

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

Estudos e Experimentos com Técnicas de
Seleção de Atributos Lasy

Aluno: Lincoln Pires
Matricula: 07.1.4044

Orientador: Luiz Merschmann

Ouro Preto
16 de abril de 2011

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

Estudos e Experimentos com Técnicas de Seleção de Atributos Lasy

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia I (BCC390).

Aluno: Lincoln Pires
Matricula: 07.1.4044

Orientador: Luiz Merschmann

Ouro Preto
16 de abril de 2011

Sumário

1	Introdução	1
2	Justificativa	2
3	Objetivos	3
3.1	Objetivo geral	3
3.2	Objetivos específicos	3
4	Metodologia	4
5	Cronograma de atividades	5

Lista de Figuras

Lista de Tabelas

1 Cronograma de Atividades. 5

1 Introdução

A quantidade de informações que as empresas de hoje armazenam são enormes. Este dados podem gerar informações relevantes para a organização. Como esta quantidade de informações é muito grande para se analisar, houve a necessidade por ferramentas computacionais capazes de analisar esses dados. A área da Ciência da Computação que realiza este trabalho é denominada Mineração de Dados. A Mineração de dados é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes que possam representar informações úteis. Estes padrões podem ser expressos na forma de regras, fórmulas e outras.

Na verdade, mineração de dados faz parte de um processo mais amplo denominado KDD (*Knowledge Discovery in Databases*) - processo de descoberta de conhecimento em bases de dados. Composto pelas seguintes etapas: Limpeza dos dados, Integração dos dados, Redução de dados, Transformação dos dados, Mineração, Avaliação ou Pós-processamento, Visualização dos Resultados.

Dentre as tarefas da área de mineração de dados, destaca-se a tarefa de classificação, que tem sido alvo de grande esforço de estudo e pesquisa devido à sua aplicabilidade e sucesso em diversos domínios. Desse modo, um dos grandes desafios dessa área de pesquisa é o desenvolvimento de classificadores precisos e eficientes que sejam capazes de lidar com bases de dados grandes em termos de volume e dimensão. Um aspecto importante para o bom desempenho das técnicas de classificação é a qualidade dos dados da base de treinamento. Atributos redundantes e/ou irrelevantes nas bases de dados de treinamento podem prejudicar a qualidade do classificador e, além disso, tornar o processo de construção do classificador mais lento.

O que ocorre é que geralmente os dados disponibilizados para a análise possuem informações inconsistentes, inválidas e até mesmo ruídos impossibilitando a aplicação dos algoritmos de mineração de dados.

Desta forma justifica-se uma preparação destes dados. Neste trabalho vamos focar na redução de atributos que é realizada a partir da seleção de um subconjunto dos atributos existentes de modo a manter a integridade original dos dados.

Neste trabalho propõe-se avaliar a estratégia de seleção de atributos *lazy* onde a ideia é adiar a seleção de atributos até o momento em que tivermos uma instância para ser classificada. Iremos realizar testes e análises em uma quantidade significativa de bases de dados e se possível sugerir uma melhoria nesta estratégia de seleção. Este trabalho é uma continuação do trabalho do aluno Marcus Vinicius Silva Soares intitulado "AVALIAÇÃO DE UMA ABORDAGEM LAZY DE SELEÇÃO DE ATRIBUTOS BASEADA NA MEDIDA DE CONSISTÊNCIA".

2 Justificativa

Existem algumas técnicas de seleção de atributos cuja abordagem considera cada atributo individualmente, enquanto outras avaliam subconjuntos de atributos com o objetivo de encontrar aquela que maximiza a acurácia preditiva do classificador.

Tradicionalmente, técnicas de seleção de atributos são executadas na fase de pré-processamento – ou preparação – dos dados e suas decisões são definitivas para a fase de construção do modelo ou classificação propriamente dita. Estas são classificadas como técnicas de seleção prévia (*eager*).

Diferentemente da abordagem *eager*, neste trabalho propõe-se uma avaliação e talvez uma implementação desta estratégia de seleção de atributos *lazy* onde a ideia é adiar a seleção de atributos até o momento em que tivermos uma instância para ser classificada. Nesse caso, para cada instância a ser classificada, diferentes atributos poderão ser escolhidos para a execução da classificação. Esperamos que essa nova abordagem de seleção de atributos contribua para melhorar a acurácia preditiva do classificador e identificar em quais condições ela é ou não favorável.

3 Objetivos

3.1 Objetivo geral

- O objetivo deste trabalho é analisar e testar a estratégia de seleção de atributos *lazy*. Tentaremos também sugerir uma alteração nesta forma de seleção com base na medida de correlação.

3.2 Objetivos específicos

- Testar a abordagem *lazy* de seleção de atributos utilizando a métrica de consistência com uma grande variedade de bases de dados.
- Realizar uma análise dos resultados experimentais para identificar as limitações da técnica de seleção de atributos *lazy* que utiliza a métrica consistência.
- Avaliar uma abordagem *lazy* de seleção de atributos com base na medida de correlação.

4 Metodologia

A seguinte metodologia deve ser realizada para o desenvolvimento do projeto:

1. Estudo da teoria de Mineração de Dados
2. Levantamento do estado da arte de técnicas e algoritmos de seleção de atributos.
3. Revisão bibliográfica sobre as heurísticas utilizadas pelas técnicas de seleção de atributos.
4. Estudo da técnica de seleção de atributo que será utilizada para a abordagem lazy e da abordagem lazy propriamente dita.
5. Realização de testes com uma grande variedade de base de dados.
6. Análise dos resultados obtidos com os testes realizados
7. Implementação de uma abordagem lazy de seleção de atributos com base na medida de correlação.

5 Cronograma de atividades

As atividades a serem desenvolvidas nesse projeto estão distribuídas conforme a tabela abaixo:

1. Leitura de livros, artigos e dissertações sobre mineração de dados e a etapa de pré-processamento dos dados.
2. Leitura de livros, artigos e dissertações sobre técnicas e algoritmos de seleção de atributos.
3. Revisão bibliográfica para identificação do estado da arte das heurísticas utilizadas pelas técnicas.
4. Seleção de instâncias e realização dos testes computacionais.
5. Análise dos resultados.
6. Estudo e implementação do algoritmo de técnica de seleção de atributos Correlation based Feature Selection de acordo com a abordagem lazy.
7. Elaboração da monografia para se apresentar os resultados e as implementações.

Etapa	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1	X	X	X	X						
2	X	X	X	X						
3			X	X	X					
4						X				
5							X	X		
6							X	X		
7									X	X

Tabela 1: Cronograma de Atividades.

Referências