A Comparative Study of Data Mining Classification Techniques using Lung Cancer Data

1Er.Tapas Ranjan Baitharu, 2Dr.Subhendu Kumar Pani

12Associate Prof., Dept. of CSE, OEC, BPUT, Odisha

Abstract.—Now-a-days the most important cause of death for both men and women is the cancer Lung Cancer is a disease of uncontrolled cell growth in tissues of the lung. Detection of Lung Cancer in its early stage is the key of its cure. Data classification is an important task in KDD (knowledge discovery in databases) process. It has several potential applications. The performance of classifiers is strongly dependent on the data set used for learning. 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. In this paper, a comparative analysis of data classification accuracy using lung cancer data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.

Keywords—Classification, Data Mining, Lung Cancer, Naive Bayes.

1.Introduction

Data and information have become major assets for most of the organizations[1.4]. The success of any organisation depends largely on the extent to which the data acquired from business operations is utilised. In other words, the data serves as an input into a strategic decision making process, which could put the business ahead of its competitors[16]. Also, in this era, where businesses are driven by the customers, having a customer database would enable management in any organization to determine customer behaviour and preference in order to offer better services and to prevent losing them resulting better business[13,14].

Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analysing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system[3,7]. Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major 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 presents data to provide useful information.

2.TECHNIQUES AND ALGORITHMS

Researchers find two important goals of data mining: prediction and description. First, the Prediction is possible by use of existing variables in the database in order to predict unknown or future values of interest. Second the description mainly focuses on finding patterns describing the data the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differs with respect to the underlying application and technique.

2.1 Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis[2]. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification[2,17]. 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[5,6]. 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:

a) Classification by decision tree induction
b) Bayesian Classification
c) Neural Networks
d) Support Vector Machines (SVM)

2.2 Clustering: Clustering is a technique for identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Some commonly used clustering methods are:

a) Partitioning Methods
b) Hierarchical Agglomerative (divisive) methods
c) Density based methods
d) Grid-based methods
e) Model-based methods

2.3 Association rules

An Association Rule is a rule of the form milk and bread =>butter,where 'milk and bread' is called the rule body and butter the head of the rule. It associates the rule body with its head.In context of retail sales data, our example expresses the fact that people who are buying milk and bread are likely to buy butter too. This association rule makes no assertion about people who are not buying milk or bread.We now define an association rule:Let D be a database consisting of one table over n attributes {a1, a2, . . . , an}. Let this table contain k instances.The attributes values of each ai are nominal1. In many real world applications (such as the retail sales data) the attribute values are even binary (presence or absence of one item in a particular market basket)[8,9]. In the following an attribute-value-pair will be called an item. An item set is a set of distinct attribute-value-pairs.Let d be a database record. d satisfies an item set X _ {a1, a2, . . . , an} if X _ d.An association rule is an implication X) Y where X, Y _ {a1, a2, . . . , an}, Y 6= ; and X \ Y = ;.The support s(X) of an item set X is the number of database records d which satisfy X.Therefore the support s(X) Y) of an association rule is the number of database records that satisfy both the rule body X and the rule head Y . Note that we define the support as the number of database records satisfying X \ Y , in many papers the support is defined as s(X\Y) k . They refer to our definition of support as support count.The confidence ˆc(X) Y of an association rule X) Y is the fraction ˆc(X) Y) = s(X\Y) s(X) .From a logical point of view the body X is a conjunction of distinct attribute-value-pairs

and the head Y is a disjunction of attribute-value-pairs where X \ Y = ;.Coming back to the example a possible association rule with high support and high confidence would be i1) i2 whereas the rule i1) i3 would have a much lower support value.

2.4 Class association rules

The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for nominal attributes. However, association rules in their general form cannot be used directly. We have to restrict their definition. The head Y of an arbitrary association rule X) Y is a disjunction of items. Every item which is not present in the rule body may occur in the head of the rule. When we want to use rules for classification,we are interested in rules that are capable of assigning a class membership. Therefore we restrict the head Y of a class association rule X) Y to one item. The attribute of this attribute-value-pair has to be the class attribute.According to this, a class association rule is of the form X) ai where ai is the class attribute and X _ {a1, . . . , ai−1, ai+1, . . . , an}.

The Apriori algorithm [8,14] has become the standard approach to mine association rules. We have adapted it to mine class association rules in the way explained by Liu et al. [9,13].The second algorithm, Predictive Apriori, has been recently proposed by Scheffer [10,12].Both algorithms have their first step in common. They generate frequent item sets in the same way. An item set is called frequent when its support is above a predefined minimum support.

3.Experimental Study and Analysis

3.1 WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

3.2 Dataset Description

We performed computer simulation on a lung cancer dataset available UCI Machine Learning Repository [11,15]. . The features describe different factor for cancer reoccurrence.. The dataset contains 768 instances shown as no reoccurrence-events while 85 instances as reoccurrence-events.

3.3 Results Analysis

We run selected classifiers in different scenarios of the dataset.

Classifier Performance

Classifier	Confusion Matrix	Avg. Precision	Avg. Recall	Avg. F-Mean
NB	a b <-- classified as 5 4 | a = no-occurrence-events 3 20 | b = occurrence-events	0.625	0.556	0.590
MULTILAYER PERCEPTRON	a b <-- classified as 3 6 | a = no-occurrence-events 5 8 | b = occurrence-events	0.375	0.333	0.354
1Bk	a B <-- classified as 3 6 | a = no-occurrence-events 4 19 | b = occurrence-events	0.429	0.333	0.381
J48	a B <-- classified as 4 5 | a = no-occurrence-events 2 21 | b = occurrence-events	0.667	0.444	0.555

4. CONCLUSION

In this paper we conducted an experiment to find the impact of lung cancer data on the predictive performance of different classifiers. We select five popular classifiers considering their qualitative performance for the experiment. After analysing the quantitative data generated from the computer simulations, we find that the general concept of improved predictive performance of all above classifiers but 1Bk performance is not significant.

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 Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), 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, Proceeedings of the 20th International Conference on Very Large Data Bases (VLDB'94), pages 475–486, Santiago de Chile, Chile, Sept 1994 . Morgan Kaufmann.
9. Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. Unpublished manuscript.
10. Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. In L. De Raedt and A. Siebes, editors, Proceeedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), pages 424–435, Freiburg, Germany, September 2001. Springer-Verlag.
11. UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/.
12. SAS Institute Inc., Lie detector software: SAS Text Miner (product announcement), Information Age Magazine, [London, UK], February 10 (2002), Available at: http://www.sas.com/solutions/fraud/index.html.
13. Berry M J A and Linoff G S, Data mining techniques: for marketing, sales, and relationship management, 2 nd edn (John Wiley; New York), 2004.
14. Delmater R and Hancock M, Data mining explained: a manager's guide to customer-centric business intelligence, (Digital Press, Boston), 2002.
15. Fuchs G, Data Mining: if only it really were about Beer and Diapers, Information Management Online, July 1, (2004), Available at: http://www.information-management.com/ news/1006133-1.html.
16. Subhendu Kumar Pani and Satya Ranjan Biswal and Santosh Kumar Swain, A Data Mining Approach to Identify Key Factors for Systematic Reuse (October 31, 2012). The IUP Journal of Information Technology, Vol. VIII, No. 2, June 2012, pp. 24-34. Available at SSRN: http://ssrn.com/abstract=2169262
17. Subhendu Kumar Pani and Amit Kumar and Maya Nayak," Performance Analysis of Data Classification Using Feature Selection" (October 24, 2013). The IUP Journal of Information Technology, Vol. IX, No. 2, June 2013, pp. 36-50.