Implementation and Classification of Anomalous Detection with Varying Parameters

Manvi ChaharMs. Savita M.Tech, Ass. Prof.
(CSE Dept.) DPGITM Gurugram, India

Abstract

Classification is a classic data mining technique based on machine learning. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. We consider the problem of discovering attributes, or properties, accounting for the a-priori stated abnormality of a group of anomalous individuals (testing data) with respect to an overall given population (training data). To this aim, we use the notion of gain ratio. Gain ratio is an attribute selection method and has been used to rank the attributes of the datasets. For this we found that the attributes which have high gain ratio will have high classification accuracy and those attributes which have lower gain ratio can be neglected which helps in reduction of the attributes. This thesis shows that if we apply gain ratio on attributes to rank them and classify our data with small no of attributes and get the high accuracy rate. The results in the report on this dataset also show the efficiency and accuracy of Naïve bayes classifier.

Keywords - *Data Mining, Gain Ratio, Naive Bayes Classifier*

I. INTRODUCTION

Data mining and knowledge discovery in databases, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a few technical approaches, such as clustering, data summarization, classification, finding dependency networks, analysing changes, and detecting anomalies.

Data retrieval in its usual sense in database attempts to retrieve data that is stored explicitly in database and present it to the user in a way that user can understand it. It does not attempt to extract implicit information.

Data mining is the search for the relationships and global patterns that exist in the large databases but are hidden among vast amount of data, such as the relationship between patient data and their medical diagnosis. This relationship represents valuable knowledge about the database and the objects in database, if the database is a faithful mirror of the real world registered by the database.

II. DATA MINING TECHNIQUES

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in database to predict unknown or future values of interest and description focuses on finding patterns describing the data and the subsequent presentations for user interpretation There are several data mining techniques fulfilling these objectives. Some of these are association, classification, sequential patterns and clustering[1]. Another approach of the study of the DM technique is to classify the techniques as [2]:

1. User guided or verification- driven data mining: In this process of data mining the user makes a hypothesis and tests the hypothesis on the data to verify its validity. The emphasis is on the user who is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis.
2. Discovery driven or automatic discovery of rules: The discovery model differs in its emphasis in that it is the system automatically discovering information hidden in the data.

   The data is sifted in search of frequently occurring patterns trends and generalization about the data without intervention or guidance from the user. Following is the Data mining techniques:

   – Classification

   – Clustering

A data mining (machine learning) technique used to predict group membership for data instances. Several major kinds of classification method including

1. – Decision tree induction,
2. – Bayesian networks,
3. – k-nearest neighbor classifier,
4. – case-based reasoning,
5. – Genetic algorithm and fuzzy logic techniques.

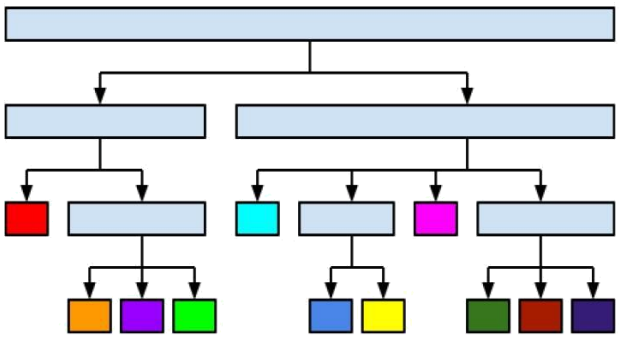

Fig 1.1 One type of classification, here in the form of a decision tree.

III. DATA MINING ALGORITHMS

Data mining also known as Knowledge Discovery in databases is very often utilized in the field of medicine. The process of supporting medical diagnoses by automatically searching for valuable patterns undergoes noticeable improvements in terms of precision and response time.
All this shortly describes the most common data mining algorithms and explains the use cases of each of them. The usefulness of the following methods was verified by medical personnel and confirmed by independent experts. The selection of the data mining algorithms was made after a complete in-depth analysis of the scientific articles on topic [3].

A. Decision Trees

Decision trees are one of the most frequently used techniques of data analysis. The advantages of this method are unquestionable. Decision trees are, among other things, easy to visualize and understand and resistant to noise in data. Figure 1.2 shows a sample decision tree. [4]

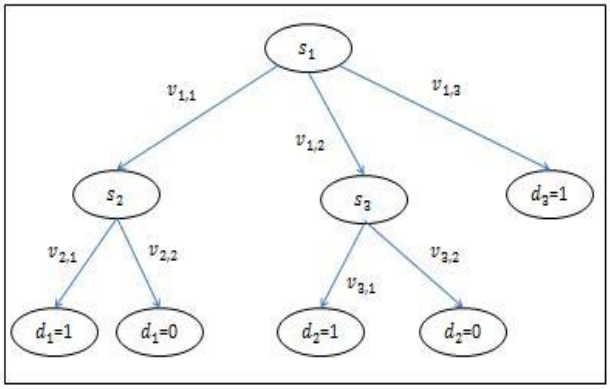

Figure 1.2 Sample decision tree

B. Naïve Bayes

The Naïve Bayes is a simple probabilistic classifier. It assumes about mutual independency of attributes (independent variable, independent feature model). Usually this assumption is far from being true and this is the reason for the naivety of the method.
The probabilities applied in the Naïve Bayes algorithm are calculated according to the Bayes' Rule: the probability of hypothesis H can be calculated based on the hypothesis H and evidence about the hypothesis E according to the following formula:

$$\mathbf{P(H|E) = \frac{P(E|H)\ X\ P(H)}{P(E)}}$$

Depending on precision of the probability model, the Naïve Bayes may give a model with high effectiveness for a supervised learning problem. Frequently the Naïve Bayes uses a method of maximum likelihood (particularly in practical applications). In practice the Naïve Bayes method works effectively in various real-world situations.
The structure of a Naïve Bayes model forms a Bayesian network of nodes with one node for each attribute. The nodes are interconnected with directed edges and form a directed acyclic graph [5].

C. Gain Ratio

The features of the data set are identified as either being significant to the process, or redundant. This process is known as feature selection. Choosing a good set of features helps in improving the performance of the system. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain measure prefers to select attributes having many values. The gain ratio is defined as:
Gain Ratio (A) =Gain (A)/split Info A(S)
The gain is calculated as:

$$\mathbf{GR\ (S,\ A) = \frac{GAIN(S,A)}{IntI(S,A)}}$$

Where S is a set consisting of s data samples and A is an attribute having distinct values. Also, 'I' is the expected information needed to classify a given sample.

IV. CONCLUSION

A layered classification approach is presented to classify the bug tracking systems dataset in terms of probability of the risk associated with attributes. In this work we examined various techniques of attribute selection and classification. In the result of the classification we found that the classification accuracyof those attributes which have highest gain ratio is high as compare to other combinations of the attributes. This thesis shows that if we apply gain ratio

on attributes to rank them so, we can classify our data or bug report with small no of attributes and get the high accuracy.

ACKNOWLEDGMENT

Our thanks to the experts who have contributed towards development of the template.

REFERENCES

[1] Beckerman R, "Distributional Word Clusters vs. Words for Text Categorization," Journal of Machine Learning Research, vol 3, pp 1183– 1208, 2003.
[2] I.H. Witten, E. Frank and M.A. Hall, Data mining practical machine learning tools and techniques, Morgan Kaufmann publisher, Burlington 2011.
[3] H. Grosskreutz and S. Ruping, "On subgroup discovery in numerical domains," Data Mining and Knowledge Discovery, vol. 19, no. 2, pp. 210– 226, 2009.
[4] J. Han And M. Kamber, Data Mining: Concepts and Techniques. San Francisco, Morgan KauffmannPublishers (2001).
[5] Irina Rish," An empirical study of the naïve bayes classifier", IBM Research Report, pp. 1-7, 2001.