Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

DESENVOLVIMENTO DA TÉCNICA "CONSISTENCY-BASED FEATURE SELECTION" COM ABORDAGEM LAZY

Aluno: Marcus Vinicius Silva Soares
Matricula: 07.1.4131

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
2 de outubro de 2010

Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

DESENVOLVIMENTO DA TÉCNICA "CONSISTENCY-BASED FEATURE SELECTION" COM ABORDAGEM LAZY

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 II (BCC391).

Aluno: Marcus Vinicius Silva Soares
Matricula: 07.1.4131

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
2 de outubro de 2010

Resumo

A quantidade de dados disponível em ambientes computacionais tem aumentado consideravelmente a cada dia. Os bancos de dados relacionais são responsáveis por armazenar e recuperar dados de forma eficiente. No entanto, atualmente, somente estas atividades não garantem a continuidade dos negócios. Cada vez mais é necessário que se tire um proveito maior dos dados.

Dados armazenados podem esconder diversos tipos de padrões e comportamentos relevantes que a princípio não podem ser descobertos utilizando-se linguagens de consulta. Neste contexto aconteceu o surgimento da área conhecida como mineração de dados. Processos de mineração de dados permitem a transformação de dados, uma matéria bruta, em informação e conhecimento úteis em diversas áreas de aplicação, tais como administração, finanças, saúde, educação, marketing, entre outras.

De forma simples, tarefas em mineração de dados podem ser definidas como processos automatizados de descoberta de novas informações a partir de grandes massas de dados armazenadas. O processo de descoberta de informação (conhecimento) é composto das seguintes fases: Limpeza dos dados, integração dos dados, redução de dados, transformação de dados, mineração, pós-processamento, visualização dos resultados.

O desenvolvimento de novas técnicas que aumentem a eficácia do processo de descoberta de informação trará contribuições importantes para a área de mineração de dados, a proposta em questão visa justamente isto.

Palavras-chave: Bancos de dados. Mineração de dados. Redução de dados.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Atualmente empresas e organizações estão cada vez mais coletando e armazenando grandes quantidades de dados devido à queda dos preços de meios de armazenamento e computadores. A popularização na utilização de armazém de dados, ou *data warehouse*, que são grandes bancos de dados criados para análise e suporte à decisão, tende a aumentar ainda mais a quantidade de informações disponíveis. Os métodos tradicionais de análise de dados, como planilhas e consultas SQL não são apropriados para tais volumes de dados, pois podem criar relatórios informativos sobre os dados, mas não conseguem analisar o conteúdo destes relatórios a fim de obter diversos tipos de padrões e comportamentos relevantes.

Diante deste contexto, a necessidade por ferramentas computacionais capazes de analisar esses dados motivou o surgimento da área de pesquisa e aplicação em ciência da computação conhecida como Mineração de Dados [2]. Mineração de dados é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis. Estes padrões podem ser expressos na forma de regras, fórmulas e funções, entre outras.

Mineração de dados tem como tarefa explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou seqüências temporais, para detectar relacionamentos sistemáticos entre variáveis.

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. O mesmo é composto das seguintes etapas:

1. **Limpeza dos dados:** etapa onde são eliminados ruídos e dados inconsistentes.

2. **Integração dos dados:** etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.

3. **Redução de Dados** etapa onde são selecionados os atributos considerados relevantes para as etapas seguintes.

4. **Transformação dos dados:** etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).

5. **Mineração:** etapa essencial do processo consistindo na aplicação de técnicas inteligentes afim de se extrair os padrões de interesse.

6. **Avaliação ou Pós-processamento:** etapa onde são identificados os padrões interessantes de acordo com algum critério do usuário.

7. **Visualização dos Resultados:** etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao usuário o conhecimento minerado.

O projeto proposto objetiva a pesquisa e estudos focados na etapa de redução de dados, mais especificamente nas técnicas de seleção de atributos.

2 Justificativa e relevância

A classificação é uma das tarefas mais importantes da Mineração de Dados. 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 redundates 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.

Normalmente os dados disponíveis para análise não estão num formato adequado para a etapa de mineração de dados, ou seja, é muito comum existirem bases de dados contendo ruídos, dados inconsistentes e instâncias com valores de atributos desconhecidos. Além disso, em virtude de limitações como tempo de processamento ou recursos computacionais, em muitas situações não é possível a aplicação direta dos algoritmos de mineração de dados aos dados disponíveis. Desse modo, uma fase de preparação dos dados (pré-processamento) pode ser utilizada com o intuito de melhorar a qualidade dos mesmos. Na etapa de pré-processamento podem ser realizadas transformações nos dados como: limpeza, integração, seleção, transformação e redução de dados. A redução de dados pode envolver a redução do número de instâncias, de atributos e de valores de um atributo [8].

A redução do número de atributos é realizada a partir da seleção de um subconjunto dos atributos existentes de modo a manter a integridade original dos dados. Existem algumas técnicas de seleção de atributos cuja abordagem considera cada atributo individualmente, enquanto outras avaliam subconjuntos de atributos com o objetivo de encontrar aquele que maximiza a acurácia preditiva do classificador. A técnica *Information Gain Attribute Ranking* [1, 9] avalia os atributos individualmente, já *Correlation based Feature Selection* (CFS) [3, 4] e *Consistency-based Feature Selection* [6] avaliam os subconjuntos de atributos [5].

A técnica de seleção de atributos abordada neste projeto será *Consistency-based Feature Selection*. A mesma avalia os subconjuntos de atributos e utiliza a consistência como medida de avaliação dos mesmos. A técnica em questão busca por combinações de atributos cujos valores dividam os dados em subconjuntos associados a uma classe majoritária. Normalmente, a busca privilegia subconjuntos de atributos menores com alta consistência. A medida de consistência, proposta em [6], é dada pela Equação 1.

$$Consistncia_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N}, \tag{1}$$

onde S é um subconjunto de atributos, J é o número de combinações distintas de valores dos atributos de S, |Di| é o número de ocorrências da i-ésima combinação de valores de atributos, |Mi| é a cardinalidade da classe majoritária para a i-ésima combinação de valores de atributos e N é o número total de instâncias da base de dados. Adotando alguma heuristica, a técnica *Consistency-based Feature Selection* percorre o espaço de soluções em busca de um bom subconjunto de atributos, cuja

avaliação é feita segundo a Equação 1.

Tradicionalmente, técnicas de seleção 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*). Neste trabalho, estamos propondo uma nova estratégia de seleção de atributos, onde a idéia 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. A expectativa é de que essa nova abordagem de seleção de atributos contribua para melhorar a acurácia preditiva do classificador.

A importância do tema abordado neste trabalho justifica a proposta de realização de um estudo que apresente uma avaliação sobre essa nova abordagem de seleção de atributos, a qual certamente será uma contribuição para as pesquisas da área de Mineração de Dados.

3 Objetivos

3.1 Objetivos gerais

A principal característica da abordagem lazy é adiar a seleção dos atributos relevantes ao momento em que uma nova instância for submetida ao processo de classificação, ao invés de se fazer a seleção previamente. A hipótese para essa proposta é de que conhecer os valores dos atributos da instância a ser classificada pode contribuir para a identificação dos melhores atributos para aquela instância em particular. Desse modo, para diferentes instâncias submetidas à classificação, diferentes atributos (customizados para cada instância) podem ser selecionados para a realização dessa tarefa.

Então o principal objetivo deste projeto é adaptar o método *Consistency-based Feature Selection* para realizarmos seleção de atributos de acordo com a abordagem lazy, e provar que o uso do método com esta abordagem é viável.

3.2 Objetivos específicos

- Desenvolvimento do algoritmo da técnica *Consistency-based Feature Selection* adaptada para a abordagem *lazy*.
- Implementação computacional da técnica desenvolvida.
- Realização de testes do algoritmo implementado.
- Analisar os testes feitos e estudar a viabilidade da utilização da técnica desenvolvida.

4 Metodologia

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

1. Levantamento do estado da arte de técnicas e algoritmos de seleção de atributos.
2. Revisão bibliográfica sobre as heurísticas utilizadas pelas técnicas de seleção de atributos.
3. Estudo da técnica de seleção de atributo que será adaptada para a abordagem *lazy* e da abordagem *lazy* propriamente dita.
4. Implementação computacional da seleção de atributos de forma *lazy*.
5. Seleção de instâncias para realização de experimentos computacionais.
6. Realização de estudo experimental do algoritmo proposto e implementado utilizando-se as instâncias selecionadas.
7. Redação de monografia com os resultados obtidos nestes estudos.

5 Cronograma de atividades

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

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

Atividades	**Ago**	**Set**	**Out**	**Nov**	**Dez**
1	X				
2	X	X			
3		X	X	X	
4			X	X	
5			X	X	
6				X	X
7					X

Tabela 1: Cronograma de Atividades.

Referências

[1] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. pages 149–155. In Proceedings of the International Conference on Information and Knowledge Management, 1998.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. pages 17(3):37–54. AI Magazine, 1996.

[3] Mark A. Hall. Correlation-based feature selection for machine learning. PhD thesis, Departament of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[4] Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the 17th International Conference on Machine Learning, 2000.

[5] Mark A. Hall and Geoffrey Holmes. Benchmarking attribute selection techniques for discrete class data mining. pages 15(6):1437–1447. IEEE Transactions on Knowledge and Data Engineering, November-December 2003.

[6] H. Liu and R. Setiono. A probabilistic approach to feature selection: A filter solution. pages 319–327. In Morgan Kaufmann, editor, Proceedings of the 13th International Conference on Machine Learning, 1996.

[7] S. M. Weiss and N. Indurkhya. Predictive data mining: A practical guide. Morgan Kaufmann Publishers, San Francisco, CA, 1998.

[8] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. pages 412–420. In Proceedings of the Fourteenth International Conference on Machine Learning, 1997.