International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Classification of Medical Dataset using Hybrid Feature Selection & Enhanced Decision Table Classification Approach

Parneet Kaur[1], Deepak Aggarwal[2]

[1]Scholar at Department of CSE, BBSBEC, Punjab, India

[2]Professor, Department of CSE, BBSBEC, Fatehgarh Sahib, Punjab, India

Abstract: ***Data mining is a emerging area of research that has been used for classification, clustering and association. In this paper classification approaches have been discussed that has been used for prediction of a class label to a instance available in the dataset. In this paper main concern of classification has been based on medical data classification. This classification of dataset can be used for prediction of various diseases to different patients on the basis of initial test values. In this paper diabetes dataset has been classified for prediction of accuracy of classification approach that use decision tables based classification using support and confidence computed for single instance available in dataset. In this paper purposed approach provides much better classification.***

Keywords: Fuzzy KNN, Naive Bayes, classification approaches, Decision Tree

1. Introduction

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem. Table 1 below illustrates the training and prediction sets of such database.

1.1 Dataset

The experiment1 dataset consists of 600 samples and experiment2 dataset consists of 298 samples. Its attributes represents the CBC features.Some features; such as the sex, age, and some others features which are dropped due the privacy of the blood sample's owner, and finally it contain diagnoses attribute which represent the target label of the sample, it has several labels: Suggestive of anaemia of chronic disorder, Eosinophilia, Microcytic hypochromic anaemia, Normocytic anaemia, Neutrophil leucocytosis, Neutrophilia, Non-specific findings, High ESR, and Other which represented any other haematological data comments. In the pre-processing of the dataset we eliminate useless attributes, refill the missing values and remove/refill the outlier values on the outlier samples.

1.2 Naive Bayes``

Naive Bayes implements the probabilistic Naïve Bayes classifier. Naïve Bayes Simple uses the normal distribution to model numeric attributes. Naïve Bayes can use kernel density estimators, which develop performance if the normality assumption if grossly correct; it can also handle numeric attributes using supervised discretization. Naïve Bayes Updateable is an incremental version that processes one request at a time. It can use a kernel estimator but not discretization.

1.3 Fuzzy KNN (k nearest neighbour)

A "Fuzzy KNN" algorithm utilizes strength of test sample into any class called fuzzy class membership and thus produces fuzzy classification rule. An example of Fuzzy set is the set of real numbers much larger than zero, which can be defined with a membership function as follows:

Numbers less than zero are not in the set because value of membership function for those is zero. While numbers larger than zero are in the set based on strength of numbers with respect to zero. This makes Fuzzy Set a useful tool for classification of samples having imprecise boundary. Fuzzy Set gives degree of presence of any sample into specific class.

2. Review of Literature

Tzung-Pei Hong explained in his paper "**GA-based item partition for data mining" w**hen a mining procedure is directly executed on very large databases, the computer memory may not allow the processing in memory. In the past, we adopted a branch-and-bound search strategy to divide the domain items as a set of groups. Although it works well in partitions the items, the time is quite time consuming. In this paper, we thus propose a GA-based approach to speed up the partition process. A new encoding

representation and a transformation scheme are designed to help the search process. Experimental results also show that the algorithm can get a proper partition with good efficiency.

Wang, Guoyin described in the paper " Granular computing based data mining in the views of rough set and fuzzy set" in this data mining is performed at granual level using rough set as fuzzy sets.Data mining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In our data-driven data mining model, knowledge is originally existed in data, but just not understandable for human. Data mining is taken as a process of transforming knowledge from data format into some other human understandable format like rule, formula, theorem, etc. In order to keep the knowledge unchanged in a data mining process, the knowledge properties should be kept unchanged during a knowledge transformation process. Many real world data mining tasks are highly constraint-based and domain-oriented. Thus, domain prior knowledge should also be a knowledge source for data mining. The control of a user to a data mining process could also be taken as a kind of dynamic input of the data mining process. Thus, a data mining process is not only mining knowledge from data, but also from human. This is the key idea of Domain-oriented Data-driven Data Mining (3DM).

Jagannathan, included in the paper **"**Seventh IEEE International Conference on Data Mining Workshop"The following topics are dealt with: data mining in Web 2.0 environment; knowledge-discovery from multimedia data and multimedia applications; mining and management of biological data; data mining in medicine; optimization-based data mining techniques; high performance data mining; mining graphs and complex structures; data mining on uncertain data; data streaming mining and management; spatial and spatio-temporal data mining.

Tzung-Pei Hong investigated in the paper "Using divide-and-conquer GA strategy in fuzzy data mining" that Data mining is most commonly used in attempts to induce association rules from transaction data. Transactions in real-world applications, however, usually consist of quantitative values. This work thus proposes a fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions. A GA-based framework for finding membership functions suitable for mining problems is proposed. The fitness of each set of membership functions is evaluated using the fuzzy-supports of the linguistic terms in the large 1-itemsets and the suitability of the derived membership functions. The proposed framework thus maintains multiple populations of membership functions, with one population for one item's membership functions. The final best set of membership functions gathered from all the populations is used to effectively mine fuzzy association rules.

Mikhailov, L described in the paper **"Method for fuzzy rules extraction from numerical data"** The main idea consists in separation of the input space into activation rectangles, corresponding to different output intervals. The generation of fuzzy rules and membership functions is based on these activation rectangles, whereas an appropriate fuzzy rules inference mechanism is proposed. The method formalizes the synthesis of the fuzzy system and could be used for function approximation, classification and control purposes. An illustrative example for implementation of the method for synthesis of traffic fuzzy control is given.

3. Methodology

Data mining is process that has been used for extraction of different patterns from raw information for classification. In this process various rules have been generated from the training dataset for classification of instances available in testing dataset.

Advances in Data mining classification used for enhancement of medical diagnosis based on different attributes computed from patient health.

In process of classification diabetes prediction has been done form various instances on the basis of values of parameters of previous predicted patient dataset.

In the purposed work classification of diabetes dataset has been done on the basis of classification approach that has been used for enhancing decision based classification approach. In this approach classification has been done using enhanced decision table classifier. Decision table classifier predicts different instances attributes that has been used for computation of class count and global count. These values for particular instance have been used for classification that can be used for computation of support and confidence. On the basis of maximum support and confidence value classification label has been added to particular class. In the purposed work various phases have been used for selection of dataset classification. These various processes have been discussed for classification of a dataset.

- **Dataset Accusation**

In this step dataset has been loaded to system. Dataset contains various attributes and class labels for all the instances available in the dataset. In the purposed work diabetes dataset has been loaded to the system for classification process. Dataset loaded to the system can be in arff, csv, txt or in db format that can be used for classification and preprocessing process. Diabetes dataset contains 768 instances and 8 attribute and one is class attribute.

Table 4.1: Dataset description of different attributes of diabetes dataset

Dataset Attribute	*Description*
PREG	Represents value of number of times pregnant
PLAS	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
PRES	Diastolic blood pressure (mm Hg)
SKIN	Triceps skin fold thickness (mm)
INSU	2-Hour serum insulin (mu U/ml)
MASS	Body mass index (weight in kg/(height in m)^2)
PEDI	Diabetes pedigree function
AGE	Age (years)
CLASS	Test positive and test negative

Table 4.1 represents description of the all the attributes available in the dataset. These attribute are essential for classification process that can be used for classification to a testing dataset.

- **Preprocessing of dataset**

In this phase dataset that has been loaded to the system has been preprocessed using various filters. In the process of preprocessing instances values have been checked for anomalies and missing values. Missing values available in the dataset has been preprocessed using various filters. These filters are used for conversion of dataset value to nominal, numeric or null values.

- **Training of Dataset**

After preprocessing of dataset each attribute has been trained by using dataset. This dataset has been used for training of the classifier. In the process of training dataset has been used for extraction of different rules based on different attributes of the dataset. These rules have been generated on the basis of different attributes combination and on the basis of these rules dataset classification has been done for testing dataset.

- **Testing Dataset**

After generation of different rules from training dataset, rules have been implemented on testing dataset. On the basis of these rule dataset has been classified into different class labels. After prediction of class labels dataset classification parameters are analyzed for performance evaluation of classifier.

4. Results

In the process of data mining different classification approaches have been used for classification process. In the purposed work enhanced decision table classifier has been used for classification of diabetes dataset. Decision table classifier generates a table of support and confidence on the basis of structured dataset using two or more than two attributes. Attribute selection has been done using different attribute selection approaches. These approaches computes best attributes from the dataset that can be used for classification of dataset on the basis of generation of rules. These attributes are selected on the basis of association and interdependency between different attributes that are responsible for prediction of class label for a single class. In the purposed work various parameters have been analyzed for performance evaluation of purposed classifier.

a) **AUC**: AUC stands for Area under the Curve. The ROC curve shows the sensitivity of the classifier by plotting the rate of true positives to the rate of false positives. The perfect classifier that makes no mistakes would hit a true positive rate of 100
b) **Confusion Matrix:** A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix.
c) **Accuracy**: the proportion of the total number of predictions that was correct.
d) **Precision** the proportion of positive cases that were correctly identified.
e) **Negative Predictive Value:** the proportion of negative cases that were correctly identified.
f) **Sensitivity or Recall**: the proportion of actual positive cases which are correctly identified.
g) **Specificity:** proportion of actual negative cases which are correctly identified.

In the purposed work diabetes testing dataset has been used for classification using improved decision table based classifier. In this process of classification 500 instances dataset has been used for classification process. In the purposed various parameters have been analyzed for performance valuation of purposed work.

These parameters have been computed for classification and represented in tabular form for validation of purposed work.

Table 5.1: Compassion table for different classifiers using various parameters

Parameter	*Improved DTB*	*DTB*
Correctly Classified	405	396
Incorrectly Classified	95	104
Precision	0.807	0.789
Recall	0.81	0.792
F-measure	0.807	0.79
TP Rate	0.81	0.792
FP Rate	0.256	0.264
ROC Area	0.78	0.848

This table represents various parameters for performance evaluation of purposed system. These parameters are important for performance evaluation of a classifier.

5. Conclusion & Future Scope

Data mining is major part that has been used for data warehousing process. Data mining is used for extraction of various features or pattern from historical dataset that can be used for further classification of raw information. Medical dataset contain a large number of instances that have records of different patients attributes. in this paper a better classification approach has been purposed that use support and confidence computation based on prediction of values using previous patients information. In this paper a classification approach has been predicted that provides better classification accuracy than that of previous used approaches for classification.

References

[1] C. M. Velu "Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE Conf. on Advance Computing Conference (IACC), 2013, pp. 1070 – 1075.
[2] E.W.T. Ngai"Application of data mining techniques in customer relationship management: A literature review and classification" Expert Systems with Applications , ELSEVIER, Volume 36, Issue 2, Part 2, March 2009, Pages 2592–2602.
[3] Jagannathan, G. "Seventh IEEE International Conference on Data Mining Workshops", IEEE Conf. on Data Mining Workshops, 2007, pp. 1-3.
[4] Mikhailov, L "Method for fuzzy rules extraction from numerical data" IEEE Conf. on Intelligent Control, 1997, pp 61 – 66
[5] Nedaabdelhamid, Aladdin Ayesh and FadiThabtah

"Emerging trends in associative classification data mining" International journal of electronics and electrical engineering Volume 3, Issue 1, Feb 2015.

[6] Robert E. Marmelstein "Application of Genetic Algorithms to Data Mining" MAICS-97 Proceedings,1997 AAAI , pp. 53-57

[7] Sankaranarayanan, S. "Diabetic Prognosis through Data Mining Methods and Techniques", International Conf. on Intelligent Computing Applications (ICICA), 2014, pp. 162 – 166.

[8] S. UmmugulthumNatchiar "Customer Relationship Management Classification Using Data Mining Techniques" International Conf. on Science Engineering and Management Research (ICSEMR), 2014, pp. 1 – 5.

[9] Tzung-Pei Hong "Using divide-and-conquer GA strategy in fuzzy data mining" IEEE Conf. on Computers and Communications, 2004, pp. 116 - 121 Vol.1.

[10] Tzung-Pei Hong "GA-based item partition for data mining" IEEE Conf. on Systems, Man, and Cybernetics (SMC), 2011, pp. 2238 – 2242.

[11] Wang, Guoyin "Granular computing based data mining in the views of rough set and fuzzy set" IEEE Conf. on Granular Computing, 2008, pp. 67.