

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

2016

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**

Proposta de pesquisa apresentada como um dos requisitos 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

2016

LISTA DE ILUSTRAÇÕES

| | | |
|---|--|----|
| 1 | Arquitetura da rede neural ART | 9 |
| 2 | Cronograma de Atividades | 10 |

LISTA DE TABELAS

| | | |
|---|-------------------------------------|----|
| 1 | Observações Seleccionadas | 9 |
| 2 | Atividades Previstas | 10 |

SUMÁRIO

| | |
|---|----|
| Identificação | 4 |
| 1 Introdução | 5 |
| 1.1 Contextualizacao do Problema | 5 |
| 1.2 Justificativa | 5 |
| 1.3 Objetivo | 5 |
| 1.3.1 Objetivos Específicos | 6 |
| 1.4 Resultados Esperados | 6 |
| 2 Revisão da Literatura | 7 |
| 3 Materiais e Métodos | 9 |
| 3.1 Procedimentos Experimentais | 9 |
| 3.2 Análise dos Dados | 9 |
| 3.3 Cronograma | 9 |
| 3.4 Recursos | 10 |
| 3.4.1 Recursos Humanos | 10 |
| 3.4.2 Recursos Físicos | 10 |
| 4 Considerações Finais | 11 |
| REFERÊNCIAS | 12 |

IDENTIFICAÇÃO

1. Aluna(o):
2. Orientação:
3. Coorientação:
4. Curso: Mestrado em Computação Aplicada
5. Data de Ingresso: Janeiro/2016
6. Linha: Computação, Automação e Gestão de Dados em Agricultura.
7. Aluna(o) Regular

1 INTRODUÇÃO

1.1 Contextualizacao do Problema

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1.2 Justificativa

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1.3 Objetivo

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.

1.3.1 Objetivos Específicos

1.4 Resultados Esperados

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 análise 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 iterações. Para obter estas configurações para a RNA foram realizados 56 análises diferentes demandando de cerca de 56 horas para cada análise. 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 demonstraram que a utilização de técnicas de computação paralela podem melhorar significativamente o tempo de resposta das atividades de mineração.

Código 2.1: Exemplo de Código em Java

```
1  /**
2   * This is a doc comment.
3   */
4  package com.ociweb.jnb.lombok;
5
6  import java.util.Date;
7  import lombok.Data;
8  import lombok.EqualsAndHashCode;
9  import lombok.NonNull;
10
11  @Data
12  @EqualsAndHashCode(exclude={"address","city","state","zip"})
13  public class Person {
14      enum Gender { Male, Female }
15
16      // another comment
17
18      @NonNull private String firstName;
19      @NonNull private String lastName;
20      @NonNull private final Gender gender;
21      @NonNull private final Date dateOfBirth;
22
23      private String ssn;
24      private String address;
25      private String city;
26      private String state;
27      private String zip;
28  }
```

3 MATERIAIS E MÉTODOS

3.1 Procedimentos Experimentais

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 auxílio 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

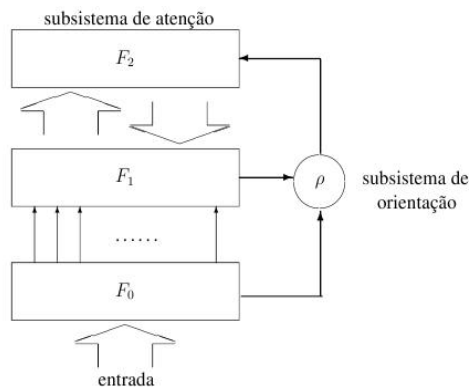


Figura 1: Arquitetura da rede neural ART

3.2 Análise dos Dados

3.3 Cronograma

Visando atingir os objetivos propostos apresenta-se um cronograma de atividades a ser realizado no âmbito do Departamento de Automação e Sistemas (DAS/UFSC). Estas atividades e o cronograma estão ilustrados nas tabelas 2 e 2, respectivamente.

Tabela 2: Atividades Previstas

| Atividades | Descrição |
|------------|---------------------------------|
| A | Revisão bibliográfica. |
| B | Estudo de novas representações. |
| C | Aplicação dos algoritmos. |
| D | Desenvolvimento da interface. |
| E | Validação dos resultados. |
| F | Elaboração da monografia. |
| G | Defesa. |

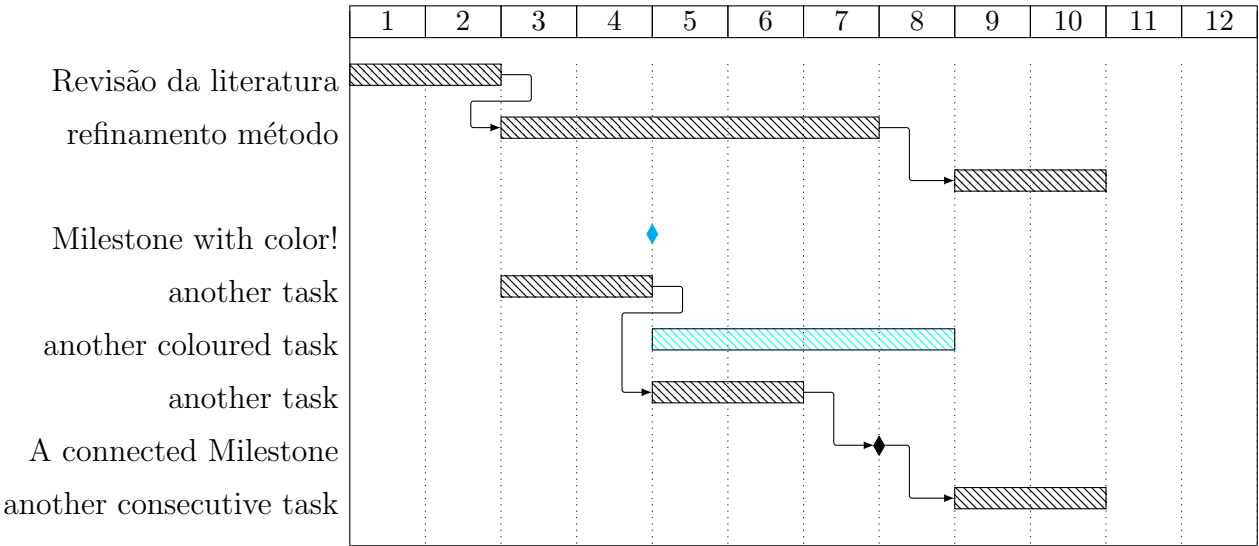


Figura 2: Cronograma de Atividades

3.4 Recursos

Descrever os recursos necessários para o desenvolvimento da pesquisa.

3.4.1 Recursos Humanos

3.4.2 Recursos Físicos

4 CONSIDERAÇÕES FINAIS

Fique atento aos principais problemas que podem ocorrer ao elaborar o projeto de dissertação, os quais podem comprometer o mesmo:

- O objetivo é muito amplo, geral ou vago;
- O projeto ignora conhecimentos anteriores já acumulados sobre o objeto de investigação;
- A revisão de literatura não apresenta o estado da arte efetivo ou ela é inadequada e não se articula com o objeto de investigação;
- O objeto de investigação não é interdisciplinar se enquadra nos campos de atuação da linha de pesquisa na qual o aluno se insere;
- O objeto de investigação não se enquadra nos campos de atuação da linha de pesquisa na qual o aluno se insere;
- O objeto de investigação é pouco relevante;
- A descrição da metodologia ou inadequada ao objeto de investigação;
- O impacto científico/tecnológico da investigação proposta não é enfatizado ou é pouco relevante;
- Cronograma é irrealista ou não contempla as atividades necessárias.

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.