An Efficient Prediction Model for Liver Disorder Database using Data Mining Techniques

R. Thangarajan[1] S. Manoranjitha[2] C. Nandhini[3] V. Nav Krishna[4]
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]Kongu Engineering College, Perundurai., India

Abstract— Data Mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, predictions, forecasting and estimation. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. This research has developed a Decision Support in Liver Disorder Prediction System using various data mining modeling techniques, like, Decision tree, Naive Bayes, SVM, KNN, Random Forest algorithm to classify these diseases and compare the effectiveness, correction rate among them. Detection of Liver disease in its early stage is the key of its cure. It leads to better performance of the classification models in terms of their predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. The predictive performances of popular classifiers are compared quantitatively.

Key words: Classification; Data Mining; J48; Naive Bayes; SVM; KNN; Random Forest

I. Introduction

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases bur are hidden among large amounts of data.

In this research work, J48, Naive Bayes, SVM, KNN, Random Forest algorithm classifier algorithms are used for liver disease prediction. There are various numbers of liver disorders that required clinical care of the physician [3]. The main objective of this research work is to forecast liver diseases such as Cirrhosis, Bile Duct, Chronic Hepatitis, Liver Cancer and Acute Hepatitis from Liver Function Test (LFT) dataset using above classification algorithms [2]. The liver is the second largest inside organ in the human body, playing a key role in metabolism and serving several imperative functions, e.g. Decomposition of red blood cells, etc. Its weight is around three pounds. The liver does many essential functions related to digestion, metabolism, immunity, and the storage of nutrients within the body. These functions formulate the liver as an important organ, without this, body tissues would rapidly die from lack of energy and nutrients. There are number of factors which boost the risk of liver disease. Data mining is regarded as an emerging technology that has made radical change in the information world. The term `data mining' (often called as knowledge discovery) refers to the method of analyzing data from different perspectives and summarizing it into valuable information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, ³data mining is the method of finding correlations or patterns among dozens of fields in large relational databases´. Therefore, data mining consists of key functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyze data using application tools and techniques, and meaningfully present data to provide useful information [5,6].

II. Technique

Classification: Classification is the most commonly applied data mining technique, which employs a set of preclassified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Some well-known classification models are as follows.

A. Decision trees J48

J48 [14] is an important decision tree classifier. Decision tree is a predictive machine-learning representation that makes decisions the target value (dependent variable) of a fresh sample based on various attribute values of the available data. The internal nodes of a decision tree indicate the different attributes, the branches between the nodes gives the probable values that these attributes can have in the observed samples, while the terminal nodes generates the final value (classification) of the dependent variable. The attribute that is to be predicted is recognized as the dependent variable, since its value depends upon, or is chosen by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.Figure1 shows the decision tree J48 implementation using Liver Disorder Dataset.[15]

B. Naive Bayes

A Naïve Bayesian model is very simple to build and can be implemented for very huge datasets. Naive Bayesian classifier often achieves well than more sophisticated classification techniques. The posterior probability, P (c|x) is computed from P(c), P(x), and P (x|c). The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is named class conditional independence [9,10].

C. k-Nearest Neighbour

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.[16]

D. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set[17]

E. Support Vector Machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier[18]

III. Experimental Study and Analysis

A. Methodology

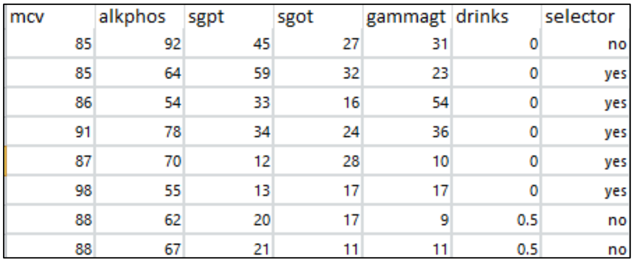
This section is comprised of the dataset description, the preprocessing procedure, and the classification algorithm. All the experimental processes have been completed using the RSTUDIO.

1) Dataset Description:

The BUPA liver disorders Dataset consists of information on 345 patients each instance is comprised of 7 attributes, which are all numeric. The detailed attributes in the dataset are listed as follows, and Table 1 shows some samples extracted from the dataset[7].

Attribute information:

1) mcv mean corpuscular volume
2) alkphos alkaline phosphotase
3) sgpt alamine aminotransferase
4) sgot aspartate aminotransferase
5) gammagt gamma-glutamyl transpeptidase
6) drinks number of half-pint equivalents of alcoholic beverages drunk per day
7) selector field used to split data into two sets



mcv	alkphos	sgpt	sgot	gammagt	drinks	selector
85	92	45	27	31	0	no
85	64	59	32	23	0	yes
86	54	33	16	54	0	yes
91	78	34	24	36	0	yes
87	70	12	28	10	0	yes
98	55	13	17	17	0	yes
88	62	20	17	9	0.5	no
88	67	21	11	11	0.5	no

Table 1: Samples of Dataset

2) Data Pre-Processing:

Data goes through a series of steps during preprocessing [20]:

– Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
– Data Integration: Data with different representations are put together and conflicts within the data are resolved.
– Data Transformation: Data is normalized, aggregated and generalized.
– Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
– Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

B. Classifier Selection

We select six commonly used classifier for prediction classification in our work based on their qualitative performance.

C. Results Analysis

We run selected classifiers in different scenarios of the dataset. By analysing the results, SVM gives the overall best classification result than other classifiers.

IV. Conclusion

An experiment is conducted to get the impact of liver disorder on the predictive performance of different classifiers. We select five popular classifiers considering their qualitative performance for the experiment. After analyzing the quantitative data generated from the computer simulations, we find that the general concept of improved predictive performance of all above classifiers but Naive Bayes performance is not significant. However more experiments with different datasets are required to support the findings. Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This research work used classification algorithms for liver disease prediction. Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time.[15]

References

[1] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2] Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.
[3] Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.
[4] Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.
[5] Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html
[6] Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.
[7] Smyth P, Breaking out of the BlackBox: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD2001), held on May 20 (2001), Santra Barbara, California, USA.
[8] Agrawal R. and Srikant R. Fast Algorithms for Mining Association Rules. In M. Jarke J. Bocca and C. Zaniolo, editors, Santiago de Chile, Chile, Sept 1994. MK
[9] J. Su, H. Zhang, C.X. Ling, S. Matwin, Discriminative parameter learning for Bayesian networks, in: Proceedings of the Twenty-Fifth International Conference on Machine Learning, ACM Press, Helsinki, Finland, 2008, pp. 1016±1023.
[10] A.A. Balamurugan, R. Rajaram, S. Pramala, et al., NB+: An improved Naïve Bayesian algorithm, Knowledge-Based Systems 24 (5) (2011) 563±569.
[11] Sta´nczyk, U.. Establishing relevance of characteristic features for authorship attribution with ANN. In: Decker, H., Lhotska, L., Link,S., Basl, J., Tjoa, A., editors. Database and Expert Systems applications; vol. 8056 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013, p. 1±8.
[12] Sta´nczyk, U.. Rough set and artificial neural network approach to computational stylistics. In: Ramanna, S., Jain, L.C., Howlett, R.J.,editors. Emerging Paradigms in Machine Learning; vol.
[13] of Smart Innovation, Systems and Technologies. Springer Berlin Heidelberg; 2013, p. 441± 470. 13.Sta´nczyk, U.. Decision rule length as a basis for evaluation of attribute relevance. Journal of Intelligent and Fuzzy Systems 2013; 24(3):429± 445.
[14] Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M., Strachan, R.. Hybrid decision tree and Naive Bayes classifiers for multi-class classification tasks. Expert Systems with Applications 2014; 41(4, Part 2):1937±1946.
[15] Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset by Tapas Ranjan Baitharua, Subhendu Kumar Panib
[16] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[17] https://en.wikipedia.org/wiki/Random_forest
[18] https://en.wikipedia.org/wiki/Support_vector_machine
[19] https://en.wikipedia.org/wiki/R_(programming_language)
[20] https://www.techopedia.com/definition/14650/data-preprocessing.