ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

A Literature Survey on Healthcare Data using Data Mining Classification Methods

V. Murali [1] Dr. T. Balasubramanian[2]

[1]Assistant Professor, Dept. of Computer Science, MGR College,Hosur.

[2]Assistant Professor & HEAD, Dept. of Computer Science, Sri Vidhya Mandhir, Uthangaria, Krishnagiri.

Abstract

The role of data mining in healthcare is to discover useful and perceivable patterns by analyzing large sets of data. These data patterns help forecast and then determine what to do about them. In the healthcare industry specifically, data mining can be used to decrease costs by increasing efficiencies, improve patient quality of life, and perhaps most importantly, save the lives of more patients. Classification is a data mining function that allocates items in a collection to target groups or classes. The aim of classification is to precisely predict the target class for each case in the data. Due to this we have evaluate various papers involved in the field of data mining methods, algorithms and theirs results. This survey paper shows some of the recently reviewed papers in order to the methods, tasks and results. Performances and accuracy of each method are discussed for selected papers and a summary table of the finding is shown to conclude the paper.

Keywords: Data mining Techniques, SVM, Bayes Classifier, Classification, Healthcare, Tools

I. Introductions

Data mining is depicted as the process of detecting correlations, patterns and trends to search through a huge amount of data stored in repositories, databases, and data warehouses. Humans in that sensitivity are restricted by data overload so there are new tools and methods are being improving to solve this problem through computerization. Data mining accepts a series of pattern detection technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Electronic health records (EHR) are rapidly becoming more common among healthcare services. With increased access to a large amount of patient data, healthcare providers can now optimize the efficiency and quality of their organizations using data mining. In healthcare, data mining has proved the effectiveness in areas such as predictive medicine, customer relationship management, finding of fraud and abuse, management of healthcare and determining the effectiveness of certain treatments. The role of data mining in healthcare is to discover useful and perceivable patterns by analyzing large sets of data. These data patterns help forecast and then determine what to do about them. In the healthcare industry specifically, data mining can be used to decrease costs by increasing efficiencies, improve patient quality of life, and perhaps most importantly, save the lives of more patients.

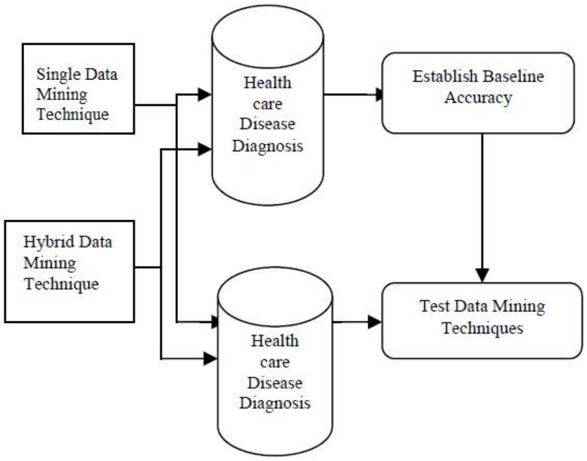
Single Data Mining Technique
Hybrid Data Mining Technique
Health care Disease Diagnosis
Health care Disease Diagnosis
Establish Baseline Accuracy
Test Data Mining Techniques

Fig. Healthcare Disease Diagnosis using Data Mining

Classification techniques in data mining are capable of processing a huge amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any background in which some choice or predict is made on the basis of currently available information. Classification method is recognized system for repeatedly making such decisions in new situations. The aim of classification is to precisely predict the target class for each case in the data. In the training process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for detecting relationships. These relationships are reviewed in a model, which can then be applied to a dissimilar data set in which the class assignments are unfamiliar.

Here if we assume that problem is a unease with the construction of a procedure that will be applied to a continuing progression of cases in which each new case must be allocated to one of a set of predefined classes on the basis of experimental features of data. Creation of a classification procedure from a set of data for which the precise classes are known in advance is termed as pattern recognition or supervised learning. Several data mining techniques are used in the Healthcare management system such as CART, Artificial Neural Network, K nearest neighbour method, Support Vector Machine, Naïve Bayes Classifier and C4.5 showing different levels of accuracies.

II. Data Mining Tasks

Data mining used in different type of techniques to extract the knowledge from the data, the techniques are: [3]

- Anomaly detection (Outlier/ deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.(ex-bank fraud identification) Three type of anomaly detections: [15]

- Unsupervised Anomaly Detection
- Supervised Anomaly Detection
- Semi-Supervised Anomaly detection

➢ Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

➢ Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

➢ Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

➢ Regression – attempts to find a function which models the data with the least error.

➢ Summarization – providing a more compact representation of the data set, including visualization and report generation.

III. Data Mining Classification Algorithms

1. Decision Tree

Decision tree uses the simple divide - and conquer algorithm. In these tree structures, leaves represent classes and branches signify conjunctions of features that lead to those classes. The attribute that most effectively splits samples into different classes is chosen, at each node of the tree. A path to a leaf from the root is found depending on the assessment of the predicate at each node that is visited, to predict the class label of an input. Decision tree is fast and easy method since it does not require any domain information. In the decision tree inputs are divided into two or more groups continue the steps till to complete the tree.

1.1 C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier One limitation of ID3 is that it is overly sensitive to features with large numbers of values. C4.5 uses a metric called "information gain," which is defined by subtracting conditional entropy from the base entropy; that is, Gain (P|X) =E (P)-E (P|X). This computation does not, in itself, produce anything new. However, it allows you to measure a gain ratio. Gain ratio, defined as Gain Ratio (P|X) =Gain (P|X)/E(X), where(X) is the entropy of the examples relative only to the attribute. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too- much details in the training data set. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute. Decision trees are built in C4.5 by using a set of training data or data sets as in ID3. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sub lists.

Pseudocode: C4.5 General algorithm for building decision trees is:

```
1. Check for any base cases
2. For each attribute a
3. Find the normalized information gain from splitting on a
4. Let a_best be the attribute with the highest normalized information gain
5. Create a decision node that splits on a_best
6. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node
```

2. Nearest Neighbor

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. [9] The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified.[15].

Unlike other techniques, there is no learning process to create a model. The data used for learning is in fact a model. When the new data shows up, the algorithm analyzes all the data in the database to find a subset of instances that are the best fit and based on that it is able to predict the outcome.

3. Bayesian Classification

A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. [1]A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. [8] Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. [14] Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Bayesian classifier is a statistical classification approach based on the Bayes theorem.

Theorem:

To calculate probability of A given B, P (B given A) = P (A and B)/P (A) the algorithm counts the number of cases where A and B occurs simultaneously and divides it by the number of cases where A alone occurs.

Let X be a data tuple, X is considered “Evidence”, in Bayesian terms.

Let H be some hypothesis, such that the data tuple X belongs to class C.

P (H|X) is posterior probability, of H conditioned on X.

P (H) is the prior probability of H in contract.

4. Neural Networks

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interrelated group of artificial neurons and processes information using a connectionist approach to computation. Neurons are structured into layers. The input layer consists of the original data, while the output layer nodes represent the classes. There may be several hidden layers. A main feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained.

A main concern of the training phase is to focus on the interior weights of the neural network, which is used according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output.

Neural network represent a brain image or symbol for Information processing.[1][3]

These models are biologically inspired rather than an exact replica of how the brain actually functions.[8] Neural networks have been shown to be very talented systems in many forecasting applications and business classification applications due to their ability to “learn” from the data, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize. Neural computing refers to a pattern recognition methodology for machine learning.[11] The resulting model from neural computing is often called an artificial neural network (ANN) or a neural network. Neural networks have been used in many business applications for pattern recognition, forecasting, prediction, and classification.

The human brain possesses distract capabilities for information processing and problem solving that modern computer cannot compete with in many aspects. [1]It has been predicate that a model or a system that is open minded or liberal and supported by the results from brain research, with a structure similar to that of biological neural networks, could exhibit similar intelligent functionality. Based on this bottom-up guess, ANN (also known as connectionist models, parallel distributed processing models, neuromorphic systems, or simply neural networks) have been developed as biologically inspired and plausible models for various tasks. [12] Biological neural networks are composed of many massively interconnected primitive biological neurons. Each neuron possesses axons and handwrites finger-like projections that enable the neuron to communicate with its neighboring neurons by transmitting and receiving electrical and chemical signals. More or less resembling the structure of their counterparts, ANN are composed of interconnected, simple processing elements called artificial neurons.ANN possess some desirable traits similar to those of biological neural networks, such as the capabilities of learning, self-organization, and fault tolerance.

5. Support Vector Machine

SVM has develop as one of the most popular and useful techniques for data classification [7]. It can be used for classify the both linear and nonlinear data. [2] The objective of SVM is to produce a model that predicts the target value of data occurrence in the testing set in which only attributes are given.[8] The classification goal in SVM is to separate the two classes by means of a function prepare from available data and their by to produce a classifier that will work well on further unseen data. [8] The simplest form of SVM classification is the maximal margin classifier. It is used to solve the most basic classification problem, namely the case of a binary classification with linear separable training data. [1] The aim of the maximal margin classifier is to find the hyperplane with the largest margin, i.e., the maximal hyperplane, in real-world problems, training data are not always linear separable. In order to handle the nonlinearly separable cases some slack variables have been introduced to SVM so as to tolerate some training errors, with the influence of the noise in training data thereby decreased. This classifier with slack variables is referred to as a soft-margin classifier.

TABLE – Performance of Various Classification algorithms and their data sets with tools – Reviewed papers

Author	Year	Tool	Diseases	Techniques	Accuracy
Gaganjot Kaur and Amit Chhabra, 2014 [4]	2014	WEKA	Diabetes Mellitus	Modified J48 Classifier	99.87%
P.Radha and Dr. B. Srinivasan, 2014 [5]	2014	Tanagara	Diabetes Mellitus	C4.5	86%
Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, 2014 [6]	2014	Matlab	Diabetes Mellitus	Bayesian Networks	90.4%
M. Lavanya, P.M. Gomathi, 2016 [9]	2016		Heart diseases	Naïve Bayes	85.92%
				KNN	100%
				J48	91.85%
				ANN	99.25%
Hlaudi Daniel Masethe, Mosima Anna Masethe,2014 [10]	2014		Heart diseases	J48	99.0741%
				Naïve Bayes	97.222%
				Simple CART	99.074%
A.Priyanga , S.Prakasam,2013[13]	2013	WEKA	Breast Cancer	J48	98.16%
				ID3	100%
				Naïve Bayes	86.23%
A.Priyanga , S.Prakasam,2013[13]	2013	WEKA	Lung Cancer	J48	98.3%
				ID3	100%
				Naïve Bayes	89.03%
A.Priyanga , S.Prakasam,2013[13]	2013	WEKA	Skin Cancer	J48	80%
				ID3	100%
				Naïve Bayes	78.3%
Dr. S. Vijayarani, Mr.S.Dhayanand, 2015 [16]	2015	MATLAB	Kidney Diseases	Naïve Bayes	70.96%
				SVM	76.32%
Dr. S. Vijayarani, Mr.S.Dhayanand, 2015[17]	2015	MATLAB	Kidney Diseases	ANN	87.7%

IV. Conclusion

This paper deals with various classification techniques including Decision tree classification, Bayesian Classification, nearest neighbour classification, Neural Network, and Support Vector Machine and a study on each of them. Data mining is a wide area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large volumes of data. Classification methods are typically strong in modelling interactions Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. These classification algorithms can be implemented on different types of healthcare data sets like Diabetes Mellitus, Heart attack, Lung cancer according to performances. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. Each technique has got its own pros and cons as given in the paper. Based on the needed Conditions each one as needed can be selected On the basis of the performance of these algorithms and further improvement can be made.

V. References

[1] Jiawei Han, Micheline Kambar, Jian Pei, "Data Mining Concepts and Techniques" Elsevier Second Edition.

[2] Galit Shmueli, Nitin R.Patel, Peter C.Bruce, "Data Mining Business Intelligence" Wiley India Edition. [7] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method", VOL. 23, NO. 3, AUGUST 2008.

[3] Margaret H.Dunham, "Data Mining Introductory and Advanced Topics" Pearson Education.

[4] Gagajot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the prediction of Diabetes", International Journal of Computer Applications(0975-8887) vol.98 No.22, July 2014.

[5] P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339

[6] Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes" , MAGNT Research Report, vol.2(5), pp.892-902.

[7] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method", VOL. 23, NO. 3, AUGUST 2008.

[8] Vipin Kumar , J. Ross Quinlan, Joy deep Ghosh ,Qiang Yang ,Hiroshi Motoda, Geoffrey J. McLachlan , Angus Ng , Bing Liu, "Survey paper on Top 10 Algorithms in Data Mining", 4 December 2007© Springer-Verlag London Limited 2007.

[9] M. Lavanya, P.M. Gomathi, "Prediction of Heart Diseases using Classification Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). Volume 5, Issue 7, July 2016.

[10] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Diseases using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II. WCECS 2014, 22-24 October, 2014, San Francisco, USA

[11] Guoqiang Peter Zhang, "Neural Networks for Classification: A Survey" IEEE VOL. 30, NO. 4, NOVEMBER 2000.

[12] Yashpal Singh, Alok Singh Chauhan, "Neural Networks in Data Mining", Journal of Theoretical and Applied Information Technology.

[13] A.Priyanga , S.Prakasam, Ph.D "Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS) ", International Journal of Computer Applications (0975 – 8887) Volume 83 – No 10, December 2013

[14] J.Aroba ,J.A. Grande, J. M. Andu´jar,M. L. de la Torre, J. C. Riquelme, "Application of fuzzy logic and data mining techniques as tools for qualitative interpretation of acid mine drainage processes", 8 December 2006 Springer-Verlag 2007.

[15] www.wikipedia.org

[16] Dr. S. Vijayarani, Mr.S.Dhayanand, "Data Mining Classification Algorithms for Kidney Diseases Prediction", International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015

[17] Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Diseases Prediction Using SVM and ANN Algorithms", International Journal of Computing and Business Research (IJCBR) ISSN (Online): 2229-6166, Volume 6 Issue 2 March 2015