International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Medical Diagnosis for Liver Cancer using Classification Techniques

Reetu[1], Narender Kumar[2]

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

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. This paper encapsulates various review and research articles on liver cancer. The main goal of this review paper is to study the related works on cancer especially liver cancer. In this paper we present an overview of the current research being carried out using the data mining techniques. 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.***

Keywords: Data Mining, Liver cancer, Decision tree, C4.5, Association rule, Bayesian networks, Support vector Machine, K-NN, Neural networks, CART.

1. Introduction

Medical data mining is capable to showing the hidden patterns in data sets of medical field. The available medical data are distributed, different in nature and more complex in nature [1]. There is more precious information and knowledge "hidden" in such database; and without any automatic methods for extracting or explore this information it is practically impossible to determine or mine that data. During the years many algorithms were created to find out what is called bunch of knowledge from large set of data.

2. Data Mining

Data mining is an important process where intelligent methods are applied to find out data patterns. It is the process of 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. 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 [2].

3. Overview of 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 [3]. Liver cancer is one of the more dangerous and threatening diseases at global level with more than one million cases diagnosed each year [4]. 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 [3]. Several types of risk factor like cirrhosis, obesity, smoking, hepatitis B and hepatitis C or alcoholism are highly linked to the liver cancer [4] [5]. The medical term that is used for the liver cancer is Hepatocellular Carcinoma. It is most harmful to life and it is more common for men than women [6].

Medical diagnosis is important and more complicated task. It needs to be executed exactly or strictly and efficiently. As, there is scare of resources such as expertise to handle these types of life threatening diseases, an automated system would be taken as better option to put all resources persons and expertise all together to create value and benefits for society at minimum cost and maximum efficiency [1]. The implementation of an automated system needs a detailed study of various techniques available.

4. Related Work

The medical diagnosis needs proficiency as well as experience in dealing with uncertainty. Although in these days, boundaries of medical science have extremely expanded. To overcome this uncertainty [7] provide a framework to construct the model using fuzzy theory towards medical diagnosis improvement. They designed a

fuzzy system for learning, analysis and diagnosis of liver disorders. This system is faster, cheaper, and also more liable and more accurate than other traditional diagnostic system. This system is compound of an expert and a fuzzy system and they are known as hybrid system. The verification or accuracy of proposed system is 91%. This system also appoints the rate of diseases intensity.

In recently ANN have been extensively used or applied in engineering, medicine, business, education, manufacturing and so on. ANN is also used in the medical diseases diagnosis. In ANN, for a high efficiency, selection of an appropriate architecture and learning algorithm is very important and it is very complex task. To evolve the ANN learning and accuracy, a new meta-heuristic algorithm, centripetal accelerated particle swarm optimization (CAPSO) is applied [8]. The hybrid learning of CAPSO and multi-layer perceptron (MLP) network are used to classify the data. The efficiency of the methods is evaluated based on mean square error, accuracy, sensitivity, specificity, and area under the receiver operating characteristics (ROC) curve. The result shows that this method gives better performance than other methods in terms of testing data and data sets with high missing values.

[9] Presented a method i.e. unsupervised feature learning can be used for cancer detection and cancer type analysis from gene expression data. This method is used for detection and classification of cancer types. This method is better than the traditional methods because of applying data from various types of cancer to automatically form feature to enhance the detection and diagnosis of a specific. The result shows that it improves the accuracy of classification problem and also provides more general and scalable approach to deal with gene expression data across different cancer types.

Traditionally data sets are single-value, single-labeled. But nowadays, there is multi-valued and multi-labeled data in real world. The traditional classifiers are not capable of handling the multi-valued and multi-labeled data. So to handle or solve this problem, Shihchieh Chou, Chang-Ling Hsu designed a decision tree classifier named MMC (multi-valued and multi-labeled classifier). But due to the problem of overfitting or to improve the accuracy they designed another classifier named MMDT (multi-valued and multi-labeled decision tree). It differs from MMC in terms of attribute selection. The result shows that MMDT has improved accuracy than MMC and gives better result. It can also apply for semi-structured data and object-oriented data. It also could be applied to the handling of more complicated data [10].

The machine learning technique is used to develop classifiers for detection or diagnosis of diseases but in clinically validated diagnostic technique so far limited and the method is prone to be overfitting. So to remove this problem, Kenneth R Foster, Robert Koprowski and Joseph D Skufca [11] focus on use of SVM, a computationally intensive statistical technique. In this model leave-one-out cross validation method is used.

[12] Developed a fuzzy MLP (Multilayer Perceptron) model to handle the uncertainty or impreciseness in input and output. This model is used as a connectionist expert system for diagnosis hepatobiliary disorders. Inputs are modeled in term of linguistic properties using pi-function. In case of missing information or partial inputs the model was capable for querying the user. The output or decision could be produced in form of rule i.e. If-Then rule form. The model is designed using logical operators based on AND/OR functions. This model can also be designed using the sigmoid-function.

In the real world, there is a highly uncertain and noisy data. To deal with highly uncertain and noisy data e.g. biochemical laboratory examinations a classifier is required. So, I-Jen Chiang, Ming-Jium Shiem, Jane Yung-Jen Hsu, Jau-Min [13] proposed a fuzzy classifier to classify the data with noise and uncertainties. Instead of determining a single class for a given instance, Fuzzy classification predicts the degree of possibility for every class.

In the biochemical laboratory examinations data, fuzzy classification tree (FCT), which combine or integrate decision tree technique and fuzzy classification, provide the efficient way to classify the data in order to generate the model for polyp screening. In FCT an instance has a membership value at each node. There is no need to generate multiple classification trees, because it uses the information-based measures. Therefore, it require less time and space. The decision tree algorithm C4.5 is useless for polyp screening. FCT is more sensitive than C4.5. The experimental result shows that, a much better prediction can be made by FCT than C4.5.

Bing-Yu Sun, Zhi-Hua Zhu, Jiuyong Li and Bin Linghu developed L1-L2 norm SVM (Support Vector Machine) as an effective classification technique for automatic feature selection. The model L1-L2 norm SVM is also used for the regression analysis with automatic feature selection. It is also an algorithm to utilize the information of censored data. In comparison with other prognosis method, this method performs feature selection and model building simultaneously. The result or performance of this method is obtain on three data sets namely, WPBC (Wisconsin Prognosis Breast cancer), DLBCL (Diffuse Large B-Cell Lymphoma), NSCLC (Non Small Cell Lung Cancer). It performs consistently better than the medium performance. It is more efficient than other algorithms with similar performance [14]. At that time we can improvement the efficiency and automation of parameter setting of the methods. Varun Kumar, Luxmi Verma [15] used the binary classification tasks to diagnosis of medical testing to know if a patient has certain diseases or not. They used the various classification algorithms namely ID3, K-NN, C4.5 and SVM on the breast cancer database and the performance of K-NN gives a promising classification result with accuracy rate and robustness. They used the "TANAGRA" tool (a data mining tool) for practical work. They used the large database "Wisconsin Breast Cancer Database "for their work.

Traditional model uses statistical methods and analyze linear data. They are thus less capable of handling massive and complicated nonlinear and dependent data. So to handle non linear and dependent data Rong-Ho Lin, Chun-Ling Chuang [16] developed an intelligent liver Diagnosis model (ILDM).

This model integrates artificial neural networks (ANN), analytic hierarchy process (AHP) and case-based reasoning (CBR) methods. These methods determine if patients suffer from liver diseases and to determine the type of the liver diseases. Here ANN is used not only for simplifies the diagnosis but also help to diagnosis the existence with greater accuracy and confidence. AHP and CBR outperform the pure CBR in terms of accuracy and discover the types of liver diseases. This model can be made by or explore by other machine learning technique, such as SVM, Bayesian Belief networks, Decision tree and Genetic algorithm.

5. Proposed Work

There are many strains (a group of organism within a species) of any group of virus which are transmitted by various arthropods. But our research focuses on a particular area. The main goal or objective of the research is to first understand or deep studies of the data mining algorithms and techniques or analyze the data from the surveys (or take live data from the hospitals) and to find which technique gives the better performance than other methods. The comparison of the techniques or methods is done on the basis of various metrics like accuracy, error rate, specificity, sensitivity, confusion matrices and ROC curve.

At the end, an effort is to be made to give some important or useful medical knowledge extracted by the methods. The analyses performed within this research are based on the data surveys and filled out by patients and cards filled out by doctors from different hospitals. Data is extracted by using a standardized data collection form and is analyzed using the WEKA tool.

In summarizing way we can separate the following methodological steps [17]:

1. Studied of various classification algorithms.
2. Collect the data for making a data set.
3. Select or separating the best or appropriate algorithms which is suitable for data sets.
4. Testing the full data sets on selected number of classification algorithms.
5. Selecting the best algorithms to use for further experiments.
6. By removing the attribute that seems to be resemblance in building the decision tree, training the selected algorithms on reduced data sets.
7. Using the most useful data in data sets identified in the step 6 modifying the algorithms' default parameter values.
8. Calculate or evaluating the result.
9. Randomizing the data sets.
10. On randomized data sets perform steps 6 and 7.
11. Calculate and evaluate results as well as algorithms performance.

These are the main steps that we are planned or used to perform with our data mining environment and data sets as well.

The choice of the WEKA tool is 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. Also, the new machine learning schemes can be developed with this package or software. WEKA is open source software issued under General Public License [18].

6. Conclusion

This paper provides or gives a study of various researches or review paper on liver cancer and explores data mining techniques that offer uncover pattern hidden in data that can help in decision making process. The presented discussion on find out the best algorithm on the liver cancer datasets is merely a short summery of the ongoing efforts in this area. It does, however point to interesting area of our research. Implementations of the techniques are highly acceptable and can help in the medical field and also help in decision making process at early diagnosis.

References

[1] Jyoti Soni, Ujma Ansari, Dijesh Sharma, Sunita Soni," Predictive Data Mining for Medical Diagnosis: Overview of Heart Diseases Prediction," *IJCST (0975-8887)*. Vol. 17, No. 8, pp, 43-48, March 2011.

[2] 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*, pp. 3465-3469, 2009.

[3] Sharifah Hafizah Sy Ahmad Ubaidillah, Roselina Sallehuddin, Noorfa Haszlinna Mustaffa," Classification of Liver Cancer Using Artificial Neural network and Support Vector Machine," *Elsevier Science Proc. Of Int. Conf on Advance in Communication Network, and Computing, CNC*, pp. 488-493, 2014.

[4] Lam, Yee Hong Brian," Proteomic Classification of Liver Cancer using Artificial Neural Network," May, 2005.

[5] Jung Hun Oh and 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. pp. 1522-1534, Nov/Dec 2011.

[6] P. Thangaraju, R. Mehala," Novel Classification Based approaches over Cancer Diseases," *IJARCCE*, Vol. 4, Issue 3, pp. 294-297, March 2015.

[7] M. Neshat, M. Yaghobi, M.B. Naghibi, A. Esmaelzadeh," Fuzzy Expert System Design for Diagnosis of Liver Disorders," *IEEE International Symposium on Knowledge Acquisition and Modeling*, pp. 252-256, 2008.

[8] Zahra Beheshti. Siti Mariyam Hj. Shamsuddin. Ebrahim Beheshti. Siti Sophiayati Yuhaniz," Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis," *Springer Science*, Vol. 18, pp. 2253-2270, 15 December 2013.

[9] Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber," Using deep learning to enhance cancer diagnosis and classification," *Proceeding of the 30th International Conference on Machine Learning, Atlanta, Georgia USA*, Vol. 28, 2013.

[10] Shihchieh Chou, Chang-Ling Hsu," MMDT: a multi-valued and multi-labeled decision tree classifier for data mining," *Elsevier Science Expert System with Application 28*, pp. 799-812, 2005.
[11] Kenneth R Foster, Robert Koprowski and Joseph D Skufca," Machine learning, Medical diagnosis, and biomedical engineering research-commentary," *Biomedical engineering online* 2014.
[12] Sushmita Mitra," Fuzzy MLP based expert system for medical diagnosis," *Elsevier Science Fuzzy Sets and System 65*, pp. 285-296, 1994.
[13] I-Jen Chiang, Ming-Jium Shiem, Jane Yung-Jen Hsu, Jau-Min Wong," Building a Medical Decision Support System for Screening by Using Fuzzy Classification Trees," *Springer Science Applied Intelligence 22*, pp. 61-75, 2005.
[14] Bing-Yu Sun, Zhi-Hua Zhu, Jiuyong Li and Bin Linghu, "Combined Feature Selection and Cancer Prognosis Using Support Vector Machine Regression," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8. NO. 6. pp. 1671-1677, Nov/Dec 2011.
[15] Varun Kumar, Luxmi Verma,"Binary Classifier for Heath Care Databases: A Comparative study of Data mining Classification Algorithms in the Diagnosis of Breast Cancer*," IJCST* Vol. 1. Issue 2, pp. 124-129, December 2010.
[16] Rong-Ho Lin, Chun-Ling Chuang, " A hybrid diagnosis model for determining the types of the Liver Diseases," *Elsevier Science Computers in Biology and Medicine 40*, pp, 665-670, 2010.
[17] Mertik M, Kokol P, Zalar B.Gaining ," Features in Medicine Using Various Data Mining Techniques," *Computational Cybernetics ICCC 2005,IEEE International Conference*, 2005, pp. 21–24, 2005.
[18] 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, pp. 520-523, 2007.