Volume-6, Issue-3 (Oct-16)
ISSN (O) :- 2349-3585

A Review on Various approaches for Diabetes dataset classification

Abhinav Kathuria[1]
[1]abhinav.kathuria90@gmail.com
[1]DCSA, Panjab University, Chandigarh

Abstract-Data Mining is the process for extraction of valuable information from raw dataset so that information can be used for decision making process. In the process of data mining classification, clustering and attribute selection has been done for information extraction. Association rules in classification have been used for extraction of inter-dependency between different attributes of dataset Classification has been done on the basis of feature selection from raw data. In this paper different tree based, rule based and distance based classifier has been discussed. On the basis of this discussion best classifier for dataset classification has been evaluated.

Keywords: Data Mining, KNN, Naive Bayes, Tree structure

INTRODUCTION

1.1 DATA MINING

It is the process of fetching hidden knowledge from a wide store of raw data. The knowledge must be new, and one must be able to use it. Data mining has been defined as "It is the science of fetching important information from wide databases". It is one of the tasks in the process of knowledge discovery from the database. Data Mining is used to discover knowledge out of data and present the data in a easy and understood able form. It is a process to examine large amounts of data routinely collected. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. Two goals of data mining are prediction and description. Prediction tells us about the unknown value of future variables. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans. The medical data mining has the high potential in medical domain for fetching the hidden patterns in the datasets. These patterns are used for clinical diagnosis and prognosis. The medical data are widely distributed, heterogeneous, voluminous in nature. The data should be integrated and collected to provide a user oriented approach to novel and hidden patterns of the data. A major problem in medical science or bioinformatics analysis is in attaining the correct diagnosis of certain important information.

1.2 Classification in Data Mining:

2 Text Classification Text classification (TC) is an important part of text mining looked to be that of manually building automatic TC systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules that convert expert knowledge on how to classify documents under the given set of categories. For example would be to automatically label each incoming news story with a topic like "sports", "politics", or "art". A data mining classification task starts with a training set D = (d1….. dn) of documents that are already labeled with a class C1, C2 (e.g. sport, politics). The task is then to determine a classification model which is able to assign the correct class to a new document d of the domain Text classification has two flavors as single label and multi-label .single label document is belongs to only one class and multi label document may be belong to more than one classes .

3 The stages of TC are discussing as following points

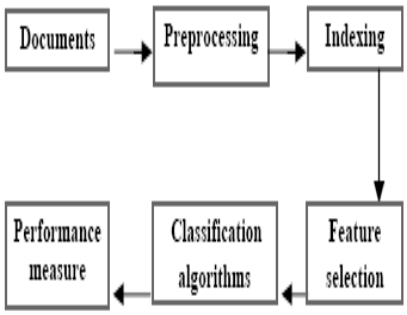

Figure 1.1: Process of classification

1.3 Probabilistic Classifiers

Probabilistic classifiers are designed to use an implicit mixture model for generation of the underlying documents. This mixture model typically assumes that each class is a component of the mixture. Each mixture component is essentially a generative model, which provides the probability of sampling a particular term for that component or class. This is why this kind of classifiers is often also called generative classifier.

1.4 DATA MINING TECHNIQUES

Data mining technique is linked with data processing, identifying patterns and trends in information. Or we can say that data mining simply means collection and processing data in systemic manner by using computer based programs and subsequent formation of disease prediction or patient management system aid. With the invention of information technology, now these days it is even more prevalent. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages.

1.4.1 Benefits of Developments of Disease Prediction System Using Data Mining Techniques

1. **Prevention and diagnosis:** Data mining technique made prediction system plays a vital role in strategy preparation for prevention of communicable as well as non communicable diseases in located area. Lifestyle related diseases like hyper tension, diabetes mellitus, cardiovascular diseases; stroke etc can be easily and accurately classified and possible to locate their etiological area cluster patterns. These techniques are also useful in disease diagnosis. Ms. Is take et al. developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bays and Neural Network.

2. Workout of treatment plan: The data mining techniques play an important role in treatment plan workout, surgical procedures, rehabitation, chronic diseases management plan etc. Long term follow up plan may be easily guided and keen supervision is possible. Example, a patient of hypertension can be long term manage and back through record of number of patients will guide in implementing future strategies.

3. **Reduction of cost of patient management:**These systems may definitely helpful in reduction of cost of patient management by avoiding unnecessary investigations and patients follow up.These prediction systems will add accuracy and time management.

4. **Discovery of hidden etiological factors**: This is most excitable objective planed by using these systems. This will be helpful for confirmation of geographical variations. Most of our health strategies are planned on the basis of data interpretations from developed countries. We can formulate our own systems and can avoid geographical errors.

REVIEW OF LITERATURE

Subbaiah (2013) in the paper "Extracting Knowledge using Probabilistic Classifier for Text Mining" explains extraction of knowledge using probabilistic classifier for text mining. Text mining is a process of extracting knowledge from large text documents. A new probabilistic classifier for text mining is proposed in this paper. It uses ODP taxonomy and Domain ontology and datasets to cluster and identify the category of the given text document. The proposed work has three steps, namely, preprocessing, rule generation and probability Calculation. At the stage of preprocessing the input document is split into paragraphs and statements. In rule generation, the documents from the training set are read. In probability Calculation, positive and negative weight factor is calculated. The proposed algorithm calculates the positive probability value and negative probability value for each term set or pattern identified from the document. Based on the calculated probability value the probabilistic classifier indexes the document to the concern group of the cluster.

Mahender et al. (2012) in the paper "Text Classification and Classifiers a Survey" describes classifiers to classify the text and then mining is performed. As most information (over 80%) is stored as text, text mining is believed to have a high commercial Potential value. knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. Text classification is that which classifies the documents according to predefined categories. In this paper they are tried to give the introduction of text classification, process of text classification as well as the overview of the classifiers and tried to compare

ISSN:-2349-3585 |www.ijrdt.org

the some existing classifier on basis of few criteria like time complexity, principal and performance**.**

Wijeirckremaet al. (2012) in the paper "An enhanced text classifier for automatic document classification" used for automatic classification the enhanced classifiers. Automatic classification has become an important research area due to the exponential growth of digital content in the modern world. Evidently, manual classification of documents is very painstaking and labor-intensive task. It takes much time to organize a collection of documents according to the subject area. This research has developed a computer program that can automatically classifying a given text document. Therefore, the user gets correct classification results just after feeding the document to the new system. For the process of classification, they use a new algorithm developed by enhancing basic form of an existing text classifier called tf-idf. The results were obtained for classification accuracy of the new text classification algorithm. They were compared with the results obtained for the basic tf-idf classifier. The research revealed that, the newly developed classifier algorithm can obtain better classification accuracy than the basic tf-idf classifier.

Teije et al. (2012) in the paper "Knowledge Engineering and Knowledge Management" describes knowledge engineering and management in the context of text mining. This paper focuses on the management of the knowledge which has been extracted by the information extraction process. Further tools are introduced and used for the management of the information.

Silwattananusarn et al. (2012) in the paper "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012" explores the applications of data mining techniques which have been developed to support knowledge management process. Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization over coming decades. The journal articles indexed in Science Direct Database from 2007 to 2012 are analyzed and classified. The discussion on the findings is divided into 4 topics: (i) knowledge resource; (ii) knowledge types and/or knowledge datasets; (iii) data mining tasks; and (iv) data mining techniques and applications used in knowledge management. The article first briefly describes the definition of data mining and data mining functionality. Then the knowledge management rationale and major knowledge management tools integrated in knowledge management cycle are described. Finally, the applications of data mining techniques in the process of knowledge management are summarized and discussed.

Mooney et al. (2010) in the paper"Mining Knowledge from Text Using Information Extraction" used information extraction process ids for information extraction. An important approach to text mining involves the use of natural-language information extraction. Information extraction (IE) distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. IE systems can be used to directly extricate abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analyzed with traditional data-mining techniques to discover more general patterns. The methods and implemented systems for both of these approach and summarize results on mining real text corpora of biomedical abstracts, job announcements, and product descriptions.

APPROACHES USED

K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. In an unpublished US Air Force School of Aviation Medicine report in 1951, Fix and Hodges introduced a non-parametric method for pattern classification that has since become known the k-nearest neighbor rule (Fix & Hodges, 1951). Later in 1967, some of the formal properties of the k-nearest-neighbor rule were worked out; for instance it was shown that for k=1 and $n \rightarrow \infty$ the k nearest neighbor classification error is bounded above by twice the Bayes error rate (Cover & Hart, 1967). A basic rule in classification

ISSN:-2349-3585 |www.ijrdt.org

analysis is that class predictions are not made for data samples that are used for training or learning. If class predictions are made for samples used in training or learning, the accuracy will be artificially biased upward. Instead, class predictions are made for samples that are kept out of training process.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithmfor training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Decision tables are a precise yet compact way to model complex rule sets and their corresponding actions. Decision tables,like flowcharts and if-then-else and switch-case statements, associate conditions with actions to perform, but in many cases do so in a more elegant way.

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data miningand machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

CONCLUSION

Data classification is used for prediction of raw information to different classes. In this process features have been evaluated from different data instances and used for prediction of class labels to raw information instances.In this paper various approaches of data classification has been reviewed that has been used for labeling to dataset instances. On the basis of discussion defined in above chapter's tree based classifier provide better accuracy in classification due to division of different dataset attributes into different purring tree structure. Tree based structure developed from dataset classification provides full description about dataset attributes association with other attributes. This process make and easy process to predict a class label for data instance.

REFRENCES

[1] S.subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22 2013.

[2] Suneetha K.R "Data Preprocessing and Easy Access Retrieval of Data through Data Ware House" Proceedings of the World Congress on Engineering and Computer Science, 2009, Vol. I, October 20-22, 2009, San Francisco, USA.

[3] Raymond J. Mooney "Mining Knowledge from Text Using Information Extraction" Department of Computer Sciences University of Texas at Austin 1 University Station C050.

[4] Maria Vargas Vera "Knowledge Extraction by using an Ontology based Annotation Tool" Knowledge Media Institute (KMi), The Open University,Walton Hall, Milton Keynes, MK7 6AA, United Kingdom.

[5] C Namrata Mahender "Text Classification and Classifiers a Survey" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[6] Wijewickrema "An enhanced text classifier for automatic document classification" Journal of the University Librarians Association of Sri Lanka, vol. 6, no 2, 2012.

[7] Annetietenteije "Knowledge Engineering and Knowledge Management" 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012

[8] TipawanSilwattananusarn "Data Mining and Its Applications for Knowledge Management:A Literature Review from 2007 to 2012" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, and September 2012.

[9] Hejab M. Alfawareh "Resolving Ambiguous Preposition Phrase for Text Mining Applications" Computer application technology (ICCAT) international conference (2013), pp. 1-5.

[10] ShaidahJusoh and Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[11] NingZhong, Yuefeng Li, and Sheng-Tang Wu, " Effective Pattern Discovery for Text Mining" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

[12] Mrs. SayantaniGhosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2012