Healthy Diet Recommendation System using Apriori Algorithm Decision Rules for Breast Cancer Data

K.Geetha
School Computer Science, Application and Engineering,
Bharathidasan University,Trichy.

Dr.M.Manimekalai,
Department Of M.C.A.,
Shrimati Indira Gandhi College, Trichy

Abstract—
Medical science has discovered that people set a bigger possibility of countering free radicals and warding off illness by consumption of healthy foods and by increasing their resistant system. We adopt Apriori Algorithm to explore the relationship between treatment preferences, healthy food and survival of cancer patient based on their medical attributes. The publice-use data 2011 is used in this research. After the preprocessing of the data set, we apply Apriori algorithm of Association Rules and Decision Rule mining. As a result, we obtain a great deal of Association Rules related and Decision Rule supported. We pick up some easy understandable and comparable rules to discuss and show that data mining technique is efficient method to explore the relation between Cancer treatment preferences, food and survivability.

KEY WORDS: ID3 decision tree, Granular Network, uncertainty, consistent Classification and SPSS Clementine

I. INTRODUCTION

Cancer has become one of the major cause of mortality around the world and research into cancer diagnosis and treatment has become an important issue for the scientific community. Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Knowledge discovery in databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Some people treat data mining as a synonym for KDD . Recent progress in data mining research has led to the developments of numerous efficient methods to mine interesting patterns and knowledge from large databases.

One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data. Machine learning is an adaptive process that enables computers to learn from experience, learn by example, and learn by analogy. The use of machine learning tools in medical diagnosis is increasing gradually.

This is mainly because of the effectiveness of classification and recognition systems to help medical experts in diagnosing diseases. In this paper, three neural network based classification models are evaluated for their suitability for clinical cancer data classification. The objective of classification is to determine whether the outcome (class) would be ‘Benign’ or ‘Malignant’.

II. DATA MINING

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and the simplest analytical step in data mining is to describe the data. Data mining is summarized its statistical attributes, review it visually using charts and graphs. Another task, look for potentially meaningful links among variables. Collecting, exploring and selecting the correct data are critically important. But, data description

cannot alone provide an action plan. You must build a predictive model based on patterns which are determined from known results, and then test that model on results outside the original sample. A good model should never be confused with reality, but it can be a useful guide to understand your business. The final step is to verify the model empirically. There are two keys to success in data mining. First, is coming up with a precise which you are formulation of the problem trying to solve. A focused statement usually results in the best payoff. The second key is using the right data. After choosing some data from the available data, or buying external data, you may need to transform and combine them in significant ways. Neural networks are of particular interest because they offer means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. Actual biological neural networks are incomparably more complex. Neural nets may be used in classification problems (where the output is a categorical variable) or for regressions

III. APRIORI ALGORITHM

The association rule is an implication of the form LHS RHS, where LHS and RHS are both item sets.

Item sets could be an attribute or the combination of attributes, which in our application refer to treatment preferences, survivability and other medical attributes**.**

The support and confidence are both interest measure of the association rule, which respectively reflect the usefulness and certainty of the discovered rules. The definition is introduced as below.

$$Support\ (A \Rightarrow B) = P(A \cup B)$$

$$Confidence\ (A \Rightarrow B) = P(B/A)$$

IV. DATA MINING PROCEDURE

In this work, we applied Apriori algorithm to explore the relation o£ treatment preferences and survival of breast cancer patient based on data set.

We have used **SPSS** Clementine **11.1** to experiment with Apriori algorithm. **SPSS** Clementine is a data mining software tool by **SPSS** Inc. which contains the tools for data preparation, classification, clustering and visualization. It was renamed PASW Modeler 1**3** on March 11**, 2011** by SPSS.

To identify the survivability of the patient, we have adopted Abdelghani's method to preprocess (he data and obtained the attribute 'survivability'**.** As Abdelghani's method, the value of survivability attribute is to 'survived* if STR ≥ **20** months and VSR is alive, and is 'not-survived- if STR < **20** months and COD is breast cancer.

After reading the Dr.K.Shantha Breast Cancer Foundation(SBCF) USE RECORD DESCRIPTION, we found that some important attributes like 'Tumor Size', 'ROD Extension', and 'Lymph Nonde Involv' in the record of 2003-2007 should be obtained from the '4-Digit Extent of Disease'. So we spited it into three values and fill them into the corresponded filed. After this step, the records with missing information in the above attributes arc removed from the data set.

For the selection of input attributes, we applied the feature selection algorithm. Feature selection is a process, wherein the best subset of the attributes of the dataset is selected; the best subset discards the important attributes. And ***then*** we consulted with medical experts for the attribute selection.

V. DECISION RULE MINING

Recommendation systems are used to predict the desire value. By applying the data mining algorithm on data set in recommendation system predict the data according to the user preference. Prediction can be categorized into: classification, density estimation and regression. In classification, the predicted variable is a binary or categorical variable. Various well-liked decision tree classification methods include decision trees, logistic regression and support vector machines. We defined decision tree is a tree in which each branch node symbolize a preference between a number of substitute, and each leaf node correspond to a decision. Decision tree are generally used for gaining information for the reason of decision -making. It starts with a root node on which it is for users to acquire actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. There are various decision tree classification algorithm are used like 11.3, C4.5, C5.0 etc we work on ID3 and C4.5 the basic decision tree learning algorithm used for classify data.

Apply Decision Tree Rule Mining on Recommendation System

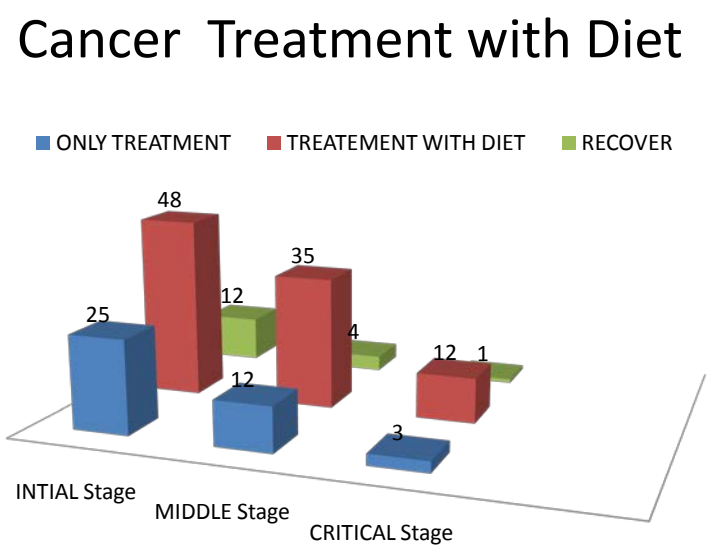
The performance of healthy diet recommendation system used the 1D3 and C4.5 decision tree classification algorithm for classify the healthy diet data set. First the content base filters analysis the user access pattern. Content base filter analyzed the user profile whether the user vegetarian or non vegetarian, suffering from some kind of diseases etc are analyzed.

Then according to the user profile healthy diet data set is classified by the decision rule mining. It trains the data set and generate rule according to the user access pattern. In recommendation system we use the ID3 decision rule mining for mining the data and generate rule. These rules are applied on healthy diet data set and suggest food which is beneficial for your health. For performance analysis we calculate the accuracy of the system with ILX3 and then compare the accuracy of ID3 with C4.5. For improving the performance of the system we apply bagging with ID3.

VI. Result Analysis

In the performance analysis of healthy diet recommendation system decision tree first get the data from content base filter. In the implementation phase we first select the data set then the generated

rule. Then these rules arc applied into the healthy diet recommendation data set. After applying the rule admin selects the profile where we want to apply rule. Once the profile selected the rules arc applied and according to the user profile the food is suggested. Then we apply the rules on and analysis the system. In given chart 1.1, analysis result Breast Cancer Data



VII. Conclusion

My Research work is concerned about the usage of a better approach known as Apriori Algorithm and decision rules . This is new type of research process providing a better solution to the problem as compared to the existing one. Apriori algorithm optimization algorithms have been applied to many combinatorial optimization problems, ranging from quadratic assignment to protein folding or routing vehicles and a lot of derived methods have been adapted to dynamic problems in real variables, stochastic problems, multi-targets and parallel implementations. They have an advantage over simulated annealing and genetic algorithm approaches of similar problems when the graph may change dynamically; the Apriori algorithm can be run continuously and adapt to changes in real time. This is where the Decision rule and Apriori algorithm proves to be better than the genetic algorithm.

Acknowledgements

This work was supported in part by Dr.K.Shantha Breast Cancer Foundation ,Trichy. We would like to thank all my guide and colleague.

References

[1] Alinia, S.H. and Delavar, M.R. (2010). Granular computing model for slving data quality from process to decision, pp. 132-133

[2 Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd Jr. CE. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995;196:817–22.

[3] Bishop CM, Neural networks for pattern recognition. New York: Oxford University Press; 1995.

[4] Breiman L, Friedman J, Olshen R, Stone C, Classification and regression trees. Belmont: Wadsworth International Group; 1984.

[5] Chen D, Chang RF, Huang YL. Breast cancer diagnosis using self-organizing map for sonography.

Ultrasound in Med Biol 2000;26:405–11.
[6] Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999;31:97–109.
[7] Efron B, Tibshirani RJ, An Introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.
[8] Tom M. Mitchell, (1997). Machine Learning, Singapore, McGraw-Hill.
[9] Paul E. Utgoff and Carla E. Brodley, (1990). 'An Incremental Method for Finding Multivariate Splits for Decision Trees', Machine Learning: Proceedings of the Seventh International Conference, (pp.58). Palo Alto, CA: Morgan Kaufmann
[10] http://www.health360.info, "The Role of Food and Nutrition in Cancer"