A LITERATURE REVIEW IN DATA MINING MODELS USED FOR SURVIVABILITY PREDICTION OF CANCER PATIENTS

Saumya T M D *(t.m.d.saumya@gmail.com)*
Department of Industrial Management, University of Kelaniya, Sri Lanka

Rupasinghe T *(thashika@kln.ac.lk)*
Department of Industrial Management, University of Kelaniya, Sri Lanka

Abeysinghe P *(prasadabeysinghe@hotmail.com)*
National Cancer Institute, Maharagama, Sri Lanka

ABSTRACT

The research in the medical domain is clinical in its nature but with the advancement of information technology the trends of researchers in health care sector has been moving towards medical informatics. Usage of data mining techniques plays a major role in medical informatics. Especially when it comes to predicting or forecasting the survivability of a disease which is known as medical prognosis, data mining plays a major role. With the time medical prognosis is becoming highly important to increase the morality of patients especially who are diagnosed as cancer victims. Although the importance is increasing in the cancer prognosis, the methods that are in the practice for predicting still need to be improved and refined. In this paper we present an overview of the current research being carried out using the data mining techniques for prognosis of cancers. The goal of this study is to identify the well-performing data mining algorithms used on medical databases in order to predict survivability of cancer patients. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial Neural Networks, Naïve Bayes and Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for cancer prognosis. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

Keywords: *Prediction, Survival, Cancer, Data Mining*

INTRODUCTION

Medical domain is an area where the most clinical and biological researchers are exploiting new knowledge through their studies but with the advancement of the information technology things have changed rapidly. Research in the medical domain has refined with information technology and now the researchers in the information technological domain has started their research on how to use information technology to improve the medical domain and the health care sector and this area of research has popularized as health informatics research.

These researchers in health informatics have paid their especial attention to the cancer related studies since it is one of the most death causing illnesses and if they are capable of improving the lives of the cancer patients or if they can give any aid to improve the cancer prevention; that would be a worthwhile effort to be taken into consideration.

This study has its main objective to review the previous related studies that have been carried out in order to predict survivability of cancer patients and summarize and analyze the data mining techniques used in those studies to predict the cancer survivability

Survivability or the life expectancy is an important factor when it comes to treating the patient in the most relevant way according to the state of their cancer, so the results of this review will help future researchers to continue with the studies to develop systems to predict survivability using the well performing data mining algorithms revealed through this review.

As previously mentioned this study will incorporate the previous studies that have been used data mining techniques to predict survivability of different types of cancer. Due to lack of the previous efforts that have been put into cancer survivability prediction in the Sri Lankan context, the attention of this paper will be paid for overall cancer survivability through the studies carried out in different context using different datasets like SEER (Surveillance, Epidemiology, and End Results) data set.(Lakshmi, 2013)

METHODOLOGY

The methodology used for this paper was selected through the survey of journals and publications in the fields of Computer Science, Engineering and Health Care. In order to obtain a general overview on the previous studies and the available knowledge in the related fields, book chapters, dissertations, working papers and conference papers are also included in the study. The research is mainly focused on the most recent publications which will cover the years from 2007 to 2013, in order to improve the relevance of the results of this study to the current context.

Initially, this paper will summarize the previous studies that have been carried out in Health Care using data mining techniques. This will only consider the most recent studies which will cover the last five to six years. Depending on the way that data mining techniques have been used for those studies, those have been categorized with examples.

Then the paper will narrow down its focus to the most recent previous studies that have been carried out to predict cancer survivability using data mining techniques. Then those used data mining techniques will be analyzed and compared with each other to find the best data mining technique in cancer survivability prediction using performance measures in those studies.

Data Mining in Health Care

Advancement of information technology has also resulted in low cost hardware and software, with this low cost hardware and software the amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically in the last decade. As a result, traditional data analysis techniques have become inadequate for processing such volumes of data, and new techniques have been developed. The most prominent area of development is called knowledge discovery in databases (KDD).

KDD is well equipped with variety of statistical analysis, pattern recognition and machine learning techniques. In general, KDD can be defined as a formal process which contains the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and visualization and demonstration of knowledge discovered which are employed to exploit the knowledge from large amount of recorded data .The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining.

Data mining techniques have been used in different ways in the healthcare sector, if we consider the medical domain as a whole there are different kinds of studies done using data mining techniques in medical databases. Analyzing of those studies resulted in several categories. Those categories are as follows:

Table1 Study Ares of Data Mining in Healthcare Sector

Category of the study	Examples for the category
Studies that summarize reviews and challenges in medical data mining using medical data in general	(Canlas, 2009), (Satyanandam et al., 2012), (Hosseinkhah et al., 2009), (Dakheel et al., 2012), (Wasan et al., 2009)
Studies of data mining techniques used for diagnosing of a specific diseases	(Kumari,2011),(Soni et al.,2011),(Prasad et al.,2011),(Dangare, 2012),(Aftarczuk, 2007)
Studies of data mining techniques used for disease prognosis	(Srinivas et al., 2010),(Aljumahet al., 2011),(Delen, 2009), (Osofisanet al., 2011), (Kika et al., 2010),(Floyd, 2007)
Studies of data mining techniques used for both diagnosis and prognosis	(Guptaet al., 2011),(Huanget al., 2007)
Studies to investigate factors which have higher prevalence of the risk of a disease using data mining	(Karaolis et al., 2009), (Yanget al., 2010)
Studies that present new data mining technologies and algorithms	(McCormiket al., 2009),(Chao, 2009),(Chao, 2011),(Habrardet al., 2003),(Ullah, 2012),(Kusiak, 2001)
Studies that present new data mining techniques improving old ones	(Kavithaet al., 2010),(Ha, 2009),(Parvathi, 2011),(Gaoet al., 2005)
Studies that present new data mining frameworks, tool and applications in medicine and healthcare systems	(Patil, 2009),(Shukluet al., 2009),(Kumar, 2011),(Duanet al., 2011),(Sakthimurugan, 2012),(Palaniappn, 2011)

Above categorization illustrate that medical prognosis is a main research area of data mining related to the medical domain. Medical prognosis is a field in medicine that encompasses the science of estimating the complication and recurrence of disease and to predict the survival of the patient or group of patients (Ohno-Machado, 2001).

In other words, medical prognosis involves prediction modeling where the different parameters related to patients' health could be estimated. The importance of these estimates is that they can help to design treatment as per the expected outcomes.

Data Mining in Survivability Prediction

Survival analysis is a field in medical prognosis that deals with the application of various methods to estimate the survival of a particular patient suffering from a disease over a particular time period. ‘‘Survival’’ is generally defined as a patient remaining alive for a specified period of time after the diagnosis of disease.

Traditionally, conventional statistical techniques such as Kaplan-Meier test and Cox-Propositional hazard models (Cox, 1984) were used for modeling survival. These techniques are conditional probability based models that provide a probability estimate of survival. With advances in the field of knowledge discovery and data mining a new stream of methods have come into existence. These methods are proved to be more powerful as compared to traditional statistical methods.

The next sections of this paper will be mainly focusing on reviewing and analyzing the previous studies that have been carried out for survivability prediction of the various cancer types.

Most of the researchers in this study area have used breast cancer, lung cancer and pancreatic cancer patients’ data as the data source so the results of this study will only be applicable for the prediction of the survivability of those cancer types only not for the other cancer types.

All these studies has measured the accuracy of survivability prediction using the common measurement in survivability prediction , usually this survivability prediction is given as rates, which describe the percentage of people with a certain type of cancer who will be alive a certain time after the cancer is detected. Bellow studies have used five-year relative survival rate which describes the percentage of people with cancer who will be alive five years after diagnosis, excluding those who die from other diseases. Accuracy of these rates are measured using 100 non survivors details, to get the percentage accuracy ,they have checked that how many patients’ survivability can be predicted correctly by each technique out of those 100 non survivors. Since they have used a common accuracy measure to measure accuracy of predicted survivability, we can use that to measure relative performance of each data mining technique in cancer survivability prediction.

Now the paper will pay attention to individual studies that have been carried out to predict cancer survivability using data mining techniques. A research has been conducted using three data mining techniques (decision trees, artificial neural networks and support vector machines) along with the most commonly used statistical analysis technique logistic regression to develop prediction models for prostate cancer survivability. The data set has contained around 120 000 records and 77 variables. A k-fold cross-validation methodology has been used in model building, evaluation and comparison. The results have shown that support vector machines are the most accurate predictor (with a test set accuracy of 92.85%) for this domain, followed by artificial neural networks and decision trees (William, 2009).

Another comparative research has used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset of breast cancer patients. The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.(Delen et al, 2005)

A performance comparison study to evaluate data mining techniques for prediction and diagnosis of breast cancer disease survivability has been carried out and it has evaluated popular data mining techniques like Decision trees, Artificial neural networks ,*k*-nearest neighbors' algorithm Multinomial Logistic Regression, *k*-means algorithm, Apriori algorithm, Partial Least Squares Regression using breast cancer patients' data in SEER dataset. The result of this study has shown that Partial Least Squares Regression is the most performing data mining technique for prediction and diagnosis of breast cancer disease survivability from the techniques used in the study. (Lakshmi, 2013)

Another study showed that out of Decision Trees, Artificial Neural Networks, Support Vector Machines and Logistic Regression the best technique for predicting cancer survivability is Support Vector Machines (Delen, 2009)

Another study has been carried out to evaluate the applications of machine learning in cancer prediction and prognosis using Decision Tree, *k*-Nearest Neighbor, Neural Network, and Genetic Algorithm for comparison and given the result that Neural Networks are the most dominating techniques in cancer prediction and prognosis(Cruz, 2006)

DISCUSSION

The above analysis can be summarized as graphs using the usage frequency of each data mining technique for prediction of cancer survivability studies. Also, the above analysis resulted in the performance ratings of each data mining technique for prediction of cancer survivability studies and this performance of data mining technique has been measured through the accuracy of the survivability prediction of each study. That also can be shown in a graphical representation. These two graphs are represented in Figure 1 and Figure 2 below the section.

As we can see in Figure 1, the most frequently used data mining technique for prediction of the cancer survivability is artificial neural networks, the second most used technique is decision trees and then studies have used support vector machines, genetic/fuzzy algorithms, bayesian algorithms and k- algorithms consecutively. Although the logistic regression is a traditional statistical method most of the studies have used this technique in survivability prediction in order to show the low performance in traditional statistical techniques compared to the data mining techniques in cancer survivability prediction .

As we can see in Figure 2, artificial neural networks is the best data mining technique for prediction of the cancer survivability based on performance, the second most performing technique is decision trees. This order of performance is as same as the usage frequency order. Also genetic/fuzzy algorithms are also having same performance as decision trees. Support vector machines and bayesian algorithms have poor performance compared to decision trees. As the figure is illustrated logistics regression has the lowest performance.

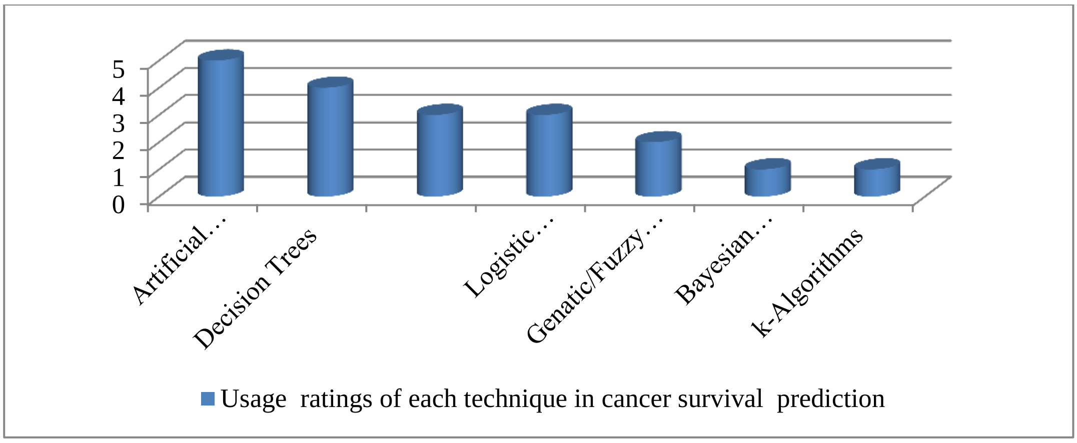

Figure 1 Usage frequency of data mining techniques in survival prediction of cancer

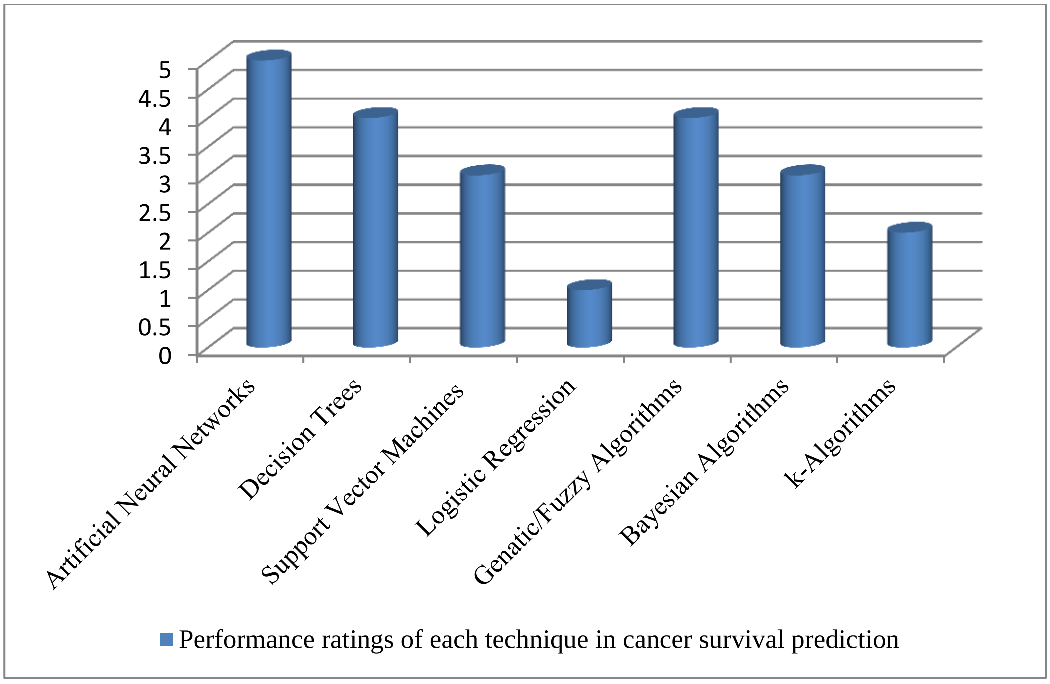

Figure 2 Performance ratings of data mining techniques in survival prediction of cancer

CONCLUSION

In this paper we identified and evaluated the most commonly used data mining algorithms as well-performing in cancer survivability prediction, based on recent studies. The following algorithms have been identified as well performing techniques artificial neural networks, decision trees, genetic/fuzzy algorithms, support vector machine. However, it is very difficult to name a single data mining algorithm as the best for predicting survivability of all the cancer types and for all the different contexts. Depending on certain situations, sometime some algorithms perform better than others, but there are cases when a combination of algorithms results more effective survival prediction.

Future researchers have opportunities to continue with the research related to survivability prediction in various cancer types using different data sets where the prognosis factors may completely differ from SEER(Surveillance, Epidemiology, and End Results)data set and the data sets used in the above studies and proposed survivability prediction model for those cancer types in those different contexts by selecting the appropriate data mining technique or combination of techniques considering the above reviewed results.

REFERENCES

Aftarczuk, K. (2007): *Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems.*

Aljumah, A.A. , Ahamad, M. G., Siddiqui, M.K. (2011): Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia, *Intelligent Information Management*,3, pp. 252-261

Canlas, R.D. (2009)*Data Mining in Healthcare: Current Applications and Issues*

Chao, S., Wong, F. (2009): *An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining*

Chao , S., Wong, F. (2011): *A Multi-Agent Learning Paradigm for Medical Data Mining Diagnostic Workbench*

Cox, D.R., 1984 : Analysis of survival data. London: Chapman & Hall; 1984.[online] Available at- http://onlinelibrary.wiley.com/doi/10.1002/bimj.4710290119/abstract

Cruz J. and Wishart S., 2006, Applications of Machine Learning in Cancer Prediction and Prognosis ,*Cancer Informatics* 2006:2 59–77Ullah, I. (2012): *Data Mining Algorithms And Medical Sciences*

Dakheel, F.I., Smko, R., Negrat, K., Almarimi, A. (2011): *Using Data Mining Techniques for Finding Cardiac Outlier Patients*

Dangare, C. S., Apte, S.S. (2012): *Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques*

Delen D, Walker G, Kadam A.2005 Predicting breast cancer survivability: a comparison of three datamining methods. *Artificial Intelligence in Medicine.* 2005 Jun; 34(2):113-27.

Delen, D. (2009): *Analysis of cancer data: a data mining approach*

Duan, L. , Street , W. N., Xu, E. (2011): *Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems*, 5:2, pp169-181 580 ICT Innovations Floyd, S. (2007): *Data Mining Techniques for Prognosis in Pancreatic Cancer*

Gao, J., Denzinger, J., James, R.C. (2005): *A Cooperative Multi-agent Data Mining Model and Its Application to Medical Data on Diabetes*

Gupta, S., Kumar, D., Sharma, A. (2011): *Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis*

Habrard, A., Bernard, M., Jacquenet, F. (2003): *Multi-Relational Data Mining in Medical Databases, Springer-Verlag* LNAI 278

Ha, S.H., Joo, S.H. (2010): A Hybrid Data Mining Method for the Medical Classification of Chest Pain, *International Journal of Computer and Information Engineering* 4:1,pp 33-38

Hosseinkhah, F., Ashktorab, H., Veen, R., Owrang, M. M. O. (2009): *Challenges in Data Mining on Medical Databases* IGI Global pp. 502-511

Huang, M.-J. , Chen, M.-Y., Lee, S.-C. (2007): Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis, *Expert Systems with Applications* 32 pp.856–867

Karaolis, M., Moutiris, J.A., Papaconstantinou, L., Pattichis, C.S. (2009): *Association Rule Analysis for the Assessment of the Risk of Coronary Heart Events*

Kavitha, K.S., Ramakrishna, K.V., Singh, M. K. (2010): Modeling and design of evolutionary neural network for heart disease detection, *IJCSI International Journal of Computer Science* Issues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814, pp. 272-283

Kika, A., Cico, B., Alimehmeti, R. (2010): *Using Machine Learning for Preoperative Peripheral Nerve Surgical Prediction*

Kumari, M., Godara, S. (2011): *Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction*, IJCST ISSN: 2229- 4333 Vol. 2, Issue 2

Kumar, D.S., Sathyadevi, G., Sivanesh, S. (2011): *Decision Support System for Medical Diagnosis Using Data Mining*

Kusiak, A.(2001) *, Decomposition in Data Mining: A Medical Case Study* , B.V. Dasarathy (Ed.), Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, Vol. 4384, SPIE, Orlando, FL, April 2001, pp. 267-277

Lakshmi K.R. ,2013 Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability, *Asian Journal Of Computer Science And Information Technology* 3 : 5 (2013) 81 - 87.

McCormick, T.H., Rudin, C., Madigan, D.(2009): *A Hierarchical Model for Association Rule Mining of Sequential Events: An Approach To Automated Medical Symptom Prediction*

Ohno-Machado, L., 2001 *Modeling medical prognosis: survival analysis techniques*. J Biomed Inform 2001; 34:428—39.

Osofisan , A.O. , Adeyemo, O.O. , Sawyerr, B.A., Eweje O. (2011): *Prediction of Kidney Failure Using Artificial Neural Networks*

Palaniappan, S., Awang, R. (2008): *Intelligent Heart Disease Prediction System Using Data Mining Techniques*

Parvathi, R. , Palaniammalì, S. (2011): An Improved Medical Diagnosing Technique Using Spatial Association Rules, *European Journal of Scientific Research* ISSN 1450-216X Vol.61 No.1 pp. 49-59

Patil, S. B., Kumaraswamy, Y.S.(2009): Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, *European Journal of Scientific Research* ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

Prasad, B.D.C.N., Prasad, P.E.S.N. K., Sagar, Y. (2011): *A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis*

Sakthimurugan, T., Poonkuzhali, S. (2012): An Effective Retrieval of Medical Records using Data Mining Techniques, *International Journal of Pharmaceutical Science and Health Care*. ISSN: 2249-5738. 2(2), pp 72-78

Satyanandam, N., Dr. Satyanarayana, Ch., Riyazuddin, Md.,Shaik, A. (2012): *Data Mining Machine Learning Approaches and Medical Diagnose Systems*: A Survey

Shukla, A., Tiwari, R., Kaur, P. (2009): Knowledge Based Approach for Diagnosis of Breast Cancer,IEEE*International Advance Computing Conference (IACC)*

Soni, J., Ansari, U., Sharma, D., Soni, S. (2011): *Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction*

Srinivas , K., Rani, B.K., Dr. Govrdhan, A. (2010),: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks (IJCSE) *International Journal on Computer Science and Engineering* Vol. 02, No. 02, pp 250-255

Wasan, S.K.,Bhatnagar, V., Kaur, H. (2006): The Impact of Data Mining Techniques on Medical Diagnostics, *Data Science Journal*, Volume 5, pp. 119-126

Yang, C., Street, W. N., Lanning, L. (2010): *A Data Mining Approach to MPGN Type II Renal Survival Analysis*