Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp. 169–174, June 2–5, 2001 - Centro Universitario da FEI, Sao Paulo - SP - Brazil

Classificação de Textos Através do Modelo Nebuloso de Formas e Métodos de Agregação

Evandro de Oliveira Araújo
UFMG - Universidade Federal de Minas Gerais
Departamento de Eletrônica
Av. Antônio Carlos, 6627 31270-901 Belo Horizonte - MG, Brasil
E-mail: earaujo@cpdee.ufmg.br

Abstract

This paper approaches the problem of using pattern recognition techniques in large text databases for documents classification. It is proposed a vector representation for documents based on the frequency of words in the document. This vector representation of the texts corresponds to two-dimensional shapes for which we can determine a fuzzy model. A document can be classified comparing its fuzzy model with the fuzzy models of the prototypes of the available classes. An unknown text belongs to a class whose fuzzy model is the most similar to the fuzzy model of the text. It is also proposed to classify a text based on the idea of clustering. The class of the text is that of the nearest prototype. Some texts have been downloaded from the web and used for classification purposes. The simulation results obtained show the validity of the approach.

1. Introdução

O grande desenvolvimento tecnológico das últimas décadas é responsável pelo fluxo cada vez maior de informações exigindo o uso de sistemas de recuperação de informação cada vez mais sofisticados. Técnicas de Aprendizado de Máquina [1, 2] têm sido empregadas para pesquisar Bases de Dados de Documentos a partir de descrições de seus conteúdos. Recentes desenvolvimentos em Aprendizado de Máquina, Recuperação de Informação bem como em Agentes Inteligentes têm oferecido soluções promissoras para os usuários de internet encontrar uma boa e rápida seleção da informação que procuram [3, 4]. A utilização de todas estas técnicas requer estudos em Representação, Raciocínio e Aprendizado. A busca de uma representação adequada para estes problemas de classificação/recuperação de textos é uma questão chave nestas aplicações de reconhecimento de padrões. A possibilidade de representar um texto como uma forma geométrica plana a ser descrita através de um modelo nebuloso permite que uma ampla variedade de técnicas de reconhecimento de padrões possam ser empregadas. A identificação de um texto com uma classe de textos versando sobre um tema corresponde à procura de um protótipo cujo modelo é o mais similar ao do texto. Uma outra abordagem deste problema, sob a ótica da Agregação, também é possível. Neste caso, a classe do texto desconhecido é a do protótipo cujo centro é o mais próximo dos pontos correspondentes ao texto. A classificação automática de textos tem aplicações variadas desde a seleção de textos para leitura e *download* na rede de computadores ao auxílio a um operador na manutenção e operação de uma planta. Neste caso, um módulo de reconhecimento de textos poderia ser incorporado a um Sistema Especialista de auxílio à tomada de decisão agilizando o trabalho do operador. Dada a descrição de um problema, o sistema buscaria o texto contendo as indicações dos procedimentos correspondentes àquela situação descrita.

Um texto pode ser caracterizado para fins de reconhecimento de padrões pelo conjunto de palavras que o compõe. Uma representação de um texto mais compacta pode ser feita levando em conta o número de ocorrências de cada palavra no texto: $\{ \ldots \ p_i \ / n_i \ \ldots \}$, onde cada palavra p_i é seguida pelo número de ocorrências n_i dela no texto. Dois textos similares devem conter, aproximadamente, as mesmas palavras com o mesmo grau de incidência. Um procedimento natural para classificação de textos pode ser estabelecido, reunindo os diversos grupos de textos e constituindo uma Base de Dados geral de palavras que servirá de guia para pesquisa das similaridades. Um certo número de assuntos é escolhido e para cada um dos assuntos, determina-se um conjunto protótipo para representar este assunto (palavras seguidas pelo número de ocorrências). Forma-se então uma Base de Dados de palavras constituída por todas as palavras empregadas nas classes, sem repetição. A escolha de um ordenamento adequado para a representação da Base de Dados de palavras é muito importante. Uma escolha simples e natural é a ordem alfabética. Esta escolha, porém, apresenta inconvenientes. Assuntos bastante distintos geralmente compartilham palavras comuns. Se distribuirmos o conjunto total de palavras ordenadas alfabeticamente espacialmente (ao longo do eixo horizontal), notaremos uma grande interseção entre os diversos conjuntos, dificultando a separação entre eles. Uma solução para este problema é apresentada na seção 3.

1.1. Organização do Artigo

O artigo é organizado da seguinte maneira. Na próxima seção é apresentado o modelo nebuloso de formas planas [5] que será usado neste trabalho. Na seção seguinte é proposta uma representação para textos visando o reconhecimento de padrões. Em seguida, esta representação é vista como uma forma geométrica plana e uma estratégia de classificação de textos baseada nesta interpretação é sugerida. Uma outra abordagem deste problema de classificação de textos, sob a ótica de Agregação, é apresentada na subseção 4.1. Resultados da aplicação destas duas visões do problema são mostrados na seção 5 . Por fim, uma análise das limitações e da aplicabilidade da metodologia proposta é realizada e conclusões são apresentadas.

2. Modelo Nebuloso de Formas Geométricas Planas

Nesta seção é feita uma revisão do Modelo Nebuloso de Formas Geométricas Planas citado acima.

Um sistema de reconhecimento de padrões deve ser capaz de determinar a similaridade entre uma forma desconhecida e os representantes das classes (protótipos). No caso de figuras geométricas planas, a similaridade entre uma forma desconhecida F e o representante de uma classe C pode ser expressa como a combinação convexa entre as similaridades horizontal S_h e vertical S_v:

$$S(F,C) = w_h.S_h + w_v.S_v \tag{1}$$

A similaridade horizontal S_h é uma medida da proximidade entre os elementos das matrizes f^x (versão horizontal do modelo) do protótipo e da forma desconhecida.

Para se estabelecer o modelo nebuloso de uma figura plana [5, 6], a figura que se quer modelar é envolvida por uma moldura Γ, retângulo mínimo que contém a figura. Esta moldura é dividida em $N_x N_y$ células. O eixo horizontal é dividido em N_x regiões, associadas aos subconjuntos nebulosos $H_i, 1 \leq i \leq N_x$, caracterizados por uma função de pertinência triangular. De forma idêntica, é feita uma partição em relação ao eixo vertical, como mostra a Fig. 1. Nesta figura, é escolhida uma partição de Γ com $N_x = 5$ e $N_y = 4$, definindo 20 células. A versões horizontal e vertical do modelo nebuloso da forma plana são representadas pelas matrizes f^x e f^y abaixo:

$$f^x_{ij} = \frac{M^x{}' M^y}{\mathbf{1}\ M^y} \tag{2}$$

$$f^y_{ij} = \frac{M^x{}' M^y}{\mathbf{1}\ M^x} \tag{3}$$

Acima, o vetor linha **1** é dado por [1,...,1] e os vetores coluna M^x e M^y são formados pelos graus de pertinência dos pontos contidos na célula ij em relação aos conjuntos nebulosos definidos no eixo horizontal e vertical, respectivamente. Os elementos destas matrizes se situam entre zero e um. O valor unitário para f^x_{ij} significa que o valor da coordenada x dos pontos da célula ij são todos iguais ao valor modal. Neste caso, não nebuloso, os pontos desta região têm abscissa bem definida. O valor $f^x_{ij} = 1/2$ corresponde ao máximo de nebulosidade (indecisão). Ele ocorre quando todos os pontos da célula ij estão "ocupados". Não temos pois, neste caso, um valor de x que caracteriza a célula. A determinação da similaridade horizontal S_h (vertical S_v) entre duas figuras é feita a partir do cálculo de uma média ponderada das proximidades entre os elementos das matrizes f^x (f_y) correspondentes às formas geométricas. A escolha dos pesos pode ser feita de forma a refletir algum conhecimento a respeito de diferentes regiões de Γ. De forma idêntica, no cálculo da similaridade entre duas formas, combinação convexa de S_h e S_v, pode-se pesar diferentemente as duas dimensões de modo a realçar uma determinada dimensão.

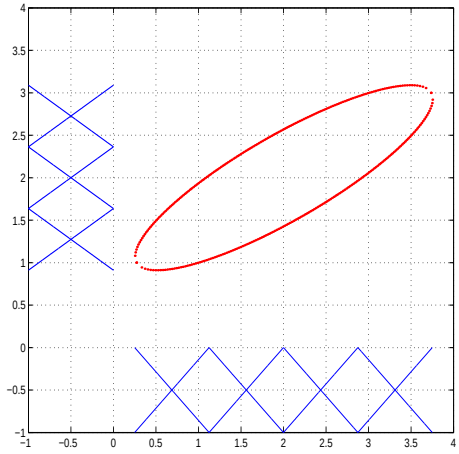

Figura 1: Partição Nebulosa de Γ.

3 Representação de Textos

Nesta seção é apresentado um procedimento para a formação de uma Base de Dados de palavras e é proposta uma representação vetorial muito simples e intuitiva para arquivos textos. Nas simulações realizadas, esta representação mostrou-se bastante eficiente para o nosso propósito de classificar um arquivo texto desconhecido numa das classes previamente escolhidas.

3.1 Formação de uma Base de Dados Geral de Palavras

Num problema de classificação de padrões, uma escolha adequada dos atributos é essencial para o êxito do sistema de classificação. Escolhidos os atributos, seus valores na amostra em particular que se quer classificar devem ser medidos. No caso de textos, estes atributos são, essencialmente, a presença de determinadas palavras ou melhor, o número de ocorrência delas num determinado texto. Dois textos são tão mais similares quanto

mais próximas forem as incidências das mesmas palavras nos dois textos. Deve ser salientado que estamos tratando os textos apenas quanto à forma, o que já não é muito simples. Comparar textos quanto ao conteúdo é uma tarefa muito mais complexa que não faz parte de nossos objetivos neste trabalho. Segundo o critério de similaridade aplicado aqui, dois textos podem ser considerados bastante similares (quanto à forma) mas apresentar conteúdos (ou argumentos) bastante distintos. Por exemplo, o advérbio de negação *não* é uma das palavras consideradas irrelevantes para a caracterização dos textos e, conseqüentemente, é sempre descartada. O objetivo deste trabalho é obter um sistema de classificação automática de textos que associe um texto desconhecido a uma das n classes conhecidas. A estratégia adotada é a seguinte: definidas as classes, monta-se uma Base de Dados geral de palavras BDG, constituída de todas as palavras relevantes utilizadas nos textos correspondentes às diferentes classes. Estas palavras contidas na BDG constituem os atributos do nosso sistema de reconhecimento de padrões. O número de ocorrência destas palavras num determinado texto é comparado com o número de ocorrência destas mesmas palavras nos protótipos das classes. Para se obter o protótipo P^i de uma classe C^i, $i = 1, \ldots, n$ é necessária uma coleção de arquivos texto t^i_j, $j = 1, \ldots, n_i$ versando sobre o tema da classe. Destes arquivos são extraídos os conjuntos de palavras relevantes para a caracterização do tema. Uma Base de palavras inicial é formada pela união das Bases de palavras relevantes de cada um dos conjuntos t^i_j, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$. O conjunto resultante é ordenado alfabeticamente. Obtida a BDG, o protótipo, representante da classe C^i, vai ser representado como

$$P^i = p_1/n^i_1, \ldots, p_{nt}/n^i_{nt} \tag{4}$$

onde nt é o número de palavras da BDG e n^i_k, representa a incidência média da palavra p_k da BDG nos conjuntos formadores da classe C^i. O número n^i_k pode ser simplesmente a média aritmética das incidências nos conjuntos individuais ou calculado como uma média ponderada das incidências individuais, onde os pesos representariam a qualidade da amostra na formação da classe. Se representarmos graficamente os protótipos, dispondo no eixo horizontal o conjunto BDG e no vertical, o número de ocorrências de cada palavra no protótipo, perceberemos uma grande superposição dos gráficos correspondentes aos protótipos, visto que as iniciais de palavras representativas das classes (assuntos) estão distribuídas aleatoriamente pelo alfabeto. Para assegurar uma maior separabilidade entre as representações gráficas das classes, torna-se necessário reordenar a Base de palavras conforme explicado abaixo. Consideremos o conjunto de palavras BD_i correspondente ao protótipo P^i (obtido a partir das palavras relevantes dos arquivos t^i_j) e o seguinte esquema para a obtenção de uma nova BDG:

- $i = 1$

$$BDG = BD_1$$

- até $i = n - 1$, faça

$$i = i + 1$$

$$Aux = BD_i - BDG, BDG = BDG \cup Aux$$

A operação de diferença de conjuntos do passo 2 garante a inclusão de apenas novos elementos na atualização da BDG corrente. Ao final do processo, teremos a reunião dos conjuntos de palavras com o máximo de separação e ordenamento possível. Se os conjuntos BD_i forem disjuntos, a representação gráfica da BDG exibirá uma total separação das representações gráficas das classes representadas pelos protótipos.

3.2 Determinação de uma Forma Geométrica para um texto

Uma representação gráfica para um texto pode ser obtida pela tradução do vetor protótipo em coordenadas cartesianas de acordo com as seguintes equações:

$$x(k) = k.K_x \tag{5}$$

$$y(k) = n^i_k.K_y \tag{6}$$

onde n^i_k é o número de ocorrências da palavra p_k da BDG no protótipo P^i. Os parâmetros K_x, K_y representam fatores de escala horizontal e vertical, respectivamente. Este conjunto de pontos define uma forma geométrica plana, cujo modelo nebuloso pode ser determinado como descrito na seção 2.

4 Estratégia para Classificação de Textos

O procedimento para classificação automática de textos envolve a escolha de um grande número de parâmetros. Alguns destes parâmetros são: os tópicos (classes) tratados, o número de arquivos texto necessários para caracterizar cada classe, o tamanho (número de palavras) de cada texto e a partição lingüística do modelo nebuloso da forma geométrica plana associado ao texto. A primeira etapa do processo de classificação de textos compreende a determinação dos protótipos das classes. Um estratégia geral para a classificação de textos é delineada abaixo.

1. Definir os parâmetros iniciais: as n classes, o número e o tamanho das amostras;
2. Escolher as amostras e eliminar as palavras irrelevantes;
3. Obter o protótipo de cada classe e determinar a BDG;
4. Obter um novo protótipo para cada classe considerando a BDG obtida;
5. Determinar um modelo nebuloso para cada protótipo;
6. Obter o modelo nebuloso de um texto desconhecido e classificá-lo.

Assegurar uma boa qualidade dos protótipos é fundamental para o sucesso do sistema de classificação. Para validar um protótipo P^i, podemos testar os modelos nebulosos dos conjuntos t^i_j que lhe deram origem em relação aos modelos nebulosos de todas as classes. Se um texto t^i_j usado para formar a classe C^i mostrar-se mais próximo de uma classe C^k, $k \neq i$, então considera-se que este texto não é representativo da classe C^i. Neste caso, ele deve ser substituído. Isto é comum de acontecer quando escolhemos um tema muito amplo e os textos escolhidos como amostras são pouco focados.

4.1 Agregação e a Classificação Automática de Textos

Técnicas de Agregação (*clustering*) [7] desempenham um papel de grande importância em Classificação de Padrões. O FCM [8, 9, 10], um dos algoritmos de agregação mais usados, procura os centros dos vetores protótipos de modo a maximizar a coesão dos protótipos, minimizando a soma das distâncias intra agregados. Visualizando os protótipos de cada classe como agregados de pontos, podemos classificar um texto desconhecido de acordo com a soma das distâncias de seus pontos aos centros destes agregados. Considerando que um texto t^k produziu r pontos, ele recebe o rótulo da classe C^j se

$$\sum_{l=1}^{r} d(t^k_l, P^j) = \min_i \sum_{l=1}^{r} d(t^k_l, P^i) \qquad (7)$$

5 Apresentação dos resultados

Para testar a metodologia proposta, foram escolhidos alguns temas e extraídos vários arquivos texto na internet versando sobre os temas. Alguns destes arquivos foram usados para formar os protótipos, e os outros foram separados para integrar o conjunto de testes. Os arquivos texto de teste foram classificados segundo a maior similaridade com algum protótipo e também pela soma das distâncias de seus pontos aos agregados representados pelos protótipos. Os resultados de classificação foram coincidentes na grande maioria dos casos. O índice de acerto geral ficou em torno de 87%.

5.1 Escolha de parâmetros

Para cada tema, foram selecionados sete textos contendo de 100 a 200 palavras como conjuntos de treinamento. Os tópicos escolhidos foram os seguintes: artes, ciências, esportes e economia. O tema artes representa um grande conjunto de manifestações humanas como cinema, música e literatura, entre outras. Os textos escolhidos versaram principalmente sobre cinema e literatura. O segundo tema, ciências, também é bem amplo. Encontra-se nele uma variedade muito grande de assuntos, de ciências físicas a biológicas, incluindo aí, temas mais atuais como transplantes, clonagem, etc. O tema esportes incluiu apenas reportagens sobre futebol. As amostras de textos para a caracterização da última classe, economia, foram também bastante diversificadas. Destes textos foram formados os protótipos das classes. A busca das palavras relevantes de cada tema foi realizada através de diversas operações sobre os textos, reduzindo-se assim o número de palavras representativas de cada classe. Foram criados diversos operadores com a função de modificar, substituir e eliminar palavras. Como exemplo de modificação podemos citar a supressão de desinências de flexão verbal, de gênero e número. Foi estabelecida uma equivalência entre formas verbais de modo a concentrar numa única forma (radical) diferentes conjugações de um verbo. Foi montado um pequeno dicionário temático que permitiu a substituição de determinadas palavras características de um assunto por um sinônimo. Foram eliminadas as palavras com menos de quatro letras ou mais de 16 letras por julgá-las pouco representativas. Foram eliminadas palavras de diferentes classes gramaticais como advérbios, numerais, preposições, conjunções. Para ilustrar o processamento efetuado, consideremos o seguinte grupo de palavras: exames, examinar, examinou, exame, examinará e examinado. Todas elas são substituídas pelo radical "exam". Neste exemplo, vemos que o operador que trata da flexão em número foi empregado estabelecendo a equivalência entre as palavras "exames" e "exame". Também as flexões verbais foram percebidas, resultando na ocorrência do radical "exam". Após esta filtragem inicial, foram consideradas as cem primeiras palavras de cada texto por ordem de aparição. Este procedimento visa homogenizar as diferentes amostras. Após a obtenção da BDG e a formação dos protótipos, um modelo nebuloso para cada classe foi obtido. A partição lingüística adotada foi de três termos para cada um dos eixos, $N_x = N_y = 3$.

5.2 Classificação de acordo com a similaridade

Para validar a metodologia, foram selecionados diversos textos do conjunto de testes e classificados segundo a maior similaridade com os representantes das classes, os protótipos. A Tabela 1 mostra os resultados obtidos com dez amostras. Os resultados são expressos em termos do grau de similaridade dos textos de teste nas seguintes classes: artes, ciências, esporte e economia. O valor unitário representa a similaridade total e zero, a ausência de similaridade. Como se pode observar, as classificações foram corretas em nove dos dez casos.

5.3 Classificação pela técnica de agregação

As mesmas amostras foram apresentadas ao classificador que utiliza uma abordagem baseada em agregação. As classes foram colocadas em ordem alfabética (na mesma ordem de formação da BDG) e representadas graficamente no sistema de eixos cartesianos. Como o ordenamento explicado anteriormente privilegia uma boa separação das classes, pudemos visualizar nitidamente a formação dos agregados. Um algoritmo que implementa *fuzzy c-means* foi executado para determinar os vetores protótipos correspondentes às quatro classes. Obtidos os vetores protótipos, procurou-se determinar para

cada vetor representativo das dez amostras acima o vetor protótipo mais próximo. A Tabela 2 mostra os resultados obtidos. Houve dois erros de classificação. A amostra de número sete, texto sobre esportes, foi novamente rotulada como pertencente à classe economia. O outro erro ocorreu exatamente onde não se esperava. A amostra de número quatro, texto sobre economia, apresentou uma grande similaridade (0,73) com o protótipo desta classe e bastante superior àquela relativa à classe esporte (0,27). No entanto, a classificação pelo critério de agregação indicou a classe esportes como sendo a classe deste texto de teste. A soma das distâncias dos pontos aos centros foi ligeiramente inferior quando se considerou esta classe. Isto pode ser explicado pela observação de que a classe economia foi colocada como a classe mais à direita na representação gráfica e, em particular, esta amostra continha poucos pontos e um deles, característico da primeira classe (artes) está bastante distante do centro do protótipo da classe economia e mais próximo do centro do protótipo da classe esporte. Como eram poucos pontos, este fato pesou bastante, o suficiente para tornar a soma das distâncias ao centro representando a classe economia maior que o valor correspondente à classe esportes.

Tabela 1: Tabela de resultados de classificação através do cálculo de similaridades.

	art	cie	esp	eco
texto artes	0,40	0,33	0,20	0,23
texto ciências	0,38	0,68	0,40	0,53
texto esportes	0,17	0,33	0,43	0,33
texto economia	0,17	0,47	0,27	0,73
texto artes	0,38	0,30	0,31	0,35
texto ciências	0,32	0,77	0,34	0,53
texto esportes	0,19	0,49	0,58	**0,65***
texto economia	0,11	0,33	0,19	0,68
texto esportes	0,22	0,48	0,52	0,36
texto economia	0,25	0,41	0,19	0,65

6 Aplicabilidade e Limitações da Metodologia Proposta

O método proposto se insere na área conhecida como *text retrieval* e pode usado em aplicações diversas como em classificação de assuntos obtidos esparsamente numa grande Base de Dados como a internet; responder de forma automática às perguntas propostas, associando um texto pergunta ao texto resposta cujo modelo nebuloso é o mais próximo do texto pergunta. Este tipo de aplicação pode auxiliar os operadores / técnicos a encontrar o diagnóstico de uma falha a partir de uma base de Dados constituída de pares problemas / solução. As principais limitações desta proposta são relativas à indefinição, subjetividade e às sutilezas da linguagem natural escrita. Na procura de identificação de padrões, encontramos situações em que diferentes palavras ou expressões (e até gírias) são empregadas de forma exclusiva para descrever determinadas situações. As regiões de fronteira são, geralmente, difíceis de tratar. Mesmo em casos onde a classificação através da comparação de similaridades entre o texto desconhecido e os padrões das classes é segura, a utilização da idéia de agregação para classificar o mesmo texto pode levar a um resultado inesperado como aconteceu no exemplo mostrado. Algumas questões importantes relativas à parametrização do procedimento devem ser resolvidas. Qual é o número ideal de amostras para bem caracterizar uma classe e como escolher as amostras? Considerando um grande número de amostras, conseguimos filtrar mais eficientemente determinados termos muito particulares de um certo assunto, muito freqüentes num certo momento. A questão temporal desempenha um papel relevante na obtenção das amostras. A BDG é dinâmica, ela é um retrato de um certo momento. Um possível inconveniente de considerarmos um número muito elevado de amostras é obter um protótipo com valores de ocorrência associados às palavras muito baixo devido à inevitável expansão de base de palavras. Como o número de palavras do nosso vocabulário é finito, evidentemente, este problema seria atenuado a partir de um certo número de amostras devido às repetições das palavras. O tamanho de cada amostra é também um parâmetro importante. O melhor é escolher textos maiores e limitar o número de palavras relevantes durante o processamento. Um texto pequeno, comum em problemas do tipo pergunta / resposta, é mais difícil de caracterizar. Uma outra questão relevante é a definição do número de palavras que caracterizam a amostra. Uma primeira idéia para resolver este problema é considerar todas as palavras, excluídas aquelas classes de palavras já citadas. Isto implicaria um grande ônus computacional sem assegurar necessariamente a contrapartida de um melhor desempenho. Uma forma de limitar o número de palavras características de uma classe é fixar um número mínimo de ocorrências de uma palavra para ela fazer parte do conjunto de palavras da classe. Um número máximo de ocorrências deve também ser fixado de forma a evitar uma predominância atípica, particular para um texto, de uma certa palavra. É interessante também analisar a estabilidade do protótipo considerando diferentes números de amostras. Neste trabalho, foram tomadas sete amostras (textos) de cada tema textos contendo de 100 a 200 palavras. As palavras com incidência inferior a 5 (nos sete textos) foram descartadas na formação do protótipo. É interessante observar que não é possível estabelecer uma correspondência segura entre os ordenamentos mostrados nas linhas das duas tabelas. A primeira linha da Tabela 1 mostra que o arquivo de testes sobre artes é mais similar ao protótipo da classe artes e em segundo, terceiro e quarto lugar, aos protótipos das classes ciências, esportes e economia, respectivamente. Já a primeira linha da Tabela 2 mostra o seguinte ordenamento: artes, ciências, economia e esportes. Para o arquivo teste sobre ciências, segunda linha, já se observa uma discordância quando se compara a segunda classe mais similar (economia) com a segunda menor soma de distâncias (artes).

Tabela 2: Tabela de resultados de classificação usando a técnica de agregação.

	art	cie	esp	eco
texto artes	27,48	31,91	45,12	62,51
texto ciências	22,36	13,09	30,57	51,39
texto esportes	31,96	19,13	13,55	22,32
texto economia	26,42	18,11	**14,08***	14,15
texto artes2	28,14	38,18	41,42	35,22
texto ciências	35,43	19,12	33,35	28,91
texto esportes	39,87	26,32	15,38	**14,22***
texto economia	34,49	20,84	6,71	5,76
texto esportes3	54,73	38,93	30,39	48,11
texto economia	31,97	24,08	26,69	16,71

7. Conclusões

O tratamento de textos é um assunto muito complexo pois envolve as dificuldades relacionadas à linguagem natural escrita. A abordagem deste problema deve ser encarada sob a ótica do raciocínio aproximado, procurando estabelecer alguns limites de erro, algumas tolerâncias e sobretudo tentando encontrar uma representação mais adequada para discriminar bem os padrões. Os estudos aqui apresentados estão ainda numa fase bastante preliminar. Os resultados, contudo, são bastante encorajadores. A abordagem do problema por análise de agregados é bastante simples e fácil de implementar. Ela parece, entretanto, ser menos confiável e flexível do que a abordagem via modelos nebulosos. Uma idéia interessante de se explorar é utilizar uma outra dimensão nesta representação onde se poderia atribuir um peso às palavras de acordo com a sua importância para a sua classe. Um especialista examinaria cada protótipo e atribuiria os pesos. Poderia ser também idealizado um esquema de aprendizado destes pesos, usando, talvez, uma abordagem com elementos de redes neurais. Neste esquema, uma mesma palavra teria certamente diferentes pesos em diferentes classes. Nesta representação tridimensional, o modelo nebuloso poderia também ser obtido como mostrado em [6].

Referências

[1] R. Forsyth and R. Rada. *Machine Learning Applications in Expert Systems and Information Retrieval*. Ellis Horwood, New York, 1986.

[2] P. Edwards, D. Bayer, C. L. Green, and T. R. Payne. Experience with learning agents which manage internet-based information. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, CA, 1996.

[3] D. Mladenic. *Machine Learning on non Homogeneous, distributed text data*. Phd Thesis, University of Ljubjana, Slovenia, 1998.

[4] A. P. Ladeira. *Desenvolvimento de um Agente Inteligente Moderador de Listas de Discussão de Cursos à Distância: Um Estudo de Casos para o curso Programação em Linguagem C*. Dissertação de Mestrado UFMG, Belo Horizonte, 2001.

[5] B. Lazzerini and F. Marcelloni. A fuzzy approach to 2-d shape recognition. 9(1):5–16, 2001.

[6] E. Araújo. Modelo nebuloso de formas geométricas planas: Propriedades e aplicações. In *Anais do XI Congresso Brasileiro de Automática*, volume I, pages 39–60, Natal - RN, 2000.

[7] B. S. Everitt. *Cluster Analysys*. John Wiley, New York, 3rd edition, 1993.

[8] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.

[9] J. C. Bezdek and S. K. Pal. *Fuzzy models for Pattern Recognition: methods that search for structure in data*. IEEE, New York, 1992.

[10] E. Araújo. A heuristic adjustment to the calculation of the dissimilarity in the fcm algorithm. In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, volume I, pages 25–30, Vancouver, Canada, 2001.