso, como explicado anteriormente, estabelecemos um limite mínimo aceitável como a menor predominância possível para que um termo seja considerado predominante em pelo menos uma classe. Se um termo não atingir o limite mínimo em nenhuma das classes em um dado ano, então ele não é considerado predominante naquele ano. Assim, realizamos a transformação das bases de dados usando valores para esse limite mínimo de tolerância da variação da predominância de $0,1$ a $0,9$ (em intervalos de $0,1$). Apresentamos a seguir algumas características das bases de dados geradas.

Novamente, para cada uma das bases, analisamos o número de termos que foram considerados predominantes em uma classe em pelo menos um ano. Esses resultados estão apresentados na Figura 5.38, em que o eixo x do gráfico representa o limite mínimo do limiar de predominância variável utilizado para gerar a nova base de dados, e o eixo y representa o número de termos predominantes em uma classe em pelo menos um ano, encontrados quando cada um dos limites mínimos de predominância variável

é utilizado.

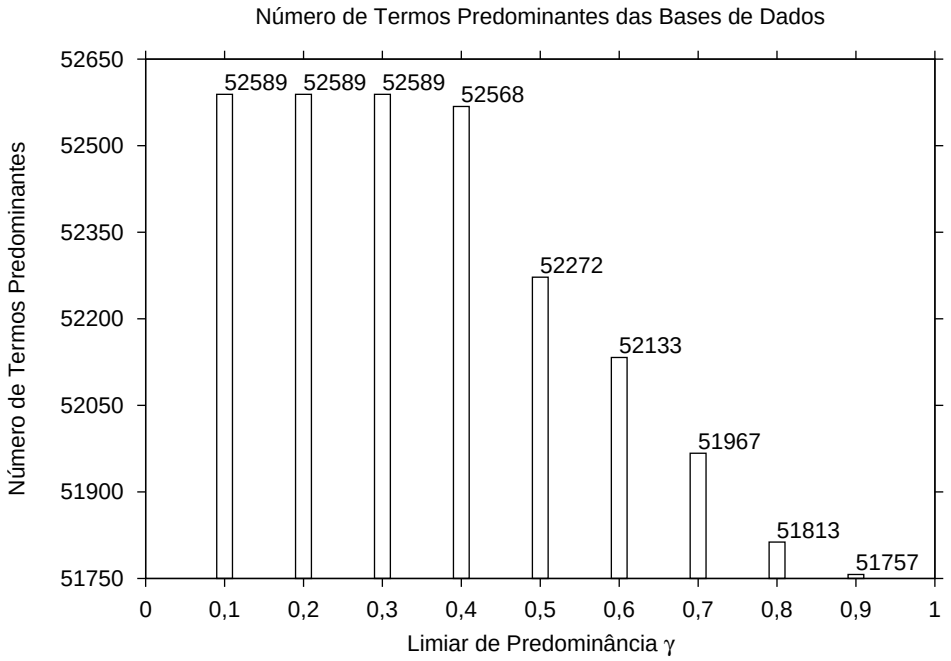

Figura 5.38. Número de Termos Predominantes

Ao se comparar os dados obtidos com os resultados da estratégia anterior, temos que o número de termos considerados predominantes foi praticamente idêntico. Isso se deve ao fato de que realmente não mudamos a forma de contabilizar os termos da base de dados e sim, a forma de determinar como os novos rótulos serão atribuídos a eles. Dessa forma, o número de termos considerados predominantes realmente deve ser similar. É importante lembrar ainda que, nessa estratégia, um termo pode ser considerado predominante em apenas uma determinada classe a cada ano e, portanto, não faz sentido apresentar o gráfico com o número médio de classes a que um termo está associado, como fizemos na estratégia anterior.

A Figura 5.39, por sua vez, mostra uma análise dos termos pertencentes aos documentos do conjunto de teste, uma vez que eles foram novamente considerados de forma especial no tratamento de substituição por novos rótulos. O eixo x do gráfico representa o limite mínimo do limiar de predominância variável utilizado para gerar a nova base de dados, e o eixo y representa o número de termos. Dessa forma, contabilizamos entre os termos pertencentes aos documentos do conjunto de teste, aqueles que foram considerados predominantes e aqueles que não foram considerados predominantes. Além disso, entre os termos de teste considerados predominantes em uma dada classe, contabilizamos também quantos foram considerados predominantes na classe certa e quantos foram considerados predominantes, mas associados à classe errada.

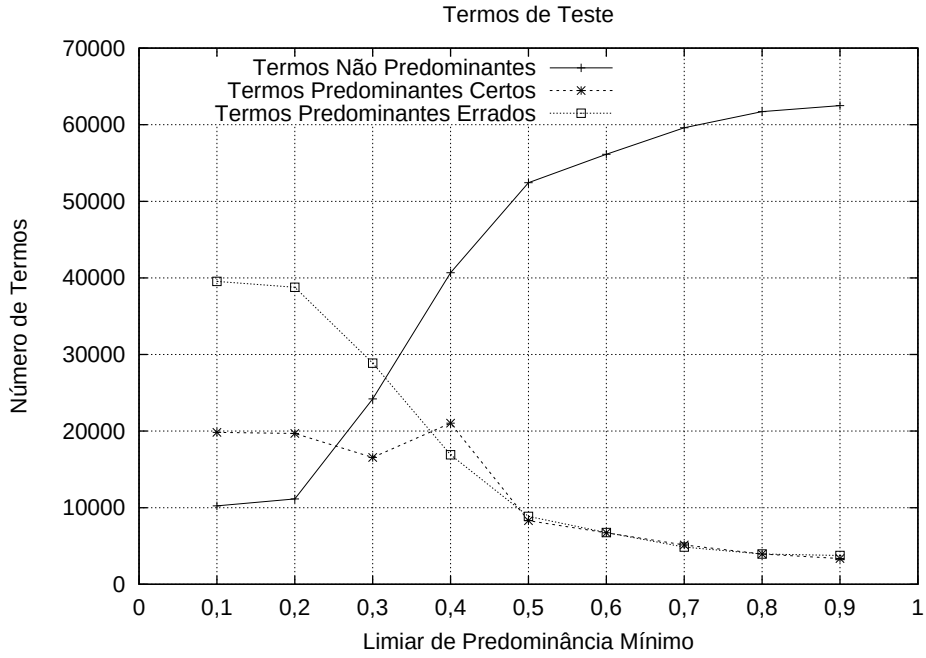

Figura 5.39. Análise do Número de Termos de Teste

Fazendo-se uma análise comparativa com a estratégia anterior, temos que o comportamento do gráfico é muito semelhante ao observado previamente. Isso ocorre uma vez que a forma de tratar os termos pertencentes ao documento de teste é a mesma nas duas estratégias. Em ambas, associamos os termos dos documentos de teste à classe em que ele ocorreu com mais freqüência no ano considerado, uma vez que não temos a informação de à qual classe o documento de teste pertence. Assim, podemos perceber que a mudança do rótulo dos termos que pertencem aos documentos de treino é o que realmente diferencia as duas estratégias.

Por fim, após o estudo das características das bases de dados geradas, analisaremos o impacto na acurácia do processo de classificação que as transformações proporcionam. Assim, realizamos o processo de *3-fold cross-validation* para cada uma das bases e calculamos a média das acurácias obtidas. Os resultados obtidos estão apresentados na Figura 5.40, em que o eixo x do gráfico representa o limite mínimo do limiar de predominância variável utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas.

Através do gráfico, podemos perceber que o limiar de variação mínimo da predominância não exerce a mesma influência na acurácia do algoritmo que o limiar de predominância fixo. Na verdade, o melhor resultado obtido foi quando esse limiar de variação foi o menor considerado, em que conseguimos uma acurácia de 73,60% (re-

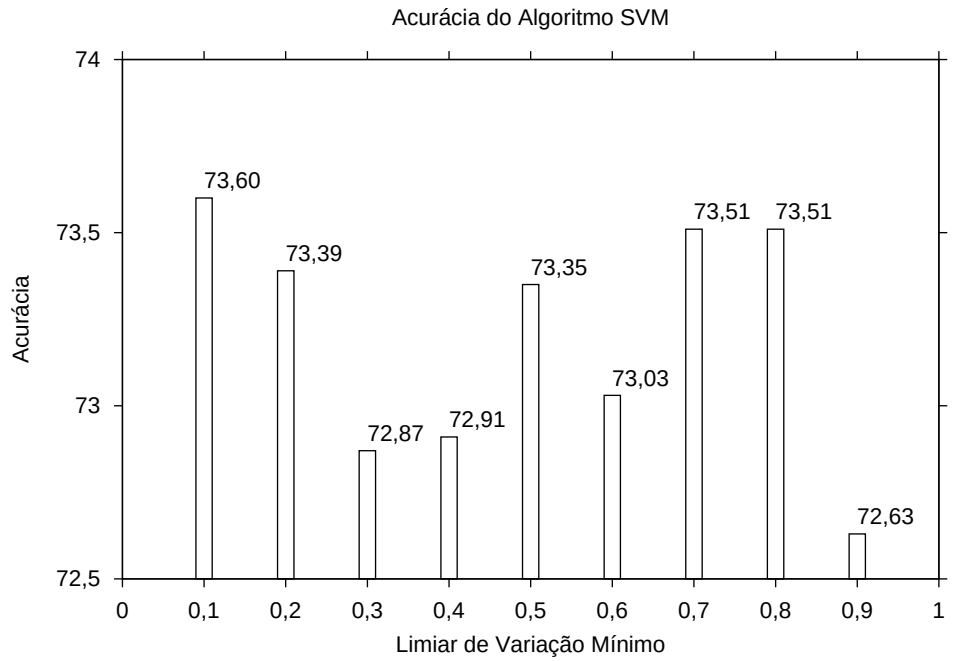

Figura 5.40. Análise da Acurácia - Limiar de Predominância Variável

presentando um ganho de mais de 5% em relação à acurácia obtida quando não há transformação da base de dados). Ou seja, quando permitimos uma variação grande do limiar mínimo para que um termo seja considerado predominante, obtemos uma melhor configuração da base de dados. Isso nos mostra que é importante considerar cada classe de forma específica, permitindo limiares mínimos de predominância diferentes para cada uma delas. Além disso, nota-se que é também importante permitir essa variação ao longo do tempo (foi permitido que uma mesma classe também tivesse limiares de predominância mínimos diferentes em anos diferentes), já que constatamos que existe uma evolução temporal das classes.

Novamente, vamos também analisar o impacto em cada uma das classes, objetivando entender melhor a transformação realizada na base de dados, e como ela contribuiu para uma melhora da eficácia do classificador. Para ilustrar esse impacto, utilizamos a base de dados que proporcionou o maior ganho da acurácia, gerada quando o limite inferior da variação de predominância foi estabelecido como 0,1. A Figura 5.41, a Figura 5.42 e a Figura 5.43 mostram, respectivamente, as métricas de precisão, revocação e F_1 obtidas para essa base de dados, em comparação a essas mesmas métricas para a base de dados original. O eixo x desses gráficos representa então cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se como o

limite mínimo do limiar de predominância um valor igual a $0,1$).

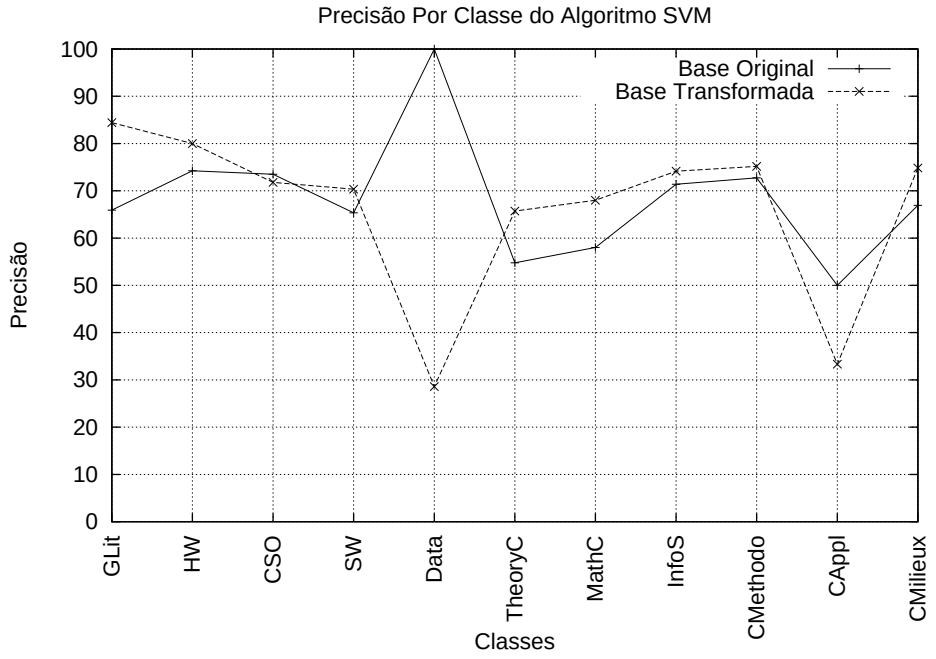

Figura 5.41. Análise da Precisão por Classe

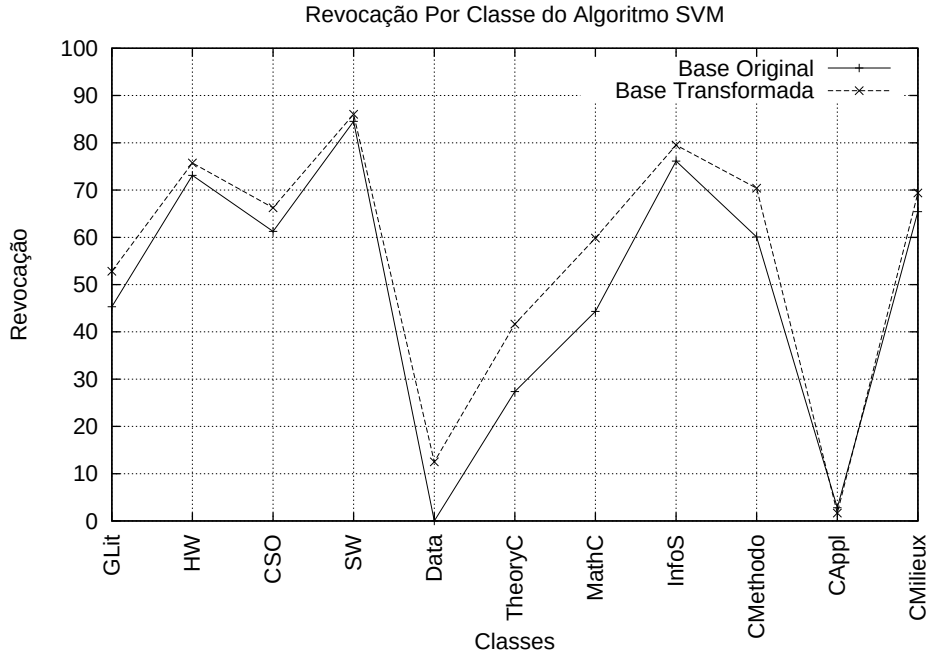

Figura 5.42. Análise da Revocação por Classe

No gráfico que apresenta a métrica de precisão, podemos observar que, mesmo obtendo-se uma precisão mais baixa na classe *CAppl*, ao compararmos com o resultado

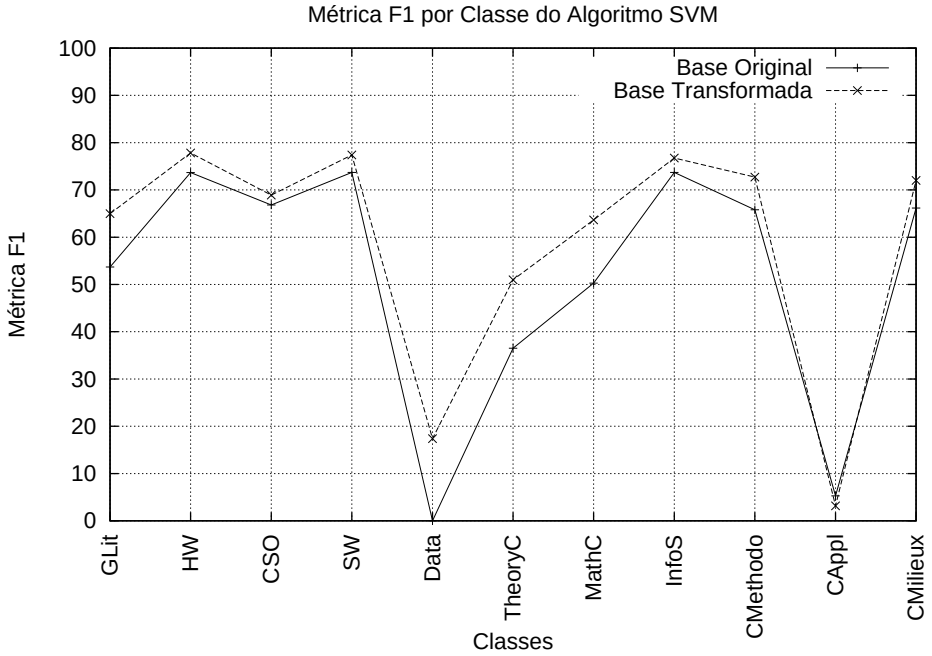

Figura 5.43. Análise da F1 por Classe

obtido na estratégia anterior, constatou-se ganhos consideráveis na precisão para as classes *GLit* e *Data*. Assim, o ganho obtido nessas classes foi o suficiente para compensar a perda na classe *CAppl*, já que obtivemos uma acurácia maior. A métrica de revocação, por sua vez, se apresentou melhor para todas as classes na base de dados transformada (com exceção da classe *CAppl*, em que foi um valor muito próximo), quando comparamos com a base original. Por fim, a F_1, que retrata a ponderação dessas duas métricas, apresenta um comportamento similar ao descrito para a métrica revocação. Além disso, podemos explicar os baixos valores dessas métricas obtidos para as classes *Data* e *CAppl*, devido ao fato de que, novamente, apenas um pequeno número de documentos foram associados a elas.

Além da análise dessas métricas por classe, pode-se notar que conseguimos uma acurácia melhor do que a obtida utilizando a estratégia anterior. Uma vez que verificamos que a diferença principal entre essas duas estratégias se concentra no tratamento dos termos que pertencem aos documentos do conjunto de treino, podemos concluir o seguinte fato: é mais interessante associar um termo à classe em que ele se apresenta mais predominante em um dado ano (mesmo que o documento a que ele pertence seja associado a uma outra classe), do que tentar associá-lo somente à classe associada ao documento em que ele ocorre. Assim, espera-se que esse documento contenha termos suficientes que serão mais predominantes na classe à qual ele pertence, fazendo com que ele seja associado a classe correta, mesmo contendo um ou mais termos predominantes

em outras classes.

5.6 Localidade Temporal na Agregação do Ano e da Classe a Termos Predominantes

As estratégias propostas até o momento evoluíram da seguinte forma: começamos com a transformação da base de dados apenas agregando a informação dos anos aos termos. Em seguida, adicionamos o conceito de termos predominantes aos experimentos. Consideramos, ainda, a distinção entre classes dentro de cada ano, fazendo experimentos com um limiar de predominância fixo e variável. Por fim, nesta seção, iremos propor uma nova estratégia, que incorpora um novo aspecto às estratégias previamente desenvolvidas.

Esse novo aspecto introduzido é a localidade temporal. Até então, analisamos a predominância de um termo (em uma dada classe) dentro de um dado ano. Nesse experimento, utilizaremos também os anos próximos ao ano do documento que contém o termo analisado, para estabelecer a predominância desse termo. Ou seja, se o termo não é predominante no ano do documento que o contém, ao invés de automaticamente considerá-lo não predominante, será feita uma análise dos anos próximos ao ano do documento em questão, para realmente estabelecer se esse termo será ou não considerado predominante.

A adaptação para considerar anos próximos ao ano do documento que contém o termo analisado é diferente para a estratégia com o limiar de predominância fixo e para a estratégia com o limiar de predominância variável. A seguir, descreveremos cada uma delas.

5.6.1 Limiar de Predominância Fixo

A estratégia desenvolvida aqui é muito similar à estratégia proposta na Subseção 5.5.2. O que as diferencia é o fato de que, quando um termo não é considerado predominante em uma dada classe e em um dado ano, analisamos os anos próximos a ele. Dessa forma, temos que analisar separadamente dois casos nos quais um termo não é predominante. No primeiro caso, o termo não é predominante e pertence a um documento de treino. No segundo caso, o termo não é predominante e pertence a um documento de teste.

Quando temos o primeiro caso, procedemos da seguinte forma. Suponha que um termo t, que pertence a um documento da classe C e do ano de 1993, não é predominante. Iremos analisar então se, no ano de 1992, esse termo t foi predominante na classe C. Caso não tenha sido, analisamos ainda o ano de 1994 e assim sucessivamente,

analisando sempre anos no máximo até J de distância (tanto no passado quanto no futuro) do ano original. Esse valor J é obtido como parâmetro e equivale ao tamanho da janela temporal que será permitida. Se o termo for predominante em pelo menos 50% dos anos dentro da janela temporal, então ele é considerado predominante no ano em questão.

No segundo caso, por sua vez, estamos analisando um termo t que pertence a um documento de teste do ano 1986, por exemplo, cuja classe é desconhecida. Temos ainda que esse termo não é predominante em nenhuma classe nesse dado ano. Nesse caso, investigamos se t é predominante em alguma classe no ano 1985. Caso seja, associamos t a classe em que ele possui sua maior predominância nesse ano. Se t também não for predominante em nenhuma classe em 1985, analisamos sua predominância no ano 1987 e assim sucessivamente, analisando sempre anos até no máximo J de distância (tanto no passado quanto no futuro) do ano original. Portanto, utilizamos novamente uma janela temporal, cujo tamanho foi passado como parâmetro.

Definimos, então, a seguinte configuração para realizar os experimentos. Variamos o tamanho da janela temporal assumindo os valores de 1 a 10 (de 1 em 1). É importante notar que esse valor do tamanho da janela temporal representa o número de anos que serão analisados no passado e no futuro, a partir do ano em questão. Assim, quando $J = 1$, estão sendo analisados três anos: o ano em questão, o ano anterior e o ano posterior a ele. Quando $J = 2$, o número de anos sendo analisados é cinco (2 anteriores, 2 posteriores e o ano em questão) e assim por diante. Além disso, para cada tamanho da janela temporal, variamos o parâmetro γ entre $0,1$ e $0,9$ (de $0,1$ em $0,1$), e geramos uma nova base para cada uma dessas configurações. Em seguida, calculamos a média das acurácias do processo de classificação obtidas através do *3-fold cross-validation* para cada uma das novas bases. Os resultados desse experimento estão apresentados na Figura 5.44, em que o eixo x do gráfico representa o limiar de predominância mínimo γ utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas. Além disso, cada curva do gráfico corresponde a um determinado tamanho da janela temporal utilizado na realização desses experimentos.

Através do gráfico, pode-se perceber que para todos os casos, ou seja, para qualquer tamanho de janela temporal, a acurácia do classificador aumentou à medida que aumentamos o limite mínimo de predominância. Isso pode ser explicado pelo fato de que, ao aumentarmos o limite mínimo de predominância, estamos considerando realmente só os termos mais discriminantes para aquela classe e diferenciá-los (como fazemos nessa estratégia e em estratégias anteriores) pode contribuir para uma melhor classificação

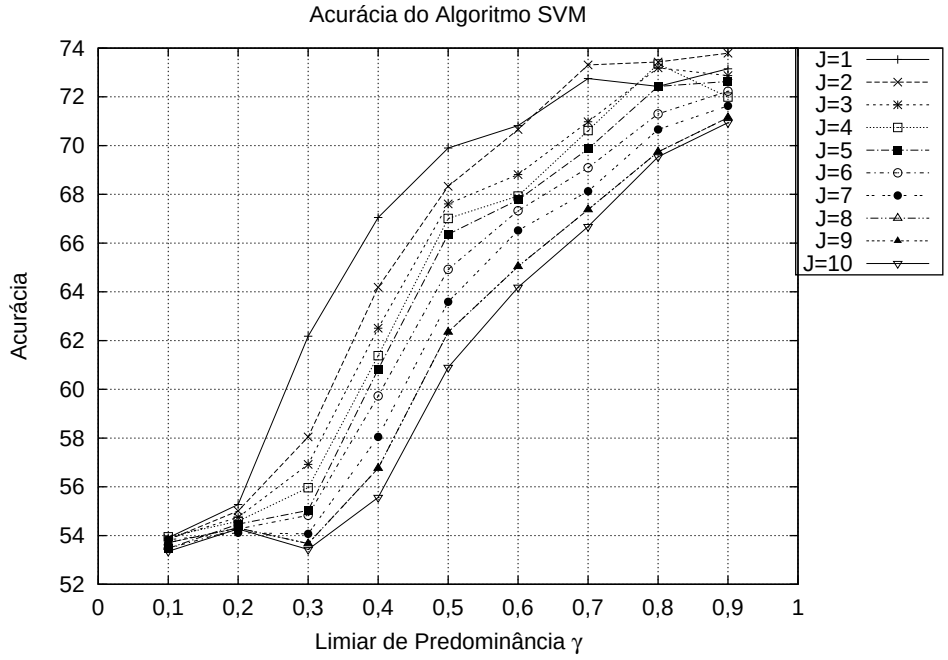

Figura 5.44. Análise da Acurácia - Limiar Fixo com Janela Temporal

dos documentos. Além disso, nota-se que a janela temporal $J = 1$ tende a apresentar melhores acurácias até atingir o limite mínimo de predominância $\gamma = 0,7$. A partir desse valor, a janela temporal $J = 2$ passa a apresentar as melhores acurácias. É interessante observar que isso está de acordo com a análise feita utilizando-se a filtragem de dados (Seção 5.2), em que o tamanho da janela ótima para essa base de dados equivale a 5 anos. Assim, com essa janela temporal e o limite mínimo de predominância igual a $0,9$ conseguimos obter uma acurácia de $73,79\%$, o que representa um ganho de mais de $5,3\%$ em relação à acurácia obtida sem a transformação da base de dados.

Para analisar o impacto da transformação em cada uma das classes, vamos utilizar, agora, a base de dados obtida com a janela temporal $J = 2$ e o limiar de predominância mínimo $\gamma = 0,9$. A Figura 5.45, a Figura 5.46 e a Figura 5.47 mostram, respectivamente, as métricas de precisão, revocação e F_1 obtidas para essa base de dados, em comparação a essas mesmas métricas para a base de dados original. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se um limiar de predominância mínimo igual a $0,9$ e uma janela temporal $J = 2$).

Novamente, ao compararmos com o resultado obtido na estratégia apresentada na Seção 5.5.2, temos uma melhora na precisão para as classes *GLit* e *Data*. É interessante

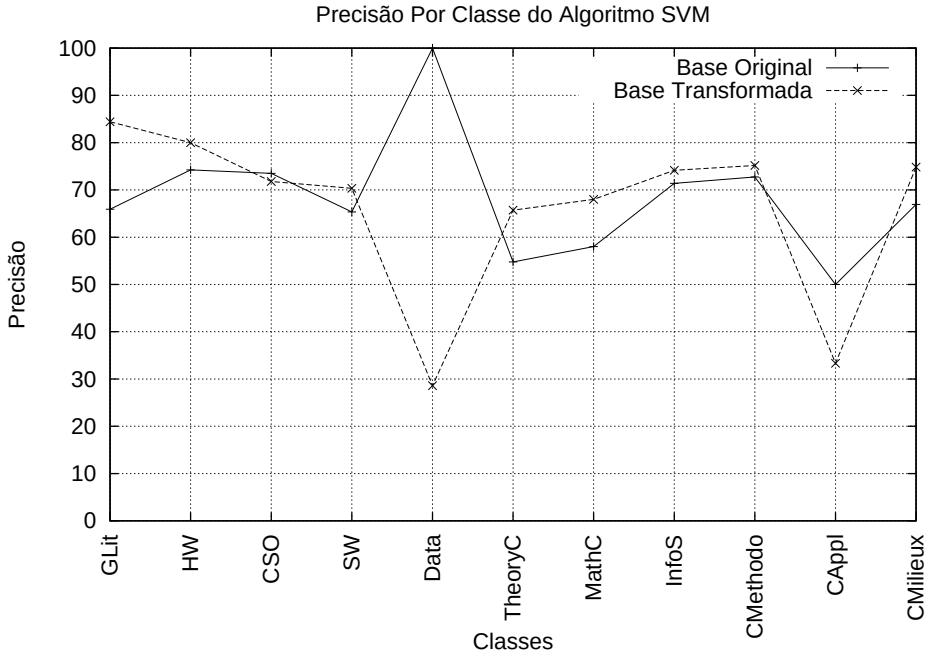

Figura 5.45. Análise da Precisão por Classe

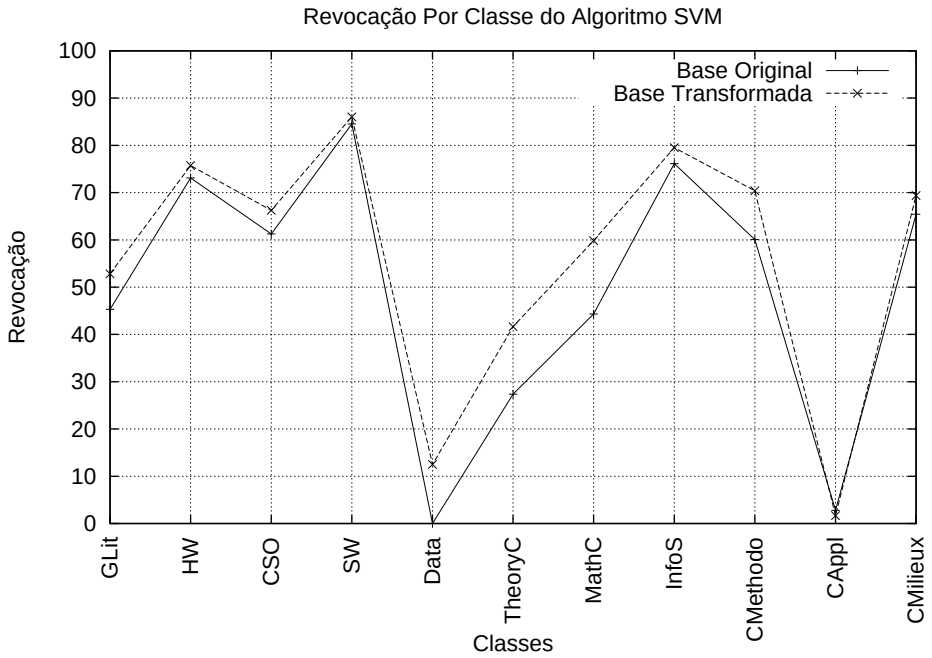

Figura 5.46. Análise da Revocação por Classe

lembrar, como mencionado anteriormente, que essas duas classes estão exatamente entre as que mais sofrem com a evolução temporal, tanto na evolução de seus termos, quanto na sua distribuição ao longo dos anos (como mostrado na Figura 4.13 e na Figura 4.9, respectivamente). Mais uma vez, a métrica de revocação se apresentou

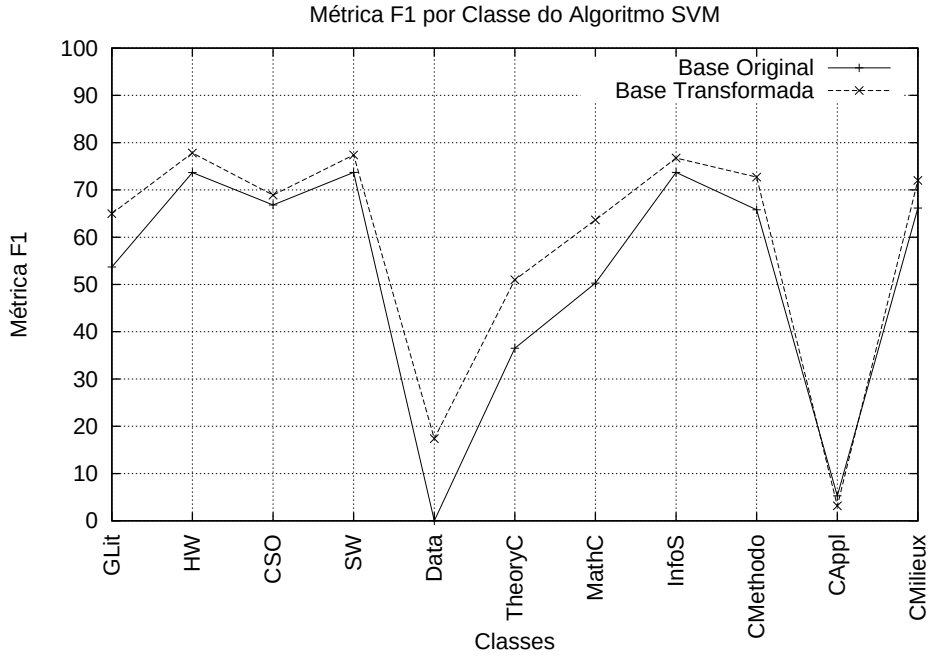

Figura 5.47. Análise da F1 por Classe

melhor para todas as classes na base de dados transformada (com exceção da classe *CAppl*), quando comparamos com a base original. A métrica F_1, por sua vez, que retrata a ponderação dessas duas métricas, se apresenta melhor para as mesmas classes que a revocação. Lembramos, novamente, que os baixos valores das métricas obtidos para as classes *Data* e *CAppl* é devido ao pequeno número de documentos que foram associados a elas.

5.6.2 Limiar de Predominância Variável

Nesta subseção, vamos de novo considerar a localidade temporal, entretanto, utilizando agora um limiar de predominância variável. Esse experimento é muito semelhante ao da Subseção 5.5.3, com a incorporação desse novo conceito de também avaliar os anos próximos ao ano do documento em questão.

Para incorporar a localidade temporal nessa estratégia, procedemos da forma descrita a seguir. Após ser feita a contabilização do número de vezes que cada termo aconteceu em cada classe e em cada ano, verifica-se, para cada ano, em qual classe cada um dos termos foi mais predominante, associando apenas uma classe a cada um deles. Se um termo t, entretanto, não for predominante em nenhuma classe no ano de 1999, por exemplo, na estratégia da Subseção 5.5.3 ele não recebia um novo rótulo. Nesta estratégia, contudo, analisamos se t é predominante em alguma classe no ano de 1998 e, caso seja, t recebe um novo rótulo de acordo com a classe em que ele se apre-

senta mais predominante em 1998. Caso t também não seja predominante em 1998, analisamos então o ano de 2000, e assim sucessivamente, analisando sempre anos até no máximo J de distância (tanto no passado quanto no futuro) do ano original. Como feito anteriormente, esse valor J é passado como parâmetro e equivale ao tamanho da janela temporal que será permitida.

Utilizamos novamente a mesma configuração do experimento anterior, ou seja, variamos o tamanho da janela temporal assumindo os valores de $J = 1$ até $J = 10$. Para cada tamanho da janela temporal, variamos o parâmetro γ entre $0,1$ e $0,9$ (de $0,1$ em $0,1$), e geramos uma nova base para cada uma dessas configurações. Em seguida, calculamos a média das acurácias do processo de classificação obtidas através do *3-fold cross-validation* para cada uma das novas bases. Os resultados desse experimento estão apresentados na Figura 5.48, em que o eixo x do gráfico representa o limite inferior do limiar de predominância mínimo variável utilizado para gerar a nova base de dados, e o eixo y representa a média da acurácia obtida pelo SVM (utilizando-se o processo de *3-fold cross-validation*) ao classificar os documentos que compõem cada uma das novas bases geradas. Além disso, cada curva do gráfico corresponde a um determinado tamanho da janela temporal utilizado na realização desses experimentos.

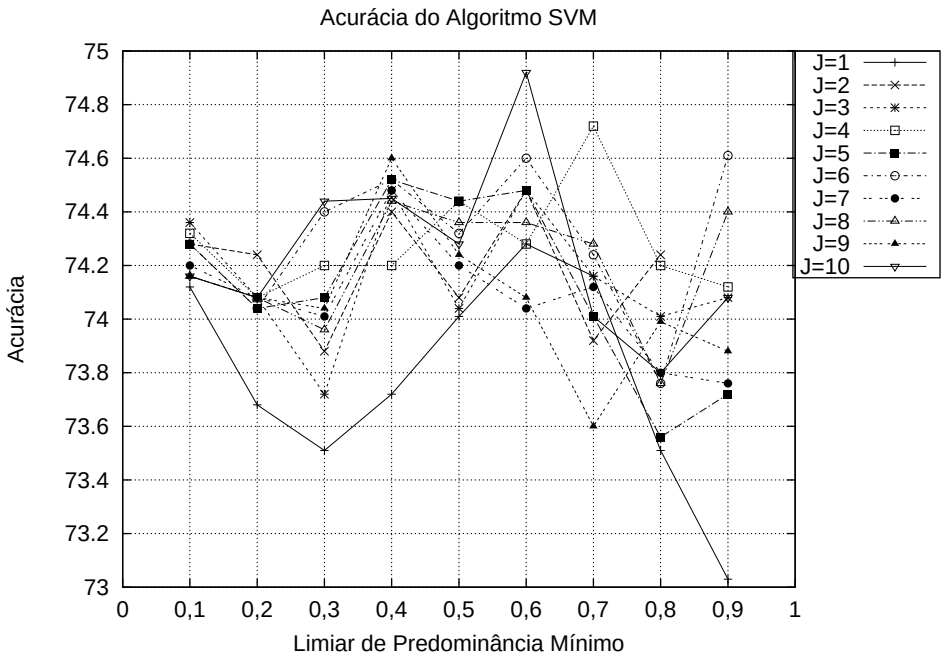

Figura 5.48. Análise da Acurácia - Limiar Variável com Janela Temporal

Através do gráfico, pode-se perceber que, assim como aconteceu no experimento da Seção 5.5.3, o limiar mínimo de variação da predominância não exerce a mesma influência na acurácia do algoritmo que o limiar de predominância fixo. Além disso,

constatamos que o melhor resultado obtido foi quando esse limiar de variação assumiu o valor $0,6$. Isso nos mostra que, ao mesmo tempo que é importante permitir uma variação do limiar mínimo de predominância entre as classes e ao longo dos anos, é também importante não deixar que esse limiar assuma valores muito baixos, para não haver introdução de ruídos na base de dados e para possibilitar realmente a distinção dos termos que têm um maior poder discriminativo.

Outro fato a ser considerado é que a melhor acurácia obtida foi para uma janela temporal de tamanho $J = 10$, ou seja, o tamanho máximo considerado. Entretanto, isso não significa que a base de dados inteira deve ser considerada. Quando vamos analisar se um termo é predominante ou não, partimos do ano em questão do documento considerando, primeiramente, os anos mais próximos a ele e, se necessário, vamos afastando gradualmente, conforme foi explicado. Assim, estamos permitindo janelas temporais de tamanho variado para cada termo, uma vez que, para cada um deles, podemos caminhar um número específico de anos para achar em qual classe ele se apresenta mais predominante. Dessa forma, quando $J = 10$, estamos apenas permitindo uma variação máxima dessa janela temporal estabelecida para cada termo. Podemos concluir, portanto, que é importante permitir a variação dessa janela temporal, uma vez que cada termo sofre um impacto diferente ao longo do tempo, modificando o seu poder discriminativo de forma diferente. Por fim, conseguimos obter uma acurácia de $74,92\%$, o que equivale a um ganho de aproximadamente $6,5\%$ em relação à acurácia obtida sem alteração da base de dados.

Novamente, para entendermos melhor a transformação realizada na base de dados e como ela contribuiu para uma melhora da eficácia do classificador, vamos também analisar o impacto em cada uma das classes. Para ilustrar esse impacto, utilizamos a base de dados que proporcionou o maior ganho da acurácia, isto é, a base gerada com a janela temporal $J = 10$ e com o limiar mínimo de variação da predominância igual a $0,6$. As métricas de precisão, revocação e F_1 obtidas para essa base de dados, em comparação a essas mesmas métricas para a base de dados original, estão apresentadas na Figura 5.49, na Figura 5.50 e na Figura 5.51, respectivamente. Assim, o eixo x desses gráficos representa cada uma das classes presentes na coleção da ACM, e o eixo y representa a respectiva métrica (precisão, revocação ou F_1) obtida pelo SVM ao classificar os documentos de cada classe (tanto para a base original, quanto para a base transformada, utilizando-se um limite inferior do limiar de predominância mínimo variável igual a $0,6$ e uma janela temporal $J = 10$).

Através dos gráficos, pode-se perceber que, apesar de apresentar uma precisão menor para as classes *Data* e *CAppl* (que são duas classes que devem ser consideradas como casos especiais, devido ao pequeno número de documentos associados a elas), a

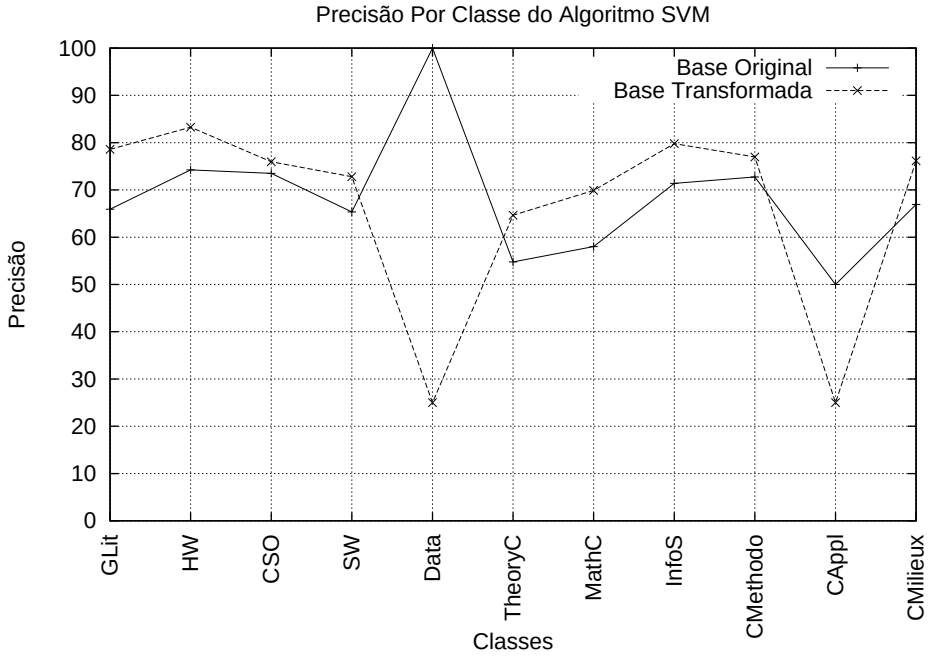

Figura 5.49. Análise da Precisão por Classe

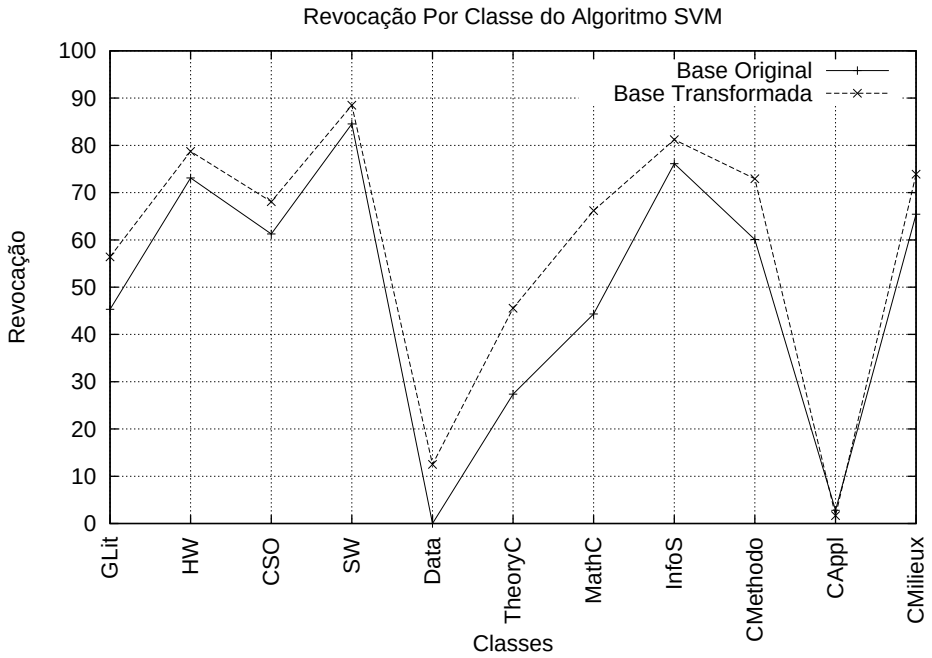

Figura 5.50. Análise da Revocação por Classe

base de dados transformada apresenta uma melhora considerável na precisão das outras classes quando comparadas à base original. A métrica de revocação também se apresentou melhor para todas as classes (com exceção da classe *CAppl*), obtendo uma melhora de até 17% na classe *TheoryC*. A métrica F_1, por sua vez, apresenta também

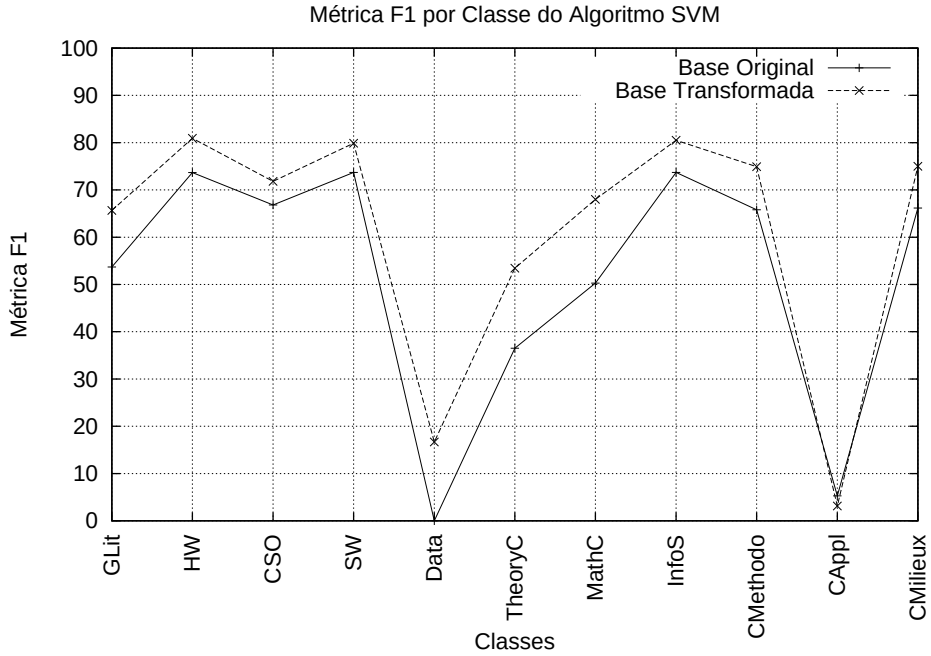

Figura 5.51. Análise da F1 por Classe

melhoras.

Assim, podemos concluir que a transformação de termos predominantes por classe e por ano, considerando um limiar de predominância mínimo variável e incorporando o conceito de localidade temporal dos documentos, foi a estratégia que nos proporcionou a maior melhora da eficácia do processo de classificação.

Capítulo 6

Conclusões e Trabalhos Futuros

Neste trabalho, conseguimos mostrar evidências de que o tempo é realmente um fator importante a ser considerado nas técnicas e algoritmos de classificação. Além disso, mostramos ser possível melhorar a eficácia da tarefa de classificação, quando são considerados os efeitos temporais na construção de classificadores automáticos. Assim, para lidar com a evolução temporal, propomos estratégias que envolvem engenharia de dados, através de processos de filtragem e/ou transformação dos mesmos.

Para comprovar e mensurar o impacto dos aspectos temporais no processo de classificação, definimos e aplicamos uma nova metodologia de análise. Nessa metodologia, constatamos a existência de um compromisso entre o efeito amostral e a evolução temporal da coleção. Nós mostramos também que a evolução temporal pode ser explicada por três fatores: distribuição de classes, distribuição de termos e similaridade de classes. Além disso, lidamos com cada um desses fatores separadamente, ressaltando as causas que resultam na degradação da eficácia do classificador.

Após a caracterização da evolução temporal na CAD, compreendendo melhor os seus aspectos, propomos estratégias de engenharia de dados para tratar os efeitos temporais.

Com o processo de filtragem de dados, nossa estratégia objetivou otimizar o compromisso entre o efeito amostral e a evolução temporal e, conseqüentemente, otimizar a acurácia do classificador, utilizando uma seleção de documentos para o conjunto de treinamento. Ou seja, para cada documento sendo testado, selecionamos como o nosso conjunto de treinamento somente os documentos que fossem temporalmente próximos a ele. Essa proximidade é definida por uma janela temporal que pode crescer simetricamente em ambas as direções, passado e futuro, partindo do ano do documento de teste. Utilizando uma janela temporal ótima para cada ano que maximiza a acurácia do classificador, atingimos uma acurácia de $89,76\%$ para a coleção da ACM e $87,57\%$ para a coleção da MedLine, o que significa um ganho de aproximadamente 20% na

acurácia, com um conjunto de documentos de treino muito menor. Entretanto, apesar de a estratégia exaustiva de filtragem ser apropriada para uma avaliação analítica, certamente não é apropriada para cenários reais. Por outro lado, ela mostra que há espaço para melhoras em métodos de classificação melhorando sua acurácia.

Com o processo de transformação de dados, por sua vez, objetivamos lidar com os efeitos temporais, agregando aos dados informações que podem contribuir com o tratamento da evolução temporal que existe na coleção. Para tal, desenvolvemos estratégias que envolvem a atribuição de novos rótulos aos termos dos documentos de forma que eles incorporem novas informações. Assim, essas estratégias foram evoluindo gradativamente: a primeira envolve a agregação do ano aos termos dos documentos; a segunda estratégia desenvolve o conceito de termos predominantes por ano; a terceira lida com esses termos predominantes, separando-os também por classes; e a quarta estratégia incorpora o conceito de localidade temporal nos aspectos considerados anteriormente. Com exceção da primeira, em todas as outras estratégias propostas, conseguimos obter uma melhor eficácia do algoritmo. Esse ganho foi maior na última estratégia, em que obtivemos uma acurácia de $74,92\%$, o que equivale a um ganho de aproximadamente $6,5\%$ em relação à acurácia obtida sem alteração da base de dados.

Para prosseguir com os estudos realizados neste trabalho, explorando as contribuições fornecidas pelo mesmo, propomos os seguintes trabalhos futuros:

- Aplicação da metodologia de caracterização dos efeitos temporais a outras coleções de documentos (coleções Web, por exemplo), para mostrar que a evolução temporal também ocorre.

- Realização de um estudo comparativo (através da aplicação da metodologia de caracterização dos efeitos temporais), verificando quais tipos de bases de dados sofrem um maior impacto com a evolução temporal.

- Realização de um estudo sobre o impacto na eficiência computacional do classificador, ocasionado pela aplicação das estratégias propostas (tanto de filtragem quanto de transformação dos dados).

- Utilização de outros algoritmos de classificação para avaliar como as estratégias apresentadas neste trabalho afetarão no seu desempenho, assim como, para comprovar a extensibilidade da solução proposta.

Referências Bibliográficas

Allan, J. (1996). Incremental relevance feedback for information filtering. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 270–278, New York, NY, USA. ACM.

Allan, J.; Carbonell, J.; Doddington, G.; Yamron, J. e Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, Lansdowne, VA, USA. 007.

Belkin, N. J. e Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12):29–38.

Borko, H. e Bernick, M. (1963). Automatic document classification. *J. ACM*, 10(2):151–162.

Brieman, L. e Spector, P. (1992). Submodel selection and evaluation in regression: The x-random case. *International Statistical Review*, 60:291–319.

Caldwell, N. H. M.; Clarkson, P. J.; Rodgers, P. A. e Huxor, A. P. (2000). Web-based knowledge management for distributed design. *IEEE Intelligent Systems*, 15(3):40–47.

Cohen, W. W. e Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Trans. Inf. Syst.*, 17(2):141–173.

Dumais, S.; Platt, J.; Heckerman, D. e Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pp. 148–155, New York, NY, USA. ACM.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 252–259, New York, NY, USA. ACM.

Friedman, J. H.; Kohavi, R. e Yun, Y. (1996). Lazy decision trees. In Shrobe, H. e Senator, T., editores, *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pp. 717–724, Menlo Park, California. AAAI Press.

Haykin, S. (1999). Adaptive filters. In *Signal Processing Magazine*. IEEE Computer Society.

Joachims, T. (1999). Making large-scale svm learning practical. In *Schölkopf B., Burges C.J.C., and Smola A.J. (Eds.), Advances in Kernel Methods-Support Vector Learning*, pp. 169–184, Cambridge, MA. MIT Press.

Joachims, T. (2006). Training linear svms in linear time. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 217–226, New York, NY, USA. ACM Press.

Jones, R. e Diaz, F. (2007). Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14.

Klinkenberg, R. e Renz, I. (1998). Adaptive information filtering: Learning in the presence of concept drifts. In *Learning for Text Categorization*, pp. 33–40, Menlo Park, California, USA. AAAI Press.

Lam, W. e Ho, C. Y. (1998). Using a generalized instance set for automatic text categorization. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 81–89, New York, NY, USA. ACM Press.

Lanquillon, C. e Renz, I. (1999). Adaptive information filtering: detecting changes in text streams. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pp. 538–544, New York, NY, USA. ACM.

Lawrence, S. e Giles, C. L. (1998). Context and page analysis for improved web search. *IEEE Internet Computing*, 2(4):38–46.

Lewis, D. D. (1995). Evaluating and Optimizing Autonomous Text Classification Systems. In Fox, E. A.; Ingwersen, P. e Fidel, R., editores, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 246–254, Seattle, Washington. ACM Press.

Liu, R.-L. e Lu, Y.-L. (2002). Incremental context mining for adaptive document classification. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 599–604, New York, NY, USA. ACM.

McCallum, A. e Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48. AAAI Press.

Mladenic, D. e Grobelnik, M. (1998). Feature selection for classification based on text hierarchy. In *Conference on Automatic Learning and Discovery*.

Mourão, F.; Rocha, L.; Araújo, R.; Couto, T.; Gonçalves, M. e Meira, Jr., W. (2008). Understanding temporal aspects in document classification. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pp. 159–170, New York, NY, USA. ACM.

Salton, G. e McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Scholz, M. e Klinkenberg, R. (2007). Boosting classifiers for drifting concepts. *Intell. Data Anal.*, 11(1):3–28.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Veloso, A.; Meira, Jr., W.; Cristo, M.; Gonçalves, M. e Zaki, M. (2006). Multi-evidence, multi-criteria, lazy associative document classification. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 218–227, New York, NY, USA. ACM.

Yang, Y. e Kisiel, B. (2003). Margin-based local regression for adaptive filtering. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pp. 191–198, New York, NY, USA. ACM.