

Forecasting Financial Schedules for Cancer Patients

P.M.U.A.Jayatilaka

149214T

Faculty of Information Technology, University of Moratuwa, Sri Lanka

Forecasting Financial Schedules for Cancer Patients

P.M.U.A.Jayatilaka

149214T

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

May 2017

Declaration

I declare that this research is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and the list of references is given

Name of the Student

Signature of the student

P.M.U.A.Jayatilaka

Date:

Supervised by

Signature of the supervisor

Mr. S.C.Premarathne

Date:

Dedication

I dedicate this thesis to my parents who have always been my nearest and reverse nearest neighbors and have been so close to me that I found them with in me whenever I needed. It is their unconditional love that motivates me to set higher targets.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Mr.S.C.Premarathne, Senior Lecturer in University of Moratuwa ,Sri Lanka whom expertise in the field of Information and Technology, encouraged me in understanding and applying knowledge in practice considerably throughout my research period. I do appreciate his vast knowledge and skills in many areas and his assistance and patience on me.

Furthermore, I would like to thank the other lecturers of Department of Information Technology, University of Moratuwa especially Prof.Asoka Karunanada who taught us Research Methodology and Literature Review and Thesis writing subjects which were the basis for this research.

More over a very special thanks goes to Dr. B. I. Kurukulasuriya Deputy Director, National Cancer Institute Maharagama and his staff members for assisting me in collecting data sets. And also Dr. Y. Ariyaratne for granting me the permission to access his patients medical records , Dr.(Mrs.) K.K. Murthi for helping me to translating the medical records in to readable format.

I would also like to thank all the batch mates of MSc. in IT degree program who gave their valuable feed backs to improve the results of the research .Finally I would like to thank my family for the support they provide me through my entire life in particular.

Abstract

During the time a patient is admitted in hospital rich sources of clinical, bio medical, contextual, and environmental data about patients have been available in medical and health sciences. These clinical sources of information are marked increasing in both volume and variety. Due to continuous increasing of the size of health care data, certain complexity raised in it. The hidden patterns among patient data can be extracted by applying different techniques. The techniques and tools are very helpful as they provide health care professionals with significant knowledge towards a decision.

This research has been conducted to analyze cancer patient's medical records in Sri Lanka in an effective manner to predict the future cost estimation for medicine. It is hypothesized that analyzing cancer patient's medical record can be done through machine learning data mining techniques. By doing so we can help the patients to schedule their financial matters in coming years. In Sri Lanka almost 23,000 individuals diagnosed with a new invasive cancer each year. The costs of the diagnosis and treatment for the cancer are considerable important. These costs have increased over the past years and are expected to increase in the future. It has been claimed that most of the families face difficulties with a diagnosis of cancer have financial issues. To respond to this challenge, we developed an interactive application which utilizes a multiple linear regression model to forecast financial schedule for cancer patients

This solution take data set collected from Apeksha Hospital -Maharagama as the input and predict the factors associate with the question. Having received the input this approach preprocess the data set to remove the anomalies. Then build the data model to predict the future cost for the patient. Linear regression technique is used to develop for prediction of future cost estimation. This total solution is build using WEKA data mining software. (WEKA GUI and WEKA API). According to the evaluation of the predictive model the correlation coefficient value is 0.5725 and the root mean squared error is Rs. 553183.25.

Table of Content

Declaration	
Dedication	II
Acknowledgement	III
Abstract	IV
List of Figures	IX
List of Tables	X
Chapter 1 Introduction	1
1.1 Prolegomena	1
1.2 Background and Motivation	1
1.3 Problem statement	4
1.4 Hypothesis	5
1.5 Objectives	5
1.6 Data mining-based approach for analyze cancer patient records.	5
1.7 Structure of the Thesis	6
1.8 Sunnary	6
Chapter 2 Developments and Challenges in Patient Record Analysis	7
2.1 Introduction	7
2.2 Early Developments	7
2.3 Modern Trends in Patient Record Analysis	8
2.4 Future challenges	12
2.5 Sunnary	12
Chapter 3 Technology Adapted-Data mining Tools and Techniques	13
3.1 Introduction	13
3.2 What is Data mining?	13
3.3 Data Mining Techniques	14
3.3.1 Multiple Linear Regression	15

3.3.2 SMO Regression	16
3.4 Data Mining Tools	16
3.4.1 WEKADataMining Tool	16
3.5 Java	17
3.6 Surmmary	17
Chapter 4 Approach to Forecasting Financial Schedules For Cancer Patients	18
4.1 Introduction	18
4.2 Hypothesis	18
4.3 Users	18
4.4 Input	18
4.5 Out put	19
4.6 Process	19
4.6.1 Data Selection	20
4.6.2 Data Pre processing	20
4.6.3 Data Transformation	20
4.6.4 Data mining	21
4.6.4.1 Data mining - Classification vs Regression	21
4.6.5 Evaluation/interpretation	21
4.7 Features	22
4.8 Surmmary	22
Chapter 5 Design of FFSCP	23
5.1 Introduction	23
5.2 Top level Architecture ofFFSCP	23
5.3 Data Model ofFFSCP	24
5.4 Cost Estimation Module	25
5.5 User Interfaces	26
5.6 Surmmary	26

Chapter 6	Implementation of FFSCP	27
6.1	Introduction	27
6.2	WEKA	27
6.3	Data Collection	27
6.4	Pre-processing of the Dataset	27
6.4.1	Missing Value Imputation	27
6.4.2	Feature Selection	28
6.5	Data Model using WEKA	30
6.5.1	Regression Models for Future Cost Estimation	30
6.5.1.1	Multiple Linear Regression	30
6.5.1.2	SMO Regression	30
6.6	Implementation of the Application	31
6.7	Summary	31
Chapter 7	Evaluation	32
7.1	Introduction	32
7.2	Data Model Testing on Regression Models	32
7.2.1	Correlation Coefficient (CC)	32
7.2.2	Root Mean Squared Error	32
7.3	Data model Evaluation	33
7.4	Summary	33
Chapter 8	Conclusion and Further Work	34
8.1	Introduction	34
8.2	Overview of the research	34
8.3	Problem encountered and limitations	34
8.4	Further Work	35
8.5	Summary	35

References	36
Appendix- A	41
Appendix- B	42
Appendix- C	48

List of Figures

Figure -3.1 - Steps of data mining process	15
Figure -3.2 -Data mining techniques	15
Figure 4.1 - Steps in knowledge discovery process	19
Figure- 5.1- Top level architecture ofFFSCP	23
Figure- 5.2- Data model ofFFSCP	24
Figure 5.3- Cost estimation module	25
Figure 5.4-User Interface	26
Figure 6.1-Replace missing values	28
Figure 6.2 -Attribute selection	29
Figure 6.5- Application Interface with results	31
Figure 6.4.1 -Applying Numeric to Nominal filter before attribute selection.	42
Figure 6.4.2- Apply NominalToBinary Filter	42
Figure 6.3- After applying Linear regression model	43
Figure 6.4- After applying SMO regression model	44
Figure 6.6 – Graphical view of the application	45
Figure 7.1 - Evaluation results of the multiple linear regression model	46
Figure 7.2- Evaluation results of the SMO regression model	47

List of Tables

Table 7.1 -Summary of the evaluation results

36