UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
CRISTIAN COSMOSKI RANGEL DE ABREU

TÉCNICAS DE COMPUTAÇÃO PARALELA PARA MELHORAR O TEMPO DA MINERAÇÃO DE DADOS:
Uma análise de Tipos de Coberturas Florestais

Ponta Grossa

2013

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
CRISTIAN COSMOSKI RANGEL DE ABREU

TÉCNICAS DE COMPUTAÇÃO PARALELA PARA MELHORAR O TEMPO DA MINERAÇÃO DE DADOS:
Uma análise de Tipos de Coberturas Florestais

Dissertação apresentada para obtenção do título de Mestre em Computação Aplicada na Universidade Estadual de Ponta Grossa, Área de concentração: Computação para Tecnologias Agrícolas.

Orientador: Prof. Dr. Luciano José Senger

Ponta Grossa

2013

Aos meus pais

Agradecimentos

A Universidade Estadual de Ponta Grossa, pela oportunidade e estrutura cedida para os estudos e a Capes pela bolsa de estudos.

Resumo

O objetivo deste trabalho é investigar a utilização da computação paralela para reduzir o tempo de resposta da mineração de dados na agricultura. Para esse fim, uma ferramenta, chamada de *Fast Weka* foi definida e implementada. Essa ferramenta permite executar algoritmos de mineração de dados e explorar o paralelismo em computadores multi-núcleos com o uso de *threads* e em grades computacionais empregando redes *peer-to-peer*. A exploração do paralelismo ocorre por meio do paralelismo de dados inerente ao processo de validação cruzada (*folds*). A ferramenta foi avaliada por meio de experimentos de mineração de dados utilizando algoritmos de redes neurais artificiais aplicados em um conjunto de dados de tipos de coberturas florestais. A computação multi-*thread* e a computação em redes *peer-to-peer* permitem reduzir o tempo de resposta das atividades de mineração de dados. Os melhores resultados são obtidos quando empregados um número múltiplo de *threads* ou pares em relação ao número de *folds* da validação cruzada. Observou-se uma eficiência de 87% quando utilizadas 4 *threads* para 24 *folds* e 86% de eficiência também com 24 *folds* utilizando redes *peer-to-peer* com 11 pares.

Palavras-chave: Computação Paralela, Mineração de Dados, *Peer-to-Peer*, Tipos de Coberturas Florestais

Abstract

The objective of this study is investigate the use of parallel computing to reduce the response time of data mining in agriculture. For this purpose, a tool, called Fast Weka been defined and implemented. This tool allows running data mining algorithms and explore parallelism in multi-core computers with the use of threads and in computational grids employing peer-to-peer networks. The exploration of parallelism occurs through the data parallelism inherent to the process of cross-validation (folds). The tool was evaluated through experiments using artificial neural networks data mining algorithms applied to a data set of forest cover types. The multi-thread computing and computing on peer-to-peer networks allow to reduce the response time of data mining activities. The best results are achieved when employed a multiple number of threads or pairs in the number of folds of cross validation. It was observed an efficiency of 87% when used 4 threads to 24 folds and 86% efficiency also in peer-to-peer networks using 24 folds with 11 pairs.

Keywords: Parallel Computing, Data Mining, Peer-to-Peer, Forest Cover Types

Sumário

Capítulo

1

Introdução

O objetivo principal deste trabalho é investigar a utilização da computação paralela afim de melhorar o tempo de resposta das atividades de mineração de dados agrícolas, avaliando e comparando os resultados obtidos nas diferentes abordagens.

Capítulo

2

Revisão da Literatura

El-Telbany, Warda e El-Borahy (2006) desenvolveram um trabalho utilizando MD no qual o objetivo foi desenvolver modelos para classificação de doenças do arroz egípcio. Um dos algoritmos de aprendizagem utilizado foi a RNA. A RNA foi construída e treinada utilizando uma configuração de 52 entradas, 33 neurônios na camada oculta, 5 saídas, taxa de aprendizagem de 0.3, momento de 0.2 e 500 iterações. O modelo obtido para a previsão de doenças de arroz atingiu um índice de acerto de 96,4% para o conjunto de dados de teste. Este resultado demonstra a grande eficiência da aplicação de RNAs.

Blackard e Dean (1999) realizaram a comparação entre RNA e analise discriminante para criação de classificadores para tipos de coberturas florestais a partir de variáveis cartográficas. O RNA construída utilizou as configurações de 54 entradas, 120 neurônios na camada oculta, 7 classes de tipos de coberturas florestais, com uma taxa de aprendizagem de 0.05, taxa de momento de 0.5 e 1000 interações. Para obter estas configurações para a RNA foram realizados 56 analises diferentes demandando de cerca de 56 horas para cada analise. Após a comparação das técnicas, as RNAs obtiveram uma maior precisão, chegando a 70,58%.

Gradecki (2002) observou que quando o processo de aprendizado de um modelo por meio de algoritmos de aprendizagem é iniciado, este com um grande conjunto de dados, ou com um alto número de repetições, demanda de um alto custo computacional e tempo

de execução. Por meio destas observações, Guimarães desenvolveu um aplicativo para distribuir o processamento da construção de seu modelo, por meio do qual o processamento poderia ser realizado por vários computadores, utilizando o algoritmo de aprendizagem Algoritmos Genéticos (AG). Este aplicativo obteve bons resultados conseguindo reduzir seu tempo de execução estimado para construção do modelo de 1450 horas para 84 horas.

Senger, Souza e Foltran (2011) aplicaram uma ferramenta de MD em paralelo para construção de um modelo de classificação utilizando RNA para produção de soja, afim de observar a relação existente entre os atributos químicos do solo e a produção. Utilizando apenas um computador eles reduziram o tempo de processamento de 280 para 80 segundos. Esses resultados demostraram que a utilização de técnicas de computação paralela podem melhorar significativamente o tempo de resposta das atividades de mineração.

Capítulo

3

Materiais e Métodos

3.1 Considerações Iniciais

A partir da classe com menor representatividade na base de dados, classe 4 - Algodão Americano/Salgueiro com 2.747 observações, sendo todos os dados selecionados, foram extraídas das outras seis classes de tipos de cobertura florestal 2.747 observações, selecionadas aleatoriamente com o auxilio do software R, totalizando o conjunto de teste com 19.229 observações, conforme a tabela 1.

Tabela 1: Observações Selecionadas

Tipo de Cobertura Florestal	Total de Observações	Observações Selecionadas	Porcentagem por Tipo de Cobertura
Classe 1	211.840	2747	1,29%
Classe 2	283.301	2747	0,97%
Classe 3	35.754	2747	7,68%
Classe 4	2.747	2747	100%
Classe 5	9.493	2747	28,94%
Classe 6	17.367	2747	15,82%
Classe 7	20.510	2747	13,39%
Total	**581.012**	**19.229**	**3,31%**

Fonte: O autor

Capítulo

4

Resultados e Discussões

De maneira geral, a medida que aumenta-se o número de pares também aumenta a eficiência (Tabela 2). O *rank* 0, que não realiza processamento, começa a representar uma porcentagem cada vez menor para o calculo da eficiência. Considerando 3 pares, o *rank* 0 representa 33% do calculo da eficiência, já considerando 5 pares o mesmo representa 20%, assim a medida que aumentam-se os pares a eficiência aumenta. Está melhoria ocorre até certo ponto, pois a comunicação entre processos aumenta e medida que adicionamos novos pares. É importante ressaltar que esta particularidade ocorre devido ao *rank* 0 não processar nenhuma atividade e participar do calculo de eficiência, e com o aumento do números de pares a eficiência diminua devido as taxas de comunicação. Se desconsiderado o *rank* 0, a eficiência observada inicia elevada e na medida que aumentam-se os pares a eficiência deve diminuir devido à comunicação entre processos da rede.

Tabela 2: Resumo dos resultados modo P2P - 10 *Folds*

Grupo	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Experimento	Sequencial	3-Pares	5-Pares	7-Pares	9-Pares	11-Pares
Média	301,92	146,07	91,22	61,44	60,98	31,97
Desvio Padrão	1,05	0,88	1,10	0,69	0,75	0,82
SpeedUp		2,07	3,31	4,91	4,95	9,44
Eficiência		0,69	0,66	0,70	0,55	0,86

Fonte: O autor

Utilizando as técnicas de mineração de dados, RNA e o método de validação cruzada, aplicados ao conjunto de dados de tipos de coberturas florestais foram obtidos os resultados presentes na tabela 3, expressos pelos índices de acerto, erro de classificação e índice Kappa.

Tabela 3: Precisão do Classificador

Resultado	10 *Folds*	24 *Folds*
Índice de Acerto	15833 - 82.33%	15921 - 82.79%
Índice de Erro	3396 - 17.66%	3308 - 17.20%
Índice Kappa	0.794	0.7993

Fonte: O autor

Observando os percentuais de acerto na utilização das duas diferentes configuração de *folds*, não se verifica uma grande diferença entre os resultados. Com 24 *folds* o classificador acertou 88 instancias a mais que utilizando 10 *folds*, que corresponde cerca de 0,45% a mais de precisão.

Blackard e Dean (1999), utilizando os mesmo parâmetros da RNA, obtiveram uma precisão de 70,58% com o modelo, porém não pode-se dizer que os resultados obtidos neste trabalho são melhores que os observador por Blackard e Dean, devido aos fatores: seleção de dados e método de validação empregado.

Além do percentual de acerto, também foi observado o coeficiente estatístico denominado índice Kappa ou Estatística K, definido como uma medida de concordância em escalas nominais. Neste contexto de classificação, verificando os altos valores do índice Kappa, podemos verificar o elevado nível de concordância entre a classificação do modelo e a classificação de referência, ou seja, o quão os dois estão de acordo quanto à classificação, novamente sendo observado um maior valor para 24 *folds*.

Observando a área ROC, tabela 4, que retrata o desempenho de um classificador sem levar em conta os custos de distribuição ou de classe de erro, os resultados são muito expressivos, tendo em vista que todos os valores para área ROC estão superiores a 0.9.

Tabela 4: Detalhes da Precisão por Classe

Tipo de Cobertura	Área Roc		Precisão	
Florestal	10 *Folds*	24 *Folds*	10 *Folds*	24 *Folds*
Classe 1	0.941	0.942	0.749	0.761
Classe 2	0.924	0.925	0.687	0.704
Classe 3	0.963	0.965	0.805	0.798
Classe 4	0.995	0.996	0.927	0.926
Classe 5	0.987	0.985	0.876	0.861
Classe 6	0.973	0.974	0.776	0.798
Classe 7	0.994	0.994	0.927	0.929
Média	**0.968**	**0.969**	**0.821**	**0.825**

Fonte: O autor

Capítulo

5

Conclusões

Texto de conclusões

REFERÊNCIAS

BLACKARD, J. A.; DEAN, D. J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, v. 24, p. 131–151, 1999.

EL-TELBANY, M.; WARDA, M.; EL-BORAHY, M. Mining the classification rules for egyptian rice diseases. *The International Arab Journal of Information Technology*, v. 3, p. 303 – 306, 2006.

GRADECKI, J. D. *Mastering Jxta: Building Java Peer-to-Peer Applications*. Tese (Doutorado), New York, NY, USA, 2002.

SENGER, L. J.; SOUZA, M. A.; FOLTRAN, D. C. J. Towards a peer-to-peer framework for parallel and distributed computing. In: *Computer Architecture and High Performance Computing (SBAC-PAD), 2010 22nd International Symposium on*. [S.l.: s.n.], 2011.