International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 4 Issue 10, October-2015

A Decision Tree and Bayesian Network Approach for Zoo Data Classification

Gautam[1], Punam Rani[2], Anil[3]

[1]M.Tech. Scholar, Dept. of Computer Science, UIET, MDU Rohtak, Haryana
[2]M.Tech. Student, Dept. of Computer Science, UIET, MDU Rohtak, Haryana

ABSTRACT

Outlier is defined as the impurities or irrelevancy of some data item from the rest of available data. To improve the significance of dataset it is required to identify these critical objects from the dataset. In this present work, a generic model is been defined to identify the outlier over the dataset. This model will use different statistical measures to depict the outlier, it will be used as the weighted value for the analysis. In final stage, this weighted dataset will be trained under Bayesian network and decision tree methods separately. As the work is based on the statistical and predictive probabilistic model, an accurate detection of outlier is expected. The work will be implemented in weka integrated java environment. The work is about to improve the recognition rate.

KEYWORDS: DAG, Outlier, Bayesian Network, Decision Tree.

1. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. It is the process of finding patterns in large amount of data, to produce useful information from it. Data Mining has vital applications in various fields. The great possibilities of improvement in Health Care through Data Mining only further justify the need to apply data mining principles to clinical data.

Data mining is also known as Knowledge Discovery in Data (KDD). The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

However, prior to applying data mining techniques to collect information from data, the data has to be 'prepared' to ensure the veracity of the information obtained. 'Preparing' the data involves removal of incorrect information or 'noise' from the data and ensuring that the data mining principles are applied on real data. This document gives a detailed description of the purpose, design and implementation of the Data Mining Framework. The primary purpose of the Data Mining Framework is to help determine trends in patient records to improve Health Care.

1.1 DATA

Data are any facts, numbers, or text that can be processed by a computer. Today, there are vast and growing amounts of data in different formats and different databases. This includes sales, cost, inventory operational or transactional data, and accounting. Nonoperational data, such as industry sales, forecast data. Meta data - data about the data itself, such as logical database design or data dictionary definitions

1.2 INFORMATION

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when. Knowledge Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Scope of Mining Data mining is about to finding the similarities between searching the valuable business information from the large database systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of valuable dataset. Both kind of processes required either shifting through an immense amount of material, or to perform the search intelligently so that exactly match will be performed. Data mining can be done on a database whose size and quality are sufficient. The technology of data mining can generate new business opportunities by providing these capabilities:

- Automated prediction and analysis of various trends and behaviors - Data mining itself automate the process by obtaining the predictive information from large databases. It first setup the questions and then provides the relative solutions. A typical example of such a predictive system is in the marketing field.

- Another application of data mining is automated discovery of historical patterns dynamically. The presented Data mining system is able to sweep over the databases to identify the hidden patterns. One of such example of pattern discovery is the analysis of retail sales data to identify the seemingly unrelated products so that the effective purchase can be done. It also includes the detection of the transactions to identify the anomalous data.

Data mining techniques are able to get the benefits of automation on existing software and hardware platforms that can be implemented on new systems which can be upgraded and new products can be developed. When data mining tools are defined on high performance parallel systems, they can be analyzed with large amount databases in minutes. Faster processing is required in such system to derive the effective results from complex systems. High speed processing and accurate outcome from the system makes it possible for users to analyze large set of information.

More columns-The column here signify the object features or the attributes. Analysts sometimes defines the limits on number of variables so that examine to the system can be done effectively. There are number of attributes that are discarded because they seem unimportant may carry information about unknown patterns. This step is basically called data cleaning.

More rows- Larger number of record samples, lower errors and variance in the system will be.

1.3 CLASSIFICATION

Classification is the most commonly used data mining technique. It generally refers to mapping of data item into predefined groups and classes. The data classification process involves learning and classification. In learning, the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. It is also termed as supervised learning because classes are determined before examining the data. Example: Identifying of credit risk and determining whether to make bank loan or not.
In classification following techniques are made in use like that of:

1. Decision tree
2. Bayesian Classification
a. Bayes' theorem
b. Naïve Bayesian classification
c. Bayesian belief network
3. Naïve bayes tree
4. Support vector machine
5. K-nearest neighbor
6. Case based reasoning

The classification methods like decision tree, Bayesian classification, support vector machine, are all example of eager learners. Eager learner when given a set of training tuples will construct a generalization model before receiving new tuples to classify.

The classification method like k-nearest neighbor and case based reasoning classifiers are examples of lazy learners. Lazy approach is one in which the learner instead waits until the last minute before doing any model construction in order to classify a given test tuple. That is when given a training tuple, lazy learner simply stores it and waits until it is given a test tuple. Only when it sees the test tuple does it perform generalization in order to classify the tuple based on its similarity to the stored training tuple.

2. PROPOSED METHOD

In this present work, a statistical method is defined to perform outlier detection. The work is here defined as the generic model that can be applied on different dataset. The work is here divided in two main layers. In first layer, the statistical analysis on dataset will be performed under different parameters. In second stage, this statistical information will be trained under Bayesian network and decision tree approach for outlier recognition.

2.1 FLOWCHART OF THE PROPOSED METHOD

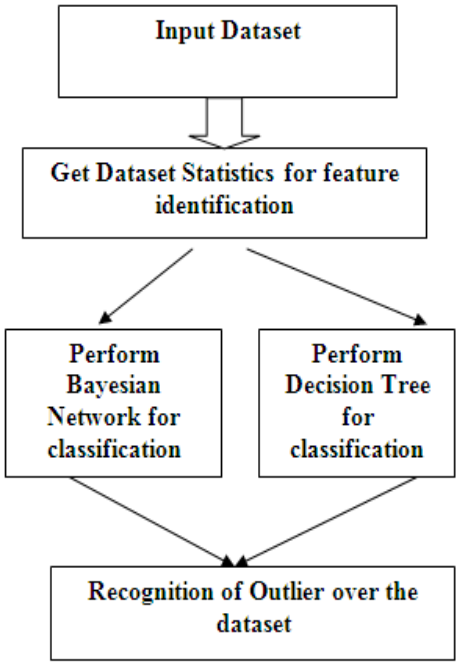

Fig. 1: Flowchart of the proposed method

2.2 BAYESIAN NETWORK

BNs are probabilistic graphical models that encode probabilistic dependence relations among variables. A Bayesian network, Bayes network, directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via DAG. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies.

2.3 DECISION TREE

Decision tree is a flow chart like tree structure, consisting of internal node branches and leaf nodes. Internal nodes are denoted by rectangles and leaf nodes are denoted by ovals. Internal nodes may have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it. Decision tree also requires an attribute selection method. There are three measures for attribute selection method namely info gain, gain ratio and gini index. And for our research work we select the gain ratio attribute selection method. Decision tree involves the use of J48 algorithm. J48 acts as a java implementation of C4.5 algorithm.

2.4 TASK 1: DATASET COLLECTION

The work is here defined to perform the detection of outlier over the breast cancer dataset. The work required the authenticated dataset. Such a dataset is here collected from external web source. The dataset properties are given here under:

Table 1: Dataset properties

Properties	Values
URL	UCI Repository
DB Name	Zoo data
Total Instances (Training Dataset)	86
Number of Attributes	18
Number of Classes	7
Format	Arff

2.5 TASK 2: GUI GENERATION

The work is here defined to perform intrusion detection with subsequent number of stages. These stages includes the processing of training dataset processing, test dataset processing, feature extraction, classification etc. To represent these stages in effective way, a graphical interface is defined in this work.

2.6 TASK 3: IMPLEMENTATION

The code is defined for training data processing. The code is defined for :

1. Accepting Training Data
2. Accept Test Data
3. Data Visuzalization
4. Feature Identification
5. Navie Bays Implementaiton

3. RESULTS AND CONCLUSION

As the work is defined as the generic model so that it can be applied on different datasets including medical and financial datasets. The work has combined two classification approaches so that more accurate results are expected. The comparative analysis results are shown for the proposed approach respective to the existing tree approach. In this work, both of the methods are implemented:

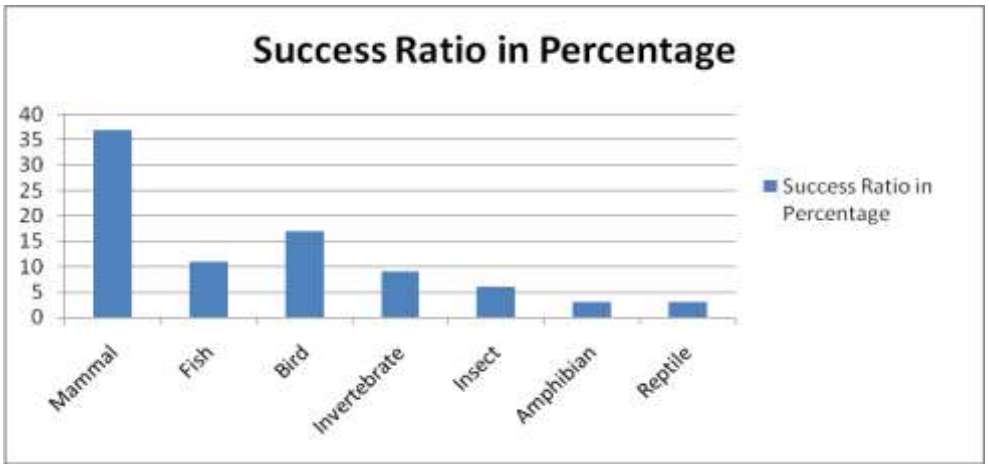

Fig. 2: Success Rate Analysis of Proposed Approach

Here the figure is showing the comparative prediction results. The prediction results are taken for the testing dataset. The testing dataset is having 86 instances and as the improved decision tree algorithm is applied over it the recognition rate obtained here is100 % in case of proposed approach. The result shows that the presented approach performed more accurate analysis.

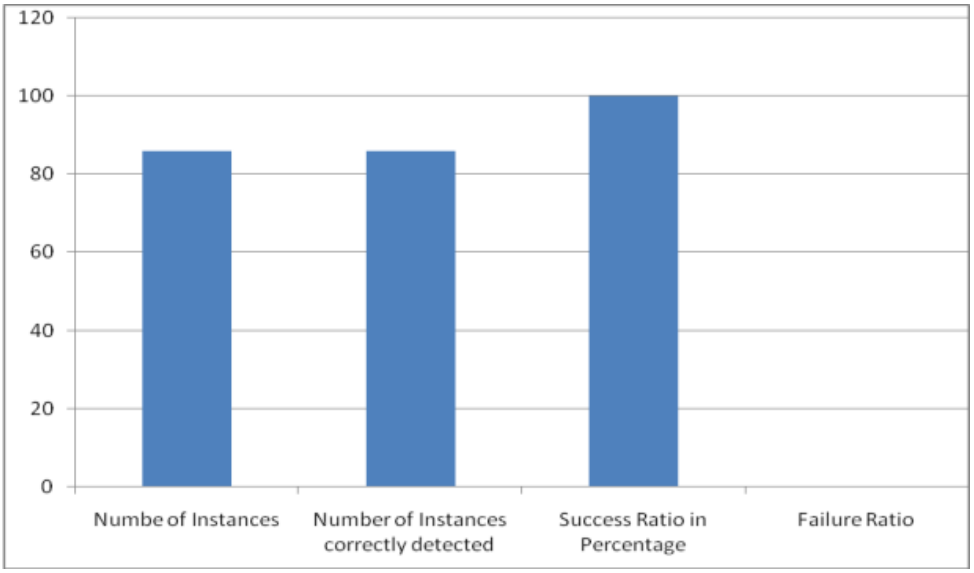

Fig. 3: Recognition Rate Analysis of Proposed Approach

Here figure 3 is showing the results The testing dataset is having 86 instances and as the improved decision tree algorithm is applied over it the recognition rate obtained here is 100 % in case of proposed approach.

REFERENCES

[1]. H.E. Bassil, "Detection and Correction of Outliers in Foetal Heart Rate Time Series".
[2]. Qi Fin "A Robust Adaptive Estimator of Rate for Cardiac Arrhythmia Detection", 0-7803-1254-6/93@1993 IEEE
[3]. E. Munevar," Detection of Abnormalities in the Signal Averaged Electrocardiogram: A Subspace System Identification Approach", Proceedings of the 38Conference on Decision & Control Phoenix, Arizona USA 0-7803-5250-5/99/@1999 IEEE
[4]. J. A. Jo," Detection of Autonomic Abnormality In Obstructive Sleep Apnea Using A Nonlinear Model of Heart-Rate Variability", Proceedings of the Second Joint EMBSBMES Conference Houston, TX, USA -0-7803-7612-9/02@ 2002 IEEE
[5]. Norhashimah Mohd Saad," Detection of Heart Blocks in ECG Signals by Spectrum and Time-Frequency Analysis", 4th Student Conference on Research and Development (SCOReD 2006), 1-4244-0527-0/06 2006 IEEE
[6]. A Muller," Integrating Model Based and Data Driven Approaches for the Automatic Segmentation of Cardiac Short-Axis Cine MRI Recordings", Computers in Cardiology 2006, ISSN 0276−6547
[7]. P. S.Vikhe," Heart Sound Abnormality Detection Using Short Time Fourier Transform and Continuous Wavelet Transform", Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09, 978-0-7695-3884-6/09 2009 IEEE
[8]. Zhinan Li," Dimensionality Reduction for Anomaly Detection in lectrocardiography: A Manifold Approach", 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks 978-0-7695-4698-8-2012
[9]. Xiaowei Zhou," Automatic Mitral Leaflet Tracking in Echocardiography by Outlier Detection in the Low-rank Representation", 978-1-4673-1228-8/12 2012 IEEE
[10]. Yuanjing Yang," Outlier Detection in Heart Rate Signal using Activity information", Proceedings of the 10th World Congress on Intelligent Control and Automation 978-1-4673-1398-8/12 2012 IEEE
[11]. Ishanka S. Perera," Automated Diagnosis of Cardiac Abnormalities using Heart Sounds", 2013 IEEE Point-of-Care Healthcare Technologies (PHT) 978-1-4673-2767-1/13 2013 IEEE