International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 4 Issue 6, June-2015, pp: (45-49), Impact Factor: 1.296, Available online at: www.erpublications.com

A literature review on Data Mining Algorithm for Classification of Cancer Dataset

Nidhi Dangi[1], Mr. Prashant Ahlawat[2]

[1]Department of CSE, Gurgaon Institute of Technology & Management, Gurgaon, Haryana
[2]Asst. Professor, Department of CSE, Gurgaon Institute of Technology & Management, Gurgaon, Haryana

ABSTRACT

Different algorithms based on soft computing and hard computing has been developed to apply on medical datasets. Hybrid techniques find their way out to search for the solution of cancer databases. Some are inspired by nature and some are inspired by biological phenomenon. Biological inspired methods are successfully developed and applied for different medical problems. In current research, an intelligent technique Naive based decision tree have been applied and evaluated successfully to classify lung cancer based datasets. A comparison has been made with other techniques to check the effectiveness of the proposed method. TP rate, ROC and Precision is highest for proposed method amongst other method.

Keywords: data mining, algorithm, naive bayse, dataset.

INTRODUCTION

Data mining is the process of digging data for discovering latent patterns which can be translated into valuable information. Data mining usage witnessed unprecedented growth in the last few years, the usefulness of data mining techniques has been realized in Healthcare domain. This realization is in the wake of explosion of complex medical data. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on. The data mining techniques that have been applied to medical data include .There are several major data mining techniques that have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction, which is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers and therefore they can put beers and crisps next to each other to save time for customer and increase sales.

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

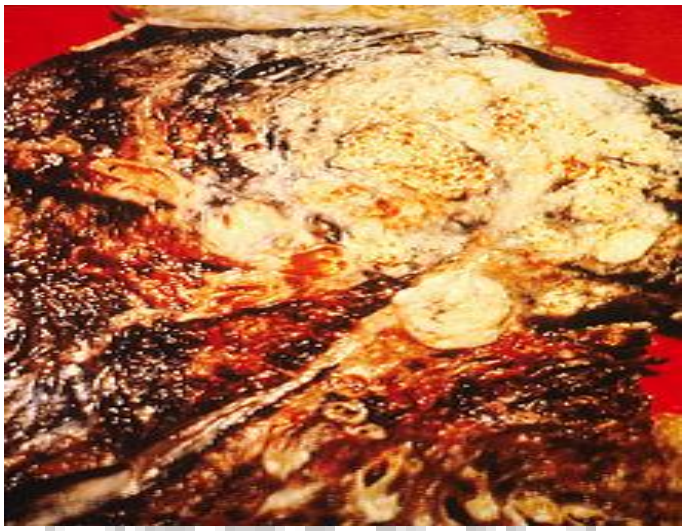

Fig. 1

Cross section of a human lung: The white area in the upper lobe is cancer; the black areas are discoloration due to smoking. Smoking, particularly of cigarettes, is by far the main contributor to lung cancer. Cigarette smoke contains at least 73 known carcinogens, including benzo [a] pyrene, NNK, 1,3-butadiene and the radioisotope polonium-210. Across the developed world, 90% of lung cancer deaths in men during the year 2000 were attributed to smoking (70% for women).[19] Smoking accounts for 80–90% of lung cancer cases. Passive smoking—the inhalation of smoke from another's smoking—is a cause of lung cancer in nonsmokers. A passive smoker can be defined as someone living or working with a smoker. Studies from the US, Europe and the UK[have consistently shown a significantly increased risk among those exposed to passive smoke. Those who live with someone who smokes have a 20–30% increase in risk while those who work in an environment with second hand smoke have a 16–19% increase in risk. Investigations of sidestream smoke suggest it is more dangerous than direct smoke Passive smoking causes about 3,400 deaths from lung cancer each year in the USA. Marijuana smoke contains many of the same carcinogens as those in tobacco smoke. However the effect of smoking cannabis on lung cancer risk is not clear. A 2013 review did not find an increased risk from light to moderate use. A 2014 review found that smoking cannabis doubled the risk of lung cancer.

LITERATURE SURVEY

Goodwin et al. applied data mining techniques for birth outcomes. Data mining is useful in medical applications such as medications, medical tests, prediction of surgical procedures, and discovery of relationships between pathological data and clinical data.

Evans et al. stated that hereditary syndromes can be detected automatically using data mining techniques. Data mining plays an important role in IT as it discovers knowledge from historical data of various domains. For instance data mining can be used to mine medical data as Healthcare domain produces huge amount of data about patients, diseases, diagnosis, medicine and so o.n

Doron Shalvi and Nicholas De Claris, discussed medical data mining through unsupervised neural networks besides a method for data visualization. They also emphasized the need for preprocessing prior to medical data mining. By applying data mining techniques in Healthcare domain, the administrators can improve the QoS (Quality of Service) by discovering latent potentially useful trends required by medical diagnosis

Krzysztof J. Cior , bioengineering professor, identified the need for data mining methods to mine medical multimedia content. They used the image segmentation to segment the breast tissue corresponding to the tumor and used the discrete wavelet transform (DWT) as a feature extraction method to extract various features from the segmented images.

Tsumoto identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS).

they also used SVM classifier to classify the breast tissue corresponding to the features and achieved an accuracy of 88.75%.

Brameier and Banzhaf explored and analyzed two programming models such as neural networks, and linier genetic programming for medical data mining. . Apriori and FP Growth are the most widely used frequent pattern mining algorithms.

Abidi and Hoe proposed and implemented a symbolic rule extraction workbench for generating emerging rule-sets. They also emphasized the need for pre-processing prior to medical data mining They used K-means clustering algorithm for image segmentation and gray level co-occurrence matrix to describe and analyze the texture of segmented structures in the image.

Abidi et al. explored the usage of rule-sets as results of data mining for building rule-based expert systems. multimedia content. They identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS).

Olukunle and Ehikioya proposed an algorithm for extracting association rules from medical image data. The association rule mining discovers frequently occurring items in the given dataset. an association rule learner which is based on the criteria collected from past breast cancer patients. The rule learner is used in a tool by name "Clinical Trial Assignment Expert System"

Shim and Xu proposed a classification method based on Bayesian Ying Yang (BYY) which is a three layered model. They applied this model to classify liver disease through automatic discovery of medical trends. The classification
of these structures was achieved through Support Vector Machines, which separate them into two groups; using shape and texture descriptors: masses and non-masses

Brunie et al. proposed architecture for mining geno-medical data in heterogeneous and grid-based distributed infrastructures. They used Particle Swarm Optimization technique as well. The results revealed that, their approach is capable of performing surgery candidate selection process effectively in epilepsy

Mahmud Khan et al. focused on decision tree data mining algorithm for medical image analysis. Especially they studied on lung cancer diagnosis through classification of x-ray images.

PROBLEM FORMULATION

CHALLENGES

1) Tedious and time consuming Mathematical Formulation

2) Memory consumption

3) In previous researches, researchers have used methods which are dependent on each other. Thus, error in any one of the feature result in malfunction of the complete classification system.

ADVANTAGES OF PROPOSED METHOD

- Fast Convergence
- Linearize the non linear data
- Intelligent Decision Making

OBJECTIVES

1. To develop a NBT algorithm for the classification of cancer data sets.

2. To evaluate the developed algorithm on cancer data sets.

3. Compare the proposed method with other algorithms.

OTHER ALGORITHMS

BAYESIAN NETWORKS

A Bayesian network (also referred to as Bayesian belief network, belief network, probabilistic network, or causal network) consists of a qualitative part, encoding existence of probabilistic influences among a domain's variables in a directed graph, and a quantitative part, encoding the joint probability distribution over these variables. Each node of the graph represents a random variable and each arc represents a direct dependence between two variables. The directed graph is a representation of a factorization of the joint probability distribution. As there can be many graphs that are capable of encoding the same joint probability distribution.

SUPPORT VECTOR MACHINE

SVMs are inspired by the Structural Risk Minimization principle from statistical learning theory. In their basic form, SVMs attempt to perform classification by constructing hyper planes in a multidimensional space that separates the cases of different class labels. It backs both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models like linear, polynomial. In the last years, SVMs have been widely investigated and used in a lot of different fields and for various classification tasks, due to their good performances. Learning algorithms such as neural network & SVMs, both trained with different parameters and input features, showed that SVMs produce the most robust results.

MULTILAYER PERCEPTRON TREE

A supervised multilayer perceptron tree (SMLPT) is a trained feed forward neural network model that maps sets of input data onto a set of appropriate outputs. A SMLPT consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. SMLPT utilizes a supervised learning technique called back propagation for training the network through attribute selection feature. SMLPT is a modification of the standard multilayer perceptron and can distinguish data that are not linearly separable.

CONCLUSION

According to the literature survey and analysis, the proposed NBT based approach evolved as optimal approach to classify the cancer datasets with a remarkable accuracy and fast computation time as compared to MLP techniques and other classification methods. With such high accuracy in proposed method, it will be easy to identify the cancer and non cancer patients from different attributes of people for large data chunks where other decision tree algorithms fail to achieve high accuracy.

REFERENCES

[1] Carson Kai-Sang Leung, Christopher L. Carmichael and BoyuHao. (2007). “Efficient Mining of Frequent Patterns from Uncertain Data”,. Proceedings of IEEE ,pp.489-494.
[2] Shariq Bashir, Zahid Halim, A. Rauf Baig. (2008).,” Mining Fault Tolerant Frequent Patterns using Pattern Growth Approach”. Proceedings of IEEE ,pp.172-179.
[3] Sunil Joshi and Dr. R. C. Jain.(2010). “A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database”, Proceedings of IEEE, pp.498-501.
[4] Thanh-Trung Nguyen. (2010).”An Improved Algorithm for Frequent Patterns Mining Problem”, Proceedings of IEEE, pp.503-507.
[5] Xiaoyong Lin and Qunxiong Zhu. (2010). “Share-Inherit: A novel approach for mining frequent patterns”, Proceedings of IEEE, pp.2712-2717.
[6] Markus Brameier and Wolfgang Banzhaf. (2001).”A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining”, Proceedings of IEEE ,pp.1-10.
[7] Safwan Mahmud Khan Md. Rafiqul Islam Morshed U. (2006).“Medical Image Classification Using an Efficient Data Mining Technique”, Proceedings of IEEE, pp.1-6.
[8] Yanwei Xing, Jie Wang and Zhihong Zhao (2007).“Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease”.Proceedings of IEEE.pp.1-5.

[9] Tsang-Hsiang Cheng, Chih-Ping Wei, Vincent S. Tseng (2009). "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches". Proceedings of IEEE.pp.1-6.

[10] Mohammad Saraee, George Koundourakis, Babis Theodoulidis. (2007). "EasyMiner: Data Mining In Medical Databases", Proceedings of IEEE, pp.1-3.

[11] Sam Chao(2009) "An Incremental Decision Tree Learning Methodology Regarding Attributes In Medical Data Mining". Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, pp.101-105.