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
13 de junho 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

Relatório de atividades desenvolvidas apresentado 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
13 de junho de 2011

Resumo

Vivemos hoje na era da informação digital. A nossa capacidade de produzir conteúdo e de armazenar conteúdo é absurdamente maior do que a nossa capacidade de interpretar o conteúdo, extrair informação e conhecimento deste contéudo. Diante desta situação surgiu a área da ciência da computação denominada Mineração de Dados. Esta área consiste em obter informações relevantes contidas intrisicamente nestas bases de dados. Entretanto, este processo não é trivial e dependendo da base de dados analisada este processo pode ser extremamente custoso. Para tentar otimizar este processo, existem técnicas de pré-processamento de dados, dentre elas a seleção de atributos, que visa identificar atributos relevantes na base de dados. Este trabalho propõe-se a avaliar um novo paradigma de seleção de atributos - "lazy" - que adia a escolha dos atributos até o momento em que uma instância da base é submetida à classificação, diferentemente da forma tradicional, que seleciona os atributos utilizando todas as instâncias da base. Um segundo objetivo seria sugerir uma nova técnica de seleção usando este paradigma.

Palavras-chave: Banco de Dados. Mineração de Dados. Seleção de Atributos.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Alguns autores utilizam o termo mineração de dados como sinônimo de KDD (Knowledge Discovery in Database) - processo de descoberta de conhecimento em bases de dados -, outros consideram que a mineração de dados representa a etapa central desse processo maior denominado KDD. As outras etapas tratam, basicamente, do pré-processamento dos dados e pós-processamento da informação minerada (visualização e análise).

Os problemas tratados em mineração de dados são resolvidos por dois grandes grupos de soluções:

- Tarefas descritivas: têm como objetivo encontrar padrões que descrevam os dados, permitindo sua análise. As principais tarefas descritivas são: Extração de Regras de Associação e Agrupamento (Clustering).
- Tarefas preditivas: realizam inferências sobre os dados existentes para predizer o comportamento de novos dados. As principais tarefas preditivas são: Classificação e Regressão.

Para realizar estas tarefas é de extrema importância a qualidade dos dados. Neste trabalho iremos estudar e avaliar algumas formas de selecionar e medir a qualidade destes dados visando otimizar a tarefa descritiva de classificação.

2 Justificativa

Dentre os problemas tratados na mineração de dados, existe o problema da classificação, que tem como objetivo predizer o comportamento de novos dados. Esta tarefa se destaca devido à sua aplicabilidade e sucesso em diversos domínios. Para que uma técnica de classificação seja aplicada e retorne informações consistentes, a qualidade dos dados analisados é de extrema importância. Base de dados de treinamento com atributos reduntantes e/ou irrelevantes podem prejudicar a qualidade do classificador e tornar o processo de classificação muito custuso.

Desta forma justifica-se uma preparação dos dados, etapa denominada de pré-processamento dentro do processo de KDD. Nesta etapa podemos realizar a limpeza, integração, seleção, transformação e redução dos dados. A redução de dados pode envolver a redução do número de instâncias, de atributos e de valores de um atributo[Weiss e Indurkhya (1998)].

O processo de seleção de atributos visa identificar e remover informações irrelevantes e redundates na base de dados. Uma abordagem comumente utilizada para esta tarefa considera cada atributo individualmente, ordenando-os de acordo com as suas capacidades preditivas e selecionando os melhores atributos para compor o subconjunto que será utilizado pelo algoritmo de mineração de dados. Um exemplo desta abordagem é técnica é a Information Gain Attribute Ranking. Utiliza o cálculo da entropia do atributo classe antes e depois de se observar um determinado atributo, associando um valor denominado ganho de informação para o atributo analisado. Assim os atributos associados aos maiores ganhos de informação serão os selecionados.

Outra abordagem avalia os subconjuntos de atributos com o objetivo de encontrar aquele que maximize a acurácia preditiva do classificador. O problema ocorre devido ao grande número de subconjuntos que podemos ter utilizando está abordagem. Para uma base com n atributos, tem-se 2^n subconjuntos de atributos possíveis. Desta forma utilizamos métodos heurísticos, para explorar o espaço de soluções e selecionar o subconjunto de atributos. Esses métodos geralmente são gulosos no sentido de que a busca pelo espaço de soluções é sempre conduzido pela melhor escolha a cada momento. Neste caso a melhor escolha é o valor da avaliação dos subconjuntos de atributos.

Temos o método Consistency-based Feature Selection e o Correlation-based Feature Selection para a avaliação dos subconjuntos de atributos. O primeiro utiliza a consistência como medida de avaliação dos subconjuntos de atributos. Ele busca por combinações de atributos cujos valores dividam os dados em subconjuntos associados a uma classe majoritária. O segundo método leva em consideração a capacidade de discriminação dos atributos com relação às classes e o grau de correlação entre eles.

Estes métodos funcionam analisando todas as instâncias da base de dados. Pretendemos realizar alterações nestes métodos para que eles avaliem um subconjunto de instâncias da base dados - abordagem "lazy". O método Consistency-based Feature Selection já foi adaptado no trabalho [2], a princípio utilizaremos esta implementação, possivelmente podemos realizar alguma alteração. Iremos avaliar e comparar os resultados obtidos com o método tradicional - abordagem 'eager'.

Usaremos o software Weka - Waikato Environment for Knowledge Analysis (Ambiente Waikato para análise do conhecimento) para realizar as alterações e avaliações destes métodos. Este software começou a ser escrito em 1993 na Universidade Waikato e foi adquirido por uma empresa no final de 2006. O Weka encontra-se licenciado ao

abrigo da General Public License sendo portanto possível estudar e alterar o respectivo código fonte.

3 Objetivos

3.1 Objetivo geral

Através deste trabalho pretendemos descobrir onde a abordagem lazy de seleção de atributos apresenta-se vantajosa em relação a abordagem tradicional. Tentaremos também sugerir uma nova forma de seleção de atributos com a abordagem lazy para disponibilizar bases de dados reduzidas para a aplicação de algoritmos de mineração de dados, especialmente o de classificação.

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 identifcar 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. Estudo do funcionamento básico do Weka - Waikato Environment for Knowledge Analysis - e do seu código fonte.
3. Levantamento do estado da arte de técnicas e algoritmos de seleção de atributos.
4. Revisão bibliográfica sobre as heurísticas utilizadas pelas técnicas de seleção de atributos.
5. Estudo da técnica de seleção de atributo que será utiliza para a abordagem lazy e da abordagem lazy propriamente dita.
6. Realização de testes com uma grande variedade de base de dados.
7. Análise dos resultados obtidos com os testes realizados.
8. Implementação de uma abordagem lazy de seleção de atributos com base na medida de correlação.

5 Desenvolvimento

Durante este período o trabalho desenvolvido foi bastante teórico. Primeiramente o foco foi na leitura de artigos, teses e livros sobre Mineração de Dados, seus conceitos básicos e sua aplicação. O segundo passo foi o estudo de como utilizar a ferramente Weka utilizando a sua interface gráfica. Realizamos várias simulações de mineração de dados para verificar o funcionamento dos algoritmos implementados na ferramenta. Vimos o seu funcionamente para as tarefas de classificação, sumarização, regressão, clusterização e regras de associação. Analisamos também a parte de pré-processamento de dados implementada na ferramenta. Após está familiarização com a ferramenta iniciou-se o estudo do seu código fonte. Começamos pelo estudo da classe ConsistencySubsetEval que é a responsável pelo cálculo da medida de consistência de um subconjunto de atributos. A primeira abordagem foi feita estudando as classes que a ConsistencySubsetEval importava bem como as que ela herda e as interfaces que ela implementava. Esta abordagem não foi interessante devido a grande quantidade de classes importadas, resultando em uma imensa quantidade de código para se entendido. A segunda abordagem se mostrou mais eficiente. Partindo da classe ConsistencySubsetEval, foi feito o "rastreamento" do código e através de várias mensagens de texto foi possivel acompanhar o fluxo e chamada dos métodos e classes utilizados. Este estudo continua pois é fundamental entender bem o código para poder dar inicio as próximas atividades propostas neste trabalho.

6 Trabalhos Futuros

Como trabalho futuro poderia ser desenvolvido um sistema de mineração de dados que utilizasse a biblioteca do software Weka juntamente com as abordagens de seleção de atributos propostas. Este software especifico analisaria a base de dados antes da classificação e automaticamente escolheria a abordagem ideal para a seleção de atributos de forma transparente para os usuários do sistema.

7 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 da documentação do código do software Weka e estudo do seu código
3. Leitura de livros, artigos e dissertações sobre técnicas e algoritmos de seleção de atritutos.
4. Revisão bibliográfica para identificação do estado da arte das heurísticas utilizadas pelas técnicas.
5. Seleção de instâncias e realização dos testes computacionais.

6. Análise dos resultados.

7. Estudo e implementação do algoritmo de técnica de seleção de atributos Correlation based Feature Selection de acordo com a abordagem lazy.

8. 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					
3			X	X	X	X				
4						X	X	X		
5							X	X		
6							X	X		
7									X	X
8										X

Tabela 1: Cronograma de Atividades.

Referências

[1] Merschmann-Luiz. CLASSIFICAÇÃO PROBABILÍSTICA BASEADA EM ANÁLISE DE PADRÕES. Tese de Doutorado em Otimização Combinatória. Universidade Federal Fluminense. Brasil. 117pp.

[2] Marcus V. S. Soares. AVALIAÇÃO DE UMA ABORDAGEM LAZY DE SELEÇÃO DE ATRIBUTOS BASEADA NA MEDIDA DE CONSISTÊNCIA. Monografia apresentada a Universidade Federal de Ouro Preto.

[3] Rafael B. Pereira and Alexandre Plastino and Bianca Zadrozny and Luiz Merschmann and Alex A. Freitas. Seleção Lazy de Atributos - Uma Nova Perspectiva. Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados. IV Workshop em Algoritmos e Aplicações de Mineração de Dados, em conjunto com o XXIII Simpósio Brasileiro de Banco de Dados, Campinas, SP, p. 1-9, Outubro 2008

[4] Weiss, S. M. e Indurkhya, N. (1998). Predictive data mining: A practical guide. Morgan Kaufmann Publishers, San Francisco, CA.