

FINAL YEAR PROJECT REPORT

Canny Canker Detector – A Cancer Diagnosis System

By

Nimra Sikandar

7151

BSE 02-133042-002

Syed Samad Ahmed Bukhari

7159

BSE 02-133042-004

Haris Vohra

7144

BSE 02-133042-036

Supervised by

Faiz - ul - Haque Zeya



Bahria University (Karachi Campus)

2008

Acknowledgments

We want to thank – Head of Bahria University, Karachi Campus and Head of Department (CS) for allowing us to use the resources at the institute, Mr.Rauf Shams for providing us insight into the biological aspects of cancer related issues, Mr.Faiz-ul-Haque Zeya – Project supervisor for guiding us throughout the development period of this project and Mr. Usman Waheed for his guidance in writing this report. We also want to thank the computer lab staff for allowing us to use the project lab whenever we required it.

Abstract

Canny Canker Detector is a Cancer Diagnosis tool designed to distinguish cancerous from non-cancerous tissue samples. The aim of the project is to apply algorithms from the area of machine learning and statistics to the gene expression data acquired by using microarray technology to classify tissue samples as either cancerous or non-cancerous.

The cancer dataset consists of rows of values where each row represents a unique tissue sample and each column represents a unique gene. The last column in the set represents the class label representing the cancerous or non-cancerous tissue samples.

Two methods have been implemented in classification of the tissue samples namely the decision tree and the neural network. The results of both the methods have been compared to conclude which method can distinguish samples more accurately.

The tool is a research based application which will hopefully provide other researchers valuable information through the results of the classification.

Table of Contents

1. INTRODUCTION.....	1
2. BACKGROUND AND LITERATURE REVIEW.....	2
2.1 BIOLOGICAL BACKGROUND INFORMATION	2
2.1.1 <i>Genes</i>	3
2.1.2 <i>Nucleotide</i>	4
2.1.3 <i>Nucleic Acid</i>	4
2.1.4 <i>RNA</i>	4
2.1.5 <i>mRNA</i>	5
2.1.6 <i>Transcription</i>	5
2.2 GENE EXPRESSION AND DNA MICROARRAY TECHNOLOGY.....	6
2.2.1 <i>Gene Expression</i>	6
2.2.2 <i>DNA Microarrays</i>	7
2.3 MACHINE LEARNING.....	9
2.3.1 <i>Decision Trees</i>	9
2.3.2 <i>Neural Networks</i>	11
2.3.3 <i>Work on Various Cancer Datasets</i>	14
3. AIM AND STATEMENT OF PROBLEM.....	16
3.1 AIM.....	16
3.2 PROBLEM STATEMENT	16
3.3 PROJECT SCOPE	17
4. ANALYSIS AND DESIGN	18
4.1 THE CANCER CLASSIFICATION PROBLEM	18
4.1.1 <i>The Challenges</i>	18
4.2 PUBLICLY AVAILABLE CANCER DATA SETS FROM cDNA MICROARRAY.....	19
4.3 CANCER CLASSIFICATION METHODS	20
4.3.1 <i>Machine Learning</i>	20
4.3.2 <i>Supervised Learning ~ Learning Decision Trees</i>	21
4.4 SUPERVISED LEARNING ~ BACK PROPAGATION NEURAL NETWORKS.....	21
4.4.1 <i>Neural Networks Architectures</i>	21
4.5 SYSTEM OVERVIEW / STRUCTURE.....	22
4.5.1 <i>Context Diagram</i>	23
4.6 ACTOR USE CASE DIAGRAM.....	24
4.7 DESIGN USE CASES.....	25
4.7.1 <i>Design Use Case 1</i>	25
4.7.2 <i>Design Use Case 2</i>	26
4.7.3 <i>Design Use Case 3</i>	28
4.7.4 <i>Design Use Case 4</i>	29
4.7.5 <i>Design Use Case 5</i>	30
4.7.6 <i>Design Use Case 6</i>	32
4.7.7 <i>Design Use Case 7</i>	34
4.7.8 <i>Design Use Case 8</i>	35
4.8 CLASS DIAGRAMS	38

4.8.1	<i>Generic Tree</i>	38
4.8.2	<i>Neural Network</i>	38
4.8.3	<i>Iterative Dichotomiser</i>	39
5.	IMPLEMENTATION	41
5.1	DECISION TREES.....	41
5.1.1	<i>The Generic Tree Class Library</i>	41
5.1.2	<i>The Iterative Dichotomiser Class Library</i>	41
5.1.3	<i>The 'Set' Class</i>	41
5.1.4	<i>The 'Attrib' Class</i>	42
5.1.5	<i>The 'Dichotomiser' Class</i>	43
5.1.5.1	<i>The 'ConstructTree' Function</i>	44
5.1.5.2	<i>The 'InternalMounting' Function</i>	44
5.1.6	<i>The Splitting Criteria</i>	46
5.2	NEURAL NETWORKS.....	46
5.2.1	<i>The 'Input' Class</i>	47
5.2.2	<i>The 'Weight' Class</i>	47
5.2.3	<i>The 'DeltaW' Class</i>	48
5.2.4	<i>The 'Neuron' Class</i>	48
6.	TESTING	51
6.1.1	<i>Windows Compliance Test Scripts</i>	51
6.1.2	<i>Control Testing Check List</i>	53
6.1.3	<i>Interface Testing Check List</i>	53
6.1.4	<i>Behavioral Analysis</i>	56
6.2	TEST CASE 2:.....	56
6.2.1	<i>Windows Compliance Test Scripts</i>	57
6.2.2	<i>Control Testing Check List</i>	58
6.2.3	<i>Interface Testing Checklist</i>	59
6.2.4	<i>Behavioral Analysis</i>	61
6.3	TEST CASE 3.....	62
6.3.1	<i>Windows Compliance Test Scripts</i>	62
6.3.2	<i>Control Testing Check List</i>	63
6.3.3	<i>Interface Testing Checklist</i>	64
6.3.4	<i>Behavioral Analysis</i>	66
6.4	TEST CASE 4:.....	67
6.4.1	<i>Windows Compliance Test Scripts</i>	67
6.4.2	<i>Control Testing Check List</i>	68
6.4.3	<i>Interface Testing Checklist</i>	69
6.4.4	<i>Behavioral Analysis</i>	72
6.5	TEST CASE 5.....	72
6.5.1	<i>Windows Compliance Test Scripts</i>	73
6.5.2	<i>Control Testing Check List</i>	74
6.5.3	<i>Interface Testing Checklist</i>	74
6.5.4	<i>Behavioral Analysis</i>	77
6.6	TEST CASE 6.....	77
6.6.1	<i>Windows Compliance Test</i>	77

6.6.2	<i>Control Testing Check List</i>	79
6.6.3	<i>Interface Testing Check List</i>	80
6.6.4	<i>Behavioral Analysis</i>	82
7.	RESULTS	84
7.1	RESULTS OF THE DECISION TREE CLASSIFIER.....	84
7.2	RESULTS OF THE NEURAL NETWORK CLASSIFIER	85
7.3	COMPARISON BETWEEN DECISION TREE AND THE NEURAL NETWORK CLASSIFIER	
	85	
8.	DISCUSSION	87
8.1	ACQUIRING BACKGROUND BIOLOGICAL KNOWLEDGE	87
8.2	ACQUIRING A DATASET	87
8.3	STUDY OF DIFFERENT CLASSIFICATION ALGORITHMS.....	87
8.4	CONVERSION OF THE ACQUIRED DATASET INTO A FORMAT ACCEPTABLE BY THE	
	CLASSIFICATION METHODS.....	87
8.4.1	<i>Format for the Decision Tree Classifier</i>	87
8.4.2	<i>Format for the Neural Network Classifier</i>	88
8.5	IMPLEMENTATION OF THE METHODS	88
8.6	GENERATION OF GRAPHICAL VIEW OF THE DECISION TREE AND ITS	
	CORRESPONDING RULE SET.....	88
8.7	COMPARISON OF RESULTS	88
9.	CONCLUSIONS	89
10.	FUTURE WORK	90
10.1	STUDY OF OTHER ALGORITHMS	90
10.2	ADAPTER FOR GENERAL FORMAT.....	90
	APPENDIX A – CLASSIFICATION ALGORITHMS	91
	APPENDIX B – UML	95
	REFERENCES:	98