Uma nova metodologia para seleção de atributos no processo de extração de conhecimento de base de dados baseada na teoria de *rough sets*

ANDERSON TERUYA

Dissertação apresentada ao Departamento de Engenharia Elétrica (DEL) da Universidade Federal de Mato Grosso do Sul (UFMS) como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Campo Grande / MS (Brasil)
2008

Uma nova metodologia para seleção de atributos no processo de extração de conhecimento de base de dados baseada na teoria de *rough sets*

ANDERSON TERUYA

Dissertação apresentada ao Departamento de Engenharia Elétrica (DEL) da Universidade Federal de Mato Grosso do Sul (UFMS) como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Linha de Pesquisa:
Inteligência Computacional – Teoria e Aplicações em Sistemas de Energia

Orientador:
Profº Dr. João Onofre Pereira Pinto

Campo Grande / MS (Brasil)
2008

Só sei que nada sei.

(Sócrates)

Nunca repreenda uma pessoa vaidosa; ela o odiará por isso.

Mas, se você corrigir uma pessoa sábia, ela o respeitará.

Qualquer coisa que você ensina a uma pessoa sábia torna-a mais sábia ainda.

(Provérbios 9,8-9)

A meus pais Nelson e Anita,
tio Oswaldo e tia Edite,
pela educação.

A minha esposa Laura Helena,
pelo amor, dedicação, paciência e incentivo.

A avó Kamé (in memorian),
pelo exemplo de vida.

A meus irmãos, primos, tios, professores e amigos,
pela companhia e incentivo.

Agradecimentos

Em primeiro lugar, a Deus por tudo.

À minha família por toda formação, apoio e incentivo, de onde destaco: minha esposa e fonte de inspiração, a quem digo constantemente que a amo e que vou envelhecer ao seu lado e cuidar sempre de você, e que estes quase dois anos, foram e está sendo uma experiência maravilhosa viver ao seu lado, mas temos muito ainda para crescer e sermos cada vez melhores; meus pais pela vida e educação, e meus tios Oswaldo e Edite pela educação, pois toda minha formação sempre foi, é e será sustentada nessas quatro pessoas. Em oração peço a Deus que abençoe a todos, e nos façam cada vez melhores. Amém.

Ao Amigo e Professor João Onofre Pereira Pinto, que é um dos responsáveis pela conclusão desta Dissertação de Mestrado. Esse homem é fantástico, considero uma referência para mim, não só pelo intelecto, mas pela pessoa extremamente simples e eclética. "Valeu Professor João, Obrigado!", e que Deus abençoe você e sua família.

Aos professores que participaram da banca examinadora, tanto da qualificação como da defesa desta Dissertação de Mestrado, dos quais relaciono: Profª Dra. Valguima Victoria Viana Aguiar Odakura, Profº Dr. José Demísio Simões da Silva, Profº Dr. Milton Ernesto Romero Romero e Profº Dr. João Onofre Pereira Pinto, registro os meus agradecimentos especiais pelas sugestões e críticas a este trabalho, que foram de extrema importância para sua conclusão.

Aos professores do Departamento de Engenharia Elétrica da Universidade Federal de Mato Grosso do Sul, quero externar os meus agradecimentos por fazerem parte da conclusão desta fase de minha vida; e em momento anterior por contribuírem pela minha atual profissão de Engenheiro Eletricista. Desses destaco novamente o Profº João Onofre, Profº Flávio, Profº Jéferson e o Profº Paulo Koltermman.

À COPEL, meu atual local de trabalho, que sempre permitiu e incentivou o desenvolvimento e conclusão desta Dissertação de Mestrado. Deste convívio destaco o Engº Teófilo, pela amizade e incentivo; a Engª Ana Rita, pela compreensão e incentivo; o Engº Ambrósio, pela compreensão e incentivo.

À ENERSUL, onde tudo começou. Lá foi o laboratório de onde surgiu a idéia para o tema desta Dissertação de Mestrado, e onde o contato com o Profº João Onofre foi fortalecido com o projeto de P&D "Obtenção de Curvas de Demanda para Consumidores

de Baixa Tensão Segundo Padrão de Consumo Utilizando Técnicas de Inteligência Artificial". Deste convívio destaco o Engº Sérgio Cerchiari, pela amizade e incentivo;

Aos amigos, são tantos, gostaria de agradecer pela companhia, apoio e compreensão.

Resumo

Nesta dissertação de Mestrado, propõe-se uma nova metodologia de Seleção de Subconjuntos de Atributos, a ser utilizada no processo de extração de conhecimento de base de dados.

As bases de dados, dimensionadas para diversos fins, possuem em sua essência, o conhecimento intrínseco ao sistema de sua aplicação. Esse conhecimento é muito valioso e importante para tomadas de decisões estratégicas nesse sistema. Assim, a proposta da Inteligência Artificial, através da subárea Mineração de Dados, é extrair esse conhecimento de bases de dados de forma automática. Com isso, introduziu-se o conceito de KDD, que implica em um processo de extração de conhecimento de base de dados. Uma das etapas do KDD é a Seleção de Subconjuntos de Atributos (SSA) que tem por objetivo analisar uma base de dados e eliminar atributos não importantes para o conhecimento a ser extraído, assim reduzindo o volume de dados a ser analisado, sem que haja alterações significativas no seu conteúdo.

Então, analisando as metodologias de SSA existentes, em especial, Redutos na Teoria de *Rough Sets*, FOCUS e FOCUS-2, verificou-se que em Redutos selecionam-se atributos condicionais sem considerar o atributo de decisão que é o objeto do conhecimento a ser extraído. E na FOCUS e FOCUS-2, que aplica conceitos semelhantes à metodologia Redutos, implicando em análise de todas as combinações de exemplos (dois a dois), verifica-se que a aplicação ocorre para pares de exemplos pertencentes a classes diferentes, dessa forma considerando o atributo de decisão.

A partir dessa análise, elaborou-se a metodologia proposta neste trabalho, que utiliza os conceitos introduzidos na Teoria de *Rough Sets*, com um diferencial na composição da Matriz de Discernimento. Esse diferencial considera o atributo de decisão na composição dessa matriz, como em FOCUS e FOCUS-2, indo mais além, por prover um tratamento diferenciado para exemplos pertencentes a mesma classe. Pois, criou-se a hipótese de um subconjunto de atributos, apontado por essa metodologia de SSA, conseguir distinguir todos os exemplos pertencentes a classes diferentes e não conseguir concluir que um exemplo pertence a mesma classe de outro exemplo, por ter todos os seus atributos condicionais diferentes entre si.

Para viabilizar a implementação da proposta foi necessário introduzir uma simplificação nas matrizes de operação, pois suas dimensões, por definição, são muito grandes. Com isso, concluiu-se a sua implementação, e na seqüência a avaliação.

Os resultados das avaliações, no geral, foram satisfatórios, com exceção de alguns pontos que são expostos e discutidos nos capítulos 7 e 8 deste trabalho.

Abstract

In this dissertation, a new Feature Selection Subsets methodology is proposed, to be used in the Knowledge Discover in Database process.

The databases, dimensioned for specific purposes, own in its essence, the intrinsic knowledge to the system of its application. This knowledge is very valuable and important to take strategical decisions in this system. Thus, the Artificial Intelligence's proposal, through of the Data Mining, is to extract this knowledge of databases with automatic form. With this, the KDD concept was introduced, that implies in a knowledge extraction's database process. One of the stages of the KDD is the Feature Selection Subsets (FSS) that it has for objective to analyze a database and to eliminate attributes not important for knowledge to be extracted, thus reducing the data's volume to be analyzed, without it has significant alterations in its content.

Then, analyzing the existing methodologies of FSS, in special, Reducts in the Theory of Rough Sets, FOCUS and FOCUS-2, were verified that in Reducts selects conditional attributes without considering the decision attribute, that it is the object of the knowledge to be extracted. In FOCUS and FOCUS-2, that applies similar concepts to the Reducts methodology, implying in analysis of all combinations of examples (two by two), verifies that the application occurs to pairs of examples belonging to the different classrooms, of this form considering the decision attribute.

From this analysis, it was elaborated the methodology proposal in this work, that uses the concepts introduced in the Theory of Rough Sets, with a differential in the Discernibility Matrix's composition. This differential considers the attribute decision in the composition of this matrix, as in FOCUS and FOCUS-2, and additionally, providing a differentiated treatment to examples belonging to the same classroom. Well, a hypothesis was created that implies in an attributes subset pointed by a FSS, to obtain to distinguish all examples belonging the different classrooms and not to obtain to conclude that an example belongs the same classroom of another example, for having all its different conditional attributes between itself.

To make possible the implementation of the proposal, it was necessary to introduce a simplification in the operation matrices, therefore its dimensions, for definition, are very great. With this, it was concluded its implementation, and in the sequence, the evaluation.

The evaluations results, in the generality, had been satisfactory, with exception of some points that are displayed and argued in chapters 7 and 8 of this work.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM.....Aprendizado de Máquina (*machine learning* - ML)

BD..... Base de Dados (database - DB)

CBF.....Consistency-Based Filter (Algoritmo Filtro de SSA baseado na Consistência)

CFS ..... Correlation-based Feature Selection (Algoritmo de SSA baseado na Correlação)

IA .....Inteligência Artificial (artificial inteligent - AI)

IC ..... Indução Construtiva (*constructive induction* - CI)

KDD ..... Knowledge Discovery in Databases (descoberta de conhecimento em BDs)

MD.....Mineração de Dados (data mining - DM)

TRS..... Teoria de *Rough Sets*

SA ..... Seleção de Atributos (feature selection - FS)

SD ..... Sistema de Decisão

SI.....Sistema de Informação

SSA.....Seleção de Subconjunto de Atributos (feature subsets selection - FSS)

SGBD .....Sistema de Gerenciamento de Base Dados

Capítulo 1

Introdução

1.1 Contextualização

Com o progresso da tecnologia de coleta e armazenamento de dados, que implica em maiores capacidades e menores custos, verifica-se que Bases de Dados (BDs), dimensionadas para fins específicos, tornaram-se práticas comuns no dia-a-dia. Devido a esse fato, a quantidade de dados armazenados vem crescendo a uma velocidade muito alta. Segundo estimativas de Lyman *et al* [2003], no período de 1999 a 2002, os dados armazenados em impressos, filmes, mídias magnéticas e ópticas cresceram 30% ao ano, ou seja, em 3 anos eles dobraram; e 92% desses dados estavam contidos em mídias magnéticas, sendo a maior parte em discos rígidos.

Além dos seus propósitos funcionais, as BDs acumulam o conhecimento intrínseco ao sistema de sua aplicação, que depende fortemente da qualidade dos dados armazenados[1]. Esse conhecimento pode ser: inovador perante a "expertise" de um especialista; considerado extremamente estratégico para tomadas de decisões; utilizado para comprovar hipóteses; e agregado em sistemas automáticos de tomadas de decisões [Frawley *et al*, 1992; Fayyad *et al*, 1996; Fayyad *et al*, 1996b; Mitra, 2002; Lee, 2005; Rezende, 2005]. Em Romão [2002] é citado um exemplo de distinção entre informação e conhecimento, que consiste na suposição de uma BD contendo registros de clientes e mercadorias vendidas, então a partir dela pode-se obter:

- informação: "quantidade de computadores que foram vendidos para o cliente X na data dd/mm/aaaa". Pode ser extraída diretamente no Sistema de Gerenciamento de BD (SGBD) através de uma simples consulta;
- conhecimento: "*SE* (idade = '[25 a 35] anos') *E* (profissão = 'advogado') *ENTÃO* (compra = 'computador') com uma freqüência de 90%". É obtido através de análises de BDs e pode ser utilizado: pelo marketing (mala direta direcionada), para planejamento de estoque, para abertura de novas filiais e outras decisões estratégicas.

[1] Implica em dados confiáveis, completos e correspondentes ao conhecimento a ser extraído.

A seguir são apresentadas outras aplicações de extração de conhecimento de BDs praticadas atualmente, e são: combate a fraudes (aprovação de compras no cartão de crédito, liberação de empréstimos, venda de seguros, etc.); controle de manufatura (detecção de problemas em produtos, etc.); na medicina (em predição de diagnósticos, antecipação de tratamentos, etc.), em finanças no auxílio a investimentos (detecção de padrões nos mercados financeiros, etc.) entre outras [Dias, 2002; Rezende, 2005].

Esse conhecimento raramente é obtido de forma direta, então para obtê-lo faz-se necessário o processamento dos dados, que pode ser realizado de duas maneiras [Lee, 2000; Mitra *et al*, 2002]:

- manualmente: implica em ter disponível um analista para buscar, tabelar e consolidar informações obtidas através de: relatórios, técnicas estatísticas, questionários, debates em grupo, entrevistas com especialistas e outras abordagens. Para análise de grande quantidade de dados esse procedimento se torna inviável;
- automaticamente: implica em ter uma BD que contenha informações relacionadas com o conhecimento desejado. Então com o auxílio de uma ferramenta computacional para extração de conhecimento de BDs, processa-se essa BD e no término do processo o conhecimento da BD estará disponível.

1.2 Revisão Bibliográfica

Diante dos fatos expostos, que consistem no aumento acentuado do volume de dados, associado à crescente demanda por conhecimento novo para decisões estratégicas, observou-se o interesse crescente em descobrir conhecimento em BDs, a partir das informações contidas nesses dados [Romão, 2002]. Nessa linha de raciocínio destacam-se: KDD (*Knowledge Discovery in Databases*, ou seja, Descoberta de Conhecimento em BDs) proposto em 1989, MD (Mineração de Dados – *data mining*) proposta na década de 60 e AM (Aprendizado de Máquina – *machine learning*) proposto em 1950 [Fayyad *et al*, 1996; Gammerman, 1997]. Essas áreas estão fortemente correlacionadas entre si. Por KDD entende-se como um processo para extração de conhecimento em BDs; MD consiste em uma metodologia utilizada para "minerar" dados de uma BD, dentro do processo de KDD, buscando o seu conhecimento intrínseco; e AM como uma disciplina fornecedora de algoritmos para MD [Fayyad *et al*, 1996]. Muitos trabalhos na área de KDD são baseados em AM devido à boa performance obtida [Pérez, 1996].

Segundo Mitchell [1997], AM define-se como uma sub-área da Inteligência Artificial (IA) que pesquisa métodos computacionais relacionados à aquisição automática de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente. Desta forma, um sistema de AM consiste em um algoritmo que toma decisões, baseado em experiências acumuladas, contidas em exemplos resolvidos com sucesso. Esse aprendizado pode ser não-supervisionado cujo objetivo é buscar padrões em exemplos para agrupá-los por semelhança; ou supervisionado cujo objetivo é extrair padrões de exemplos rotulados para induzir um classificador com a capacidade de rotular novos exemplos. É importante ressaltar que os métodos de aprendizados nem sempre se encaixam completamente em uma destas classificações [Santoro, 2005].

Um sistema de AM supervisionado pode ser descrito como: dado um conjunto de m exemplos, compostos por n atributos $X = (X_1, X_2, ..., X_n)$ e cada exemplo associado a uma classe y_i onde $1 \leq i \leq m$, a tarefa é encontrar um mapeamento f, tal que $Y = f(X)$. Em AM supervisionado a tarefa de aprendizado é denominada regressão se os rótulos são contínuos; ou classificação se os rótulos são categóricos (valores discretos) [Pila, 2001]. Na Figura 1.1 é apresentado um sistema de AM supervisionado para classificação [Pila, 2001], onde inicialmente formulam-se as hipóteses (conhecimento extraído) a partir dos exemplos rotulados, que serão utilizadas para induzir o classificador (linha pontilhada); e depois com o classificador passa-se a rotular os novos exemplos (linha contínua).

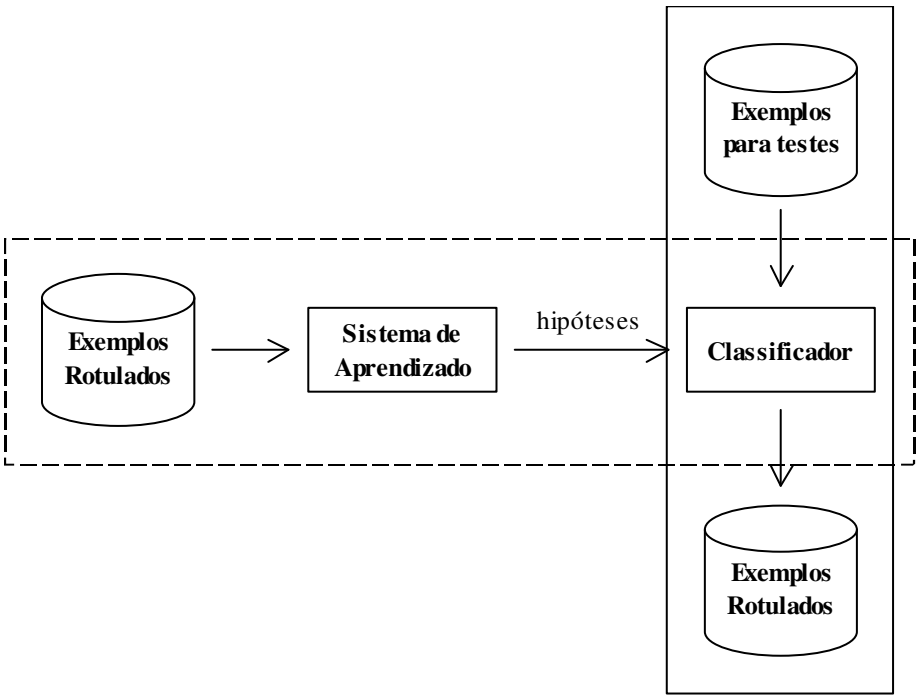


Figura 1.1: Sistema de AM para Classificação [Pila. 2001]

Conforme a definição clássica de Fayyad *et al* [1996], KDD é o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados. Foi proposto a partir da MD, e tem como objetivo, extrair informações estratégicas escondidas em grandes BDs, por meio da pesquisa dessas informações e da determinação de padrões, classificações e associações entre elas [Goebel & Gruenwald, 1999; Dias, 2002]. O grande desafio de KDD é processar automaticamente grandes quantidades de dados brutos (*raw data*), identificando os padrões significantes e apresentá-los como conhecimento apropriado para o usuário [Matheus *et al*, 1993; Pérez, 1996].

Na Figura 1.2 é apresentada a diagramação do Processo KDD.

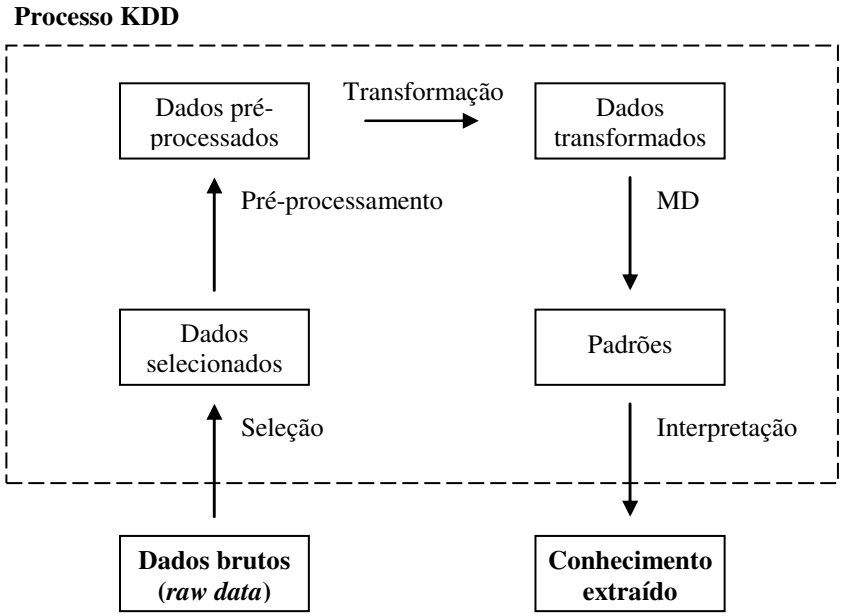


Figura 1.2: Diagramação do Processo KDD

Segundo Fayyad *et al* [1996] o processo KDD pode ser dividido em cinco etapas, sendo elas:

- Seleção de dados;
- Pré-processamento dos dados selecionados;
- Transformação dos dados pré-processados;
- Mineração de Dados (MD);
- Interpretação dos padrões encontrados.

A primeira etapa, Seleção de dados, consiste em analisar os dados disponíveis e selecionar os que serão utilizados para a extração do conhecimento.

A segunda etapa, Pré-processamento dos dados selecionados, consiste em identificar e dar um tratamento adequado aos dados faltantes, inválidos, inconsistentes e redundantes, sem que haja alterações significativas no conteúdo da BD.

A terceira etapa, Transformação dos dados pré-processados, consiste em simplificar os dados para a próxima etapa, sem que haja alterações significativas no conteúdo da BD, buscando reduzir o custo do processamento e melhorar a qualidade do conhecimento extraído [Pappa, 2002; Oliveira, 2006]. Esta simplificação pode ser feita de três maneiras: reduzindo o número de exemplos, de atributos e de valores de um atributo [Weiss & Indurkhya, 1998]. Alguns autores agrupam esta etapa e a anterior, e definem como fase de pré-processamento de dados do KDD; nessa linha, Pyle [1999] e Batista [2003] expõem que o processamento destas duas etapas consome cerca de 80% do tempo do processo de KDD.

A quarta etapa, Mineração de Dados (MD), consiste no núcleo do KDD, ou seja, é aqui que é executada a principal função do processo que consiste em analisar os dados e extrair os padrões dos dados transformados. Devido sua natureza interdisciplinar, MD tem recebido contribuições de diversas áreas como: Banco de Dados, AM, Estatística, Visualização de Dados, Recuperação de Informação, Computação Paralela e Distribuída, sendo as três primeiras citadas as que mais vêem contribuindo [Fayyad *et al*, 1996b; Zhou, 2003; Two Crows, 2005].

A quinta etapa, Interpretação dos padrões encontrados, tem como objetivo avaliar, validar os padrões encontrados, assim consolidando o conhecimento extraído; pois a MD pode gerar muitos padrões irrelevantes, redundantes [Piatetsky-Shapiro *et al*, 1994], contraditórios com o conhecimento prévio, e extremamente complexos e incompreensíveis que torna o resultado improdutivo.

Inicialmente, os esforços de pesquisadores da área de KDD estavam concentrados na fase de MD, buscando modelos precisos para a extração de padrões. Mas recentemente, essa comunidade tem-se voltado com maior atenção para as etapas anteriores e posteriores à MD [Baranauskas *et al*, 2000; Lee, 2005].

Para delinear o tema desta dissertação, destaca-se a terceira etapa do KDD (Transformação de Dados) com enfoque na redução do número de atributos de uma BD. Essa transformação consiste em buscar um Subconjunto de Atributos Ótimo da BD em análise. Esse subconjunto é formado a partir do conjunto original de atributos condicionais da BD, e pode ter várias definições distintas, das quais se apresenta duas a seguir:

- Segundo Koller & Sahami [1996], Yu & Liu [2004] e Lee [2005], um Subconjunto de Atributos Ótimo deve possuir o menor número de atributos possíveis e manter a capacidade de representar o conhecimento intrínseco da BD, considerando um erro admitido pelo usuário. Em outras palavras, encontrar um Subconjunto de Atributos Ótimo X' com n' elementos, a partir do conjunto de todos os atributos condicionais X com n elementos, tal que as distribuições de probabilidades $p(D \mid X = \vec{x})$ e $p(D \mid X' = \vec{x}')$ sejam aproximadas, onde D é o domínio dos valores de atributos de decisão, $\vec{x}$ e $\vec{x}'$ são instâncias da BD descritas em X e X', respectivamente. Em suma, têm-se: $X' \subseteq X$, $n' \leq n$, e $p(D \mid X = \vec{x}) \approx p(D \mid X' = \vec{x}')$;

- Segundo Kohavi & John [1997] um Subconjunto de Atributos Ótimo, ao ser submetido à etapa de MD, deve gerar hipóteses para induzir um classificador que proporcione a máxima precisão na classificação de novos exemplos.

Independente do conceito adotado, segundo Kohavi & John [1997] um subconjunto de atributos que atende a condição de ótimo, não é necessariamente único para uma BD, pois podem existir outros subconjuntos de atributos que atendem a mesma condição.

Uma forma de se obter a transformação descrita anteriormente é através da aplicação de metodologias de Seleção de Atributos (SA) em BDs. E como benefícios verificam-se: eliminação do custo de coleta e armazenagem de atributos não importantes; redução do volume de dados, melhorando a performance de processamento; e melhoria na compreensibilidade do conceito induzido, e por conseqüência, na precisão dos classificadores gerados.

Segundo Lee [2000], a SA pode ser realizada de duas maneiras:

- eliminando atributos não importantes, através da Seleção de Subconjunto de Atributos (SSA), que será o tema abordado e desenvolvido nesta dissertação;

- gerando um novo atributo através do correlacionamento de dois ou mais atributos, com isso todos estes atributos passarão a ser redundantes em relação ao atributo gerado e poderão ser eliminados. Esse processo é conhecido como Indução Construtiva (IC).

1.3 Objetivo

O objetivo principal desta dissertação é introduzir, discutir, aplicar e avaliar uma proposta de metodologia de SSA, que buscará apontar todos os Subconjuntos de Atributos Ótimos possíveis de uma BD composta de valores discretos.

A partir dessa premissa, agregou-se aos objetivos desta dissertação:

- Pesquisar métodos de SSA existentes para fundamentar, buscar semelhanças e promover a competição com a nova Metodologia de SSA Proposta e reportar os desempenhos obtidos;
- Aplicar a Metodologia de SSA Proposta em vários casos de referência[2] e registrar todos os subconjuntos de atributos encontrados para futuras referências.

1.4 Organização do Trabalho

Esta dissertação está organizada em oito capítulos, incluindo o presente. Neste trabalho serão apresentados: a fundamentação teórica necessária para o seu desenvolvimento, a descrição e aplicação da nova Metodologia de SSA Proposta com o máximo de detalhamento em cada etapa, os subconjuntos de atributos obtidos para casos de referência disponíveis, e por fim, as conclusões e recomendações para futuros trabalhos.

Em suma a organização deste trabalho apresenta-se da seguinte maneira:

- Capítulo 2: Terminologia – Neste capítulo apresentar-se-ão as considerações sobre a notação empregada neste trabalho, abrangendo: o formato do conjunto de dados a ser analisado (insumo do processo) e o formato do conhecimento extraído (resultado do processo);
- Capítulo 3: Seleção de Subconjuntos de Atributos – Neste capítulo apresentar-se-á a compilação de várias referências sobre o tema SSA, abrangendo: conceitos e definições, objetivos, a finalidade de sua aplicação, a importância dos atributos condicionais em relação ao conhecimento a ser extraído, os

[2] BDs públicas utilizadas pela comunidade científica para avaliação e comparação de metodologias de SSA.

princípios e mecanismos existentes para o seu funcionamento e as formas de se avaliar a aplicação desta ferramenta;

- Capítulo 4: Métodos para Avaliação de SSA – Neste capítulo apresentar-se-ão as metodologias para avaliação de SSA pesquisadas, que serão utilizadas para comparar a nova metodologia proposta com outras existentes;
- Capítulo 5: Metodologias de SSA Pesquisadas – Neste capítulo apresentar-se-ão as metodologias de SSA pesquisadas, que serão utilizadas para fundamentar, assemelhar e competir com a nova metodologia proposta;
- Capítulo 6: Metodologia de SSA Proposta – Neste capítulo apresentar-se-á a nova Metodologia de SSA Proposta neste trabalho e a sua aplicação em uma BD-exemplo de forma detalhada;
- Capítulo 7: Avaliação da Metodologia de SSA Proposta – Neste capítulo apresentar-se-ão todos os subconjuntos de atributos mínimos para os casos de referência, obtidos com a aplicação da metodologia proposta neste trabalho;
- Capítulo 8: Conclusão e Sugestões – Neste capítulo apresentar-se-á a conclusão deste trabalho e sugestões para trabalhos futuros.

Capítulo 2

Terminologia

2.1 Considerações Iniciais

Neste capítulo apresentam-se as considerações sobre a notação adotada neste trabalho, aplicáveis sobre as BDs a serem analisadas (insumo do KDD) e sobre o conhecimento extraído (resultado do KDD).

2.2 Notação das BDs a serem analisadas

Quanto à notação das BDs a serem analisadas e apresentadas na entrada do KDD, admitem-se as seguintes considerações:

- Para este trabalho, a BD a ser analisada deverá ser composta por valores discretos, assim implicando em um sistema de AM Supervisionado para Classificação;
- A notação adotada da BD a ser analisada é apresentada na Tabela 2.1, e é constituída por m exemplos (registros ou instâncias) $E = \{E_1, E_2, ..., E_m\}$, por n atributos condicionais $X = \{X_1, X_2, ..., X_n\}$ e por um atributo de decisão Y. O atributo de decisão consiste em um atributo selecionado da BD a ser analisada, que será o objeto do conhecimento extraído;
- Cada exemplo é descrito como um par $E_i = < \vec{x}_i, y_i >$ onde: $1 \leq i \leq m$, $\vec{x}_i \in C_1 \times C_2 \times .... \times C_n$ e $y_i \in D$, sendo C_j $(1 \leq j \leq n)$ o domínio de valores do atributo condicional X_j e D o domínio de valores do atributo de decisão Y;
- O vetor $\vec{x}_i = <x_{i1}, x_{i2}, ..., x_{in}>$ descrito por elementos x_{ij} $(1 \leq i \leq m$ e $1 \leq j \leq n)$ representa o valor do atributo condicional X_j para o exemplo E_i;
- A tarefa de um algoritmo indutivo de AM supervisionado é induzir uma estrutura (por exemplo árvore de decisão) tal que, dado um novo exemplo, esta tenha capacidade de prever a classe Y do exemplo dado.

Na seqüência apresenta-se um exemplo de BD na Tabela 2.2, a clássica BD-exemplo – "Jogar Golfe".

Exemplos	Atributos				
	condicionais				decisão
E	X_1	X_2	⋯	X_n	Y
E_1	x_{11}	x_{12}	⋯	x_{1n}	y_1
E_2	x_{21}	x_{22}	⋯	x_{2n}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
E_m	x_{m1}	x_{m2}	⋯	x_{mn}	y_m

Tabela 2.1: Formato padrão da BD a ser analisada pelo KDD

Exemplos	Atributos				
	condicionais				decisão
E	Previsão	Temperatura	Umidade	Vento	Y
E_1	chuvoso	ameno	alta	falso	joga
E_2	chuvoso	ameno	normal	falso	joga
E_3	chuvoso	frio	normal	falso	joga
E_4	ensolarado	ameno	normal	verdadeiro	joga
E_5	ensolarado	frio	normal	falso	joga
E_6	nublado	ameno	alta	verdadeiro	joga
E_7	nublado	frio	normal	verdadeiro	joga
E_8	nublado	quente	alta	falso	joga
E_9	nublado	quente	normal	falso	joga
E_{10}	chuvoso	ameno	alta	verdadeiro	não joga
E_{11}	chuvoso	frio	normal	verdadeiro	não joga
E_{12}	ensolarado	ameno	alta	falso	não joga
E_{13}	ensolarado	quente	alta	falso	não joga
E_{14}	ensolarado	quente	alta	verdadeiro	não joga

Tabela 2.2: A clássica BD-exemplo – Jogar Golfe [Hall, 1999]

2.3 Notação do conhecimento extraído

Quanto à notação do conhecimento extraído e coletado na saída do KDD, admitem-se as seguintes considerações [Frawley *et al*, 1992]:

- Dado um conjunto de exemplos E, uma linguagem L e uma medida de certeza c, define-se um padrão como uma indicação S em L, que descreve relacionamentos entre um subconjunto E_s de E com uma certeza c, tal que S é mais simples (em algum sentido) do que a enumeração de todos os exemplos de E_s;
- Um padrão que seja interessante (de acordo com uma medida de interesse imposta pelo usuário) e certo o suficiente (também de acordo com um critério do usuário) é chamado de conhecimento;
- E esse conhecimento pode ser descrito na forma de um conjunto de regras *SE - ENTÃO* associadas a uma medida de certeza que poderão ser apresentadas em uma árvore de decisão.

Nas equações (2.1) à (2.7) é apresentado um conjunto de regras *SE - ENTÃO* associadas a uma medida de certeza, descrevendo as condições para se jogar golfe. Estas regras foram determinadas a partir da extração de conhecimento da BD-exemplo apresentada na Tabela 2.2, e na seqüência na Figura 2.1 a árvore de decisão que descreve o problema.

SE Previsão = 'chuvoso' ENTÃO Y = 'joga' (certeza = 60%) (2.1)

SE Previsão = 'ensolarado' ENTÃO Y = 'não joga' (certeza = 60%) (2.2)

SE Temperatura = 'ameno' ENTÃO Y = 'joga' (certeza = 67%) (2.3)

SE Umidade = 'normal' ENTÃO Y = 'joga' (certeza = 86%) (2.4)

SE Previsão = 'nublado' ENTÃO Y = 'joga' (certeza = 100%) (2.5)

SE Previsão = 'ensolarado' E Umidade = 'alta'
ENTÃO Y = 'não joga' (certeza = 100%) (2.6)

SE Previsão = 'chuvoso' E Vento = 'verdadeiro'
ENTÃO Y = 'não joga' (certeza = 100%) (2.7)

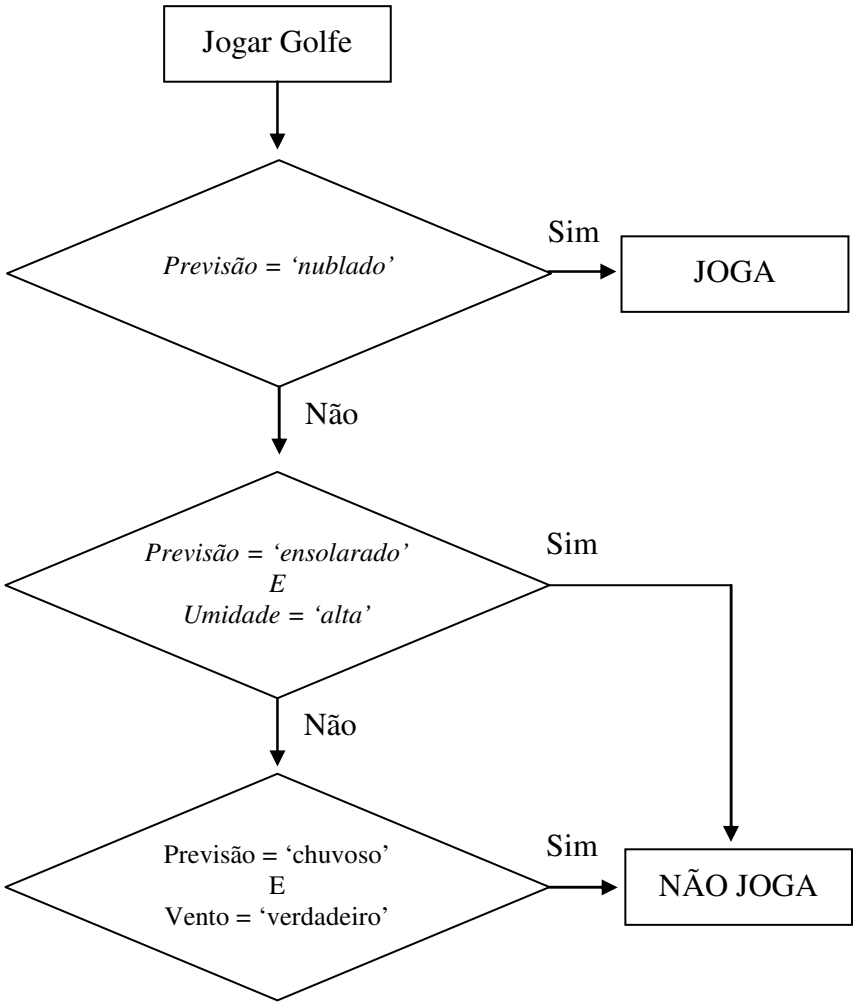


Figura 2.1: Árvore de decisão que descreve o problema "Jogar Golfe"

2.4 Considerações Finais

Neste capítulo apresentou-se o formato da BD a ser analisada e apresentada na entrada do KDD, e o formato do conhecimento extraído e coletado na saída do KDD.

Quanto ao formato da BD a ser analisada, verifica-se que é semelhante a uma matriz de duas dimensões sendo nas linhas a representação dos exemplos e nas colunas a representação dos atributos. Desses atributos seleciona-se um para ser o objeto do conhecimento extraído (atributo de decisão) e os demais atributos serão denominados atributos condicionais.

Quanto ao formato do conhecimento extraído, verifica-se que consiste em uma condição dentro da BD a ser analisada associada a um grau de certeza, ou seja, uma combinação de valores de atributos condicionais que apontará um determinado valor de atributo de decisão, associado a um grau de certeza dessa hipótese estar certa.

Capítulo 3

Seleção de Subconjuntos de Atributos

3.1 Considerações Iniciais

Conforme citado na revisão bibliográfica deste trabalho, MD é uma metodologia empregada sobre BDs com o intuito de "minerar" o seu conhecimento intrínseco.

Em busca de melhores performances para extração desse conhecimento, foi proposto o KDD, a partir da MD. Desta forma, agregou-se na etapa anterior à MD, a etapa de Transformação de Dados. Essa etapa tem por finalidade simplificar os dados, sem que haja modificações no conteúdo da BD em análise. E uma forma de se obter isso, é através da aplicação de metodologias de SSA, que buscam apontar subconjuntos ótimos de atributos da BD em análise.

Neste capítulo apresenta-se a compilação de várias referências sobre o tema SSA, abrangendo: conceitos e definições, objetivos, a finalidade de sua aplicação, a importância dos atributos condicionais em relação ao conhecimento a ser extraído e os princípios e mecanismos para o seu funcionamento.

3.2 Conceitos e definições em SSA

A SSA é um processo de busca – a ser aplicada sobre a BD em análise, na etapa de Transformação de dados do KDD – com o objetivo de identificar e eliminar os atributos condicionais não importantes para o conceito a ser aprendido, e a partir daí definir um Subconjunto de Atributos Ótimo, que não necessariamente é único.

Como benefícios da aplicação desta ferramenta verificam-se: eliminação do custo de coleta e armazenagem de atributos não importantes; redução considerável do volume de dados, melhorando a performance de processamento; melhoria na compreensibilidade do conceito induzido; e melhoria na precisão dos classificadores gerados.

Os atributos condicionais ditos não importantes para o conceito a ser aprendido, podem ser divididos em duas categorias:

- Atributos irrelevantes em relação ao conhecimento a ser extraído – se os valores de um atributo condicional não têm nenhuma correlação com os valores do atributo de decisão, este pode ser excluído, sem causar grandes impactos no conteúdo da BD. Destaca-se que um atributo irrelevante correlacionado com outro atributo pode passar a ser relevante;

- Atributos redundantes em relação a outros atributos condicionais – se os valores de um atributo condicional estão correlacionados com os valores de outro ou mais atributos condicionais, então este primeiro atributo condicional citado pode representar os outros, que podem ser eliminados sem causar grandes impactos no conteúdo da BD.

Um clássico exemplo é apresentado em Yu & Liu [2004] e Lee [2005], que consiste em buscar todos os Subconjuntos de Atributos Ótimos possíveis de uma BD com: os atributos $X = \{X_1, X_2, X_3, X_4, X_5\}$ sendo todos booleanos, as relações internas $X_2 = \overline{X}_3$ e $X_4 = \overline{X}_5$, e uma função booleana $Y = f(X_1, X_2)$ que resulta no conhecimento da BD. Desse exemplo pode-se concluir que:

- a BD é formada por 8 exemplos, sendo: {{0,0,1,0,1}, {0,0,1,1,0}, {0,1,0,0,1}, {0,1,0,1,0}, {1,0,1,0,1}, {1,0,1,1,0}, {1,1,0,0,1}, {1,1,0,1,0}};

- os Subconjuntos de Atributos Ótimos para essa BD são: $\{X_1, X_2\}$ e $\{X_1, X_3\}$;

- X_1, X_2 e X_3 são atributos relevantes em relação à Y; X_4 e X_5 são atributos irrelevantes em relação à Y; e X_2 e X_3 são atributos redundantes entre si.

3.2.1 Importância de Atributos

Nesta subseção serão apresentadas as definições de atributos importantes sugeridas nas referências pesquisadas.

1. Liu & Motoda [1998] consideram que um atributo condicional *Xi* é dito importante, se quando removido, a medida de importância considerada em relação aos atributos restantes é deteriorada.

2. A seguir apresentam-se algumas medidas de importância de atributos, segundo Liu & Yu [2002] e Lee [2005]:

   - **Medida de informação** – Determina o ganho de informação a partir de um atributo condicional *Xi*. Esse ganho é definido como sendo a diferença entre a incerteza antes e depois de inserir *Xi*, e quanto maior é o ganho mais importante é o atributo;

- **Medida de distância** – Também conhecida como medida de separabilidade, divergência ou discriminação. Para um problema de duas classes, um atributo X_i é mais importante que um atributo X_j, se X_i provê uma diferença maior que X_j entre as probabilidades condicionais das duas classes. Se a diferença é zero, então X_i e X_j são indistinguíveis;

- **Medida de dependência** – Também conhecida como medida de correlação ou associação. Determina a correlação de um atributo condicional com outro atributo que pode ser de decisão (busca-se a relevância do atributo em relação ao conceito a ser aprendido) ou condicional (busca-se a redundância do atributo em relação a outros atributos condicionais). Em relação ao conceito relevância, um atributo X_i é mais importante que um atributo X_j, se a correlação do atributo X_i com a classe D é maior que a correlação do atributo X_j com essa classe. E em relação ao conceito de redundância, se um atributo X_i está fortemente correlacionado com o atributo X_j, esses dois atributos trazem a mesma informação, assim um deles pode ser eliminado;

- **Medida de consistência** – Tem características diferentes das outras medidas, pois é fortemente dependente do conjunto de treinamento e busca subconjuntos com menor número de elementos possíveis, admitindo hipóteses com uma inconsistência definida pelo usuário. Um problema associado a esta medida, consiste no fato dela não conseguir distinguir atributos redundantes, uma vez que eles podem estar fortemente correlacionados ao conceito a ser aprendido. A inconsistência é definida como dois exemplos possuindo os mesmos valores de atributos e classes diferentes;

- **Medida de precisão** – Aponta a precisão do modelo induzido (classificador). Assim, busca-se o subconjunto que proporciona a maior precisão possível. Esta é a única medida que depende do algoritmo de aprendizagem.

Os três primeiros tipos de medidas estão fortemente relacionados, de tal maneira que é possível agrupar as medidas apresentadas em três categorias: medidas clássicas (informação, distância e dependência), medida de consistência e medida de precisão.

3. *Importância em relação ao conceito meta – Medida de consistência*: Almuallim & Diettrich [1994] definem que um atributo condicional X_i é importante em relação ao conceito meta c, se X_i aparece em todas as fórmulas booleanas que representa c, e irrelevante caso contrário. A aplicação desta definição implica em um sistema: composto por atributos condicionais e de decisão booleanos e sem ruídos.

4. *Importância em relação ao conceito meta – Medida de consistência*: Liu & Setiono [1996] e Dash *et al* [2000] definem a importância de um subconjunto de atributos através da taxa de inconsistência definida como: (1) um exemplo é considerado inconsistente se existirem pelo menos dois exemplos exatamente iguais exceto pelo valor da classe; (2) a contagem de inconsistência para um exemplo é dada pela diferença entre o número de vezes que esse exemplo aparece e o maior número de vezes que esse exemplo pertencente a uma classe diferente aparece e; (3) a taxa de inconsistência de um subconjunto de atributos é a soma de todas as contagens de inconsistência de todos os exemplos do subconjunto nos dados dividido pelo número total de exemplos.

5. *Importância probabilística – Medida de dependência*: Gennari *et al* [1989] definem que um atributo condicional X_i é importante, sss[3] existe algum x_i e y para os quais $p(X_i = x_i) > 0$ tal que: $p(Y = y \mid X_i = x_i) \neq p(Y = y \mid X_i = x_i)$.

6. *Importância probabilística – Medida de dependência*: John *et al* [1994] definem que um atributo condicional X_i é importante, sss existe algum x_i, y e s_i para $p(X_i = x_i) > 0$ tal que: $p(Y = y, S_i = s_i \mid X_i = x_i) \neq p(Y = y, S_i = s_i)$.

7. *Importância probabilística – Medida de dependência*: John *et al* [1994] definem que um atributo condicional X_i é importante, sss existe algum x_i, y e s_i para $p(X_i = x_i, S_i = s_i) > 0$ tal que: $p(Y = y \mid X_i = x_i, S_i = s_i) \neq p(Y = y \mid S_i = s_i)$.

8. *Importância em relação ao conceito meta – Medida de dependência*: Blum & Langley [1997] definem que um atributo condicional X_i é importante, sss existe um par de exemplos E_i e E_j, $i \neq j$, no espaço de exemplos tal que E_i e E_j diferem somente na atribuição de valores ao atributo X_i e $f(\vec{x}_i) \neq f(\vec{x}_j)$.

9. *Forte importância – Medida de dependência*: John *et al* [1994] definem que um atributo condicional X_i é fortemente importante, sss existe algum x_i, y e s_i para $p(X_i = x_i, S_i = s_i) > 0$ tal que: $p(Y = y \mid X_i = x_i, S_i = s_i) \neq p(Y = y \mid S_i = s_i)$.

10. *Fraca importância – Medida de dependência*: John *et al* [1994] definem que um atributo condicional X_i é fracamente importante, sss não é fortemente importante e existe $S_i' \subset S_i$ para o qual algum x_i, y e s_i' para $p(X_i = x_i,\ S_i' = s_i') > 0$ tal que: $p(Y = y \mid X_i = x_i,\ S_i' = s_i') \neq p(Y = y \mid S_i' = s_i')$.

---

[3] se e somente se.

11. *Forte importância em relação à Amostra/Distribuição – Medida de dependência*: Blum & Langley [1997] definem que um atributo condicional X_i é fortemente importante para uma amostra E, se existe um par de exemplos E_i e E_j, $i \neq j$, tal que E_i e E_j diferem somente na atribuição de valores ao atributo X_i e possuem diferentes classes (ou possuem diferentes distribuições das classes se esses aparecem múltiplas vezes em S). Similarmente, o atributo X_i e satisfazem $f(\vec{x}_i) \neq f(\vec{x}_j)$.

12. *Fraca importância em relação à Amostra/Distribuição – Medida de dependência*: Blum & Langley [1997] definem que um atributo condicional X_i é fracamente importante para uma amostra E ou para a meta f e a distribuição D se for possível remover um subconjunto de atributos tal que o atributo Xi torna-se fortemente importante.

13. *Importância em relação à Dimensão Fractal – Medida de dependência*: Traina *et al* [2000], dada a dimensão fractal calculada utilizando-se todos os atributos condicionais do conjunto de dados definem que um atributo condicional X_i é importante se a sua exclusão causar uma alteração significativa, definida pelo usuário, no valor da dimensão fractal.

14. *Importância em relação à Distância – Medida de distância*: Kira & Rendell [1992], dados dois atributos X_i e X_j, a importância de cada atributo é definida na equação (3.1):

$$W[X_i = x_i \mid Y = y] = P(X_i = x_i \mid Y = \neg y) - P(X_i = x_i \mid Y = y) \qquad (3.1)$$

    X_i será mais importante que X_j se $W[X_i = x_i \mid Y = y] > W[X_j = x_j \mid Y = y]$.

15. *Importância em relação a precisão – Medida de precisão*: Kohavi & John [1997] definem que um subconjunto de atributos X' é ótimo, se dados um algoritmo de aprendizagem I e uma BD E, formada pelo conjunto de atributos $X = \{X_1, X_2, ..., X_m\}$, têm-se: $X' \subseteq X$ e a precisão para o classificador induzido $h = I(X')$ é máxima em relação aos outros subconjuntos de atributos possíveis da BD em análise. Destaca-se que um subconjunto de atributos que atende a condição de ótimo, não é necessariamente único para uma BD, pois podem existir subconjuntos com outros atributos que podem atender a mesma condição.

16. *Incremento na precisão – Medida de precisão*: Caruana & Freitag [1994] definem que um atributo condicional X_i é importante, se dados uma BD E, um indutor I e um subconjunto de atributos X, no qual $\{X_i\} \cap X = \varnothing$, verifica-se que a precisão da hipótese gerada considerando o conjunto de atributos $\{X_i\} \cup X$ é melhor que a

precisão alcançada utilizando apenas o subconjunto *X*. Em Lee [2005] destaca-se que é interessante observar que a importância de um atributo não implica que ele esteja no Subconjunto de Atributos Ótimo quando a medida de precisão é considerada.

17. *Importância como uma Medida de Complexidade – Medida de precisão*: Blum & Langley [1997], dada uma amostra de dados *S* e um conceito meta *f*, definir *r*(*S*,*f*) como o menor número de atributos importantes para *f* de acordo com a Definição 3.7, tal que o erro sobre *S* seja o mínimo possível para o algoritmo de aprendizagem.

3.3 Princípios e mecanismos de funcionamento

Uma constatação de Lee [2005] consiste no fato de que os métodos de SSA podem selecionar os atributos por avaliação individual ou por avaliação de subconjuntos de atributos. Na avaliação individual, freqüentemente, os atributos são ordenados considerando a sua importância na discriminação das classes, isto é, tratam a relevância dos atributos. Esses métodos somente removem atributos irrelevantes, pois esperam-se que atributos redundantes tenham a mesma importância na discriminação das classes. Contudo, métodos que avaliam subconjuntos de atributos buscando por subconjuntos mínimos podem remover tanto atributos irrelevantes quanto redundantes. Assim, a maioria dos métodos existentes para a SSA que tratam tanto relevância quanto redundância de atributos, o fazem de maneira implícita por meio da avaliação de subconjuntos de atributos.

Por definição, SSA é um problema de busca que tem por objetivo encontrar o Subconjunto de Atributos Ótimo da BD em análise. Nesse contexto, define-se o espaço de busca como um conjunto contendo todas as combinações possíveis de atributos condicionais de uma BD, de onde a SSA selecionará o Subconjunto de Atributos Ótimo. Na Figura 3.1 é apresentado um exemplo de espaço de busca de subconjuntos de atributos para uma BD contendo quatro atributos, onde cada círculo não vazado representa o atributo condicional que compõe o subconjunto.

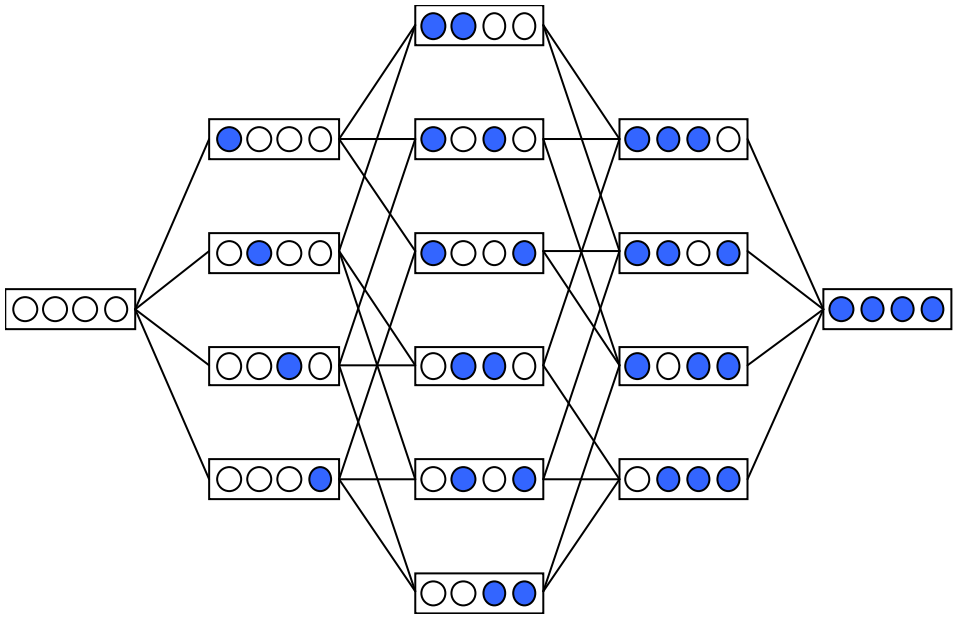


Figura 3.1: Exemplo de espaço de busca de subconjunto atributos [Langley, 1994]

Observando a Figura 3.1 nota-se que o número de elementos no espaço de busca é igual à 2^n, onde n é o número de atributos da BD em análise.

O mecanismo de funcionamento das metodologias de SSA é baseado nos tópicos apresentados nas subseções seguintes.

3.3.1 Ponto de partida e Direção de busca

É sabido que o Subconjunto de Atributos Ótimo está dentro do espaço de busca. Como em todo problema de busca, é necessário definir um ponto de partida dentro desse espaço e a direção de busca, dessa forma têm-se:

- **Seleção *forward***: determina que o sentido da busca seja do conjunto vazio para o conjunto de todos os atributos da BD. Assim, a cada iteração acrescenta-se um atributo e avalia-se o impacto;
- **Eliminação *backward***: determina que o sentido da busca seja do conjunto de todos os atributos da BD para o conjunto vazio. Assim a cada iteração elimina-se um atributo e avalia-se o impacto;
- **Seleção bidirecional**: implica que o sentido da busca pode ser alterado durante o processamento;
- **Seleção randômica**: implica que o processo irá selecionar os subconjuntos dentro do espaço através de sorteios.

3.3.2 Estratégia de Busca

Segundo Dash & Liu [1997], as estratégias de busca utilizadas pelas metodologias de SSA podem ser:

- **Busca completa**: implica em avaliar todos os subconjuntos do espaço de busca e apresentar o que alcançou a maior performance na avaliação. Essa estratégia garante a condição de subconjunto ótimo para o subconjunto apontado, mas devido à sua complexidade 2^n, esta estratégia pode ser muito custosa e até inviável;
- **Busca heurística**: emprega-se algum tipo de heurística para conduzir a busca, evitando a busca completa, porém corre-se o risco de não encontrar os subconjuntos ótimos. Essa estratégia é muito mais rápida e a complexidade do espaço de busca pode ser reduzida à n^2;
- **Busca não-determinística**: esta estratégia procura aleatoriamente os subconjuntos e o principal critério de parada é o número de iterações.

3.3.3 Medida de avaliação

Consiste em adotar uma medida de importância de atributos (as medidas de importância de atributos foram definidas na subseção 3.2.1), que será utilizada como função objetivo da SSA, ou seja, a SSA deverá buscar um subconjunto de atributos que forneça o valor ótimo para essa medida. A seguir apresentam-se as classes de medidas de importância de atributos:

- Medidas clássicas;
- Medida de consistência;
- Medida de precisão.

3.3.4 Critério de parada

Por se tratar de um processo iterativo, segue alguns critérios de parada:

- parar de remover ou adicionar atributos quando nenhuma das alternativas melhora a precisão da estimativa para a classificação;
- continuar revisando o subconjunto de atributos enquanto a precisão não se degrada;
- continuar gerando subconjuntos candidatos até que o outro extremo do espaço

de busca seja alcançado e escolher o melhor resultado encontrado;

- parar quando o subconjunto de atributos selecionado separar perfeitamente todas as classes (assumindo que não há ruídos nos dados);
- ordenar os atributos segundo alguma pontuação de importância e utilizar um parâmetro de sistema para determinar o ponto de parada, por exemplo, o número de atributos desejado para o subconjunto.

3.3.5 Abordagens para seleção de atributos

Outro aspecto importante é a abordagem da SA perante o processo de MD, que pode ocorrer de três maneiras distintas, conforme apresentadas a seguir:

- ***Embedded* (embutida)**: Nessa abordagem a tarefa da SA é realizada dentro do algoritmo de MD, ou seja, aplicam-se os dados de treinamento diretamente sobre o algoritmo de MD sem a fase de pré-processamento definida no KDD. Na Figura 3.2 é apresentado um diagrama de um esquema de abordagem *Embedded*.

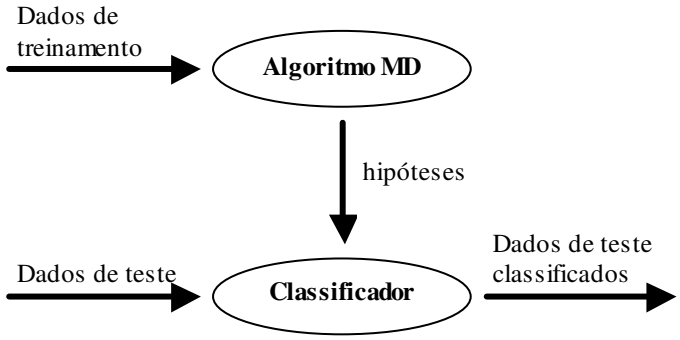


Figura 3.2: Esquema de abordagem para SA: Abordagem *Embedded*

- **Filtro**: Nessa abordagem a tarefa da SA é realizada externamente ao algoritmo de MD. A idéia é filtrar atributos irrelevantes, segundo algum critério de importância de atributos descritos na subseção 3.2.1, antes do processo de MD. Sendo assim, os métodos de filtro são independentes do algoritmo de aprendizado. Um dos esquemas mais simples de filtragem é a avaliação de cada atributo individualmente, baseada na sua correlação com o conceito meta, escolhendo o subconjunto de *n*' atributos que fornecem o melhor valor dessa correlação [Blum & Langley, 1997; Lee, 2005]. Na Figura 3.3 é apresentado um diagrama de um esquema de abordagem Filtro.

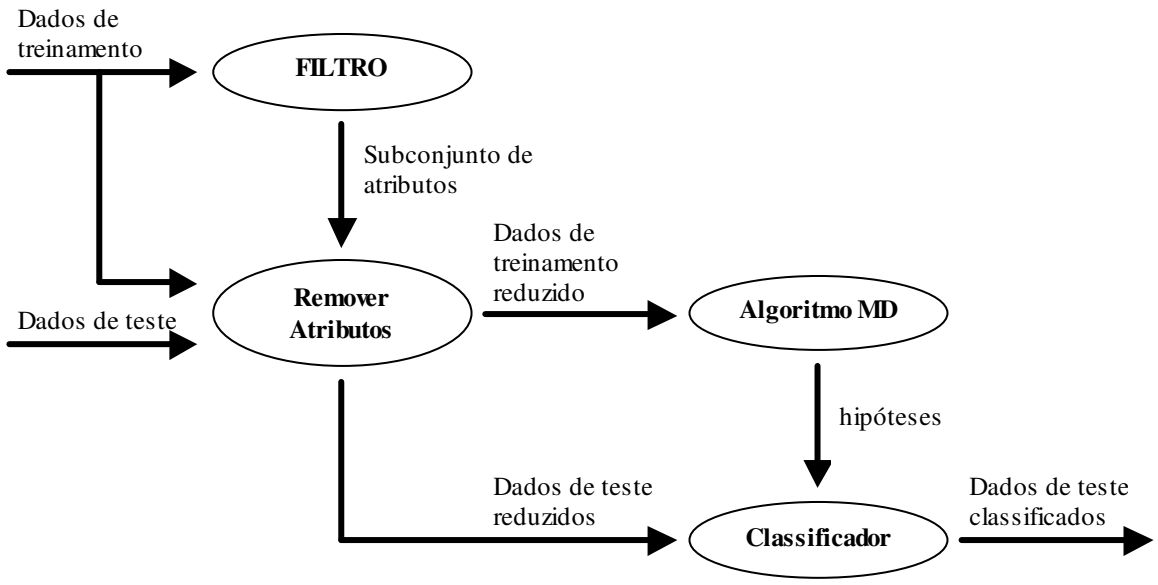


Figura 3.3: Esquema de abordagem para SA: Abordagem Filtro

- ***Wrapper* (invólucro)**: Nessa abordagem a tarefa da SA também é realizada externamente ao algoritmo de MD, porém utilizando tal algoritmo como uma caixa preta para analisar, a cada iteração, o subconjunto de atributos em questão. Em outras palavras, os métodos *wrapper* geram um subconjunto candidato de atributos selecionado do conjunto de treinamento, e utilizam a precisão resultante do classificador induzido para avaliar o subconjunto de atributos em questão. Esse processo é repetido para cada subconjunto de atributos até que o critério de parada determinado seja satisfeito. Porém, a maior desvantagem dos métodos wrapper é o custo computacional, o qual resulta da execução do algoritmo de aprendizado para avaliar cada subconjunto de atributos a ser considerado [Kohavi & John, 1997; Pila, 2001; Lee, 2005]. Na Figura 3.4 é apresentado um diagrama de um esquema de abordagem *Wrapper* (invólucro).

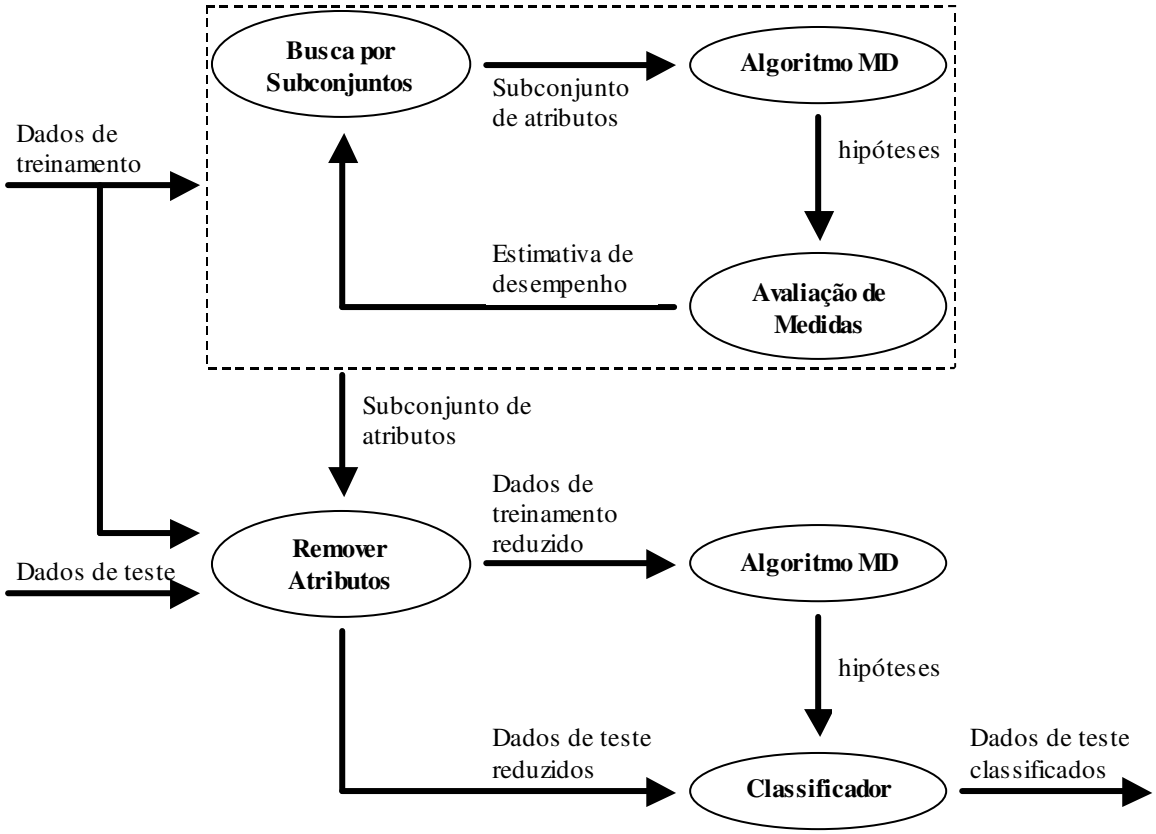


Figura 3.4: Esquema de abordagem para SA: Abordagem *Wrapper*

3.4 Considerações Finais

Neste capítulo apresentaram-se conceitos e definições aplicáveis às metodologias de SSA, cujo objeto consiste em apontar um subconjunto de atributos condicionais com menor número de elementos possíveis e contendo somente atributos importantes ao conceito a ser aprendido. Assim, introduziram-se vários conceitos de importância de atributos que avalia e atribui um grau de importância para os atributos ou subconjunto de atributos de uma BD; e a metodologia de SSA deverá selecionar os atributos buscando obter o maior grau possível.

Outro enfoque deste capítulo foram os princípios e mecanismos de funcionamento das metodologias de SSA, onde se expôs que a metodologia de SSA consiste em um problema de busca com espaço definido pelos 2^n (n - número de atributos) subconjuntos de atributos possíveis. Dentro desse contexto apresentou-se que para definição de uma metodologia de SSA é necessário determinar:

- o ponto de partida e a direção de busca dentro desse espaço;
- a estratégia de busca que consiste em analisar todos os subconjuntos existentes

ou subconjuntos apontados por heurísticas ou por sorteios;

- a medida de avaliação para referenciar a busca do subconjunto de atributos que forneça o melhor grau de importância;
- os critérios de paradas da busca iterativa;
- e as abordagens existentes de SA, que são: abordagem embedded – implicando que a SA é executada dentro da etapa de MD; abordagem filtro – implicando que a SA é executada a parte da etapa de MD, através de um algoritmo que apontará um subconjunto de atributos para ser processado na etapa de MD; abordagem wrapper – semelhante à abordagem filtro, exceto pelo fato de que dentro do algoritmo de SA existe uma etapa de MD para avaliar o subconjunto de atributos em análise, assim buscando subconjuntos de atributos que forneçam a melhor medida de precisão.

Capítulo 4

Métodos para Avaliação de SSA

4.1 Considerações Iniciais

Neste capítulo apresentam-se alguns métodos utilizados para avaliar metodologias de SSA, dos quais serão analisados e selecionados para avaliar a metodologia proposta neste trabalho.

4.2 Métodos para Avaliação de SSA

Lee [2000] cita a inexistência de um método matemático que permita a avaliação do desempenho de um modelo de SSA, e destaca a importância de que estudos empíricos sejam realizados sobre os conjuntos de dados de interesse, a fim de determinar quais são mais apropriados, ou apresentam o melhor desempenho, para esses conjuntos de dados em particular.

Nessas comparações, geralmente destacam-se análise da melhoria da medida de avaliação adotada, do número de atributos apontados e do tempo de processamento. Quanto aos dois últimos quesitos apresentados a análise é automática, ou seja, quanto menor for o número de atributos e o tempo de processamento, melhor será o modelo. Na seqüência apresentam-se Métodos para Avaliação de SSA que medem a precisão do classificador gerado por um algoritmo de MD, a partir dos subconjuntos de atributos apontados por metodologias de SSA, segundo Two Crows [2005]:

4.2.1 Validação Simples

Este método consiste em separar aleatoriamente um percentual da BD em análise e definir como BD-teste, tipicamente esse percentual está entre 5% e 33%. Então com a outra parte gera-se o modelo de representação da BD em análise, considerando o subconjunto de atributos apontados pelo método de SSA; então aplica-se esse modelo para prever as classes da BD-teste, levantando a estatística de acertos e erros. Com isso define-se: a precisão que é igual ao número de acertos dividido pelo número total de exemplos na BD-teste; e a taxa de erro que é (1 – taxa de acertos). Na Figura 4.1 apresenta-se o esquema de validação simples.

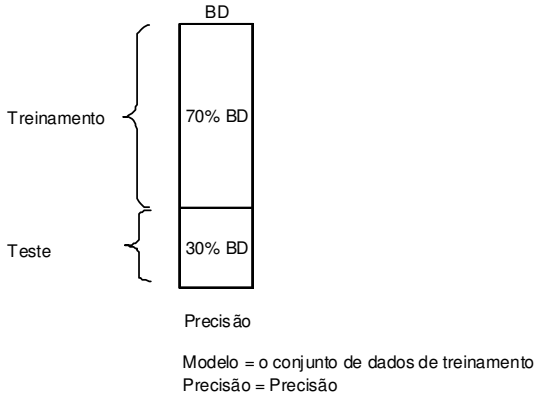


Figura 4.1: Esquema de validação simples

4.2.2 Validação cruzada

Este método é aplicado quando se possui uma BD com poucos exemplos para compor o modelo de representação da BD em análise, inviabilizando a separação proposta no método de Validação Simples. Inicialmente a BD em análise é dividida em 2 subconjuntos com a mesma quantidade de exemplos e com a mesma distribuição de classes da BD completa. Então, com o primeiro subconjunto gera-se o modelo e com segundo testa-se o modelo gerado, calculando a precisão do modelo; depois repete-se o procedimento anterior invertendo os subconjuntos. Para finalizar, gera-se o modelo com a BD completa e como precisão utiliza-se a média das duas precisões calculadas. Na Figura 4.2 apresenta-se o esquema de validação cruzada.

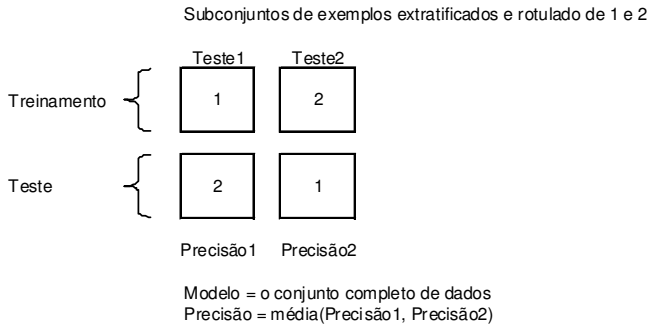


Figura 4.2: Esquema de validação cruzada

4.2.3 Validação Cruzada com 10-partições estratificadas

Neste método a BD em análise é dividida em dez subconjuntos estratificados, isto é, contendo aproximadamente a mesma proporção de classes da BD em análise e de aproximadamente do mesmo tamanho. O modelo é treinado e testado dez vezes; cada vez é testado sobre um subconjunto e treinado sobre o conjunto de dados menos esse

subconjunto. Ao final, gera-se o modelo com a BD completa e adota-se como estimativa de precisão a média das precisões estimadas em cada um dos dez subconjuntos. Na Figura 4.3 apresenta-se o esquema de validação cruzada com 10-partições estratificadas.

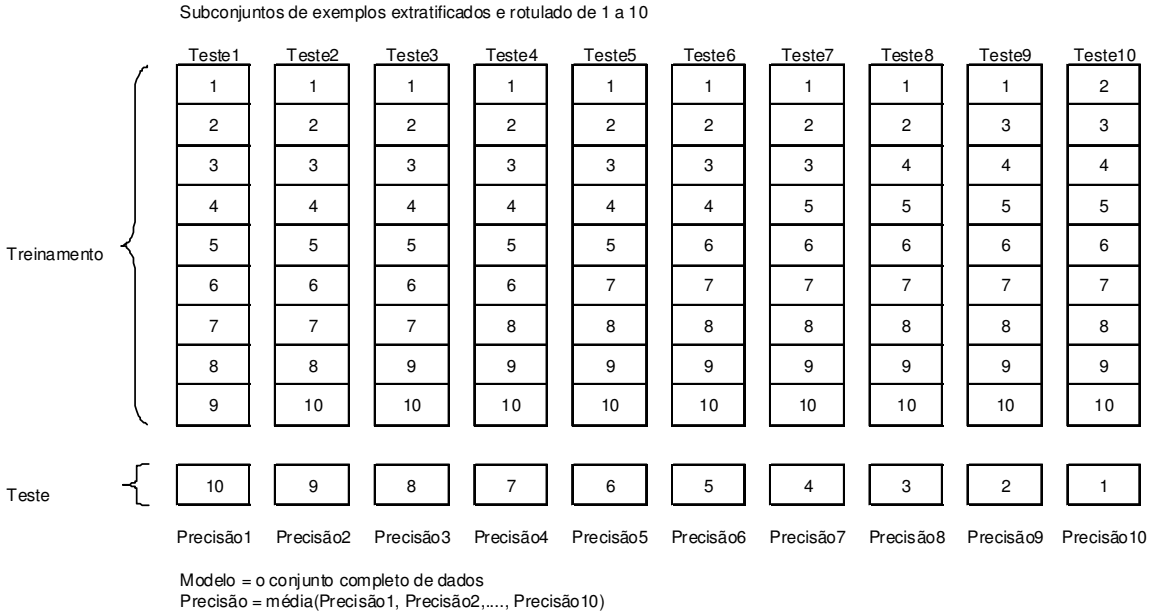


Figura 4.3: Esquema de validação cruzada com 10-partições estratificadas

4.2.4 Validação Bootstrapping

Este método é aplicado quando a quantidade de exemplos em uma BD é muito pequena. A sistemática consiste em definir aleatoriamente um subconjunto de treinamento, a partir do conjunto original de atributos; o restante dos atributos deve compor o subconjunto de teste. Calcula-se a precisão para esse caso, depois elencam-se outros subconjuntos e repete-se o processo por várias vezes. Para concluir, gera-se o modelo a partir da BD original, e adota-se como precisão a média de todas as precisões calculadas.

4.3 Considerações Finais

Neste capítulo apresentou-se que não existe nenhum modelo matemático para avaliar o desempenho de metodologias de SSA. Dessa forma, essa avaliação deve ser feita de forma experimental, através da comparação de parâmetros obtidos na aplicação de conjuntos de dados referência em metodologias de SSA diferentes. Os parâmetros geralmente analisados são: a obtenção da melhoria da medida de avaliação adotada, o número de atributos apontados e o tempo de processamento.

De forma automática, cita-se que as metodologias que obtêm o menor número de atributos condicionais e tempo de processamento são ditas melhores. E para avaliar a melhoria da medida de avaliação de importância são apresentados métodos que avaliam a precisão obtida no classificador gerado por um algoritmo de MD, a partir dos subconjuntos de atributos apontados por metodologias de SSA, e são: a validação simples, validação cruzada, validação cruzada com 10-partições estratificadas e a validação *bootstrapping*.

Capítulo 5

Metodologias de SSA Pesquisadas

5.1 Considerações Iniciais

Neste capítulo serão apresentadas as metodologias de SSA pesquisadas para fundamentar e buscar semelhanças com a nova Metodologia de SSA Proposta e promover uma competição entre elas. Nessa linha, faz-se um breve descritivo do mecanismo das metodologias pesquisadas, com apresentação do algoritmo simplificado e por fim um comentário avaliando-as quanto ao desempenho.

Por definição, todas as metodologias de SSA buscam apontar um Subconjunto de Atributos Ótimo para as BDs considerando uma definição de ótimo adotado, mas devido às simplificações aplicadas, os subconjuntos de atributos apontados poderão não ser ótimos.

5.2 Redutos em Rough Sets

Teoria de *Rough Sets* (TRS) é uma abordagem matemática para análise de dados, introduzida por Pawlak [1982], que considera e provê um tratamento adequado para as incertezas e imprecisões contidas nesses dados. Na seqüência, serão apresentados conceitos da TRS utilizados para redução da dimensão de dados sem alteração da capacidade de classificação da BD em análise; esses conceitos são aplicáveis em SSA.

5.2.1 Sistema de Informação (SI)

Consiste em um par ordenado $S = (U,A)$, onde U e A são conjuntos finitos, não-vazios, denominados Universo e Atributos, respectivamente. Os elementos do Universo são referenciados como exemplos (objetos, registros ou instâncias). Cada atributo $\alpha_i \in A$ é uma função tal que α_i: $U \rightarrow V_{\alpha i}$, onde $V_{\alpha i}$ é o conjunto dos valores permitidos para o atributo α_i (sua faixa de valores) [Pawlak, 1982; Komorowski, 1998; Pila, 2001]. Na Tabela 5.1 é apresentado o exemplo de um SI.

	Atributos		
Exemplos	*Estudos*	*Educação*	*Trabalho*
E_1	não	boa	sim
E_2	não	boa	sim
E_3	sim	boa	sim
E_4	não	pobre	não
E_5	não	pobre	não

Tabela 5.1: Exemplo de Sistema de Informação [Pila, 2001]

5.2.2 Sistema de Decisão (SD)

Consiste em qualquer SI na forma $S = (U, A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão. Os elementos de A são chamados de atributos condicionais ou simplesmente condições [Pawlak, 1982; Komorowski, 1998; Pila, 2001]. Na Tabela 5.2 é apresentado um exemplo de um SD.

	Atributos			Decisão
Exemplos	*Estudos*	*Educação*	*Trabalho*	*Renda*
E_1	não	boa	sim	alta
E_2	não	boa	sim	alta
E_3	sim	boa	sim	nenhuma
E_4	não	pobre	não	baixa
E_5	não	pobre	não	média

Tabela 5.2: Exemplo de Sistema de Decisão [Pila, 2001]

Do exemplo da Tabela 5.2 podemos concluir que:

- $U = \{E_1, E_2, E_3, E_4, E_5\}$
- $A = \{\alpha_1, \alpha_2, \alpha_3\} = \{$Estudos, Educação, Trabalha$\}$
- $d = \{$Renda$\}$
- $V_{\alpha 1} = \{$não, sim$\}$
- $V_{\alpha 2} = \{$boa, pobre$\}$
- $V_{\alpha 3} = \{$sim, não$\}$
- $V_d = \{$alta, nenhuma, baixa, média$\}$

5.2.3 Redutos

Reduto é um conceito importante dentro da TRS, e consiste em um subconjunto B, determinado a partir do conjunto de todos os atributos do SI A, tal que: $B \subseteq A$; e B possua o menor número possível de elementos (Reduto Mínimo / Ótimo) e mantenha a mesma capacidade de classificação que A [Pawlak, 1982; Komorowski, 1998; Pila, 2001]. Dessa forma, a determinação do Reduto Mínimo (Reduto Ótimo) consiste em um método de

SSA que busca o Subconjunto de Atributos Ótimo.

Um SD pode ser reduzido sem que haja alterações significativas no seu conhecimento contido, pelos seguintes fatos [Komorowski, 1998; Pila, 2001]:

- o mesmo exemplo ou exemplos indistinguíveis podem estar representados várias vezes;
- alguns atributos podem ser supérfluos.

Para cada subconjunto de atributos $B \subseteq A$, no SI $S = (U,A)$, uma relação de equivalência INDs(B) é associada, chamada de Relação de Não-Discernimento de B, e é definida na equação (5.1) [Komorowski, 1998; Pila, 2001]:

$\text{IND}_S(B) = \{(E_i, E_j) \in U^2 \mid \forall\ \alpha_i \in B,\ \alpha_i(E_i) = \alpha_i(E_j)\}$, onde $1 \leq i < j \leq m$, m é o número de exemplos da BD e E_i, E_j são respecitvamente o i-ésimo e o j-ésimo exemplo da BD *(5.1)*

Aplicando a Relação de Não-Discernimento a todos os subconjuntos de atributos do SI apresentado na Tabela 5.1, obtém-se o resultado apresentado na Figura 5.1.

U/IND({*Estudos*})	$= \{\{E_1, E_2, E_4, E_5\},\{E_3\}\}$
U/IND({*Educação*})	$= \{\{E_1, E_2, E_3\},\{E_4, E_5\}\}$
U/IND({*Trabalha*})	$= \{\{E_1, E_2, E_3\},\{E_4, E_5\}\}$
U/IND({*Estudos*, *Educação*})	$= \{\{E_1, E_2\},\{E_3\},\{E_4, E_5\}\}$
U/IND({*Estudos*, *Trabalha*})	$= \{\{E_1, E_2\},\{E_3\},\{E_4, E_5\}\}$
U/IND({*Educação*, *Trabalha*})	$= \{\{E_1, E_2, E_3\},\{E_4, E_5\}\}$
U/IND({*Estudos*, *Educação*, *Trabalha*})	$= \{\{E_1, E_2\},\{E_3\},\{E_4, E_5\}\}$

Figura 5.1: Aplicação da Relação de Não-Discernimento em um SI [Pila, 2001]

A busca do Reduto na TRS consiste em analisar o SI e eliminar todos os atributos possíveis, mantendo a mesma Relação de Não-Discernimento observada para o conjunto de todos os atributos do SI. Observando a Figura 5.1, verifica-se que o conjunto de todos os atributos do SI A = {Estudos, Educação, Trabalha} possui a seguinte Relação de Não-Discernimento U/IND(A) = $\{\{E_1, E_2\},\{E_3\},\{E_4, E_5\}\}$, então ao testar todos os subconjuntos B gerados a partir A, verifica-se que a Relação de Não-Discernimento se mantém apenas para os subconjuntos B'={Estudos, Educação} e B"={Estudos, Trabalha}, ou seja, o atributo Trabalha ou o atributo Educação pode ser eliminado sem causar nenhuma alteração na capacidade de classificação do SI.

5.2.4 Matriz de Discernimento e Função de Discernimento

Uma outra forma para se buscar os Redutos Ótimos de um SI é através da determinação da Matriz de Discernimento, que gera a Função de Discernimento, e por fim a simplificação dessa função, que resulta nos Redutos Ótimos possíveis para o SI em análise [Pawlak, 1982; Pila, 2001].

A Matriz de Discernimento consiste em uma matriz simétrica $m \times m$, onde m é o número de exemplos do SI, cujo os elementos c_{ij} são determinados pela equação (5.2) [Komorowski, 1998].

$$c_{ij} = \{ \alpha \in A \mid \alpha(E_i) \neq \alpha(E_j) \} \text{ para } 1 \leq i < j \leq m \qquad (5.2)$$

De uma maneira genérica, o preenchimento dessa matriz ocorre da seguinte forma: para cada célula c_{ij} da matriz, faz-se a comparação do exemplo E_i com o exemplo E_j (atributo a atributo), anotando na célula os atributos que são diferentes entre si, ou seja, são discerníveis. Pelas características da metodologia, nota-se que as células da diagonal principal da matriz serão sempre vazias; e pela questão de simetria o preenchimento das células acima da diagonal principal serão iguais às suas correspondentes abaixo dessa diagonal, não sendo necessário o seu preenchimento. Na Figura 5.2 é apresentada a Matriz de Discernimento para o SI da Tabela 5.1.

	Indiv1	Indiv2	Indiv3	Indiv4	Indiv5
Indiv1	-				
Indiv2	-	-			
Indiv3	Estudos	Estudos	-		
Indiv4	Educação Trabalha	Educação Trabalha	Estudos Educação Trabalha	-	
Indiv5	Educação Trabalha	Educação Trabalha	Estudos Educação Trabalha	-	-

Figura 5.2: Exemplo de Matriz de Discernimento

A Função de Discernimento de um SI é uma função booleana de *m* (número de atributos no SI) variáveis booleanas $\alpha_1^*,...,\alpha_m^*$ (correspondentes às variáveis $\alpha_1,...,\alpha_m$) conforme descrito na equação (5.3).

$$f_S(\alpha_1^*,...., \alpha_m^*) = \wedge \{ \vee c_{ij}^* \mid 1 \leq i \leq j \leq m, c_{ij} \neq \varnothing\}, \text{ onde } c_{ij}^* = \{\alpha^* \mid \alpha \in c_{ij}\} \quad (5.3)$$

De uma maneira genérica, para compor a Função de Discernimento de um SI aplica-se o operador 'OR' entre os atributos dentro das células da Matriz de Discdernimento, e depois, o operador 'AND' entre os grupos de atributos de cada célula dessa matriz, assim obtendo uma equação, que ao ser simplificada fornecerá os diversos redutos do SI em análise. Com isso, seleciona-se o Reduto Ótimo, com menor número de atributos. As equações (5.4) à (5.8) apresentam a Função de Discernimento do SI da Tabela 5.1 e a sua simplificação.

$$f(B) = Estudos \wedge Estudos \wedge (Educação \vee Trabalha) \wedge (Educação \vee Trabalha) \wedge (Educação \vee Trabalha) \wedge (Educação \vee Trabalha) \wedge (Estudos \vee Educação \vee Trabalha) \wedge (Estudos \vee Educação \vee Trabalha) \quad (5.4)$$

$$F(B) = Estudos \wedge (Educação \vee Trabalha) \wedge (Estudos \vee Educação \vee Trabalha) \quad (5.5)$$

$$f(B) = Estudos \wedge (Educação \vee Trabalha) \quad (5.6)$$

$$f(B) = (Estudos \wedge Educação) \vee (Estudos \wedge Trabalha) \quad (5.7)$$

$$f(B) = (Estudos \; \mathbf{E} \; Educação) \; \mathbf{OU} \; (Estudos \; \mathbf{E} \; Trabalha) \quad (5.8)$$

Observando a simplificação do SI exemplo da Tabela 5.1, conclui-se que este SI possui dois redutos ótimos, que podem representá-lo com a mesma capacidade de classificação observada quando se considera todos os atributos, e são eles: {Estudos,Educação} e {Estudos,Trabalha}.

Encontrar o Reduto Mínimo é um problema NP-difícil[4] [Skowron & Rauszer, 1992].

Idéias básicas da TRS e suas extensões, evoluções desses conceitos, e muitas aplicações interessantes podem ser encontradas no sítio: *International Rough Set Society* (http://www.roughsets.org).

Pela concepção dessa metodologia, verifica-se que ela fornecerá o(s) Reduto(s) Mínimo(s) possíveis da BD em análise, mantendo a mesma capacidade de classificação

[4] Tradução de *NP-hard* – abreviação de "Nondeterministic Polynomial-time hard" – que implica em um problema cuja solução ótima é extremamente difícil de se obter em um tempo computacional aceitável através de algoritmos exatos, a não ser para dimensões reduzidas. Por isso, para resolver esses problemas utilizam-se algoritmos aproximados ou heurísticas [Garey & Johnson, 1979].

do conjunto original de atributos; assim a possibilidade de melhorar a capacidade de classificação do conjunto original de atributos em relação ao atributo de decisão estaria descartada. Em sua análise não se considera o atributo de decisão, que implica na eliminação de atributos condicionais redundantes entre si.

5.3 Heurística de Zhang

Zhang *et al* [2003] propuseram uma heurística, a partir da Matriz de Discernimento introduzida na teoria da TRS, para busca do Reduto Mínimo; a proposta consiste em calcular a freqüência de ocorrência de '1's na Matriz de Discernimento para todos os atributos condicionais da BD em análise, e baseado nesse resultado busca-se os atributos condicionais com freqüência mais alta para compor o Reduto Mínimo. Para eliminar a possibilidade de selecionar atributos redundantes entre si, exclui-se da Matriz de Discernimento todas as linhas que apresentarem '1' sob o último atributo selecionado e repete-se o processo, até o Reduto proposto atender o critério de avaliação definido pelo usuário. De forma simplificada, na Figura 5.3 é apresentado o algoritmo da Heurística de Zhang.

```
Entrada: BD {BD a ser analisada}, F_Avalia {função de avaliação
  de subconjuntos de atributos} e Crit_Aval {Critério de
  Avaliação}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos que
  satisfaz o critério de Avaliação definido pelo usuário}


INICIALIZE Sub_Otimo como vazio
CALCULE M_disc {Matriz de Discernibilidade de BD}
FAÇA
  CALCULE Freq_mdisc {vetor da freqüência de '1's em M_disc para
    os atributos condicionais}
  ACRESCENTE o atributo com maior freqüência em Sub_Otimo
  ELIMINE de M_disc as linhas que possue '1' no atributo com
    maior freqüência
ENQUANTO (F_Avalia(Sub_Otimo) não atender Crit_Aval)
APRESENTE Sub_Otimo
FIM
```

Figura 5.3: Algoritmo simplificado da Heurística de Zhang

Pela concepção dessa metodologia, mantêm-se os mesmos comentários observados para o cálculo do Reduto Mínimo na TRS. O diferencial está na estratégia de busca que

deixa de ser completa e passa a ser através de uma heurística, simplificando o processamento; mas em contrapartida os redutos apontados podem não ser mínimos, por dois motivos: a) o critério de parada passa a ser uma avaliação do subconjunto proposto, e não por atender todas as linhas da matriz de discernimento (*M_disc*); b) o fato de escolher atributos com maiores freqüências pode não gerar um reduto mínimo, pois é possível haver outro reduto com menor número de atributos cujas freqüências são menores.

5.4 FOCUS

Almuallim & Dietterich [1991] propuseram uma técnica de SSA denominada FOCUS, aplicável à BDs com dados binários, que consiste em buscar um subconjunto de atributos B com i elementos, tal que: $B \subseteq \{x_1,x_2,...,x_n\}$ (conjunto de todos os atributos da BD em análise); $i = 1,2,...,n$ e menor possível; e que na comparação de todos os pares de exemplos com classes diferentes, considerando somente os atributos contidos no subconjunto B, haja em todos os pares comparados pelo menos um atributo diferente entre eles. De forma simplificada, na Figura 5.4 é apresentado o algoritmo FOCUS.

```
Entrada: BD {BD a ser analisada} e n {número de atributos da BD}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos}


FAÇA i = 1 à n
  PARA TODO Sub_Otimo ⊆ {todos atributos da BD} com tamanho i
    SE todos os pares de exemplos de BD, com classes diferentes,
      apresentarem pelo menos 1 atributo, contido em Sub_Otimo,
      diferente ENTÃO APRESENTE Sub_Otimo
FIM
```

Figura 5.4: Algoritmo simplificado FOCUS

Pela concepção dessa metodologia, busca-se um subconjunto de atributos com menor número de elementos possíveis, que seja capaz de diferenciar entre si, todos os exemplos com classes diferentes; utilizando a estratégia de busca completa. Um outro questionamento levantado na concepção desse trabalho, consiste no fato de elencar um subconjunto com mínimo número de atributos possíveis e que seja capaz de distinguir, todos os pares de exemplos com classes diferentes, mas pode ocorrer que esse subconjunto proposto, não seja capaz de mostrar as semelhanças existentes entre pares de exemplos pertencentes a mesma classe.

5.5 FOCUS-2

Almuallim & Dietterich [1992] propuseram uma evolução da ferramenta FOCUS, denominada FOCUS-2. Na nova concepção cria-se um conjunto de conflitos, que consistem em vetores binários com dimensão *n*, resultantes da comparação 2 a 2 dos exemplos com classes diferentes da BD em análise. Essa comparação é feita para cada atributo do exemplo, e se forem iguais deverá resultar em zero, caso contrário 1. Como os registros comparados possuem classes diferentes, então pelo menos um dos atributos sinalizados com '1' no vetor conflito, deverá compor o subconjunto de atributos da BD em análise apontado. A partir daí, utilizando uma lógica sugerida, compõe-se os subconjuntos de atributos de tal forma, que contemple todos os vetores do conjunto de conflitos e com mínimo número de elementos possíveis. De forma simplificada, na Figura 5.5 é apresentado o algoritmo FOCUS-2.

```
Entrada: BD {BD a ser analisada} e n {número de atributos da BD}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos}

INICIALIZE Sub_Otimo como vazio
CALCULE G {Matriz de conflitos – conjunto de vetores conflitos
  que consiste na comparação dos exemplos 2 a 2, atributo a
  atributo e que pertençam à classes diferentes. Quando os
  atributos forem iguais preencher a posição do vetor com 0, caso
  contrário com 1}
FAÇA Queue = {M∅,∅}
REPITA
  DEFINA A e B, tal que MA,B = 1° elemento de Queue
  FAÇA Queue = ∅
  FAÇA OUT = B
  SELECIONE a {vetor conflito em G}, tal que esse vetor não
    aponte o(s) atributo(s) contidos em A e |Za – B| seja o mínimo
    possível, onde Za é o conjunto de atributos apontados por a
  PARA CADA x ∈ Za – B
    SE Suficiente(A ∪ {x}) ENTÃO APRESENTE (A ∪ {x})
    SENÃO INSERIR MA∪{x},OUT em Queue; FAÇA OUT = OUT ∪ {x}
FIM REPITA
FIM
```

Figura 5.5: Algoritmo simplificado FOCUS-2

Pela concepção dessa metodologia, mantêm-se os mesmos comentários observados

para a técnica FOCUS. Um diferencial, em relação à outra citada, consiste na estratégia de busca que deixa de ser completa e passa a ser através de uma heurística, assim simplificando o processamento.

5.6 Branch and Bound

Narendra & Fukunaga [1977] propuseram uma técnica de SSA denominada *Branch and Bound*, que busca o Subconjunto de Atributos Ótimo com m elementos (definido pelo usuário), a partir dos n atributos da BD. Com isso é possível dizer que a ferramenta irá selecionar uma combinação dentro das $\binom{n}{m}$ possíveis.

Esta metodologia é baseada no princípio da monotonicidade, que implica em: dada a função $J(Z_1,Z_2,...,Z_i) \mid i=\{1,2,...,n\}$, que resulta em uma métrica de avaliação para o subconjunto de atributos $(Z_1,Z_2,...,Z_i)$ da BD em análise, então $J(Z_1) \leq J(Z_1,Z_2) \leq ... \leq J(Z_1,Z_2,...,Z_n)$. Assim, a busca inicia no topo da árvore apresentada na Figura 5.6 (exemplo para uma BD com n = 6 atributos e m = 2 atributos) e segue descendo até a sua base. De forma simplificada, na Figura 5.7 é apresentado o algoritmo *Branch and Bound*.

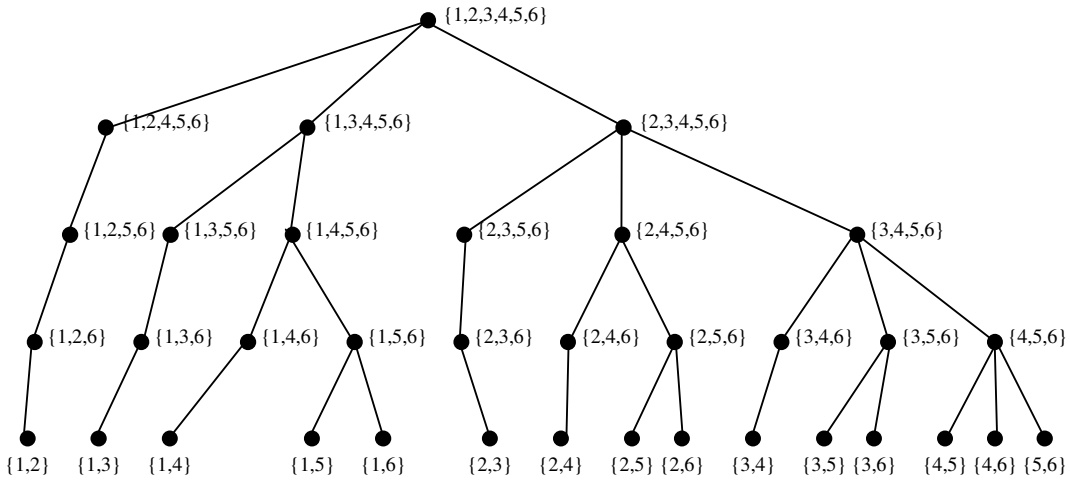


Figura 5.6: Árvore de busca da ferramenta Branch and Bound

```
Entrada: BD {BD a ser analisada}, F_Avalia {função de avaliação
  de subconjuntos de atributos} e m {número de atributos desejado
  no subconjunto ótimo de atributos definido pelo usuário}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos}

INICIALIZE B = B0 {Limite – critério de avaliação}
MONTE a árvore de busca conforme Figura 5.6, considerando m
INICIE BUSCA topo da árvore, percorrendo todos seus ramos e nós
PARA CADA Nó
  CALCULE F_Avalia(Nó em análise)
  SE F_Avalia < B ENTÃO
    NÃO ANALISE o(s) nó(s) posterior(es) {abaixo}; RETORNE para o
      nó anterior, dando seqüência na busca
  SENÃO SE Nó for base da árvore E F_Avalia > B ENTÃO
    FAÇA B assumir F_Avalia E Nó passa Sub_Otimo
FIM PARA CADA Nó
APRESENTE Sub_Otimo
FIM
```

Figura 5.7: Algoritmo simplificado Branch and Bound

Pela concepção dessa metodologia, observa-se que o autor buscou simplificar o processamento, mas a obtenção do Subconjunto de Atributos Ótimo pode não acontecer, por dois motivos: a) a metodologia forma o subconjunto considerando o número de elementos definido pelo usuário, que pode não ser o ótimo; b) o princípio da monotoniciadade pode não ocorrer na BD em análise, impedindo de avaliar subconjunto que poderiam ser o Subconjunto de Atributos Ótimo da BD em análise.

5.7 Relief

Kira and Rendell [1992, 1992b] propuseram uma técnica de SSA denominada Relief, que consiste em um algoritmo que atribui pesos para os atributos e baseado nesses valores propõe-se o subconjunto de atributos. De forma simplificada, na Figura 5.8 é apresentado o algoritmo Relief.

```
Entrada: BD {BD a ser analisada} e k_vezes {número de repetições}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos}

INICIALIZE W[i]=0 {i=1 até número de atributos da BD em análise}
LOOP PARA k = 1 até k_vezes
  SORTEIE R {um exemplo da BD em análise}
  DEFINA H {exemplo mais próximo de R com a mesma classe que R}
  DEFINA M {exemplo mais próximo de R com classe diferente que R}
  CALCULE W[i] = W[i] – diff(i,R,H) + diff(i,R,M), onde i =
    1,2,...,n e diff(i,E1,E2) é a função que calcula a diferença
    quadrática entre o atributo i dos exemplos E1 e E2
FIM LOOP
DEFINA Sub_Otimo baseado nos pesos dos atributos calculados
APRESENTE Sub_Otimo
FIM
```

Figura 5.8: Algoritmo simplificado Relief

Pela concepção dessa metodologia, observa-se que o autor buscou simplificar o processamento através da análise de exemplos sorteados aleatoriamente, assim estendendo a análise aos seus vizinhos pertencentes a mesma classe e a classes diferentes. O Subconjunto de Atributos Ótimo pode não ser alcançado devido à busca randômica; e a análise leva em consideração a comparação entre exemplos da mesma classe e de classes diferentes.

5.8 Algoritmo *hill-climbing* ou *greedy*

Kohavi & Jonh [1997] propuseram uma técnica de SSA denominada algoritmo *hill climbing* ou *greedy*, que possui a abordagem *wrapper*. De forma simplificada, na Figura 5.9 é apresentado o algoritmo *hill climbing* ou *greedy*.

```
Entrada: BD {BD a ser analisada} e F_Avalia {função de avaliação
  de subconjuntos de atributos}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos que
  satisfaz o critério de Inconsistência permitida}


INICIE Sub_Otimo com um estado inicial
REPITA Enquanto for possível GERAR subconj. a partir de Sub_Otimo
  GERE todos os subconjuntos possíveis, a partir do Sub_Otimo
  AVALIE F_Avalia(todos os subconjuntos gerados)
  FAÇA v = subconjunto que apresenta a maior F_Avalia
  SE F_Avalia(v) > F_Avalia(Sub_Otimo) ENTÃO Sub_Otimo = v
FIM REPITA
APRESENTE Sub_Otimo
FIM
```

Figura 5.9: Algoritmo simplificado *hill-climbing* ou *greedy*

Pela concepção dessa metodologia, a função de avaliação *F_Avalia* retorna a precisão do classificador gerada a partir da BD em análise, considerando os atributos contidos nos subconjuntos gerado. Dessa forma, a metodologia adota a busca *forward* ou *backward* e partir daí, avalia com *F_Avalia* todos os nós-filhos seguindo na direção do nó que apresenta a maior precisão até percorrer todo o espaço de busca. O subconjunto de atributos apontados consiste no subconjunto que apresentou a maior precisão dentro de todos os nós avaliados. O Subconjunto de Atributos Ótimo pode não ser alcançado devido à simplificação aplicada.

5.9 Consistency-Based Filter (CBF)

Liu & Setiono [1996] propuseram uma técnica de SSA, a partir do algoritmo Las Vegas, proposto por Brassard & Bratley [1996]. Esta proposição trata-se de uma aproximação probabilística, que através de *k* sorteios de subconjuntos de atributos da BD em análise, busca-se o subconjunto com menor número de atributos e que atenda a taxa de inconsistência definida pelo usuário. Essa taxa de inconsistência é calculada através da função *Calc_Incon*(), conforme descrito na subseção 3.2.1 no item 4. Na Figura 5.10 é apresentado o algoritmo simplificado CBF.

```
Entrada: k {número de repetições}, BD {BD a ser analisada}, n
  {número de atributos da BD} e Incon_Perm {Inconsistência
  permitida}
Saída: Sub_Otimo {Proposta de Subconjunto Ótimo de Atributos que
  satisfaz o critério de Inconsistência permitida}


FAÇA Cbest = n
REPITA PARA i = 1 até k
  FAÇA S = Subconjunto_Randomico(BD)
  FAÇA C = Nro_de_Atributos(S)
  SE (C < Cbest AND Calc_Incon(BD,S) <= Incon_Perm)
    FAÇA Sub_Otimo = S; FAÇA Cbest = C
FIM REPITA
APRESENTE Sub_Otimo
FIM
```

Figura 5.10: Algoritmo simplificado CBF

Pela concepção dessa metodologia, observa-se que o autor buscou simplificar o processamento através de buscas aleatórias; e o subconjunto de atributos encontrado pode não ser ótimo.

5.10 Correlation-based Feature Selection (CFS)

Hall [1999] propôs uma metodologia de SSA denominada *Correlation-based Feature Selector* (CFS), que é baseado na abordagem de filtros e classifica os subconjuntos de atributos de acordo com a correlação calculada entre os atributos condicionais e entre os atributos condicionais e a classe. Dessa forma identificando e eliminando os atributos relevantes e redundantes.

5.11 Considerações Finais

Neste capítulo apresentaram-se as metodologias de SSA pesquisadas neste trabalho, de onde se conclui:

- Redutos na TRS: utilizando a proposta de buscar redutos que mantenham a mesma capacidade de classificação que o conjunto de todos os atributos condicionais, esta metodologia fornece todos os redutos possíveis obtidos através da simplificação da Função de Discernimento; e é considerado um problema NP-difícil.
- Heurística de Zhang: consiste em uma heurística aplicada sobre a teoria da TRS; mas não há garantias quanto à otimalidade dos resultados.
- FOCUS e FOCUS-2: fornecem os mínimos subconjuntos de atributos necessários para discernir todos os pares de atributos com classes diferentes; o tempo computacional de processamento pode ser muito grande (na versão FOCUS-2 houve uma otimização nesse quesito).
- *Branch and Bound*: consiste em uma heurística baseada no princípio da monotonicidade; e não há garantias quanto à otimalidade dos resultados.
- *Relief*: através de sorteios são definidos os exemplos e vizinhanças que irão compor pesos para os atributos condicionais, e a partir desses pesos propõe-se o subconjunto de atributos importantes; nessa metodologia leva-se em consideração a análise de exemplos com a mesma classe e com classes diferentes; e não há garantias quanto à otimalidade dos resultados.
- *Hill climbing* ou *greedy*: consiste em uma metodologia que utiliza a abordagem *wrapper*, implicando em uma função de avaliação que calcula a precisão do classificador gerado a partir da BD considerando o subconjunto de

atributos em avaliação. Pela simplificação não há garantias quanto à otimalidade dos resultados.

- Algoritmo CBF: consiste em algoritmos de busca randômico; não há garantias quanto à otimalidade dos resultados.
- Algoritmo CFS: consiste em algoritmos de busca baseado no melhor valor de correlação entre atributos e a classe (atributos relevantes) e entre atributos (atributos redundantes); não há garantias quanto à otimalidade dos resultados.

Capítulo 6

Metodologia de SSA Proposta

6.1 Considerações Iniciais

Neste capítulo, expõe-se o ponto de partida e o embasamento teórico da Metodologia de SSA Proposta neste trabalho, as definições aplicadas, o mecanismo operacional e a sua aplicação de forma detalhada em uma BD-exemplo.

6.2 Ponto de Partida e Embasamento Teórico

A Metodologia de SSA Proposta baseia-se Redutos na TRS [Pawlak, 1982], que consiste em apontar subconjuntos de atributos com o menor número de elementos possíveis e que mantenham a mesma Relação de Não-discernimento do conjunto original de todos os atributos da BD. O cálculo desses Redutos é feito através da elaboração da Matriz de Discernimento, na seqüência, a Função de Discernimento e por fim a simplificação dessa função, que resulta nos Redutos da BD em análise.

Pela concepção de Redutos na TRS, verifica-se um procedimento e equacionamento para se obter todos os Redutos Mínimos possíveis de uma BD. Como foi citado é um problema NP-difícil. Não se leva em conta o atributo de decisão, logo é possível concluir que apenas os atributos redundantes são eliminados; assim levantando-se a hipótese de melhorar a capacidade de classificação do SI em relação ao atributo de decisão, esta hipótese estaria descartada.

Outras metodologias de SSA observadas foram a FOCUS e FOCUS-2. Em seus conceitos nota-se uma semelhança em relação a Redutos na TRS, que consiste na comparação de pares de exemplos, mas aplicáveis somente a pares de exemplos pertencentes a classes diferentes. Nesta situação considera-se o atributo de decisão na análise quando os exemplos pertencem a classes diferentes.

Fazendo uma analogia entre Redutos na TRS que avalia todos os pares de exemplos do SI e as metodologias FOCUS e FOCUS-2 que avalia pares exemplos pertencentes a classes diferentes, conclui-se que a filosofia da busca é a mesma diferindo apenas na seleção dos pares de exemplos avaliados. Então baseada nessa conclusão, surgiu a idéia

de realizar a implementação da metodologia de Redutos, considerando somente pares de exemplos pertencentes a classes diferentes, mas durante as avaliações criou-se um exemplo hipotético apresentado na Tabela 6.1, que consiste em um reduto que é capaz de distinguir exemplos de classes diferentes e incapaz de assemelhar[5] exemplos da mesma classe, onde observa-se que os exemplos E_1 e E_2 pertencem a mesma classe mas todos seus atributos condicionais são diferentes.

Reduto encontrado

Exemplo	Condicionais			Decis.
	A	B	C	y
E_1	'não'	'med'	2	'simples'
E_2	'sim'	'gde'	1	'simples'
E_3	'não'	'med'	1	'complex'

Tabela 6.1: Exemplo de um Reduto capaz de distinguir exemplos de classes diferentes e incapaz de assemelhar exemplos da mesma classe

Com essas colocações formulou-se a Hipótese 1, que consiste em buscar subconjuntos de atributos que sejam capazes de distinguir todos os exemplos pertencentes a classes diferentes e assemelhar todos os exemplos pertencentes a mesma classe, de tal maneira que a situação demonstrada na Tabela 6.1 não venha acontecer, pois falta um atributo condicional que possua o mesmo valor para os exemplos E_1 e E_2.

6.3 Definição adotada para Subconjunto de Atributos Ótimo

A definição adotada para Subconjunto de Atributos Ótimo de uma BD neste trabalho, consiste em um subconjunto de atributos da BD em análise, com menor número de elementos possíveis e que seja capaz, em pelo menos um atributo, distinguir todos os pares de exemplos da BD pertencentes a classes diferentes e assemelhar todos os pares de exemplos da BD pertencentes a mesma classe. Com essa definição, destaca-se que em uma BD podem existir mais de um Subconjunto de Atributos Ótimo.

[5] Capacidade de analisar dois exemplos e concluir que eles são semelhantes.

6.4 Descrição da Metodologia de SSA Proposta

A seguir apresentam-se as definições aplicadas e o mecanismo operacional da Metodologia de SSA Proposta neste trabalho.

6.4.1 Confecção da Matriz de Comparação

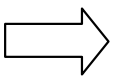
A Matriz de Comparação é uma matriz, semelhante à Matriz de Discernimento introduzida na TRS, que por definição, têm dimensões: $\binom{m}{2}$ linhas e (n + 1) colunas, onde m consiste no número de exemplos da BD. Em suas linhas encontram-se os vetores de comparação que consistem na implementação da Hipótese 1, e são gerados considerando as seguintes regras:

1. para exemplos pertencentes a classes diferentes, preencher a coluna do vetor correspondente com '1' se os atributos condicionais forem diferentes, caso contrário '0';
2. para exemplos pertencentes a mesma classe, preencher a coluna do vetor com '1' se os atributos forem iguais, caso contrário '0';
3. e para finalizar a montagem do vetor de comparação, preencher a última coluna do vetor com a freqüência de ocorrência de '1'.

Essa proposição busca registrar no vetor de comparação, os atributos necessários para distinguir ou assemelhar um par de vetores de acordo com as suas respectivas classes, ou seja, pelo menos um dos atributos apontados nesse vetor deverá compor os subconjuntos de atributos apontados. Na seqüência, como em Redutos na TRS, monta-se a Função de Comparação com a aplicação do operador lógico ***OU*** entre os atributos da célula da matriz, e o operador lógico ***E*** entre as células da matriz. Na Figura 6.1 é apresentado um exemplo de confecção da Matriz de Comparação.

BD em análise

Exemplo	Condicionais			Decis.
	A	B	C	y
E_1	'não'	'med'	2	'simples'
E_2	'não'	'gde'	2	'simples'
E_3	'não'	'med'	1	'complex'
E_4	'sim'	'gde'	1	'complex'
E_5	'sim'	'gde'	2	'simples'



matriz M_comp

(1,2)	1	0	1	2
(1,3)	0	0	1	1
(1,4)	1	1	1	3
(1,5)	0	0	1	1
(2,3)	0	1	1	2
(2,4)	1	0	1	2
(2,5)	0	1	1	2
(3,4)	0	0	1	1
(3,5)	1	1	1	3
(4,5)	0	0	1	1

Figura 6.1: Exemplo da confecção da Matriz de Comparação

No exemplo da Figura 6.1, observa-se que a relação de não-discernimento da BD exemplo, conforme definido na TRS, é U/IND({A,B,C}) = {{E_1},{E_2},{E_3},{E_4},{E_5}}, assim o reduto apontado seria {A,B,C}; considerando o atributo de decisão pode-se observar que o atributo C é capaz de classificar corretamente todos os exemplos, e sua relação de não-discernimento é U/IND({C}) = {{E_1, E_2, E_5},{E_3, E_4}}. Logo, a metodologia Redutos na TRS não conseguiria apontar o Subconjunto de Atributos Ótimo.

Por definições apresentadas, a Matriz de Discernimento é composta por vetores que são gerados a partir da comparação de todos os pares de exemplos da BD em análise, sem considerar o atributo de decisão, colocando '1' quando os atributos condicionais forem diferentes e caso contrário '0'. Na metodologia FOCUS-2, a Matriz de Conflitos segue o mesmo raciocínio, mas é aplicável a somente pares de exemplos pertencentes a classes diferentes. E na Matriz de Comparação acrescentou-se que a busca apontasse atributos condicionais necessários para assemelhar exemplos pertencentes a mesma classe, conforme a Hipótese 1 apresentada.

6.4.2 Simplificação da Matriz de Comparação

Esta etapa foi elaborada para o tratamento de BDs com muitos exemplos. Pois pela concepção da Matriz de Comparação, em BDs com aproximadamente 1400 exemplos verificam-se matrizes com mais de 1.000.000 de linhas.

A Simplificação proposta ocorre durante a confecção da Matriz de Comparação, antes de inserir um novo vetor de comparação nessa matriz. O novo vetor de comparação deverá ser comparado com todos os outros vetores existentes na matriz, com a aplicação do operador lógico ***E*** entre seus atributos condicionais. Essa comparação gera um vetor resultado que é analisado com as seguintes considerações:

1. se a freqüência de ocorrência de '1' do vetor resultado for igual à freqüência de ocorrência de '1' do vetor da Matriz de Comparação, então o vetor da matriz está contido no novo vetor e pode representá-lo, logo o novo vetor pode ser descartado;

2. se a freqüência de ocorrência de '1' do vetor resultado for igual à freqüência de ocorrência de '1' do novo vetor, então o novo vetor está contido no vetor da matriz e pode representá-lo, logo o vetor da matriz pode ser descartado. Esta varredura deve percorrer toda a Matriz de Comparação Simplificada, pois podem existir outros vetores da matriz que podem ser eliminados. E ao final da varredura, insere-se o novo vetor na matriz.

Na Tabela 6.2 é apresentado um exemplo de simplificação da Matriz de Comparação.

elementos a serem inseridos em M_comp

1	1	0	1	2
2	0	1	1	2
3	1	1	1	3
4	1	0	0	1
5	0	1	0	1

M_comp parcial passo 1

1	1	0	1	2

M_comp parcial passo 2

1	1	0	1	2
2	0	1	1	2

M_comp parcial passo 3

1	1	0	1	2
2	0	1	1	2

M_comp passo 4

2	0	1	1	2
4	1	0	0	1

M_comp passo 5 e final

4	1	0	0	1
5	0	1	0	1

Tabela 6.2: Exemplo de simplificação da Matriz de Comparação

Essa implementação consiste na aplicação da simplificação algébrica boolena sobre a Função de Comparação, de onde se conclui que, todo vetor que é igual ou está contido em outro vetor pode representá-lo, ou seja, o último vetor pode ser descartado sem que haja alteração na Função de Comparação. A demonstração é apresentada a seguir: Seja *BD_analisada* a BD em análise, $X = \{x_1, x_2, ..., x_n\}$ o conjunto de atributos condicionais de *BD_analisada*, *M_comp* a Matriz de Comparação da *BD_analisada*, os vetores v_i, v_j, v_k e $v_q \in$ *M_comp* tal que: v_i, v_j, v_k e $v_q \subset X$, $v_i \cap v_j = \varnothing$, e $v_k = v_i \cup v_j$, e a Função de Comparação $Fc = v_i \wedge v_k \wedge v_{q1} \wedge .... \wedge v_{q\alpha}$, onde $\alpha = \binom{m}{2}$ - 2, então Fc pode ser simplificada com a aplicação dos Postulados e Teoremas da Álgebra Booleana anexados no Apêndice A desta dissertação, conforme as equações (6.1) à (6.4).

$$Fc = (v_i \wedge v_k) \wedge v_{q1} \wedge .... \wedge v_{q\alpha} \qquad (6.1)$$

$$Fc = (v_i \wedge (v_i \vee v_j)) \wedge v_{q1} \wedge .... \wedge v_{q\alpha} \qquad (6.2)$$

$$Fc = (v_i(v_i + v_j)) \wedge v_{q1} \wedge .... \wedge v_{q\alpha} \qquad (6.3)$$

$$Fc = v_i \wedge v_{q1} \wedge .... \wedge v_{q\alpha} \qquad (6.4)$$

Nessa manipulação algébrica, prova-se que o vetor v_i que está contido em v_k, pode representar v_k que por conseqüência pode ser eliminado sem alterar o objeto da Função de Comparação.

6.4.3 Confecção da Matriz de Resposta

A Matriz de Resposta dispõe em suas linhas todos os Redutos possíveis da BD em análise, e terá como dimensões: (o produto da freqüência de ocorrência de '1' de todos os vetores da Matriz de Comparação Simplificada) de linhas e (n + 1) colunas, onde n é o número de atributos condicionais, e será montada conforme procedimento abaixo:

1. ordenar a Matriz de Comparação Simplificada pela freqüência de ocorrência de '1' em ordem crescente;
2. inicializar um vetor de zeros com (n + 1) colunas e denominar Matriz de Resposta parcial;
3. para cada vetor da Matriz de Comparação Simplificada, replicar a Matriz de Resposta parcial pelo número vezes igual à freqüência de ocorrência de '1' do vetor em análise; e em cada réplica inserir '1' na coluna correspondente a posição de cada '1' do vetor de comparação em análise.

Na Tabela 6.3 é apresentado um exemplo de confecção da Matriz de Resposta.

M_comp para compor M_resp

1	1	0	0	0	0	1
2	0	1	1	0	0	2
3	0	0	1	0	1	2
4	0	1	0	1	1	3

M_resp parcial passo 1

1	0	0	0	0	1

M_resp parcial passo 2

1	**1**	0	0	0	2
1	0	**1**	0	0	2

M_resp parcial passo 3

1	1	**1**	0	0	3
1	0	1	0	0	2
1	1	0	0	**1**	3
1	0	1	0	**1**	3

M_resp parcial passo 4

1	1	1	0	0	3
1	**1**	1	0	0	3
1	1	0	0	1	3
1	**1**	1	0	1	4
1	1	1	**1**	0	4
1	0	1	**1**	0	3
1	1	0	**1**	1	4
1	0	1	**1**	1	4
1	1	1	0	**1**	4
1	0	1	0	**1**	3
1	1	0	0	**1**	3
1	0	1	0	**1**	3

Tabela 6.3: Exemplo de confecção da Matriz de Resposta

Este procedimento apresentado consiste em uma forma de implementar a propriedade de distribuição da álgebra booleana na Função de Comparação. O raciocínio na forma algébrica ocorre da seguinte maneira: Seja *BD_analisada* a BD em análise,

$X = \{x_1, x_2,...., x_n\}$ o conjunto de atributos condicionais de *BD_analisada*, *M_comp* a Matriz de Comparação Simplificada de *BD_analisada*, os vetores v_i e $v_j \in$ *M_comp* tal que: $v_i = \{x_{i1}, x_{i2},..., x_{ii'}\}$, $v_j = \{x_{j1}, x_{j2},..., x_{jj'}\}$, onde *i'* e *j'* são o número de elementos dos vetores v_i e v_j, respectivamente, pelo fato da Matriz de Comparação estar simplificada têm-se $v_i \neq v_j$, $v_i \not\subset v_j$ e $v_j \not\subset v_i$, e a Função de Comparação $Fc = v_i \wedge v_j$, então *Fc* pode ser desenvolvida com a aplicação dos Postulados e Teoremas da Álgebra Booleana anexados no Apêndice A desta dissertação, conforme as equações (6.5) à (6.9).

$$Fc = v_i \wedge v_j \tag{6.5}$$

$$Fc = (x_{i1} \vee x_{i2} \vee .... \vee x_{ii'}) \wedge (x_{j1} \vee x_{j2} \vee .... \vee x_{jj'}) \tag{6.6}$$

$$Fc = (x_{i1} + x_{i2} + ... + x_{ii'})(x_{j1} + x_{j2} + ... + x_{jj'}) \tag{6.7}$$

$$Fc = x_{j1}(x_{i1} + x_{i2} + ... + x_{ii'}) + x_{j2}(x_{i1} + x_{i2} + ... + x_{ii'}) + ... + x_{jj'}(x_{i1} + x_{i2} + ... + x_{ii'}) \tag{6.8}$$

$$Fc = x_{j1}x_{i1}+x_{j1}x_{i2}+....+x_{j1}x_{ii'} + x_{j2}x_{i1}+x_{j2}x_{i2}+....+x_{j2}x_{ii'} + ... + x_{jj'}x_{i1}+x_{jj'}x_{i2}+....+x_{jj'}x_{ii'} \tag{6.9}$$

Assim a propriedade de distribuição da álgebra booleana é implementada na Função de Comparação através do mecanismo apresentado.

6.4.4 Simplificação da Matriz de Resposta

Esta etapa foi elaborada para simplificar a Matriz de Resposta, que pode ser demasiadamente grande pela sua concepção. A simplificação proposta ocorre durante a confecção da Matriz de Resposta, e utiliza o mesmo mecanismo de simplificação aplicado sobre a Matriz de Comparação. Nesta simplificação comparam-se todos os vetores da Matriz de Comparação Simplificada entre si, verificando se um contém o outro, caso seja identificado vetor que contém o outro deve ser eliminado.

Essa implementação também é baseada na aplicação da simplificação algébrica boolena sobre a Função de Comparação, de onde se conclui que, todo vetor que é igual ou está contido em outro vetor pode representá-lo, ou seja, o último vetor pode ser descartado sem que haja alteração na Função de Comparação. A demonstração é apresentada a seguir: Seja *BD_analisada* a BD em análise, $X = \{x_1, x_2,...., x_n\}$ o conjunto de atributos condicionais de *BD_analisada*, *M_resp* a Matriz de Resposta da *BD_analisada*, os vetores v_i, v_j, v_k e $v_q \in$ *M_comp* tal que: v_i, v_j, v_k e $v_q \subset X$, $v_i \cap v_j = \varnothing$, e $v_k = v_i \cap v_j$, e a Função de Comparação $Fc = v_i \vee v_k \vee v_{q1} \vee .... \vee v_{q\alpha}$, então *Fc* pode ser simplificada com a aplicação dos Postulados e Teoremas da Álgebra Booleana anexados no Apêndice A desta dissertação, conforme as equações (6.10) à (6.15).

$$Fc = (v_i \vee v_k) \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.10)$$

$$Fc = (v_i \vee (v_i \wedge v_j)) \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.11)$$

$$Fc = (v_i + (v_i v_j)) \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.12)$$

$$Fc = (v_i\,(1 + v_j)) \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.13)$$

$$Fc = (v_i\,(1)) \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.14)$$

$$Fc = v_i \vee v_{q1} \vee ....\vee v_{q\alpha} \quad (6.15)$$

Nessa manipulação algébrica, prova-se que o vetor v_i que está contido em v_k, pode representar v_k que por conseqüência pode ser eliminado sem alterar o objeto da Função de Comparação.

6.4.5 Apresentação dos Subconjuntos Obtidos

E como desfecho da metodologia, apresentam-se os vetores da Matriz de Resposta Simplificada com a menor freqüência de ocorrência de '1'. O algoritmo da metodologia proposta é apresentado no Apêndice B desta dissertação.

6.5 Aplicação da Metodologia de SSA Proposta

6.5.1 Descrição da BD-exemplo

Para a aplicação da Metodologia de SSA Proposta, de forma detalhada, elencou-se a BD-exemplo apresentada na Tabela 6.4, que possui 25 exemplos com 9 atributos condicionais (composto por somente valores discretos) e pertencentes a uma das três classes definidas.

Nro	Condicionais									Decisão
	Model	**Fuel**	**Disp**	**Weight**	**Cyl**	**Power**	**Turbo**	**Comp**	**Trans**	**Mileage**
1	USA	EFI	Medium	Light	6	High	Yes	High	Manu	High
2	Japan	EFI	Small	Medium	4	Low	No	High	Manu	High
3	Japan	EFI	Medium	Light	4	Medium	No	Medium	Manu	High
4	Japan	EFI	Small	Medium	4	High	Yes	High	Manu	High
5	Japan	2-BBL	Small	Medium	4	Low	No	Medium	Manu	High
6	Japan	2-BBL	Small	Light	4	Low	No	High	Manu	High
7	Japan	EFI	Small	Medium	4	Medium	No	High	Manu	High
8	USA	EFI	Small	Medium	4	Medium	No	High	Manu	High
9	USA	EFI	Small	Medium	4	Medium	No	High	Manu	High
10	USA	EFI	Medium	Medium	6	High	Yes	High	Auto	Medium
11	USA	EFI	Medium	Medium	6	High	No	Medium	Manu	Medium
12	USA	EFI	Medium	Medium	6	High	No	Medium	Manu	Medium
13	USA	EFI	Medium	Medium	6	High	No	Medium	Manu	Medium
14	USA	EFI	Medium	Heavy	6	High	No	High	Manu	Medium
15	USA	EFI	Medium	Medium	6	High	No	High	Manu	Medium
16	USA	EFI	Medium	Medium	6	High	No	High	Manu	Medium
17	USA	EFI	Medium	Medium	4	Medium	No	Medium	Auto	Medium
18	USA	EFI	Medium	Medium	4	Medium	No	Medium	Auto	Medium
19	USA	2-BBL	Small	Medium	4	Low	No	High	Manu	Medium
20	USA	2-BBL	Small	Medium	4	Medium	No	High	Auto	Medium
21	USA	EFI	Medium	Medium	4	High	Yes	Medium	Manu	Medium
22	USA	EFI	Medium	Medium	6	High	No	Medium	Auto	Medium
23	USA	EFI	Medium	Medium	4	High	No	Medium	Auto	Medium
24	USA	EFI	Medium	Heavy	4	High	No	Medium	Manu	Low
25	USA	EFI	Medium	Heavy	6	High	No	Medium	Auto	Low

Tabela 6.4: BD-exemplo para aplicação da Metodologia de SSA Proposta

6.5.2 Confecção da Matriz de Comparação Simplificada

Na Tabela 6.5 são apresentados alguns vetores de comparação para a BD-exemplo que irão compor a Matriz de Comparação. Esse exemplo possui $\binom{25}{2}$ = 300 vetores de comparação.

Nro	Condicionais									Decisão
	Model	**Fuel**	**Disp**	**Weight**	**Cyl**	**Power**	**Turbo**	**Comp**	**Trans**	**Mileage**
1	USA	EFI	Medium	Light	6	High	Yes	High	Manu	High
2	Japan	EFI	Small	Medium	4	Low	No	High	Manu	High
v_comp $_{1,2}$	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	

1	USA	EFI	Medium	Light	6	High	Yes	High	Manu	High
10	USA	EFI	Medium	Medium	6	High	Yes	High	Auto	Medium
v_comp $_{1,10}$	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	

1	USA	EFI	Medium	Light	6	High	Yes	High	Manu	High
24	USA	EFI	Medium	Heavy	4	High	No	Medium	Manu	Low
v_comp $_{1,24}$	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	

10	USA	EFI	Medium	Medium	6	High	Yes	High	Auto	Medium
24	USA	EFI	Medium	Heavy	4	High	No	Medium	Manu	Low
v_comp $_{10,24}$	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**1**	**1**	

12	USA	EFI	Medium	Medium	6	High	No	Medium	Manu	Medium
14	USA	EFI	Medium	Heavy	6	High	No	High	Manu	Medium
v_comp $_{12,14}$	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	

Tabela 6.5: Vetores de comparação para a BD-Exemplo

À medida que se compõe a Matriz de Comparação com os vetores calculados, o processo de Simplificação nessa matriz é aplicado conforme Tabela 6.6.

Pares Processados										
(1;2)	***v_comp*** $_{1,2}$	0	1	0	0	0	0	0	1	1
(1;3) e (1;4)	***v_comp*** $_{1,2}$	0	1	0	0	0	0	0	1	1
	v_comp $_{1,3}$	0	1	1	1	0	0	0	0	1
(1;5)...(1;10)	***v_comp*** $_{1,5}$	0	0	0	0	0	0	0	0	1
(1;11)...(1;13)	***v_comp*** $_{1,5}$	0	0	0	0	0	0	0	0	1
	v_comp $_{1,11}$	0	0	0	1	0	0	1	1	0
(1;14)...(1;20)	***v_comp*** $_{1,5}$	0	0	0	0	0	0	0	0	1
	v_comp $_{1,14}$	0	0	0	1	0	0	1	0	0

Tabela 6.6: Simplificação na Matriz de Comparação parcial da BD-exemplo

E por fim a Matriz de Comparação Simplificada é apresentada na Tabela 6.7, onde é possível verificar todos os atributos necessários para compor os subconjunto de atributos apontados. Nessa aplicação destaca-se a importância do processo de Simplificação, pois com apenas 9 vetores de comparação selecionados representa-se todo o universo de 300 vetores de comparação da BD-exemplo sem nenhuma aproximação.

	Model	Fuel	Disp	Weight	Cyl	Power	Turbo	Comp	Trans	FREQ.
1	0	0	0	0	0	0	0	0	1	1
2	0	0	0	1	0	0	0	0	0	1
3	1	1	0	0	0	0	0	0	0	2
4	1	0	0	0	0	0	0	1	0	2
5	0	1	0	0	0	1	0	0	0	2
6	0	0	0	0	1	0	0	1	0	2
7	0	0	1	0	1	1	0	0	0	3
8	1	0	1	0	1	0	1	0	0	4
9	0	0	1	0	0	1	1	1	0	4

Tabela 6.7: Matriz de Comparação Simplificada da BD-exemplo

6.5.3 Confecção da Matriz de Resposta Simplificada

Na Tabela 6.8 apresenta-se a confecção da Matriz de Resposta, onde foram processados apenas os 5 primeiros vetores da Matriz de Comparação Simplificada. Por definição essa matriz deverá possuir $(1 \times 1 \times 2 \times 2 \times 2 \times 2 \times 3 \times 4 \times 4 = 768)$ linhas.

	Model	Fuel	Disp	Weight	Cyl	Power	Turbo	Comp	Trans	FREQ.
1	0	0	0	0	0	0	0	0	**1**	1
2	0	0	0	**1**	0	0	0	0	1	2
3	**1**	0	0	1	0	0	0	0	1	3
	0	**1**	0	1	0	0	0	0	1	3
4	1	0	0	1	0	0	0	0	1	3
	1	1	0	1	0	0	0	0	1	4
	1	0	0	1	0	0	0	**1**	1	4
	0	1	0	1	0	0	0	**1**	1	4
5	1	**1**	0	1	0	0	0	0	1	4
	1	1	0	1	0	0	0	0	1	4
	1	**1**	0	1	0	0	0	1	1	5
	0	1	0	1	0	0	0	1	1	4
	1	0	0	1	0	**1**	0	0	1	4
	1	1	0	1	0	**1**	0	0	1	5
	1	0	0	1	0	**1**	0	1	1	5
	0	1	0	1	0	**1**	0	1	1	5

Tabela 6.8: Confecção da Matriz de Resposta da BD-Exemplo

À medida que se compõe a Matriz de Resposta, o processo de Simplificação nessa matriz é aplicado conforme Tabela 6.9.

	Model	Fuel	Disp	Weight	Cyl	Power	Turbo	Comp	Trans	FREQ.
1'	0	0	0	0	0	0	0	0	1	1
2'	0	0	0	1	0	0	0	0	1	2
3'	1	0	0	1	0	0	0	0	1	3
	0	1	0	1	0	0	0	0	1	3
4'	1	0	0	1	0	0	0	0	1	3
	0	1	0	1	0	0	0	1	1	4
5'	1	1	0	1	0	0	0	0	1	4
	0	1	0	1	0	0	0	1	1	4
	1	0	0	1	0	1	0	0	1	4

Tabela 6.9: Simplificação na Matriz de Resposta da BD-exemplo

E por fim a Matriz de Resposta Simplificada é apresentada na Tabela 6.10, onde é possível verificar todos os atributos necessários para compor os subconjuntos de atributos a serem apontados pela metodologia proposta. Nessa aplicação destaca-se a importância do processo de Simplificação, pois com apenas 7 vetores de comparação selecionados representa-se todo o universo de 768 vetores respostas da BD-exemplo sem nenhuma aproximação.

	Model	Fuel	Disp	Weight	Cyl	Power	Turbo	Comp	Trans	FREQ.
1	0	1	1	1	0	0	0	1	1	5
2	1	0	0	1	1	1	0	0	1	5
3	1	0	0	1	0	1	0	1	1	5
4	0	1	0	1	1	0	0	1	1	5
5	1	1	1	1	1	0	0	0	1	6
6	0	1	0	1	0	1	1	1	1	6
7	1	1	0	1	1	0	1	0	1	6

Tabela 6.10: Matriz de Resposta Simplificada da BD-exemplo

E para concluir, apontam-se todos os Subconjuntos de Atributos Mínimos encontrados pela Metodologia de SSA Proposta para a BD-Exemplo analisada e são eles:

- S_1 = {Fuel, Disp, Weight, Comp, Trans}
- S_2 = {Model, Weight, Cyl, Power, Trans}
- S_3 = {Model, Weight, Power, Comp, Trans}
- S_4 = {Fuel, Weight, Cyl, Comp, Trans}

6.6 Considerações Finais

Neste capítulo apresentaram-se as definições necessárias e mecanismos da Metodologia de SSA Proposta, que como qualquer metodologia de SSA, busca apontar Subconjuntos de Atributos Ótimos de acordo com a definição de ótimo adotada.

A metodologia apresentada surgiu de uma hipótese criada na análise da metodologia de Redutos na TRS e nas metodologias FOCUS e FOCUS-2. Assim, expõe-se que essa metodologia é semelhante a Redutos, onde se monta uma matriz de análise e resolve-se um equacionamento gerado a partir dessa matriz. Todo o equacionamento é exatamente o mesmo e resolvido sem nenhuma aproximação, sendo o diferencial a matriz montada, a qual definiu-se como Matriz de Comparação. Essa matriz é semelhante à Matriz de Discernimento introduzida na TRS, porém inseriram-se considerações sobre o atributo de decisão da BD em análise, que compreendem as metodologias FOCUS e FOCUS-2. E, adicionalmente às essas metodologias buscou-se apontar atributos necessários para assemelhar exemplos pertencentes a mesma classe. Em resumo, o subconjunto de atributos apontado pela metodologia proposta possui menor número de elementos possíveis, pode ser constituído por um subconjunto ou mais, e é capaz de distinguir qualquer par de exemplos pertencentes a classes diferentes e assemelhar qualquer par de exemplos pertencentes a mesma classe.

Com relação às definições adotadas, cita-se: a Matriz de Comparação que foi citada no parágrafo anterior, e em complemento expõe-se que cada linha dessa matriz consiste em um vetor que representa um fator da Função de Comparação e em seu conteúdo os atributos apontados estão relacionados entre si com a função booleana ***OU***; e a Matriz de Resposta é gerada através da aplicação da propriedade de distribuição da álgebra booleana na Função de Comparação, transformando a multiplicação de vetores em adição de vetores. E em complemento expõe-se que cada linha dessa matriz consiste em um vetor que representa uma parcela da Função de Comparação e em seu conteúdo os atributos apontados estão relacionados entre si com a função booleana ***E***.

E para a viabilização do processamento dessas definições citadas foram necessário implementar as Simplificações, baseadas nas simplificações da álgebra booleana, que em suma determina: se um vetor *A*, tanto da Matriz de Comparação, como da Matriz de Resposta, está contido em outro vetor *B* de sua respectiva matriz, então o vetor *A* pode representar o vetor *B* que por conseqüência poderá ser eliminado.

E por fim, com a Matriz de Resposta concluída, apresentam-se os vetores dessa matriz com menor freqüência de ‘1’ como subconjunto de atributos mínimos apontados.

Capítulo 7

Avaliação da Metodologia de SSA Proposta

7.1 Considerações Iniciais

Neste capítulo apresenta-se a aplicação da Metodologia de SSA Proposta em BDs de referência para avaliá-la em relação a outras metodologias de SSA pesquisadas. Como resultados, serão apresentados os subconjuntos de atributos apontados para a BD em análise e a precisão obtida; quanto ao tempo (custo) de processamento destaca-se que esse quesito não foi objeto na definição da metodologia, mas será apresentado.

7.2 Procedimentos para Avaliação da Metodologia de SSA Proposta

O procedimento utilizado para avaliação da metodologia proposta é descrito nos tópicos abaixo:

1. inicialmente as BDs foram pré-processadas com o intuito de eliminar os dados faltantes através da eliminação de exemplos e/ou atributos condicionais;
2. na seqüência, essas BDs foram processadas pela metodologia proposta, que foi implementada em *script Matlab Version 6.0.0.88 Release 12*, em um *notebook Core 2 Duo T5300 com 1GB de memória RAM e Windows Vista – Home Basic*; e no mesmo sistema também serão processadas pelas metodologias de Redutos na TRS, Heurística de Zhang e FOCUS;
3. essas BDs também serão processadas pelas metodologias de SSA CFS, CBF e Relief, que foram alguns critérios de avaliação de SSA utilizados por Lee [2005], através da ferramenta Weka [Witten & Frank, 2000; Silva, 2004];
4. e para finalizar, a partir dos subconjuntos de atributos encontrados nas diferentes metodologias de SSA citadas, a ferramenta Weka irá gerar um classificador através do método *C4.5* e calcular sua precisão através do método *Validação Cruzada com 10-partições estratificadas* para fornecer parâmetros para comparação entre as metodologias de SSA citadas.

7.3 Descrição das BDs referências

A seleção das BDs referências, que serão utilizadas para avaliação da Metodologia de SSA Proposta, foi feita no Repositório de Dados UCI [Asuncion & Newman, 2007], onde se levantou todas as BDs com dados discretos e utilizadas para definir classificadores. Com essas considerações, levantaram-se 21 BDs e suas características que são apresentadas na Tabela 7.1 e a descrição no Apêndice C desta dissertação.

BD em análise	Nro Exemplos	Nro Atrib. Condic.	Classes	Classes %	Dados Faltantes
Audiology (Original)	200	-	24 classes	-	Sim
Audiology (Standardized)	200	69	24 classes	-	Sim
Balance Scale	625	4	L B R	46% 8% 46%	Não
Balloons	20	4	T F	40% 60%	Não
Breast Cancer	286	9	s/Indício c/Indício	70% 30%	Sim
Car Evaluation	1728	6	unacc acc good vgood	70% 22% 4% 4%	Não
Chess (King-Rook vs. King-Pawn)	3196	36	white_win whi_nowin	52% 48%	Não
Congressional Voting Records	435	16	Democr. Republic.	61% 39%	Sim
Hayes-Roth	132	5	1 2 3	39% 39% 22%	Não
Lenses	24	4	1 2 3	17% 21% 62%	Não
Lymphography	148	18	normal metatases malign lym fibrosis	N/D	Não
MONK's Problems	432	7	0 1	62% 38%	Não
Mushroom	8124	22	edible poisonous	52% 48%	Sim
Nursery	12960	8	5 classes	-	Não
Primary Tumor	339	17	21 classes	N/D	Sim
Shuttle Landing Control	15	6	noauto auto	40% 60%	Não
Soybean (Large)	307	35	19 classes	-	Sim
Soybean (Small)	47	35	D1 D2 D3 D4	21% 21% 21% 37%	Não
SPECT Heart	267	22	0 1	50% 50%	Não
Tic-Tac-Toe Endgame	958	9	positive negative	65% 35%	Não
Trains	10	32	east west	50% 50%	Sim

Tabela 7.1: Descrição das BDs-exemplo utilizadas nesta Avaliação

7.4 Processamento e Resultado das Avaliações

Durante essa fase verificou-se que em alguns conjuntos de dados não foi possível aplicar a metodologia proposta, assim inicialmente apresentam-se esses conjuntos de dados com uma justificativa devido ao seu não processamento; na seqüência, apresentam-se os outros conjuntos de dados, relatando os tratamentos efetuados para eliminação dos dados faltantes e conflitantes durante o processamento, e por fim os subconjuntos de atributos apontados. E para concluir apresentam-se as tabelas contendo o resumo das avaliações propostas.

7.4.1 Audiology (Original)

Processamento não foi realizado devido à disposição inadequada dos dados.

7.4.2 Audiology (Standardized)

Processamento não foi concluído devido ao custo computacional que extrapolou 8 horas.

7.4.3 Balance Scale

Processamento não foi realizado, pois verificou-se que a BD possui o comportamento de uma BD com dados numérico, pois executa-se o produto dos atributos *Left-Weight* e *Left-Dist* e compara com o produto dos atributos *Right-Weight* e *Right-Dist* e em função dessa comparação define-se o comportamento da balança. E a metodologia proposta observa semelhanças e diferenças entre atributos condicionais.

7.4.4 Breast Cancer

Processamento não foi realizado por não haver dados disponíveis no repositório de dados.

7.4.5 Car Evaluation

Processamento não foi realizado, pois verificou-se que na BD existem muitos vetores pertencentes a mesma classe e com todos os atributos condicionais diferentes.

7.4.6 Chess (King-Rook vs. King-Pawn)

Processamento não foi concluído devido ao custo computacional que extrapolou 8 horas.

7.4.7 Hayes-Roth

Processamento não foi realizado, pois verificou-se que na BD existem muitos vetores pertencentes a mesma classe e com todos os atributos condicionais diferentes.

7.4.8 Lymphography

Processamento não foi realizado por não haver dados disponíveis no repositório de dados.

7.4.9 MONK's Problems

Processamento não foi realizado, pois verificou-se que na BD existem muitos vetores pertencentes a mesma classe e com todos os atributos condicionais diferentes.

7.4.10 Nursery

Processamento não foi realizado, pois verificou-se que na BD existem muitos vetores pertencentes a mesma classe e com todos os atributos condicionais diferentes.

7.4.11 Primary Tumor

Processamento não foi realizado por não haver dados disponíveis no repositório de dados.

7.4.12 Shuttle Landing Control

Processamento não foi realizado por existência de muitos dados faltantes e uma BD com pequenas dimensões.

7.4.13 Soybean (Large)

Processamento não foi concluído devido ao custo computacional que extrapolou 8 horas.

7.4.14 Tic-Tac-Toe Endgame

Processamento não foi realizado, pois verificou-se que na BD existem muitos vetores pertencentes a mesma classe e com todos os atributos condicionais diferentes.

7.4.15 Balloons

Durante a aplicação da metodologia proposta, detectou-se a seguinte particularidade: os pares de exemplos (9,16), (10,15), (11,14) e (12,13) pertencem entre si a mesma classe, porém todos os seus atributos condicionais são diferentes, e essa condição inviabiliza o processamento da metodologia proposta. Com isso, eliminaram-se os exemplos 13, 14, 15 e 16.

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 1 atributo condicional:

- S_1 = {Size}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS $\rightarrow$ S_1 = {Color, Size, Act, Age}
- Zhang $\rightarrow$ S_1 = {qualquer atributo}
- Focus $\rightarrow$ S_1 = {Size}
- CFS $\rightarrow$ S_1 = {Color, Size}
- CBF $\rightarrow$ S_1 = {Size}
- Relief $\rightarrow$ S_1 = {Size}

7.4.16 Congressional Voting Records

Inicialmente tomaram-se medidas para eliminar os dados faltantes executando: eliminação dos atributos condicionais {export-administration-act-south-africa, water-project-cost-sharing e education-spending}; e eliminação de 110 exemplos com dados faltantes.

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 9 atributos condicionais:

- S_1 = {hand_infants, adopt_budget, physician_freeze, el_salvador_aid, anti_satellite_test, immigration, synfuels_cutback, superfund_right, duty_exports}
- S_2 = {hand_infants, adopt_budget, physician_freeze, religious_schools, anti_satellite_test, immigration, synfuels_cutback, crime, duty_exports}

- S_3 = {hand_infants, adopt_budget, physician_freeze, religious_schools, anti_satellite_test, immigration, synfuels_cutback, superfund_right, crime}
- S_4 = {hand_infants, adopt_budget, physician_freeze, anti_satellite_test, mx_missile, immigration, synfuels_cutback, crime, duty_exports}
- S_5 = {hand_infants, adopt_budget, physician_freeze, anti_satellite_test, immigration, synfuels_cutback, superfund_right, crime, duty_exports}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS $\rightarrow$ S_1 = {todos atributos}
- Zhang $\rightarrow$ S_1 = {qualquer atributo}
- Focus $\rightarrow$ S_1 = {adopt_budget, physician_freeze, el_salvador_aid, religious_schools, immigration, synfuels_cutback, duty_exports}
- CFS $\rightarrow$ S_1 = {adopt_budget, physician_freeze, mx_missile, synfuels_cutback, duty_exports}
- CBF $\rightarrow$ S_1 = {physician_freeze}
- Relief $\rightarrow$ S_1 = {physician_freeze, crime, adopt_budget, synfuels_cutback, el_salvador_aid, immigration, duty_exports, mx_missile, handicapped-infants}

7.4.17 Lenses

Durante a aplicação da metodologia proposta, detectou-se a seguinte particularidade: os pares de exemplos (1,16), (1,24), (7,18), (9,24), (15,18) e (16,17) pertencem entre si a mesma classe, porém todos os seus atributos condicionais são diferentes, e essa condição inviabiliza o processamento da metodologia proposta. Com isso, eliminaram-se os registros 16, 18 e 24.

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 2 atributos condicionais:

- S_1 = {Astigmatic, Tear}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS $\rightarrow$ S_1 = {Age, Myop_Hype, Astigmatic, Tear}

- Zhang → S_1 = {qualquer atributo}
- Focus → S_1 = {Astigmatic, Tear}
- CFS → S_1 = {Tear}
- CBF → S_1 = {Age, Myop_Hype, Astigmatic, Tear}
- Relief → S_1 = {Age, Myop_Hype}

7.4.18 Mushroom

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 6 atributos condicionais:

- S_1 = {A3,A4,A5,A17,A18,A22}, S_2 = {A4,A5,A6,A7,A12,A20},
  S_3 = {A4,A5,A12,A17,A18,A22}, S_4 = {A4,A5,A12,A17,A18,A20},
  S_5 = {A4,A5,A6,A12,A17,A22}, S_6 = {A4,A5,A13,A17,A18,A22},
  S_7 = {A4,A5,A13,A17,A18,A20}, S_8 = {A4,A5,A15,A17,A18,A22},
  S_9 = {A4,A5,A7,A17,A18,A20}, S_{10} = {A4,A5,A17,A18,A19,A22},
  S_{11} = {A4,A5,A17,A18,A20,A22}, S_{12} = {A4,A5,A6,A17,A19,A22},
  S_{13} = {A3,A5,A12,A17,A18,A20}, S_{14} = {A5,A6,A7,A12,A21,A22},
  S_{15} = {A5,A12,A17,A18,A21,A22}, S_{16} = {A5,A12,A15,A17,A18,A20},
  S_{17} = {A5,A12,A17,A18,A19,A20}, S_{18} = {A5,A12,A17,A18,A20,A21},
  S_{19} = {A5,A6,A12,A17,A21,A22}, S_{20} = {A3,A5,A13,A17,A18,A20},
  S_{21} = {A5,A6,A7,A13,A21,A22}, S_{22} = {A5,A13,A17,A18,A21,A22},
  S_{23} = {A5,A13,A15,A17,A18,A20}, S_{24} = {A5,A13,A17,A18,A19,A20},
  S_{25} = {A5,A13,A17,A18,A20,A21}, S_{26} = {A5,A7,A15,A17,A18,A20},
  S_{27} = {A5,A15,A17,A18,A21,A22}, S_{28} = {A5,A15,A17,A18,A20,A22},
  S_{29} = {A5,A7,A17,A18,A19,A20}, S_{30} = {A5,A17,A18,A19,A21,A22},
  S_{31} = {A5,A17,A18,A19,A20,A22}, S_{32} = {A5,A6,A17,A19,A21,A22},
  S_{33} = {A5,A6,A7,A19,A21,A22}, S_{34} = {A5,A6,A17,A19,A20,A22},
  S_{35} = {A3,A5,A7,A17,A18,A20}, S_{36} = {A3,A5,A17,A18,A20,A22},
  S_{37} = {A5,A7,A17,A18,A20,A21}, S_{38} = {A5,A17,A18,A20,A21,A22}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS → S_1 = {A1,A2,A3,A5,A6,A7,A9,A12,A13,A14,A15,A17,A20,A21,A22}
- Zhang → S_1 = {A1, A2, A3, A5, A6, A6}
- Focus → S_1 = {A4, A5, A12, A22}

- CFS → S_1 = {A5, A7, A12, A17}
- CBF → S_1 = {A2, A3, A5, A12, A20}
- Relief → S_1 = {A5, A20, A8, A22, A19, A4}

7.4.19 Soybean (Small)

Durante o processamento verificaram-se as seguintes particularidades: os atributos condicionais {A11,A13,A14,A15,A16,A17,A18,A19,A29,A30,A31,A32,A33,A34} possuíam apenas um valor, por isso foram eliminados;

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 2 atributos condicionais:

- S_1 = {A4, A22}, S_2 = {A21, A22}, S_3 = {A22, A23} e S_4 = {A22, A28}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS → S_1 = {A1, A6, A9, A10, A24}
- Zhang → S_1 = {A1, A22}
- Focus → S_1 = {A4, A22}
- CFS → S_1 = {A2, A4, A12, A21, A22}
- CBF → S_1 = {A4, A22}
- Relief → S_1 = {A22, A21}

7.4.20 SPECT Heart

Durante a aplicação da metodologia proposta, detectou-se a seguinte particularidade: os pares de exemplos (28,50), (29,41), (29,63), (29,67), (31,49), (31,52), (31,53), (31,59), (31,60), (31,62), (31,64), (31,65), (31,66), (31,68), (31,70), (31,75), (32,49), (32,52), (32,53), (32,59), (32,60), (32,62), (32,64), (32,65), (32,66), (32,68), (32,70), (32,75) e (39,78) pertencem entre si a classes diferentes, porém todos os seus atributos condicionais são iguais, e essa condição inviabiliza o processamento da metodologia proposta. Com isso, eliminaram-se os registros 28, 29, 31, 32 e 39.

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 9 atributos condicionais:

- S_1 = {F1,F4,F6,F7,F8,F13,F18,F19,F21}

- S_2 = {F1,F4,F7,F8,F11,F13,F18,F19,F21}
- S_3 = {F4,F5,F7,F8,F11,F13,F17,F18,F21}
- S_4 = {F4,F5,F7,F8,F11,F13,F18,F19,F21}
- S_5 = {F4,F6,F7,F8,F10,F13,F18,F19,F21}
- S_6 = {F4.F7.F8.F10,F11,F13,F18,F19,F21}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS → S_1 = {F2, F3, F4, F5, F7, F8, F9, F10, F11, F13, F15, F19, F20, F22}
- Zhang → S_1 = {F1, F2, F3, F4, F5, F6, F7, F16, F21}
- Focus → S_1 = {F1, F4, F6, F7, F8, F13, F17, F21, F22}
- CFS → S_1 = {F4, F7, F8, F11, F13, F16, F17, F22}
- CBF → S_1 = {F1, F2, F4, F5, F7, F8, F11, F13, F16}
- Relief → S_1 = {F13, F8, F11, F7, F22, F16, F21, F20, F2}

7.4.21 Trains

Inicialmente tomaram-se medidas para eliminar os dados faltantes executando: eliminação dos atributos condicionais {a13, a14, a15, a16, a17, a18, a19, a20, a21, a22}; No processamento verificou-se que os atributos {a25, a30, a32} possuíam somente um valor, portanto foram eliminados também.

A relação de todos os subconjuntos de atributos apontados para a BD analisada segue abaixo contendo 3 atributos condicionais:

- S_1 = {A2,A3,A27}, S_2 = {A2,A11,A27}, S_3 = {A3,A5,A24},
  S_4 = {A3,A8,A10}, S_5 = {A3,A10,A24}, S_6 = {A3,A10,A27},
  S_7 = {A3,A10,A28}, S_8 = {A3,A10,A31}, S_9 = {A3,A24,A27},
  S_{10} = {A4,A5,A24}, S_{11} = {A5,A6,A24}, S_{12} = {A5,A11,A24},
  S_{13} = {A5,A24,A28}, S_{14} = {A5,A24,A31}, S_{15} = {A6,A10,A11},
  S_{16} = {A8,A10,A11}, S_{17} = {A10,A11,A24}, S_{18} = {A10,A11,A27},
  S_{19} = {A10,A11,A28}, S_{20} = {A10,A11,A31}, S_{21} = {A11,A24,A27} e
  S_{22} = {A23,A24,A27}

Para efetuar as comparações obtiveram-se os seguintes subconjuntos de atributos utilizando outras metodologias:

- TRS → S_1 = {A4, A10}
- Zhang → S_1 = {A10, A5, A7}
- Focus → S_1 = {A10}
- CFS → S_1 = {A10, A24}
- CBF → S_1 = {A10}
- Relief → S_1 = {A24, A1, A7}

A seguir serão apresentados os resumos das avaliações, sendo na Tabela 7.2 a comparação entre a metodologia proposta e o conjunto original de atributos; na Tabela 7.4 os resultados da competição entre a metodologia proposta e as metodologias Redutos na TRS, Heurística de Zhang e FOCUS; e na Tabela 7.4 os resultados da competição entre a metodologia proposta e as metododologias CFS, CBF e Relief. Ressalta-se que para as metodologias Heurística de Zhang e Relief, que priorizam os atributos da BD em análise, elencou-se um subconjunto de atributos considerando essa priorização e com o mesmo número de atributos obtido pela metodologia proposta.

BD em análise	Conjunto Original		Metodologia Proposta		
	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)	Tempo de process.
Balloons	4	100,0 ± 0,0	1	100,0 ± 0,0	< 1s
Congressional Voting Records	13	97,1 ± 14,2	9	97,3 ± 13,6	3 s
Lenses	4	85,0 ± 32,5	2	85,6 ± 29,4	< 1s
Mushroom	20	100,0 ± 0,0	6	100,0 ± 0,0	1h 5m
Soybean (Small)	21	98,9 ± 10,3	2	100,0 ± 0,0	1h 17m
SPECT Heart	22	73,4 ± 44,0	9	71,8 ± 42,3	8m 52s
Trains	19	80,0 ± 34,6	3	44,8 ± 56,5	5m 2s

Tabela 7.2: Comparação entre a metodologia proposta e a BD em análise ajustada

BD em análise	Metodologia Proposta		Reduto na TRS		Heurística de Zhang		FOCUS	
	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)
Balloons	1	100,0 ± 0,0	4	100,0 ± 0,0	qualquer atributo		1	100,0 ± 0,0
Congressional V.Records	9	97,3 ± 13,6	13	97,1 ± 14,2	qualquer atributo		7	94,4 ± 17,7
Lenses	2	85,6 ± 29,4	4	85,0 ± 32,5	qualquer atributo		2	85,6 ± 29,4
Mushroom	6	100,0 ± 0,0	15	100,0 ± 0,0	6	99,1 ± 6,9	4	100,0 ± 0,0
Soybean (Small)	2	100,0 ± 0,0	5	87,3 ± 29,1	2	93,1 ± 20,5	2	100,0 ± 0,0
SPECT Heart	9	71,8 ± 42,3	14	75,9 ± 41,3	9	75,5 ± 38,2	9	74,2 ± 42,4
Trains	3	44,8 ± 56,5	2	42,2 ± 58,0	3	42,8 ± 59,3	1	44,4 ± 55,6

Tabela 7.3: Comparação entre a metodologia proposta e outras metodologias semelhantes pesquisadas

BD em análise	Metodologia Proposta		CFS		CBF		Relief	
	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)	Nro Atrib. Condic.	Precisão (%)
Balloons	1	100,0 ± 0,0	2	100,0 ± 0,0	1	100,0 ± 0,0	1	100,0 ± 0,0
Congressional V.Records	9	97,3 ± 13,6	5	94,1 ± 19,1	1	94,1 ± 17,3	9	94,4 ± 17,8
Lenses	2	85,6 ± 29,4	1	76,5 ± 35,7	4	85,0 ± 32,5	2	63,1 ± 43,8
Mushroom	6	100,0 ± 0,0	4	98,1 ± 9,8	5	100,0 ± 0,0	6	99,8 ± 3,0
Soybean (Small)	2	100,0 ± 0,0	5	100,0 ± 0,0	2	100,0 ± 0,0	2	100,0 ± 0,0
SPECT Heart	9	71,8 ± 42,3	8	73,9 ± 42,8	9	73,2 ± 43,2	9	73,2 ± 42,2
Trains	3	44,8 ± 56,5	2	80,0 ± 34,6	1	44,4 ± 55,6	3	80,0 ± 34,6

Tabela 7.4: Comparação entre a metodologia proposta e outras metodologias pesquisadas

7.5 Crítica dos Resultados Obtidos

O critério de avaliação adotado aqui consiste em apontar metodologias de SSA que forneçam subconjuntos de atributos com menor número de elementos possíveis e que gerem classificadores mais precisos possíveis. Quanto ao tempo de processamento ressalta-se que não foi o foco deste trabalho.

Analisando o resumo da avaliação experimental para as BDs Balloons, Congressional V.Records, Lenses, Soybean (Small) e SPECT Heart, verifica-se que a metodologia proposta obteve um desempenho semelhante ou ligeiramente melhor que as outras elencadas para comparação.

Com relação à BDs Mushroom e Trains comentários serão feitos dentro das próximas subseções.

7.5.1 Mushroom – comentários

Analisando o resumo da avaliação experimental verifica-se que a metodologia proposta apontou um subconjunto de atributos com 6 elementos e uma precisão de 100,0 ± 0,0%; a metodologia FOCUS apontou o subconjunto S_1 = {A4, A5, A12, A22} com 4 elementos e uma precisão de 100,0 ± 0,0%; e a metodologia CBF apontou o subconjunto S_1 = {A2, A3, A5, A12, A20} com 5 elementos e uma precisão de 100,0 ± 0,0%. Observando a Tabela 7.5 e considerando os critérios da metodologia proposta, tem-se: por se tratar de dois exemplos pertencentes a mesma classe, então pelo menos um dos atributos de S = {A1, A6, A7, A10, A17, A18} deveriam compor o subconjunto de atributos mínimo.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A12	A13	A14	A15	A17	A18	A19	A20	A21	A22	Class
E1	x	s	n	t	p	f	c	n	k	e	s	s	w	w	w	o	p	k	s	u	p
E1817	x	f	g	f	f	f	c	b	h	e	k	k	n	b	w	o	l	h	y	g	p

Tabela 7.5: Exemplos E1 e E1817 da BD Mushroom

Então observa-se que estes atributos necessários não constam nos subconjuntos de atributos apontados por FOCUS e CBF, que por definição buscam apontar atributos necessários para distinguir exemplos pertencentes a classes diferentes. Em um levantamento detectou-se que a quantidade de combinações de exemplos possíveis nessa BD é da ordem de 33 milhões; e considerando o subconjunto de atributos apontado por FOCUS verifica-se 3,8% das combinações se enquadram na situação mostrada para os exemplos E1 e E1817; e para o subconjunto de atributos apontado por CBF são 5,1%.

Uma das regras do classificador gerado pelos subconjuntos de atributos apontados por FOCUS e CBF consiste em: SE A5 = 'p' OU A5 = 'f' ENTÃO Class = 'p'. Considerando a regra gerada, os exemplos E1 e E1817 passam a possuir mais um atributo condicional (A5) para assemelhá-los. O fato de um atributo condicional possuir variações de valores para exemplos pertencentes a mesma classe e esses valores não aparecerem em outras classes, pode influenciar a metodologia proposta de forma negativa.

7.5.2 Trains – comentários

Analisando o resumo da avaliação experimental verifica-se que a metodologia proposta apontou um subconjunto de atributos com 3 elementos e uma precisão de $44{,}8 \pm 56{,}5\%$; o conjunto original de atributos com 19 elementos e uma precisão de $80{,}0 \pm 56{,}5\%$; a metodologia CFS apontou o subconjunto $S_1 = \{A10, A24\}$ com 2 elementos e uma precisão de $80{,}0 \pm 34{,}6\%$; e a metodologia Relief apontou o subconjunto $S_1 = \{A1, A7, A24\}$ com 3 elementos e uma precisão de $80{,}0 \pm 34{,}6\%$.

Na Tabela 7.6 é apresentada a BD Trains que foi devidamente tratada e utilizada para avaliar a metodologia proposta. Esta BD possui apenas 10 exemplos e 19 atributos condicionais.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A23	A24	A26	A27	A28	A29	A31	Class
E1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E2	2	2	1	2	2	2	2	1	1	2	1	2	1	1	2	1	2	1	2	1
E3	2	3	1	2	1	2	3	1	1	3	1	1	1	2	1	2	2	2	2	1
E4	1	3	1	2	3	2	2	1	1	4	1	1	2	1	1	2	2	1	2	1
E5	2	2	1	2	4	2	2	2	2	5	1	2	1	1	2	1	2	1	2	1
E6	3	3	1	1	5	1	3	1	1	6	1	1	1	2	1	1	2	2	2	2
E7	2	3	1	2	4	2	3	1	1	7	1	1	1	2	1	1	2	2	2	2
E8	3	3	2	1	5	2	1	1	1	7	1	3	1	2	2	1	2	1	2	2
E9	1	3	1	2	3	2	3	1	2	8	1	2	2	2	2	1	2	1	2	2
E10	3	4	1	2	2	2	1	1	2	6	2	2	2	2	1	1	2	1	2	2

Tabela 7.6: BD Trains tratada e processada nesta avaliação experimental

Analisando a BD Trains apresentada na Tabela 7.6, conclui-se que a seguinte regra: SE A24 = 1 OU A27 = 2 ENTÃO Class = 1; é capaz de classificar todos os exemplos com precisão de 100%. Então o subconjunto de atributos ótimo para essa BD seria S_1 = {A24, A27} que não foi apontado por nenhuma das metodologias utilizadas aqui.

Entre os vários subconjuntos apontados pela metodologia proposta existe o S_9 = {A3, A24, A27}. Observando os exemplos E1 e E2 apresentados na Tabela 7.7 e considerando os critérios da metodologia proposta, tem-se: por se tratar de dois exemplos pertencentes a mesma classe, então pelo menos um dos atributos de S = {A3, A5, A8, A9, A11, A12, A23, A26} deveriam compor o subconjunto de atributos mínimo, dessa forma o subconjunto apontado não poderia ser somente S_1 = {A24, A27}.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A23	A24	A26	A27	A28	A29	A31	Class
E1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E3	2	3	1	2	1	2	3	1	1	3	1	1	1	2	1	2	2	2	2	1

Tabela 7.7: Exemplos E1 e E2 da BD Trains

7.6 Considerações Finais

Neste capítulo apresentou-se a avaliação da Metodologia de SSA Proposta, que é executada através da aplicação de BDs referências na metodologia a ser analisada e outras metodologias existentes. Então comparam-se os resultados que buscam o menor número de atributos, maior precisão nos classificadores gerados a partir dos subconjuntos apontados e por fim o tempo de processamento da metodologia. Esse último quesito não foi avaliado por não fazer parte do escopo da proposta.

O método de avaliação de SSA foi a Validação Cruzada com 10-partições estratificadas, utilizando o indutor de classificador C4.5, e as metodologias selecionadas para comparação foram: Redutos na TRS, Heurística de Zhang e FOCUS pela origem e semelhança à metodologia proposta. E, adicionalmente, conforme Lee [2005], utilizaram-se as metodologias CFS, CBF e Relief.

Nessas avaliações foi possível observar:

- as BDs a serem processadas devem ser verificadas: 1) se estão completas; 2) se em todos os pares de exemplos pertencentes a mesma classe, existe pelo um atributo condicional com o mesmo valor entre os exemplos; 3) se em todos os pares de exemplos pertencentes a classes diferentes, existe pelo um atributo condicional com o valores diferentes entre os exemplos; caso essas condições não sejam verificadas, providenciar a adequação da BD;

- o alto custo computacional que inviabilizou o processamento de algumas BDs; e conforme definição, verificou-se que o custo computacional da confecção e simplificação da Matriz de Comparação está associado ao número de exemplos; e o custo computacional da confecção e simplificação da Matriz de Resposta está associado ao número de atributos;

- a metodologia proposta não se aplica em BDs que possuem os atributos de decisão determinado, única e exclusivamente, por uma relação entre os seus atributos condicionais, como no exemplo Balance Scale, que possui o comportamento de uma BD com dados numérico, pois para determinar o atributo de decisão executa-se o produto dos atributos *Left-Weight* e *Left-Dist* e compara com o produto dos atributos *Right-Weight* e *Right-Dist*;

- a citação anterior, também se aplica ao exemplo Trains, onde o subconjunto de atributos ótimo seria $S_1 = \{A24, A27\}$, que resultaria na regra: SE A24 = 1 OU A27 = 2 ENTÃO Class = 1 com precisão de 100%. Mas a metodologia proposta inseriu A3 no subconjunto apontado;

- o fato de existir variações de valores de um atributo condicional, que classificam os exemplos em uma única e exclusiva classe, pode prejudicar a aplicação da metodologia proposta, pois se estes valores aparecem somente dentro de uma mesma classe então eles deveriam ser único. Constatação verificada no exemplo Mushroom.

Capítulo 8

Conclusão e Sugestões

A Seleção de Atributos é uma implementação importante dentro do processo de extração de conhecimentos de bases de dados, que busca melhorar o desempenho desse processo, através da identificação e eliminação de atributos, da base de dados em análise, não importantes para o conhecimento a ser extraído. Nessa linha de raciocínio, este trabalho enfocou uma modalidade de Seleção de Atributos denominada Seleção de Subconjuntos de Atributos (SSA).

Atendendo o objetivo desta dissertação, cuja idéia principal consistiu em propor uma nova metodologia de SSA, que surgiu a partir de análises executadas sobre Redutos na TRS, FOCUS e FOCUS-2. A metodologia proposta foi implementada computacionalmente utilizando as técnicas da TRS, com uma diferenciação no conceito da confecção da Matriz de Discernimento. Conforme foi exaustivamente exposto e de forma resumida, cita-se que na confecção da Matriz de Discernimento na TRS não se considera o atributo de decisão. E, na seqüência analisando a FOCUS, verificou-se que a aplicação se dava somente para pares de exemplos pertencentes a classes diferentes, assim considerando o atributo de decisão. Então, criou-se a Hipótese 1 para ser implementada e avaliada, que implica em apontar atributos condicionais que sejam capazes de distinguir exemplos pertencentes a classes diferentes e assemelhar exemplos pertencentes a mesma classe. Assim, introduziu-se o conceito de Matriz de Comparação, que é semelhante à Matriz de Discernimento, com um diferencial que consiste em: durante sua confecção, se um par de exemplos pertencer a mesma classe, os valores de ‘0’ e ‘1’ a serem inseridos na Matriz de Comparação, se invertem.

Durante a implementação da proposta, verificou-se a inviabilidade de implementação devido às dimensões das matrizes definidas. Então, durante a confecção dessas matrizes implementou-se uma simplificação, baseada na álgebra boolena, aplicável sobre a Função de Comparação, que é semelhante à Função de Discernimento na TRS. O mecanismo dessa simplificação consiste em verificar se um vetor da matriz está contido em outro vetor dessa mesma matriz, caso se confirme, elimina-se o último vetor; pois o vetor que está contido possui menos elementos e pode representar o outro vetor sem alterar as características da Função de Comparação.

Uma vez implementada a proposta, de acordo com os detalhes abordados no Capítulo 6, partiu-se para a avaliação dessa proposta através da comparação com outras metodologias existentes. As três primeiras metodologias foram Redutos na TRS, Heurística de Zhang e FOCUS, que possuem conceitos semelhantes à proposta; e utilizou-se também a CFS, CBF e Relief que possuem outras formas de concepção.

A seguir serão apresentados tópicos sobre a metodologia proposta formados durante a sua implementação e avaliação, e são eles:

- por concepção é uma metodologia aplicável às bases de dados com dados discretos, sem erros e completos, com um custo computacional que pode inviabilizar o processamento, dependendo das dimensões da BD a ser analisada;
- a efetividade da aplicação da metodologia proposta fica comprometida, para: 1) BDs que possuem seu atributo de decisão definido por relações entre os atributos condicionais; 2) BDs que possuem atributos condicionais com variações de valores para exemplos pertencentes a mesma classe, sendo que esses valores não aparecem em outras classes.

E para encerrar expõem-se sugestões para trabalho futuros, onde se recomenda implementar alternativas para contornar os problemas citados nos tópicos do parágrafo anterior.

Referências

1. Almuallim, H. & Dietterich, T.G [1991]. “Learning With Many Irrelevant Features”, in *proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91)*, vol. 2, pp. 547-552. AAAI Press, Anaheim, CA. Citado na página 35.
2. Almuallim, H. & Dietterich, T.G [1992]. “Efficient Algorithms for Identifying Relevant Features”, in p*roceedings of 9th Canadian Conference on Artificial Intelligence*, n. 92-30-03, pp. 38-45. Vancouver, BC. Citado na página 35.
3. Almuallim, H. & Diettrich, T.G. [1994]. “Learning Boolean concepts in the presence of many irrelevant features”, in *Artificial Intelligence*, vol. 69, n. 1-2, pp. 279-306. Citado na página 15.
4. Asuncion, A. & Newman, D.J. [2007]. “UCI Machine Learning Repository”. University of California, School of Information and Computer Science. Irvine, CA: (http://www.ics.uci.edu/~mlearn/MLRepository.html). Citado na página 60.
5. Baranauskas, J.A., Monard, M.C. & Batista, G.E.A.P.A [2000]. “A computational Environment for Extracting Rules from Databases”. USP - São Carlos. Citado na página 5.
6. Batista, G.E.A.P.A. [2003]. “Pré-processamento de Dados em Aprendizado de Máquina Supervisionado”, tese de doutorado em Ciências de Computação e Matemática Computacional / USP - São Carlos. Citado na página 5.
7. Blum, A.L. & Langley, P. [1997]. “Selection of Relevant Features and Examples in Machine Learning”, in *Artificial Intelligence*, vol. 97, n. 1-2, pp. 245-271. Citado nas páginas 16, 17, 18 e 21.
8. Brassard, G. & Bratley, P. [1996]. “Fundamentals of Algorithms”. Prentice Hall, New Jersey. Citado na página 41.
9. Caruana, R.A. & Freitag, D. [1994]. “How useful is relevance?”, in *AAAI Fall Symposium on Relevance*, pp. 25-29. Citado na página 17.
10. Dash, M. & Liu, H. [1997]. “Feature Selection for Classification”, in *Intelligent Data Analysis* 1, pp. 131-156. Citado na página 20.

11. Dash, M., Liu, H. & Motoda, H. [2000]. "Consistency Based Feature Selection", in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 98-109. Citado na página 16.

12. Dias, M.M. [2002]. "Parâmetros na escolha de técnicas e ferramentas de mineração de dados", in *Acta Scientiarum*, v.24, n.6, pp.1715-1725. Maringá, Brasil. Citado nas páginas 2 e 4.

13. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. [1996]. "From Data Mining to Knowledge Discovery in Databases", in *AI Magazine*, Fall 1996, pp. 37-54. Citado nas páginas 1, 2 e 4.

14. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. [1996b]. "The KDD Process for Extracting Useful Knowledge from Volumes of Data", in *Communications of the ACM*, November 1996/Vol.39, No. 11, pp. 27-34. Citado nas páginas 1 e 5.

15. Frawley, W.J., Piatetsky-Shpiro G. & Matheus, C.J. [1992]. "Knowledge Discovery in Database: An Overview", in *AI Magazine*, Fall 1992, pp.57-70. Melo Park, CA. Citado nas páginas 1 e 11.

16. Gammerman, A. [1997]. "Machine Learning: Progress e Prospects", Royal Holloway, University of London. Citado na página 2.

17. Garey, M.R. & Johnson, D.S. [1979]. "Computers and Intractability: A Guide to the Theory of NP-Completeness". W.H. Freeman. New York. Citado na página 33.

18. Gennari, J.H., Langley, P. & Fisher, D. [1989]. "Models of incremental concept formation", Artificial Intelligence. Citado na página 16.

19. Goebel, M. & Gruenwald, L. [1999]. "A survey of data mining and knowledge discovery software tools", SIGKDD Explorations, v.1, i.1, pp. 20-33. Citado na página 4.

20. Hall, M.A. [1999]. "Correlation-based Feature Selection for Machine Learning", Doctor's thesis, The University of Waikato, Hamilton, NewZealand. Citado nas páginas 10 e 42.

21. Jonh, G.H., Kohavi, R. & Pfleger, K. [1994]. "Irrelevant Features and the Subset Selection Problem", Machine Learning: Proceedings of the 11th International coference, pp. 121-129, Morgan Kaufmann Publishers, San Francisco, CA. Citado na página 16.

22. Kira, K. & Rendell, L. [1992]. "A practical approach to feature selection", in *Proceedings of th 9th International Conference on Machine Learning*, D.Sleeman & P.Edwards (eds), Morgan Kaufmann, pp. 249-256. Aberdeen, Scotland. Citado nas páginas 17 e, 39.

23. Kira, K. & Rendell, L. [1992b]. "The feature selection problem: traditional methods and new algorithm", in *Proceedings AAAI'92*, San Jose, CA. Citado na página 39.

24. Kohavi, R. & John, G.H. [1997]. "Wrappers for Feature Subset Selection", in *Journal of Artificial Intelligence*, v. 97, n. 1-2, pp. 273-324. (http://citeseer.ist.psu.edu/article/kohavi97wrappers.html). Citado nas páginas 6, 17, 22 e 40.

25. Koller, D. & Sahami, M. [1996]. "Toward Optimal Feature Selection", in *Proceedings of the 13th International Conference on Machine Learning*, pp. 284-292. Citado na página 6.

26. Komorowski, J., Pawlak, Z., Polkowski, L. &¨Skowron, A. [1998]. "Rough Sets: A Tutorial", in S.K. Pal and A. Skowron, editors, *Rough-Fuzzy Hybridization: A New Method for Decision Making*, Springer-Verlag, Singapore (in Print). (http://citeseer.ist.psu.edu/komorowski98rough.html). Citado na página 29, 30, 31 e 32.

27. Langley, P. [1994]. "Selection of Relevant Features in Machine Learning", in *Proceedings of the AAAI Fall Syposium on Relevance*. New Orleans, LA. Citado na página 19.

28. Lee, H.D. [2000]. "Seleção e Construção de *Features* Relevantes para o Aprendizado de Máquina", dissertação de mestrado em Computação e Matemática Computacional / USP - São Carlos. Citado nas páginas 2, 6 e 25.

29. Lee, H.D. [2005]. "Seleção de atributos importantes para a extração de conhecimento de bases de dados", tese de doutorado em Ciências da Computação e Matemática Computacional / USP - São Carlos. Citado nas páginas 1, 5, 6, 14, 18, 19, 21, 22, 59 e 71.

30. Liu, H. & Motoda, H. [1998]. "Feature Selection for Knowledge Discovery and Data Mining", Kluwer Academic Publishers, Massachusetts. Citado na página 14.

31. Liu, H. & Setiono, R. [1996]. "A probabilistic approach to feature selection – a filter solution", in *Proceedings of the 13th International Conference on Machine*

Learning, pp. 319-327. (http://citeseer.ist.psu.edu/321378.html) Citado nas páginas 16 e 41.

32. Liu, H. & Yu, L. [2002]. "Feature Selection for Data Mining". Department of Computer Science and Engineering, Arizona State University. Tempe, AZ. Citado na páginas14.

33. Lyman, P.; Varian, H.R. *et al* [2003], "How Much Information 2003?". (http://www2.sims.berkeley.edu/research/projects/how-much-info-2003). Citado na página 1.

34. Matheus, C.J., Chan, P.K. & Piatetsky-Shapiro, G. [1993]. "Systems for Knowledge Discovery in Databases", in *IEEE TKDE special issue on Learning & Discovery in Knowledg-Based Databases*, 1993. Waltham, MA. Citado na página 4.

35. Mitchell, T.M. [1997]. "Machine Learning", WCB/McGraw-Hill. Citado na página 3.

36. Mitra, S., Pal, S.P. & Mitra, P. [2002]. "Data Mining in Soft Computing Framework: A Survey", in *IEEE Transactions on Neural Networks*. Citado nas páginas 1 e 2.

37. Narendra, P.M. & Fukunaga, K. [1977]. "A branch and bound algorithm for feature subset selection", in *IEEE Trans. on Computer*, vol. C-26, n. 9, pp. 917 – 922. Citado na página 37.

38. Nelson, V.P., Nagle, H.T., Irwin, J.D. & Carroll, B.D. [1995]. "Digital Logic Circuit Analysis & Design", Prentice Hall, pp. 91. Upper Saddle River, New Jersey. Citado na página 81.

39. Oliveira, J.A. [2006]. "Classificação de Regiões usando Atributos de Forma e Seleção de Atributos", dissertação de mestrado em Computação Aplicada / INPE. São José dos Campos - SP. Citado na página 5.

40. Pappa, G.L. [2002]. "Seleção de Atributos utilizando Algoritmos Genéticos Multiobjetivos", dissertação de mestrado em Informática Aplicada / PUC-PR. Curitiba, PR. Citado na página 5.

41. Pawlak, Z. [1982]. "Rough sets", in *International Journal of Computer and Information Sciences*, vol. 11, no5, pp. 341-356, Plenum. New York, NY. Citado na página 29, 30, 32 e 45.

42. Pérez, A.P. & Seijas, A.R. [1996]. "An Approximation to Generic Knowledge Discovery in Database Systems". Madrid, Spain. Citado nas páginas 2 e 4.

43. Piatetsky-Shapiro, G., Matheus, C., Smyth, P. & Uthurusamy, R. [1994]. "KDD-93: Progress and Challenges in Knowledge Discovery in Databases", in *AI Magazine*, Fall 1994, pp. 77-82. Menlo Park, CA. Citado na página 5.

44. Pila, A.D. [2001], "Seleção de Atributos Relevantes para Aprendizado de Máquina Utilizando a Abordagem de Rough Sets", dissertação de mestrado em Computação e Matemática Computacional / USP - São Carlos. Citado nas páginas 3, 22, 30, 31 e 32.

45. Pyle, D. [1999]. "Data Preparation for Data Mining", Morgan Kaufmann, San Francisco, CA. Citado na página 5.

46. Rezende, S.O. [2005]. "Mineração de Dados", Mini-Curso do ENIA, São Leopoldo (RS). (http://www.addlabs.uff.br/enia_site/minicursos.htm) Citado nas páginas 1 e 2.

47. Romão, W. [2002]. "Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia", tese de doutorado em Engenharia de Produção / UFSC. Florianópolis, SC. Citado nas páginas 1 e 2.

48. Santoro, D.M. [2005]. "Sobre o Processo de Seleção de Subconjuntos de Atributos – As Abordagens Filtro e Wrapper", dissertação de mestrado em Ciência da Computação / USP – São Carlos. Citado na página 3.

49. Skowron, A. & Rauszer, C. [1992]. "The Discernibility Matrices and Functions in Information Systems", in R. Slowinski, editors, *Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory*, pp. 331-362. Kluwer Academic Publishers, Dordrecht. Citado na página 33.

50. Silva, M.P.S. [2004]. "Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka", em *ERI RJ/ES*, IV:2004, pp. 19-21. Citado na página 59.

51. Traina, C., Traina, A.J.M., Wu, L. & Faloutsos, C. [2000]. "Fast feature selection using fractal dimension, in *Proceedings of th 15th Brazilian Data Base symposium*, pp. 158-171. João Pessoa, Brazil. Citado na página 17.

52. Two Crows Corporation [2005]. "Introduction of Data Mining and Knowledge Discovery", Third Edition, Two Crows. Potomac, MD. (http://www.twocrows.com) Citado nas páginas 5 e 25.

53. Weiss, S.M. & Indurkhya, N. [1998]. "Predictive Data Mining: A Practical Guide", Morgan Kaufmann. San Francisco, CA. Citado na página 5.

54. Witten, I.H. & Frank, E. [2000]. "Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)". Morgan Kaufmann, California. (http://www.cs.waikato.ac.nz/ml/weka). Citado na página 59.

55. Yu, L. & Liu, H. [2004]. "Efficient Feature Selection via Analysis of Relevance and Redundancy", in *Journal of Machine Learning Research*, vol. 5, pp. 1205-1224. Tempe, USA. Citado na página 6 e 14.

56. Zhang, J., Wang, J., Li, D., He, H. & Sun, J. [2003]. "A New Heuristic Reduct Algorithm Base on Rough Sets Theory", in *WAIM2003*, LNCS 2762, pp. 247-253. Xi'an, P.R.China. Citado na página 34.

57. Zhou, Z.H. [2003]. "Three Perspectives of Data Mining", in *Artificial Intelligence journal*, vol. 143, n. 1, pp. 139-146. Nanjing University, China. Citado na página 5.

Apêndice A: Postulados e Teoremas de Álgebra Booleana

A seguir serão apresentados os postulados e teoremas da álgebra booleana extraídas de Nelson & *et al* [1995], utilizadas para justificação do processo de simplificação da Função de Comparação.

Expressões	Dualidade
$P2(a): a+0=a$	$P2(b): a.1=a$
$P3(a): a+b=b+a$	$P3(b): ab=ba$
$P4(a): a+(b+c)=(a+b)+c$	$P4(b): a(bc)=(ab)c$
$P5(a): a+bc=(a+b)(a+c)$	$P5(b): a(b+c)=ab+ac$
$P6(a): a+\overline{a}=1$	$P6(b): a\overline{a}=0$
$T1(a): a+a=a$	$T1(b): aa=a$
$T2(a): a+1=1$	$T2(b): a..0=0$
$T3(a): \overline{\overline{a}}=a$	
$T4(a): a+ab=a$	$T4(b): a(a+b)=a$
$T5(a): a+\overline{a}b=a+b$	$T5(b): a(\overline{a}+b)=ab$
$T6(a): ab+a\overline{b}=a$	$T6(b): (a+b)(a+\overline{b})=a$
$T7(a): ab+a\overline{b}c=ab+ac$	$T7(b): (a+b)(a+\overline{b}+c)=(a+b)(a+c)$
$T8(a): \overline{a+b}=\overline{a}\overline{b}$	$T8(b): \overline{ab}=\overline{a}+\overline{b}$
$T9(a): ab+\overline{a}c+bc=ab+\overline{a}c$	$T9(b): (a+b)(\overline{a}+c)(b+c)=(a+b)(\overline{a}+c)$
$T10(a): f(x_1,x_2,....,x_n)=x_1 f(1,x_2,....,x_n)+\overline{x}_1 f(0,x_2,....,x_n)$	
$T10(b): f(x_1,x_2,....,x_n)=[x_1+f(0,x_2,....,x_n)][\overline{x}_1+f(1,x_2,....,x_n)]$	

Apêndice B: Algoritmo da Metodologia de SSA Proposta

A seguir apresenta-se o algoritmo da Metodologia de SSA Proposta neste trabalho.

```
Entrada: BD = AttrCond ∪ AttrDec {BD em análise}
Saída: m_resp {matriz com todos Redutos possíveis de BD}
Funções internas: Soma(x) {soma todos elementos vetor x}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% MONTAGEM DA MATRIZ m_comp COM APLICAÇÃO DA SIMPLIFICAÇÃO %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

INICIALIZE nro_elem = número de linhas da BD
INICIALIZE m_comp como vazio
FAÇA PARA i = 1 até nro_elem-1
  FAÇA PARA j = i+1 até nro_elem
    SE(AttrDec_BD(i) = AttrDec_BD(j)) ENTÃO
      FAÇA vetor_aux= (AttrCond(i) == AttrCond(j))
    SENÃO
      FAÇA vetor_aux= (AttrCond(i) ~= AttrCond(j))
    FIM SE
    PARA CADA vetor vetor_k de m_comp
      SE(Soma(vetor_aux AND vetor_k)= Soma(vetor_k))
        DESCARTE o vetor_aux e saia do LOOP PARA CADA
      SENÃO SE(Soma(vetor_aux AND vetor_k)= Soma(vetor_aux))
        ELIMINE o vetor_k da matriz m_comp
      FIM SE
    FIM PARA CADA
    SE(vetor_aux não foi descartado) ENTÃO
      INSIRA vetor_aux na matriz m_comp
      REORDENE a matriz m_comp pelo número de vezes que '1'
        aparece nos vetores
    FIM SE
  FIM PARA j
FIM PARA i

CONTINUA...
```

```
...Continuação

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% MONTAGEM DA MATRIZ m_resp conjunto de redutos possíveis %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

INICIALIZE nro_mcomp = número de linhas de m_comp
INICIALIZE m_resp como vazio
FAÇA PARA i = 1 até nro_mcomp
  FAÇA m_aux = m_resp
  INSIRA 1 em todos vetores de m_resp na coluna correspondente ao
    primeiro número 1 de m_comp(i)
  FAÇA PARA os demais número 1 de m_comp(i)
    FAÇA m_aux2 = m_aux
    INSIRA 1 em todos os exemplos de m_aux2 na coluna
      correspondente ao número 1 de m_comp(i) em análise
    ACRESCENTE m_aux2 em m_resp
  END FAÇA PARA
  REORDENE a matriz m_resp pela freqüência de '1' nos vetores
  COMPARE os vetores de m_resp dois a dois, iniciando da linha 1
    até a última linha, eliminando os vetores que puderem ser
    representados pelos vetores com número de elementos menores
    ou iguais
FIM PARA i

APRESENTE m_resp
FIM
```

Apêndice C: Descrição das BDs referências utilizadas na avaliação

A seguir apresenta-se um breve descritivo para cada uma das 21 BDs pesquisadas para efetuar a avaliação experimental da Metodologia de SSA Proposta:

1. **Audiology (Original)**: incluída em 1987 no repositório de dados.
2. **Audiology (Standardized)**: incluída em 1992 no repositório de dados e consiste em uma adequação da BD Audiology (Original), e tem como intuito fornecer subsídios para avaliar a audição (atributo de decisão) em 24 classes diferentes.
3. **Balance Scale**: incluída em 1994 no repositório de dados, tendo como intuito fornecer subsídios para avaliar percepções psicológicas, quanto ao movimento esperado de uma balança.
4. **Balloons**: tem como intuito fornecer subsídios para avaliar experimentos da psicologia cognitiva, quanto ao fato de um balão estar inflado ou não. Das 4 BDs expostas utilizou-se a "small-yellow+adult-stretch".
5. **Breast Cancer**: incluída em 1988 no repositório de dados e tem como intuito fornecer subsídios para diagnosticar câncer de mama.
6. **Car Evaluation**: incluída em 1997 no repositório de dados e tem como intuito fornecer subsídios para avaliar carros de acordo com os valores de aquisição e manutenção, características técnicas como conforto e segurança.
7. **Chess (King-Rook vs. King-Pawn)**: incluída em 1989 no repositório de dados e tem como intuito fornecer subsídios para avaliar se a peças brancas de um jogo de xadrez podem ganhar o jogo ou não.
8. **Congressional Voting Records**: incluída em 1987 no repositório de dados e tem como intuito fornecer subsídios para avaliar se os congressistas americanos são republicanos ou democratas.
9. **Hayes-Roth**: incluída em 1989 no repositório de dados e tem como intuito fornecer subsídios para avaliar pessoas.

10. **Lenses**: incluída em 1990 no repositório de dados, com o intuito de fornecer subsídios para definir o tipo de lente de contato a ser recomendada para o paciente.
11. **Lymphography**: incluída em 1988 no repositório de dados e tem como intuito fornecer subsídios para avaliar uma linfografia.
12. **MONK's Problems**: incluída em 1992 no repositório de dados. Consiste em uma BD artificial criada para avaliar BDs. Para esta avaliação utilizou-se a "BD Monk-2".
13. **Mushroom**: incluída em 1987 no repositório de dados e tem como intuito fornecer subsídios para avaliar se o cogumelo é comestível ou não.
14. **Nursery**: incluída em 1997 no repositório de dados e tem como intuito fornecer subsídios para avaliar a aceitação de crianças em creches.
15. **Primary Tumor**: incluída em 1988 no repositório de dados e tem como intuito fornecer subsídios para avaliar tumores no corpo.
16. **Shuttle Landing Control**: incluída em 1988 no repositório de dados e tem como intuito fornecer subsídios para avaliar as condições para pouso automático ou manual de aeronaves.
17. **Soybean (Large)**: incluída em 1988 no repositório de dados e tem como intuito fornecer subsídios para avaliar doenças na soja.
18. **Soybean (Small)**: incluída em 1987 no repositório de dados e tem como intuito fornecer subsídios para avaliar doenças na soja.
19. **SPECT Heart**: incluída em 2001 no repositório de dados e tem como intuito fornecer subsídios para avaliar o coração humano, resultando em um parecer normal ou anormal.
20. **Tic-Tac-Toe Endgame**: incluída em 1991 no repositório de dados e tem como intuito fornecer subsídios para avaliar se o jogo Tic-Tac-Toe está vencido pelo jogador X ou não.
21. **Trains**: incluída em 1994 no repositório de dados e tem como intuito fornecer subsídios para indicar se o trem está viajando para o leste ou para o oeste. Os atributos condicionais foram denominados seqüencialmente, iniciando com 'a1' e finalizando com 'a32'.