8º Encontro Anual de Iniciação Tecnológica e Inovação
8ºEAITI
13 e 14 de setembro de 2018

MÓDULO DE OBTENÇÃO DE SUBCONJUNTO ÓTIMO PARA SOFTWARE DE SELEÇÃO DE ATRIBUTOS EM MINERAÇÃO DE DADOS

Luiz Rodolfo Machado (PIBITI/Fundação Araucária/UEPG), Rayson Bartoski Laroca dos Santos, Antônio David Viniski, Alaine Margarete Guimarães (Orientador), ~~guimaraes~~alainemg@uepg.br.

Universidade Estadual de Ponta Grossa/Departamento de Informática.

Ciências Exatas e da Terra/Ciência da Computação.
Palavras Chave: Mineração de Dados, Seleção de Atributos, dados agroindustriais

Introdução

A Mineração de Dados (MD) consiste no processo de descoberta automática de informação útil em grandes conjuntos de dados, e pode ser utilizada para predizer valores de atributos ou descrever padrões do relacionamento entre os dados.
Essa técnica permite analisar de modo eficiente bases de dados de alta dimensionalidade, ou seja, com grande quantidade de variáveis/atributos. Porém, muitas técnicas de MD apresentam queda de desempenho diante de bases com muitos atributos. É possível reduzir a dimensionalidade por meio da seleção de atributos, a qual tem o objetivo de selecionar apenas os atributos mais relevantes para a descrição do objeto de estudo.
O software WEKA é uma ferramenta de Mineração de Dados que conta com diversos algoritmos de Seleção de Atributos, os quais produzem diferentes resultados. O SynthesisFS é um software desenvolvido pela equipe coordenada pela orientadora deste projeto, o qual é capaz de fazer uma síntese desses algoritmos e organizar os atributos em um ranking de acordo com sua relevância.

Problema

O software SynthesisFS em sua versão 1.02 é capaz de fazer o ranqueamento dos atributos por ordem de relevância para o objeto de estudo. Porém, a ferramenta não é capaz de selecionar o subconjunto ótimo de atributos que deve ser utilizado para determinada tarefa de Mineração de Dados, ficando a critério do usuário a escolha dos atributos ranqueados.

Solução e Benefícios

A proposta deste trabalho foi o desenvolvimento de um módulo para o software SynthesisFS, adicionando a capacidade de selecionar o conjunto ótimo de atributos com base no ranqueamento feito pelo software.
Essa seleção é feita com base na execução iterativa de determinada tarefa de mineração de dados para diferentes subconjuntos, técnica também conhecida como *Wrapper*. Visto que testar todos os subconjuntos possíveis é computacionalmente inviável, são adotadas heurísticas como o método SFS (*Sequencial Forward Selection*), o qual consiste em começar com o subconjunto vazio e adicionar sequencialmente os atributos mais promissores.
O funcionamento do SynthesisFS e do módulo de seleção de subconjunto ótimo pelo método SFS é apresentado por meio do diagrama de fluxo de processos na Figura 1.

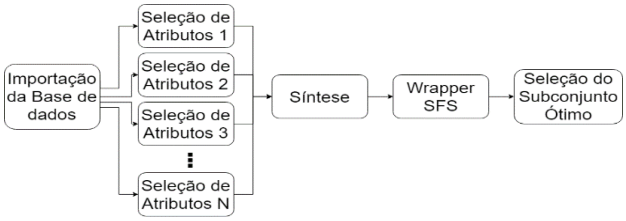

Figura 1. Diagrama de Fluxo de Processos do SynthesisFS com a incorporação do módulo de seleção de subconjunto ótimo.

Como benefício, o módulo traz a capacidade de avaliar o melhor subconjunto de atributos para determinado algoritmo de MD, com base em uma síntese de diferentes algoritmos de seleção de atributos.

Potencial de Mercado e Diferencial Competitivo

A Mineração de Dados possui diversas aplicações nos mais diferentes segmentos da indústria, do comércio, da saúde, bem como do agronegócio, especialmente devido à constante crescente da dimensão das bases de dados, e a necessidade de extrair conhecimento de nível estratégico.
O software SynthesisFS tem como mercado em potencial qualquer área que demande a extração de conhecimento em dados, principalmente de grande dimensionalidade, sendo capaz de otimizar a tarefa de MD. Seu diferencial consiste em sintetizar a contribuição de diferentes algoritmos de seleção de atributos e procurar obter o melhor resultado possível.

Considerações Finais

Tendo em vista a adição da funcionalidade de seleção de subconjunto ótimo ao software SynthesisFS, essa ferramenta possui como ponto forte a capacidade de sintetizar o conhecimento de diferentes métodos de seleção de atributos, e procurar obter o melhor resultado para o usuário. Pode-se destacar como ponto fraco o elevado custo computacional em termo de tempo do método *Wrapper*, apesar da utilização da heurística do método SFS.

Estágio de Desenvolvimento da Tecnologia

( ) Laboratório ( ) Mercado
( ) *Scale-up* (mudança de escala) (X) Protótipo

Agradecimentos

À Fundação Araucária pelo apoio financeiro.

Contato Institucional

Universidade Estadual de Ponta Grossa
Agência de Inovação e Propriedade Intelectual
www.uepg.br/agipi/
(42) 3220-3263