Técnicas de Seleção de Atributos utilizando Paradigmas de Algoritmos Disciplina de Projeto e Análise de Algoritmos

Theo Silva Lins, Luiz Henrique de Campos Merschmann
PPGCC - Programa de Pós-Graduação em Ciência da Computação
UFOP - Universidade Federal de Ouro Preto
Ouro Preto, Minas Gerais, Brasil
email: theosl@gmail.com,luizhenrique@iceb.ufop.br

***Resumo*—O processo de descoberta de conhecimento em bases de dados – processo KDD (Knowledge Discovery in Databases) – é formado basicamente por três etapas: pré-processamento dos dados, mineração de dados e pós-processamento. Uma das possíveis tarefas no pré-processamento dos dados é a redução do número de atributos das bases de dados. Essa redução do número de atributos é realizada a partir de técnicas de seleção de atributos. Algumas técnicas de seleção de atributos utilizam um dos paradigmas de projeto de algoritmos. Será feito um estudo sobre técnicas de seleção de atributos para verificar qual paradigma de programação é adotado e será detalhada pelo menos uma dessas técnicas.**

***Keywords*-algoritmos, complexidade, mineração de dados,seleção de atributos.**

I. Introdução

Mineração de dados é o processo de descoberta de informações relevantes a partir de um grande conjunto de dados [1]. Com o grande avanço da tecnologia, um grande número de informações estão sendo guardadas de formas digitais criando-se assim várias bases de dados, com informações utéis para diversas áreas.

A mineração de dados é dividida em várias tarefas e a classificação é uma das tarefas mais importantes. Ela corresponde a uma forma de análise de dados cujo objetivo é construir modelos, a partir de um conjunto de instâncias com características e classes conhecidas, capazes de classificar novas instâncias a partir de suas características [2]. Com isso um dos grandes desafios 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 [3].

Um problema que pode ocorrer nas bases de dados é o grande número de atributos. A seleção de atributos é um processo que visa identificar e remover a maior quantidade possível de informações irrelevantes e redundantes. De maneira geral, nem todos os atributos da base de dados são necessários para discriminar a classe de maneira precisa, e incluí-los no modelo de classificação pode até mesmo gerar resultados inferiores do que seriam obtidos se eles fossem removidos da base [4].

II. Justificativas

A seleção de atributos tem recebido atenção especial em aplicações que usam datasets com muitos atributos. Exemplos:

- Processamento de texto.
- Recuperação de informação em banco de imagens.
- Bioinformática.

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. 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. [3]

A seleção de atributos quase sempre melhora a precisão de modelos em problemas reais. Aspectos relevantes:

- Simplifica modelos;
- Torna os modelos mais inteligíveis;
- Ajuda a explicar melhor um problema real;

III. Objetivos

A. Objetivo Geral

A seleção de atributos busca sempre obter uma representação reduzida de base de dados, em termos de atributos, mas que produza os mesmos (ou parecidos) resultados. Com isso melhorar o desempenho dos algoritmos de aprendizado de máquina, simplificar os modelos de predição e reduzir o custo computacional para executar esses modelos. Fornecer assim um melhor entendimento dos resultados encontrados. Então o objetivo principal é realizar um estudos sobre as técnicas de seleção com os paradigmas de projetos de algoritmos.

B. Objetivos Específicos

- Fazer um estudo das técnicas de seleção de atributos
- Encontrar quais técnicas utilizam os paradigmas de algoritmos.
- Obter as análises de complexidade dos algoritmos dessas ténicas.
- Obter um comparativo entre as técnicas de seleção de atributos em relação as análises de complexidade.

IV. Metodologia

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

1) Levantamento sobre técnicas e algoritmos de seleção de atributos.
2) Revisão bibliográfica sobre os paradigmas de projetos de algoritmos utilizadas pelas técnicas de seleção de atributos.
3) Analisar as melhores abordagens de seleção de atributos que utilizam paradigma de projetos de algoritmos.
4) Estudo das técnicas de seleção de atributos que utilizam o paradigma.
5) Implementar pelo menos uma abordagem dos paradigmas de projetos de algoritmos.
6) Fazer as análises de complexidades dos algoritmos implementados.
7) Realizar experimentos com as implementações.
8) Fazer um comparativo entre as abordagens implementadas.

V. Resultados Esperados

Um estudo comparativo em termos de complexidade, dos algoritmos utilizados na seleção de atributos que utilizaram os paradigmas de projetos de algoritmos.

Referências

[1] R. B. Pereira, "Seleção lazy de atributos para a tarefa de classificação," Master's thesis, UFF - Universidade Federal Fluminense, Brazil, Julho 2009.

[2] L. H. de Campos Merschmann, "Classificação probabilística baseada em análise de padrões," PhD Thesis, UFF - Universidade Federal Fluminense, Brazil, 2007, www.ic.uff.br/ plastino/TeseLuiz.pdf.

[3] J. O. P. Yiming Yang, "A comparative study on feature selection in text categorization." *In Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, submitted in 1997.

[4] G. H. Mark A. Hall, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge ande data engineering*, vol. 6, pp. 1437–1447, 2003, submitted in 2003.