Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 6, Issue, 6, pp.4809-4813, June, 2015

International Journal of Recent Scientific Research

RESEARCH ARTICLE

MEDICAL DIAGNOSIS FOR LIVER CANCER USING CLASSIFICATION TECHNIQUES

Reetu[1] and Narender Kumar[2]

[1,2] Department of Computer Science and Engineering Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

ARTICLE INFO

Article History:

Received 14th, May, 2015
Received in revised form 23th, May, 2015
Accepted 13th, June, 2015
Published online 28th, June, 2015

Key words:

Data Mining, Liver cancer, Decision tree, J48.

ABSTRACT

The important and successful applications of data mining are in fields like business intelligence, finance, digital libraries, in other industries and sectors. One of the applications of data mining is medical diagnosis which is mostly used in research area. Medical diagnosis is the field where many researchers are concentrating. To reduce the diagnosis time and improve the diagnosis accuracy, it has become an important issue. In medical, Liver Cancer is one of the most prevalent and deadly cancers in human beings. Liver cancer is difficult to be diagnosed at an early stage due to the risk factors. Therefore, new metrologies for early Liver Cancer are needed to determine the condition of the Liver Cancer. Various Data classification techniques or algorithms are used to solve this issue. Some classification techniques or algorithms are Decision tree, C4.5, Association rule, Bayesian networks, Support vector Machine, K-NN, Neural networks etc. Classification is one of the data mining techniques or algorithm which classifies the data based on attributes given in the datasets. Classification techniques used the training and test data set to classify the data and to build a model or to find out the hidden knowledge from the datasets. This model is further used to classify the new objects. So, to classify the dataset or build a model datasets is taken from the Pt. B D Sharma Postgraduate Institute of Medical Sciences Rohtak, for the purpose of medical diagnosis or health-care research.

INTRODUCTION

According to (S. Vijayarani, S. Dhayanand *et al*., 2015) the liver is helping in numerous vital functions and playing a major role in the metabolism. In the human body liver is the second largest internal organ. The numerous vital functions which are performed or related by the liver are digestion, metabolism, immunity and storage of nutrients. If these functions are not performed properly by the liver, there is lack of energy and nutrients in the body of human. So, there are number of key factors and risks factors that cause the liver cancer.

Liver cancer is a serious problem. Until now, in the most medical research, the reasons for suffering from liver cancer are unclear. It is difficult to detect the liver cancer at starting stage and it is functioning very well or normally when it is fractionally damaged (Sharifah Hafizah *et al*., 2014). Liver cancer is one of the more dangerous and threatening diseases at global level with more than one million cases diagnosed each year (Lam, Yee Hong Brian *et al*., 2005). It is the fourth common cancer in world and third leading cause of cancer mortality. Liver cancer is difficult to detect at early stage due to the lack of symptoms. Several types of risk factor like cirrhosis, obesity, smoking, hepatitis B and hepatitis C or alcoholism are highly linked to the liver cancer (Jung Hun Oh, Jean Gao *et al*., 2011).

The medical term that is used for the liver cancer is Hepatocellular Carcinoma. Liver cancer is one of the more dangerous and threatening diseases at global level with more than one million cases diagnosed each year. Liver cancer is difficult to detect at early stage due to the lack of symptoms. Several types of risk factor like cirrhosis, obesity, smoking, hepatitis B and hepatitis C or alcoholism are highly linked to the liver cancer (Murat Karabatak, M. Cevdat Ince *et al*., 2009).

METHODOLOGY

Due to resource constraints and the nature of the paper itself, the main methodology used for this paper was through the survey of journals and publications in the fields of medicine, computer science and engineering. The research focused on more recent publications.

Data Mining In Medical Diagnosis

Data mining is an important process where intelligent methods are applied to find out data patterns. It is the process of

*Corresponding author: **Reetu**
Department of Computer Science and Engineering Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

discovering interesting pattern and important information or knowledge from large amounts of data. It is popular due to the successful applications in telecommunication, marketing and tourism. In now a day, the usefulness of the methods has been proven also in medical field. Data mining is also known as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

However, there are various accommodations to consider when choosing the relevant data mining technique to be used in a certain application. The "best" model is often found by experiment or hit and miss: trying different technologies and algorithms.

One of the algorithm is Data Classification is the process of finding a model (or function) that explain and different data classes. Data mining technology provides an easy to use approach i.e. user-oriented approach to determine the hidden patterns in data. (Murat Karabatak, M. Cevdat Ince *et al*., 2009) Classification technique has been applied in various areas of problem like medicine, social management and engineering fields. Various types of problem like diseases diagnosis, image recognition and credit evolution using classification algorithms or techniques.

This paper includes the results of the classification technique in which decision tree gives the best results in terms of correctly classified or incorrectly classified instances of the datasets.

MATERIALS AND METHODS

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Classification is a supervised machine learning procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items and based on a training set of previously labeled items.

Some classification techniques or algorithms are Decision tree, C4.5, Association rule, Bayesian networks, Support vector Machine, K-NN, Neural networks etc (Jiawei Han and Kamber *et al*., 2006). But, for this paper decision tree is used because it gives the better results than other algorithms.

Decision Tree Induction is given by Quinlan in 1993. It represents the logic method that is mostly used today's. Decision-tree algorithm is supervised-learning methods that construct decision tree from a set of input-output samples. It is hierarchical or nonparametric model (in the sense that we do not assume any parametric form for class density) for the supervised learning (Mehmed Kantardzic *et al*., 2011).

This algorithm generates the rules for the prediction of the goal variable. The large or complex datasets easily classified by the decision tree which is understandable by the human (Nadali, A; Kakhky. *et al*., 2011). The structure of the Decision tree is like a Tree structure. Like tree, a decision tree is also made of a root, internal nodes and leaf nodes. All the nodes except the root node have an incoming node. Each internal node is test node because it tests an attribute and it has an outgoing edge (S.Neelamegam, E. Ramaraj *et al*., 2013). All others nodes are leaf node except the root or internal nodes. Each leaf node assigns a class.

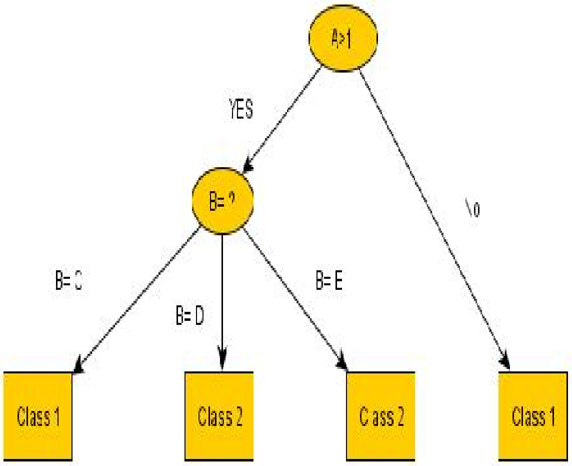


Figure 1 shows the structure of the decision tree

Figure 1 A decision Tree
In this figure A is the root node which has no incoming node and B is the internal node which has incoming node from A and test attributes C, D and E and at last Class 1 and Class 2 are leaf nodes. Decision Tree classification technique is performed in two phases:

1. Tree Building
2. Tree Pruning

In tree building phase, the tree is recursively portioned till all data objects belong to the same class and it works in top-down manner (Dharmender Kumar, Ritu Kaliraman *et al*., 2013). Tree pruning is used to improve the classification and prediction accuracy of the algorithms and works in bottom-up approach. It also minimize the over fitting problem.

For generating decision tree a well known tree-growing algorithm that is based on univariate splits is ID3 (Quinlan, 1986) and its extended version called C4.5 (Quinlan, 1993). Now C4.5 was superseded in 1997 by a commercial system C5.0. Here another algorithm of decision tree is CART an acronym for Classification and Regression Trees. C4.5 is the extension version of the ID3 and open source java implementation of the C4.5 is J48 which is used in WEKA (a data mining system tool) to build a decision tree. C4.5 algorithm is based on the Hunt's Concept Learning System (CLS) which is used to construct a decision tree from a set T of training samples.

J48 classifier is a simple C4.5 decision tree induction algorithm for classification. The feature that are added into J48 are decision tree pruning, accounting for missing values, continuous attribute value ranges, derivation of rules, etc. J48 is the open source java implementation of the C4.5 which is used in WEKA. J48 divide the data into range based on the attribute value for that item which are found in the training samples. Algorithm (Margaret H. Danham, S. Sridhar *et al*., 2006) of the J48 is as follows:

```
INPUT:
        D       //Training data
OUTPUT
        T       //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and label:
For each arc do
        D= Database created by applying     splitting predicate to D;
        If stopping point reached for this, then
        T'= create leaf node and label with appropriate class;
Else
        T'= DTBUILD (D);
        T= add T' to arc;
}
```

While building a tree, J48 disregard the missing value. For selection of attribute to be tested a criteria is used called Gain which is based on the information theory concept: Entropy. The Gain has had good results in construction of decision tree, but it has one serious deficiency: a strong bias in favor of tests with many outcomes. A solution was found in some kind of normalization (Mehmed Kantardzic *et al*., 2011) using a “Split Information” value.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right)$$

Gain ratio is defined as:
Gain-Ratio (A) = gain (A) / Split-info (A)

A similar procedure should be performed for other test in the decision tree and maximum gain ratio will be criteria for attribute selection.

This technique is applied to carry out the classification of the Liver Cancer Datasets on the basis of HCC (Hepatocellular Carcinoma). In this 100 tranining tuples are used to build the model and 50 test tules are used to test the model.

Weka (A Data Mining Tool)

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is a data mining system developed by the University of Waikato in New Zealand. It is a popular suit of machine learning software. WEKA is free software or open source software available under the GNU (General Public Licenses). It provides the user graphical user interface for easy access the functionality, also it is a collection visualization tools and algorithms for data analysis (Sudhir B. *et al*., 2013) (Svetlana S. Aksenova *et al*., 2004).

WEKA tool is used from its popularity, ease of programming and good performance. It is a collection of machine learning algorithms for data mining tasks. Using the WEKA the algorithms are applied directly to a dataset or called from own java code. Also, the new machine learning schemes can be developed with this package or software (Mohd Fauzi bin Othman *et al*., 2007). Applications written using the WEKA tool can be run on any computers with a Web browsing capability.

For the processing of data in WEKA, data is converted into the ARFF (Attribute Relation File Format) format. It described lists of instances sharing sets of attributes. WEKA has several graphical user interfaces but the main graphical user interfaces is “Explorer” that enable easy access to essential functionality (Mark Hall *et al*.).

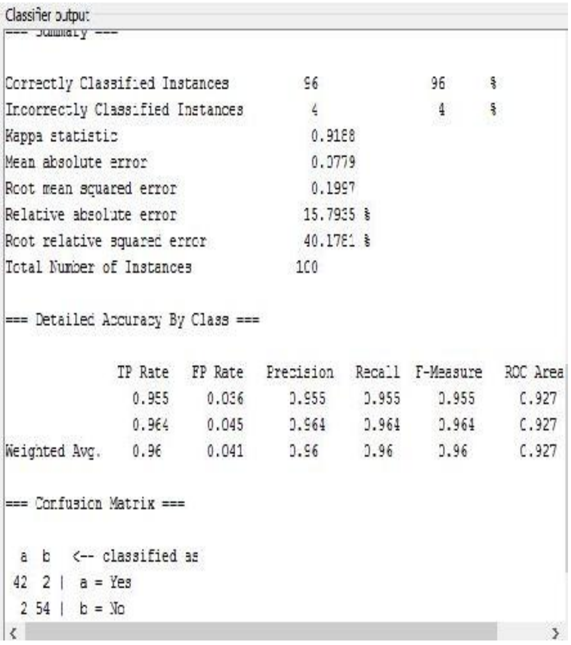
RESULTS AND DISCUSSION

We have performed classification using Decision Tree Induction (J48 Algorithm) on Liver Cancer Datasets in WEKA tool. This dataset is taken from Pt. B D Sharma Postgraduate Institute of Medical Sciences Rohtak. Description of the datasets is given in Table 1 as shown below:

Table 1 Description of datasets used for this work.

Dataset	Attribute	Class	Instances
Liver Cancer	8	2	150

First, datasets is divided into training and test datasets. For training purpose 100 instances are used which build a model and for test dataset 50 instances are used to test the model. Then, the classification techniques are applied using 10 fold cross validation method to generate the classifier for the dataset using WEKA tool. At last, the results are recorded in terms of correctly classified instance, error rate, and confusion matrix. The results are shown below in figure 2. and figure 3 shows the decision tree which classify the dataset on the basis of the age to classify the HCC.



```
=== Summary ===

Correctly Classified Instances          96               96      %
Incorrectly Classified Instances         4                4      %
Kappa statistic                          0.9188
Mean absolute error                      0.0779
Root mean squared error                  0.1997
Relative absolute error                 15.7935 %
Root relative squared error             40.1781 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.955     0.036     0.955     0.955     0.955      0.927
                 0.964     0.045     0.964     0.964     0.964      0.927
Weighted Avg.    0.96      0.041     0.96      0.96      0.96       0.927

=== Confusion Matrix ===

  a  b   <-- classified as
 42  2 |  a = Yes
  2 54 |  b = No
```

Figure 2 results of J48 decision tree with cross vaildation

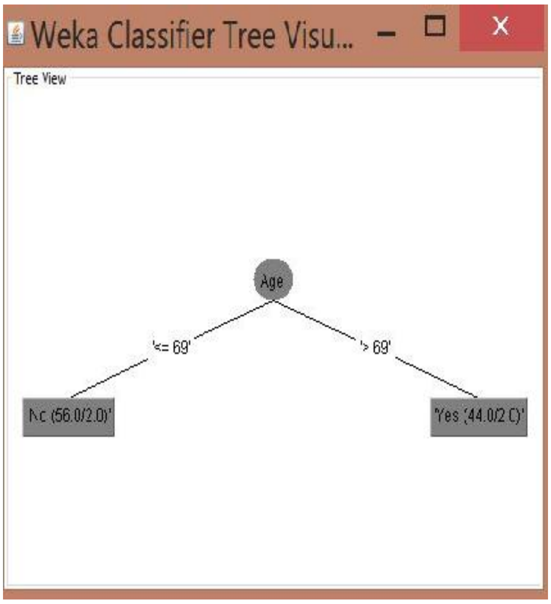
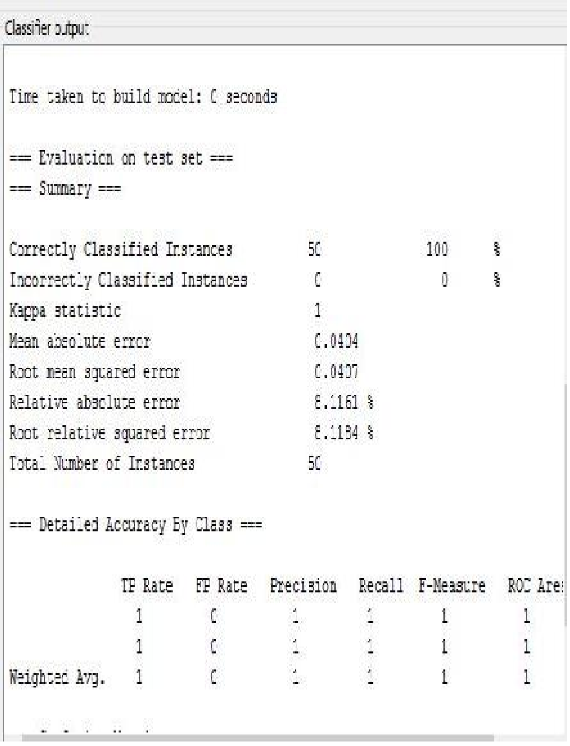


Figure 3 A Decision Tree using J48

The decision tree classifyied that the age greater than the 69 has HCC i.e. Liver Cancer and age less than 69 specifies that patient has no Liver Cancer. The results above 69 gives YES and results less than 69 gives NO.

With the comparision of the other classification algorithms 100 training tuples with 50 test tuples gives the 100 % correctly classified instance. It is shown in figure 4:



```
Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          50              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0404
Root mean squared error                  0.0407
Relative absolute error                  8.1161 %
Root relative squared error              8.1184 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 1         0         1          1         1          1
                 1         0         1          1         1          1
Weighted Avg.    1         0         1          1         1          1
```

Figure 4 J48 decision tree with test dataset

Cost/ Benefit of J48 for class YES = 44
Cost/ Benefit of J48 for class NO= 56
Classification Accuracy for YES= 56%
Classification Accuracy for NO= 44%

CONCLUSION AND FUTURE WORK

Thus, in this paper we have classified the Liver Cancer Dataset using the Decision Tree Induction (J48 Decision Tree). It is found that the performance of the J48 decision tree has better than other classification algorithms and it takes few times to build the decision tree than others algorithms. Factors that affect the classifier's performance are:

1. Data set
2. Number of tuples and attributes
3. Type of attributes
4. System configuration.

As the data available for this very small, this technique can be applied on the large dataset with large number or attribute and class. Also a combination of classification techniques will be used to improve the performance and for maximum accuracy.

Acknowledgement

We would like to acknowledge and extend our heartfelt gratitude to Head of Cancer Section, Regional Cancer Centre (RCC) department of radiation therapy from Pt. B D Sharma Postgraduate Institute of Medical Sciences Rohtak.

References

1. Dharmender Kumar, Ritu Kaliraman, "Classification of Cotton plants using Decision Tree" *J.Cottan Res. Dev. 27 (1)*, January 2013, pp. 152-156.
2. Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques" 2006.
3. Jung Hun Oh, Jean Gao, "Fast Kernel Discriminant Analysis for Classification of Liver Cancer Mass Spectra" *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8. NO. 6. Nov/Dec 2011, pp. 1531-1534.
4. Lam, Yee Hong Brian, "Proteomic Classification of Liver Cancer using Artificial Neural Network" May 2005.
5. Margaret H. Danham,S. Sridhar, "Data mining: Introductory and Advanced Topics" *Person education*, First Edition, 2006.
6. Mehmed Kantardzic, "Data Mining: Concepts, Models, Methods and Algorithms" *IEEE Second Edition.*
7. Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer" *Springer Science IFMBE Proceedings 15*, Vol. 15, 2007, pp. 520-523.
8. Murat Karabatak, M. Cevdat Ince, "An expert system for detection of breast cancer based on association rules and neural networks" *Elsevier Science Expert System with Application 36*, 2009, pp. 3465-3469.
9. Nadali, A. Kakhky, E.N. Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system" *Electronics Computer Technology (ICECT)*, 2011 3rd International Conference on , Vol. 6, No. 8, April 2011, pp. 161-165.

10. S. Neelamegam, E. Ramaraj, “Classification algorithm in Data mining: An Overview” *International Journal of P2P Network Trends and Technology (IJPTT),* Vol. 4, Issue 8, Sep 2013, pp. 369-374.
11. S. Vijayarani, S. Dhayanand, “Liver Disease Prediction using SVM and Naive Bayes Algorithms” *International Journal of Science, Engineering and Technology Research (IJSETR),* Vol. 4, Issue 4, April 2015, pp. 816-820.
12. Sharifah Hafizah Sy Ahmad Ubaidillaha, Roselina Sallehuddina, Nor Azizah Alia, “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study” *Jurnal Teknologi (Science & Engineering) 65:1,* October 2013, pp. 73-81.
13. Sudhir B. Jagtap, Kodge B. G, “Census Data Mining and Data Analysis using WEKA” *International Conference in “Emerging Trends in Science, Technology and Management” (ICETSTM),* 2013, pp. 35-40.
14. Svetlana S. Aksenova, “Machine Learning with WEKA” 2004.

How to cite this article:

Reetu and Narender Kumar., Medical Diagnosis For Liver Cancer Using Classification Techniques. *International Journal of Recent Scientific Research Vol. 6, Issue, 6, pp.4809-4813, June, 2015*

****