Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 5th Best MSc dissertation/PhD thesis contest (CTDIA'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM.
ISBN 85-87837-11-7

Seleção de Atributos Importantes para a Extração de Conhecimento de Bases de Dados

Huei Diana Lee[1,2], Maria Carolina Monard[1]

[1] Laboratório de Inteligência Computacional (LABIC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil
[2] Laboratório de Bioinformática (LABI)
Universidade Estadual do Oeste do Paraná, Parque Tecnológico ITAIPU (PTI)
Caixa Postal 1151, 85856-970 – Foz do Iguaçu, PR, Brasil
huei@unioeste.br, mcmonard@icmc.usp.br

Resumo Feature selection plays an important role in the knowledge extraction process. Its objective is to choose a subset of features that describes a data set according to some importance criterion, since irrelevant and/or redundant features may decrease data quality and reduce the comprehensibility of hypotheses induced by supervised learning algorithms. This work presents three principal contributions with the proposal of: (1) an algorithm that introduces the use of Fractal Dimension to deal with redundant features, (2) an evaluation model, which considers both accuracy and the number of selected features to evaluate the performance of feature selection algorithms, and (3) a methodology to support the automatic mapping of medical findings to structured data bases.
Keywords: Machine Learning, Feature Subset Selection, Model of Evaluation

Date of conclusion: December 16^{th}, 2005.
PhD Thesis available at http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/

1 Introdução

Em aprendizado supervisionado, a indução de um classificador **h** é influenciado pelos valores dos atributos do conjunto de exemplos. Teoricamente, o uso de um maior número de atributos para descrever os exemplos deveria fornecer um maior poder de discriminação para aproximar a verdadeira função $y = f(\mathbf{x})$ desconhecida, onde **x** representa um exemplo do conjunto de dados, composto por N exemplos, na forma $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ rotulados com os valores y, sendo $\mathbf{x}_i$ vetores da forma $(x_{i1}, x_{i2}, ..., x_{iM})$ cujos componentes são valores discretos ou contínuos relacionados aos *atributos* $X = \{X_1, X_2, ...X_M\}$. No caso dos valores y pertencerem a um conjunto discreto de N_{C_l} classes, *i.e.* $y \in \{C_1, ..., C_{N_{C_l}}\}$, a tarefa de aprendizado é chamada de *classificação*, tratada neste trabalho. Porém,

esse maior poder de discriminação pode não ocorrer na presença de atributos irrelevantes e/ou redundantes, os quais, freqüentemente, confundem o algoritmo de aprendizado. A área de Seleção de Atributos — SA —, explorada não somente em estatística mas também em Aprendizado de Máquina e Mineração de Dados [1], tem como objetivo determinar, segundo algum critério, que atributos são importantes. Os resultados obtidos, tanto teórica quanto experimentalmente, mostram que a SA melhora a predição de classificadores e reduz a complexidade do modelo **h** [1,2,3,4,5].

Além dos atributos irrelevantes, tem sido observado que atributos *redundantes*, ou seja, correlacionados parcial ou completamente, também afetam a precisão dos classificadores induzidos e, portanto, deveriam ser eliminados [4,3]. Entretanto, o número de subconjuntos de atributos cresce exponencialmente com o número de atributos em X e encontrar o subconjunto ótimo de atributos pode ser um problema de complexidade não polinomial [6].

Em geral, os métodos de SA escolhem os atributos pela *avaliação individual* ou pela *avaliação de subconjuntos* de atributos. No caso de avaliação individual, os atributos são ordenados considerando a sua importância na discriminação das classes. Esses métodos somente removem atributos irrelevantes pois espera-se que atributos redundantes tenham a mesma importância na discriminação das classes. Contudo, métodos que avaliam subconjuntos de atributos procurando por subconjuntos mínimos podem remover tanto atributos irrelevantes quanto redundantes. Assim, a maioria dos métodos existentes para a SA que trata tanto relevância quanto redundância de atributos, o fazem de maneira implícita por meio da avaliação de subconjuntos de atributos. Ainda que esses métodos geralmente apresentem melhores resultados que os métodos que não lidam com a redundância de atributos, seu alto custo computacional pode torná-los ineficientes para conjuntos de dados com alta dimensionalidade. Recentemente foi proposto o uso da abordagem filtro considerando o modelo de tratamento da relevância e da redundância de atributos como dois procedimentos separados [2]. A vantagem desse modelo sobre o modelo anterior é que, por meio da separação da análise de relevância e de redundância, é possível diminuir o custo computacional da busca por um subconjunto que aproxima o subconjunto ótimo.

Neste trabalho são descritas três das principais contribuições da Tese de Doutoramento defendida pela autora no ICMC da Universidade de São Paulo [7]. Uma das contribuições foi a proposta de um algoritmo que realiza as análises de relevância e de redundância de atributos, utilizando o conceito de Dimensão Fractal — DF —, em processos separados. Muitos dos algoritmos para SA selecionam apenas atributos relevantes, não tratando o problema da existência de atributos redundantes, os quais tem mostrado exercer grande influência sobre o classificador induzido. Dois dos aspectos mais importantes que devem ser avaliados na tarefa de SA são a precisão do classificador induzido utilizando o subconjunto de atributos selecionados e a redução proporcionada pela seleção. Porém, usualmente as análises são realizadas considerando esses dois aspectos separadamente. Uma outra contribuição foi a proposta de um modelo para avaliação de algoritmos de SA que considera simultaneamente essas duas questões.

Um outro problema comum para a ampla aplicação de métodos computacionais que possam auxiliar na análise de dados reais, é a freqüente disponibilização de dados, especialmente na área médica, no formato de laudos semi-estruturados descritos em linguagem natural. A terceira principal contribuição deste trabalho foi a proposta de uma metodologia para o mapeamento semi-automático de laudos médicos para bases de dados estruturadas, a qual foi aplicada com sucesso a um caso real.

O trabalho está organizado do seguinte modo: na Seção 2 é descrita a questão da importância de atributos. Na Seção 3 são introduzidos conceitos sobre dimensão fractal e o algoritmo proposto para SA em aprendizado supervisionado, o qual utiliza conceitos da teoria dos fractais. A avaliação da proposta é apresentada na Seção 4 por meio da realização de experimentos sobre conjuntos de dados reais e sob o modelo proposto para a avaliação de algoritmos de SA, em comparação com outros métodos representativos para SA. Na Seção 5 é apresentada a metodologia proposta para auxiliar no mapeamento de laudos médicos para bases de dados estruturadas. Na Seção 6 são apresentadas as conclusões deste trabalho e alguns trabalhos futuros.

2 Importância de Atributos

Um classificador (modelo ou hipótese) $\mathbf{h}$ é induzido com o objetivo de aproximar a verdadeira função $y = f(\mathbf{x})$ desconhecida, onde $\mathbf{x}$ representa um exemplo do conjunto de dados descrito pelos valores do conjunto de *atributos* $X = \{X_1, X_2, ...X_M\}$ que descreve os exemplos. A meta da SA é selecionar um subconjunto mínimo de atributos $X' \subseteq X$ tal que $\mathbf{P}(C|y = f'(\mathbf{x}')) \approx \mathbf{P}(C|y = f(\mathbf{x}))$, as quais são as distribuições de probabilidades das N_{C_l} possíveis classes dados os valores dos atributos em X' e X, respectivamente. Esse subconjunto mínimo X' é denominado subconjunto *ótimo* de atributos [2].

Para alcançar essa meta, o problema da SA é caracterizado sob dois aspectos: como é realizada a avaliação dos atributos e o tipo de interação entre o algoritmo de aprendizado, que posteriormente utilizará os atributos selecionados, e o algoritmo de seleção de atributos propriamente dito. Para selecionar atributos é necessário definir o significado de um atributo ser considerado importante, *i.e.* responder à pergunta: ***Importante em relação a que ?*** Ainda, a necessidade de estimativa da importância de atributos é comum, tanto à avaliação individual quanto à avaliação de subconjuntos de atributos, qualquer que seja a estratégia de busca. A questão da avaliação é complexa e multidimensional. Por exemplo, a avaliação pode ser considerada em termos de: (1) melhoria da precisão do classificador ou (2) simplificação do modelo construído de modo que ele seja mais compreensível. Assim, a importância de um atributo pode ser definida, de uma maneira geral, como:

Definição 1 ***Importância de um Atributo*** *[1]: Um atributo é dito importante se quando removido a medida de importância considerada em relação aos atributos restantes é deteriorada.*

Diferentes medidas de importância foram propostas e são, usualmente, divididas em cinco categorias: dependência, consistência, informação, distância e

precisão de classificação. Algumas dessas medidas para avaliar atributos ou determinar em relação a que são considerados importantes, são apresentadas a seguir.

Definição 2 *(Importância Probabilística — Medida de Dependência) [8] Um atributo X_j é dito importante se e somente se existe algum x_{ij}, y e s_{ij} para o qual $P(X_j = x_{ij}, S_j = s_{ij}) > 0$ tal que*

$$P(Y = y|X_j = x_{ij}, S_j = s_{ij}) \neq P(Y = y|S_j = s_{ij})$$

De acordo com essa definição, X_j é importante se a probabilidade da classe, dados todos os atributos, pode mudar com a eliminação do conhecimento sobre o atributo X_j, onde s_{ij} é o subconjunto de atributos sem x_{ij}. Outra definição de importância que permite detectar a redundância de atributos considera o conceito de dimensão fractal.

Definição 3 *(Importância em relação à Dimensão Fractal — Medida de Dependência) [9] Dada a dimensão fractal, calculada usando todos os atributos do conjunto de dados, um atributo é dito importante se a sua exclusão causar uma alteração significativa*[3] *no valor da dimensão fractal recalculada sem a presença desse atributo.*

Alguns exemplos de aplicações que usam a teoria dos fractais incluem a determinação de estruturas de indexação de alta dimensionalidade e a detecção de agrupamentos (aprendizado não supervisionado), como proposto em [9]. Deve ser observado que a DF considera o atributo classe em igualdade de condições aos outros atributos. Porém, não é de nosso conhecimento que a teoria dos fractais tenha sido utilizada no problema de seleção de atributos para algoritmos de aprendizado supervisionado, como proposta neste trabalho.

3 O Algoritmo *Fractal Dimension-Based Filter*

Muitos dos algoritmos de SA que tratam tanto relevância quanto redundância de atributos, os avaliam por meio da escolha de subconjuntos desses atributos. Embora esses métodos, em geral, apresentem melhores resultados quando comparados aos métodos que não consideram o problema da redundância de atributos, o alto custo computacional pode torná-los ineficientes para conjuntos de dados com grande número de atributos. Recentemente, foi proposto o uso da abordagem filtro sob uma estrutura que separa as análises de relevância e redundância [2]. Essa proposta apresenta a vantagem de que por meio da desassociação dessas análises, o custo computacional pela busca por subconjuntos de atributos que aproximem o subconjunto ótimo pode ser diminuído. O algoritmo proposto neste trabalho, *Fractal Dimension-Based Filter* — FDimBF) —, é baseado nessa estrutura e utiliza o conceito de dimensão fractal para realizar a análise de redundância [7,10,11,12,13].

Freqüentemente, conjuntos de dados reais comportam-se como fractais estatisticamente auto-similares. Diversos são os exemplos que podem ser encontrados na natureza, como formações de nuvens, folhas e flores, topografias e cadeias de

[3] Dependente da variação que o usuário determinar como significativa.

montanhas, entre outros. Desse modo, torna-se natural a idéia de aplicar conceitos da teoria dos fractais para a análise desses conjuntos de dados [9]. A utilização do conceito de DF está associada à existência de redundância nos conjuntos de dados e da possibilidade desses conjuntos serem bem aproximados em dimensões menores. A idéia principal é empregar a DF do conjunto de dados, a qual é relativamente não influenciada por atributos redundantes, para determinar a quantidade e quais são os atributos não redundantes segundo o critério de DF. Existem diversas medidas para a DF. Para fractais estatisticamente auto-similares, como conjuntos de dados reais, uma das maneiras de definir a DF é dada pela Dimensão Fractal de Correlação D_2, que pode ser calculada pelo método *Box-Count Plot*. A idéia nesse método consiste, primeiramente, na construção de um reticulado sobre o conjunto de dados de células de lado r. Após, o número de pontos dentro da i-ésima célula de tamanho r, denominado $C_{r,i}$, é computado. A Dimensão Fractal de Correlação D_2, denominada neste trabalho simplesmente de dimensão fractal, é definida como $D_2 = \frac{\partial log(S_2(r))}{\partial log(r)}$, onde $r \in [r_{min}, r_{max}]$ e $S_2(r) = \sum_i {C_{r,i}}^2$.

Em teoria, fractais exatamente auto-similares são infinitos. Na prática, conjuntos de dados reais, os quais possuem um número finito de elementos, são considerados fractais estatisticamente auto-similares para um determinado intervalo de escala $r \in [r_{min}, r_{max}]$ se obedecem a uma regra de construção bem definida nesse intervalo. Desse modo, a dimensão intrínseca de um determinado conjunto de dados pode ser medida como o coeficiente angular da reta que melhor se ajusta ao trecho linear do gráfico em escala logarítmica de $S_2(r)$ por r [9].

É importante ressaltar que em [9] é proposta a utilização da DF para a determinação da redundância de atributos para problemas não supervisionados diferentemente da abordagem apresentada neste trabalho, no qual é proposta a utilização da DF em conjunto com outras medidas de relevância para a seleção de atributos em domínios supervisionados. Como mencionado, o algoritmo FDimBF proposto realiza a SA em duas etapas: (1) análise de relevância para determinar os atributos importantes em relação à classe e (2) análise de redundância para determinar e remover atributos numéricos redundantes a partir do subconjunto de atributos relevantes. No algoritmo FDimBF, descrito na Figura 1, a análise de relevância (linhas 3 – 7) pode ser realizada por meio de qualquer medida de importância MI que permita quantificar quão importante um atributo é, individualmente, para a caracterização da classe. Na etapa (2) (linhas 9 – 12), os atributos não redundantes de acordo com a dimensão fractal, ou seja, quando excluídos promovem uma modificação significativa no valor da DF recalculada, denominada de Dimensão Fractal Parcial pD, são selecionados a partir do subconjunto X' de atributos relevantes escolhidos durante a etapa (1). A busca por esses atributos é realizada de modo *backward*. Primeiramente, a DF D é computada para o conjunto de dados contendo todos os atributos relevantes X' do conjunto de dados (linha 11). Após, as pDs são calculadas ignorando-se um atributo por vez. Desse modo, o atributo removido X_j que apresentar a menor influência sobre o valor da DF, será excluído do subconjunto de atributos rele-

Require: $E = \{E_1, E_2, \ldots, E_N\}$, um conjunto de dados composto por N exemplos descritos por M atributos $X = \{X_1, X_2, \ldots, X_M\}$ e rotulados com os respectivos valores $y_i, i = 1...N$, $y_i \in \{C_1, C_2, ...C_{N_{Cl}}\}$ do atributo classe Y

Ensure: $X_{otimo} \subseteq X$, subconjunto "ótimo" de atributos relevantes e não redundantes

1: *// Análise de relevância utilizando a medida de importância MI*
2: $X' = \emptyset$;
3: **for all** $X_i \in X$ **do**
4: **if** X_i é relevante em relação à Y usando a medida de importância MI **then**
5: $X' = X' \cup \{X_i\}$;
6: **end if**
7: **end for**
8: *// $X' \subseteq X$, tal que X' contém os atributos relevantes do conjunto de exemplos E*
9: L = conjunto dos N exemplos E descritos apenas pelos atributos relevantes em X' segundo a medida de importância MI, *i.e.* sem o atributo classe Y;
10: *// Calcular a dimensão fractal D do conjunto L e encontrar o conjunto de atributos não redundantes X_{otimo}*
11: $D = DimensaoFractal(L)$;
12: $X_{otimo} = AtributosNaoRedundantes(L, X', D)$;
13: **Return** X_{otimo}.

Figura 1. Algoritmo *Fractal Dimension-Based Filter* — FDimBF

vantes. Esse processo é repetido para o novo subconjunto de atributos relevantes até que não existam mais atributos redundantes de acordo com a DF (linha 12). Uma descrição detalhada dessas duas funções pode ser encontrada em [7]. A complexidade do algoritmo FDimBF é $O(N.M^3)$. Ainda que maior (ou igual) que a complexidade de outros algoritmos de SA, a sua complexidade média é competitiva com a desses algoritmos.

Neste trabalho foram implementadas duas versões do algoritmo proposto utilizando duas medidas MI de importância diferentes para análise de relevância: FDimBF(1) que usa ganho de informação e FDimBF(2) que usa distância. Ambas consideram a medida de DF para a análise de redundância, a qual é realizada com o auxílio do algoritmo *Fractal Dimension Reduction* — FDR [9].

4 Avaliação Experimental

Ambas as versões de FDimBF foram avaliadas experimentalmente e comparadas com quatro importantes algoritmos de SA, considerando 11 conjuntos de dados provenientes do repositório de dados da UCI [14] (Tabela 1). Dois desses algoritmos realizam avaliação individual de atributos: ReliefF [15] e FCBF (*Fast Correlation-Based Filter*) [2] e os outros dois selecionam atributos importantes utilizando a avaliação de subconjuntos: CFS (*Correlation-Based Feature Selection*) [3] e CBF (*Consistency-Based Filter*) [5]. Todos os algoritmos foram executados usando os parâmetros configurados com valores padrão. À exceção de FDimBF, esses algoritmos estão disponíveis no ambiente WEKA [16].

Para dados reais, usualmente, o conhecimento *a priori* sobre que atributos são importantes não está disponível. Desse modo, a precisão predi-

tiva dos modelos construídos é comumente utilizada como uma medida indireta para avaliar a qualidade dos atributos selecionados. Neste trabalho, para cada um dos 11 conjuntos de dados, o conjunto original de atributos e os subconjuntos de atributos selecionados por cada um dos algoritmos de SA foram avaliados considerando a média do erro, estimada por meio de validação cruzada com 10 partições, dos modelos induzidos por $\mathcal{C}4.5$ [17] executado usando parâmetros com valores padrão. Os resultados experimentais, descritos em detalhes em [18], foram analisados sob diversos aspectos [7]. Por restrições de espaço, são apresentados apenas os resultados referentes ao modelo por nós proposto para avaliação da performance de algoritmos de SA, o qual considera as duas principais questões relacionadas a SA: (1) precisão preditiva das hipóteses construídas usando todos os atributos do conjunto original de dados e o subconjunto de atributos selecionados por cada algoritmo de SA e (2) tamanho do subconjunto de atributos selecionados em relação ao conjunto original. Nesse modelo de avaliação (Figura 2), os algoritmos de SA são posicionados em 5 categorias: excelente ($\triangle\triangle\triangle$), muito bom ($\triangle\triangle$), bom ($\triangle$), regular ($\Diamond$) e ruim ($\nabla$), onde ESA representa o erro da hipótese construída sem SA e ECM representa o erro da classe majoritária se $\leq 50\%$, caso contrário é considerado igual a 50%. A performance dos algoritmos, de acordo com esse modelo de avaliação, para os 11 conjuntos de dados é apresentada na Tabela 1. Nessa tabela são também apresentados, para cada conjunto de dados, o número de exemplos (#Ex.), de atributos (#At.) e o ECM. Exceto para Satimage, Segment, Vehicle e Waveform, os quais possuem, respectivamente, 7, 7, 4 e 3 classes, os outros conjuntos de dados apresentam apenas duas classes.

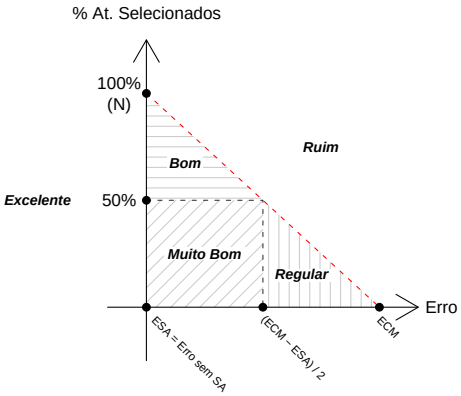

Figura 2. Modelo de avaliação de algoritmos de SA

Como pode ser observado, ambas as versões de FDimBF foram as que obtiveram o maior número total de performances excelentes e muito boas, 9 de um total de 11. CFS e CBF obtiveram, respectivamente, 8 e 7 classificações excelentes e muito boas, seguidos de FCBF com 4 performances desses tipos. Classificações regulares ocorreram de um modo uniforme entre os algoritmos considerados e apenas 3 deles apresentaram uma classificação ruim cada: ReliefF e as duas versões de FDimBF. Em relação ao número de atributos importantes selecionados, ReliefF escolheu todos os atributos do conjunto original em 8 dos 11 conjuntos de dados considerados. De fato, ambas as versões de FDimBF foram as únicas a sempre promover a redução do número de atributos para todos os conjuntos de dados — última linha da tabela. Sob o modelo de avaliação pro-

	#Ex.	#At.	ECM	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
Breast Cancer	683	9	34,48	—	—	—	△△△	△△△	△△△
Bupa	345	6	42,03	▽	◇	◇	◇	▽	△
German	1000	24	30,00	—	△△	△△△	△△△	△△△	△△△
Hungarian	261	10	36,05	—	△△△	△△△	△△△	△△	△△△
Ionosphere	34	351	35,90	△	△△	△	△△	△△	△△
Pima	769	8	34,98	—	△△	—	—	△△	▽
Satimage	4435	36	75,80	—	△△△	—	△△△	△△	△△
Segment	2310	19	85,70	△	△△	△	△△	△△	△△
Sonar	208	60	46,60	—	△△△	△△	△△	◇	△△
Vehicle	846	18	74,20	—	△	—	—	△△	△△
Waveform	5000	21	66,10	—	△△△	△△△	△	△△	△△
excellente (△△△)				0	4	3	4	2	3
muito bom (△△)				0	4	1	3	7	6
bom (△)				2	1	2	1	0	1
regular (◇)				0	1	1	1	1	0
ruim (▽)				1	0	0	0	1	1
todos at. selecionados (—)				8	1	4	2	0	0

Tabela 1. Conjuntos de dados e performance dos algoritmos de acordo com a percentagem de atributos selecionados *versus* erro da hipótese induzida

posto, dos 66 casos considerados (11 conjuntos de dados × 6 algoritmos de SA), 16 foram excelentes, 21 muito bons, 7 bons, 4 regulares e 3 ruins; 15 apresentaram o subconjunto de atributos selecionados igual ao conjunto original. Assim, $66,67\%$ dos casos foram considerados excelentes, muito bons ou bons, $22,73\%$ dos subconjuntos selecionados foram iguais aos conjuntos originais de atributos e apenas $10,61\%$ exibiram performances regulares ou ruins. Desse modo, a maioria dos algoritmos de SA contribui para a melhora, tanto em relação à quantidade de atributos selecionados quanto em relação à precisão das hipóteses induzidas, sob o modelo proposto para a avaliação de performance de algoritmos de SA.

5 Metodologia para Mapeamento de Laudos Médicos

Como mencionado, um problema freqüentemente encontrado durante a realização de processos para a extração de conhecimento de bases de dados reais, é a disponibilização desses dados, especialmente na área médica, no formato de laudos semi-estruturados descritos em linguagem natural. Neste trabalho foi realizado um estudo de caso, utilizando os algoritmos citados anteriormente, que faz parte dos projetos de Computação Aplicada à Medicina e Análise Inteligente de Dados desenvolvidos em uma parceria entre o Laboratório de Bioinformática — LABI — Universidade Estadual do Oeste do Paraná, UNIOESTE; o Laboratório de Inteligência Computacional — LABIC — Universidade de São Paulo, USP/São Carlos; o Serviço de Coloproctologia da Faculdade de Ciências Médicas — FCM — Universidade Estadual de Campinas, Unicamp e o Centro de Referência em Infertilidade Masculina — Androfert.

A metodologia proposta tem como objetivo dar suporte à construção de bases de dados estruturadas a partir de laudos médicos semi-estruturados descritos em linguagem natural [7,19]. Essa metodologia pode ser caracterizada, de um modo genérico, em duas fases: (1) identificação de padrões e construção de um dicionário e (2) mapeamento dos laudos para a base de dados. O objetivo da

fase (1) é construir um dicionário, com o auxílio de especialistas do domínio, a partir da identificação de padrões que ocorrem nos laudos. Esse dicionário é então utilizado na fase (2) para mapear os laudos médicos, por meio de casamento de padrões, para conjuntos de dados no formato atributo-valor. Deve ser observado que após o dicionário ser construído na primeira fase com o conjunto de dados disponíveis, esse dicionário pode ser armazenado e utilizado posteriormente para mapear automaticamente novos laudos, isto é, sem a necessidade de construí-lo novamente. A metodologia proposta neste trabalho foi aplicada com sucesso a uma coleção de laudos médicos relacionados à análise seminal [7]. A base de dados estruturada foi utilizada para a extração de conhecimento.

6 Conclusão

O trabalho realizado apresentou as seguintes três principais contribuições:
(1) Algoritmo FDimBF: resultados experimentais mostraram que o algoritmo proposto é comparável a outros freqüentemente utilizados para a SA, sendo capaz de selecionar subconjuntos de atributos importantes de tamanhos reduzidos que permitem a indução de hipóteses com precisão similar a algoritmos que tem apresentado boa performance tal como CFS [3]. Desse modo, consideramos que a DF é uma boa candidata para realizar a SA para algoritmos de aprendizado supervisionado, para os quais não é de nosso conhecimento que tenha sido assim utilizada. Uma limitação associada ao algoritmo proposto é a restrição ao tratamento de atributos numéricos redundantes, embora outros algoritmos apresentados na literatura, como o próprio CFS, também apresentem algum tipo de restrição, tal como a necessidade de discretização prévia dos atributos;
(2) Modelo para avaliação de performance de algoritmos de SA: permitiu avaliar os algoritmos de SA considerando duas das mais importantes questões relacionadas à SA, provendo uma visão genérica da performance desses algoritmos. Uma limitação do modelo proposto é a classificação da performance desses algoritmos em cinco classes discretas. Estamos pesquisando outros modelos multi-critério que permitam atribuir um índice único que classifica a performance desses algoritmos considerando mais de duas medidas, e
(3) Metodologia para mapeamento de laudos para bases de dados estruturadas: permitiu o mapeamento automático e padronizado de informações contidas em laudos médicos para formatos adequados para a aplicação de métodos para a extração de conhecimento. A metodologia proposta é altamente dependente do domínio, *i.e.*, a cada novo domínio o dicionário contendo os padrões a serem mapeados deve ser reconstruído, porém, uma vez realizada essa tarefa, novos laudos desse domínio podem ser mapeados automaticamente para a base de dados.

Os resultados mais importantes deste trabalho de doutorado foram publicados em [7,10,11,12,13,18,19,20]. Trabalhos futuros incluem a avaliação dos atributos selecionados do ponto de vista de especialistas do domínio, assim como a utilização de curvas ROC, ao invés do erro da hipótese induzida, dentro do modelo para avaliar a performance de algoritmos de SA.

Referências

1. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers (1998)
2. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundance. Journal of Machine Learning Research **5** (2004) 1205–1224
3. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proc. of the 17th Int. Conf. on Machine Learning, USA, Morgan Kaufmann (2000) 359–366
4. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proc. of the 13th Int. Conf. on Machine Learning, Italy (1996) 284–292
5. Liu, H., Setiono, R.: A probabilistic approach to feature selection – a filter solution. In: Proc.of the 13th Int. Conf. on Machine Learning, Italy (1996) 319–327
6. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1-2) (1997) 273–324
7. Lee, H.D.: Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, ICMC-USP. `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/` (2005)
8. John, G., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Proc. of the 11th Int. Conf. on Machine Learning, USA, Morgan Kaufmann (1994) 167–173
9. Traina, C., Sousa, E.P.M., Traina, A.J.M.: 24. In: Using Fractals in Data Mining. Volume 1 of 1. Wiley-IEEE Press (2005) 599–630
10. Lee, H.D., Monard, M.C., Wu, F.C.: A fractal dimension based filter algorithm to select features for supervised learning. In: Proc.of the IBERAMIA/SBIA 2006, Lecture Notes in Computer Science, Ribeirão Preto, SP (2006) (in print)
11. Lee, H.D., Monard, M.C., Wu, F.C.: Feature subset selection for supervised learning using fractal dimension. In: Frontiers in Artificial Intelligence and Applications. Volume 132., Japan, IOS Press (2005) 135–142
12. Lee, H.D., Monard, M.C., Wu, F.C.: Seleção de atributos relevantes e não redundantes usando a dimensão fractal do conjunto de dados. In: Anais do V Encontro Nacional de Inteligência Artificial, XXV CSBC, Brasil (2005) 444–453
13. Lee, H.D., Monard, M.C.: Seleção de atributos para algoritmos de aprendizado de máquina supervisionado utilizando como filtro a dimensão fractal. Revista Soc. Chilena de Ciencia de La Computación **4**(1) (2003) 1–8
14. Merz, C.J., Murphy, P.M.: UCI repository of machine learning datasets (1998)
15. Kira, K., Rendell, L.: A pratical approach to feature selection. In: Proc. of the 9th Int. Conf. on Machine Learning, UK, Morgan Kaufmann (1992) 249–256
16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (2000)
17. Quinlan, J.R.: $\mathcal{C}4.5$: Programs for Machine Learning. Morgan Kaufmann (1988)
18. Lee, H.D., Monard, M.C., Voltolini, R.F., Wu, F.C.: Avaliação experimental e comparação de algoritmos de seleção de atributos importantes com o algoritmo FDimBF baseado na dimensão fractal. Technical Report 264, ICMC-USP (2005)
19. Honorato, D.D.F., Lee, H.D., Monard, M.C., Wu, F.C., Machado, R.B., Neto, A.P., Ferrero, C.A.: Uma metodologia para auxiliar no processo de construção de bases de dados. In: Anais do V Encontro Nacional de Inteligência, XXV CSBC, Brasil (2005) 593–601
20. Lee, H.D., Monard, M.C., Voltolini, R., Prati, R., Wu, F.C.: A simple evaluation model for feature subset selection algorithms. In: Proc.of the Argentine Symposium on Artificial Intelligence 2006, Mendoza, Argentina (2006) (in print)