e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

SURVEY ON CLASSIFICATION ALGORITHMS USING BIG DATASET

Pooja Sharma[1], Anju Singh[2], Divakar Singh[3]
[1] Computer Science & engineering,BU-UIT
[2] Computer Science & Information Technology, UTD-BU
[3] Computer Science & engineering,BU-UIT

Abstract— Data mining environment produces a large amount of data that need to be analyzed. Using traditional databases and architectures, it has become difficult to process, manage and analyze patterns. To gain knowledge about the Big Data a proper architecture should be understood. Classification is an important data mining technique with broad applications to classify the various kinds of data used in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This paper put a light on various classification algorithms including j48, C4.5, Naive Bayes using large dataset.

Keywords - Classification, Data Mining, C4.5, J48, Naïve Bayes

I. INTRODUCTION

Data Mining is the technology to extract the knowledge from the data. Data mining refers to the analysis of the large quantities of data that are stored in computers. To discover previously unknown, valid patterns and relationships in large data set data mining involves the use of sophisticated data analysis tools [2]. These tools can include statistical models, mathematical algorithm and machine learning methods.

Data Mining is mainly used for the specific set of six activities namely classification, estimation, prediction, affinity grouping or association rules, clustering, description and visualization. This paper describes the comparison of best-known supervised techniques in relative detail [5]. Then it produces a critical review of comparison between supervised algorithms like Naïve bayes, C4.5, J48. It is not to find that which classification learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem.

II. LITERATURE SURVEY

2.1 Data Mining

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The concept of Data Mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty.

The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Stage 1: Exploration. This stage usually starts with data preparation which may involve cleaning data, data transformations, and selecting subsets of records and - in case of data sets with large numbers of variables ("fields") performing some preliminary feature selection operations to bring the number of variables to a manageable range.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples) [7]. This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Stage 3: Deployment. That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

2.2 Classification

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. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects [3]. Then, the classifier is used to predict the group of attributes of new cases from the domain based on the values of other attributes.

III. DATA MINING CLASSIFICATION METHODS

3.1 J48 Algorithm

J48 implements Quinlan's C4.5 algorithm for generating a pruned or unpruned C4.5 decision tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation [8] [10].

3.2 C4.5 Algorithm

C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993). The C4.5 algorithm generates a decision tree by recursively splitting the data. The decision tree grows using Depth-first strategy. The C4.5 algorithm considers all the possible tests that can split the data and selects a test that gives the best information gain (i.e. highest gain ratio) [11]. For each discrete attribute, one test is used to produce many outcomes as the number of distinct values of the attribute. For each continuous attribute, the data is sorted, and the entropy gain is calculated based on binary cuts on each distinct value in one scan of the sorted data. This process is repeated for all continuous attributes. The C4.5 algorithm allows pruning of the resulting decision trees. This increases the error rates on the training data, but importantly, decreases the error rates on the unseen testing data. The C4.5 algorithm can also deal with numeric attributes, missing values, and noisy data [6].

C4.5 is collection of algorithms for performing classifications in machine learning and data mining. It develops the classification model as a decision tree. C4.5 is one of the most popular algorithms for rule base classification. There are many empirical features in this algorithm such as continuous number categorization, missing value handling, etc. However in many cases it takes more processing time and provides less accuracy rate for correctly classified instances. On the other hand, a large dataset might contain hundreds of attributes. We need to choose most related attributes among them to perform higher accuracy using C4.5.The resulting decision tree is generated after classification. The classifier is trained and tested first. Then the resulting decision tree or rule set is used to classify unseen data. C4.5 is the newer version of ID3. C4.5 algorithm has many features like:

- Speed - C4.5 is significantly faster than ID3 (it is faster in several orders of magnitude)
- Memory - C4.5 is more memory efficient than ID3
- Size of decision Trees – C4.5 gets smaller decision trees.
- Rule set - C4.5 can give rule set as an output for complex decision tree.
- Missing values – C4.5 algorithm can respond on missing values by '?'.
- Over fitting problem - C4.5 solves over fitting problem through Reduce error pruning technique.

3.3 Naive Bayes Algorithm

In simple terms, a naive bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence 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 3" 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 the presence or absence of the other features [1].

For some types of probability models, naive bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive bayes models uses the method of maximum likelihood; in other words, one can work with the naive bayes model without accepting Bayesian probability or using any Bayesian methods.

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. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms [4]. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. It improves the classification performance by removing the irrelevant features and its computational time is short, but the naive bayes classifier requires a very large number of records to obtain good results and it is instance-based or lazy in that they store all of the training samples abstractly.

IV. COMPARISONS OF CLASSIFICATION ALGORITHMS [9]

S. No.	Algorithms → ↓Characteristics	J48	C4.5	Naïve Bayes
1	Proposed by	Quinlan	Quinlan	Dudo & hurt
2	Attribute type	Handle discrete & continuous data	Handle both categorical & numerical data	Handle numerical attribute
3	Missing Value	Ignore the missing value	Handle missing value	Good with missing values handling
4	Splitting criteria	Use split info and gain ratio	Used gain ratio	There is no splitting criteria
5	Pruning strategy	Used error based pruning	Used reduced error pruning	Does not support pruning
6	Outlier detection	Susceptible on outlier	Susceptible on outlier	High tolerance to outlier
7	Parameter setting	Deal with parameter	Deal with parameter	There is nothing like parameter setting
8	Learning type	eager learner	Supervised Eager learner	Eager learner
9	Accuracy	good in many domain	good in many domain	good in many domain
10	Transparency	Rules	Rules	No rules(black box)

V. CONCLUSION

In this paper the comparison of the most well-known classification algorithms like decision trees, neural network, and Bayesian network, nearest neighbor and support vector machine has been done in detail. The aim behind this study was to learn their key ideas and find the current research issues, which can help other researchers as well as students who are doing an advanced course on classification. The comparative study had shown that each algorithm has its own set of advantages and disadvantages as well as its own area of implementation. None of the algorithm can satisfy all the criteria. One can investigate a classifier which can be built by an integration of two or more classifier by combining their strength.

REFERENCES

[1] Tina R. Patil, Mrs. S. S. Sherekar “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification” International Journal of Computer Science and Applications Vol. 6 No.2, April 2013, pg. 256-261.

[2] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, Timm Euler, “YALE: rapid prototyping for complex data mining tasks”, KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pg. 935-940.

[3] A.ShameemFathima, D.Manimegalai and NisarHundewale "A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue" IJCSI International Journal of Computer Science Issues, Vol. 8 Issue 6, November 2011, pg. 322-328.

[4] Ali, M.M. , Rajamani, L "Decision tree induction: Priority classification" International Conference on Advances in Engineering, Science and Management (ICAESM), March 2012 ,pg. 668-673.

[5] A.S. Galathiya, A. P. Ganatra and C. K. Bhensdadia "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning" International Journal of Computer Science and Information Technologies, Vol. 3 (2), 2012, pg. 3427-3431.

[6] Mohammad M Mazid,A B M Shawkat Ali, Kevin Tickle, "Improved C4.5 Algorithm for Rule Based Classification" School of Computing Science, Central Queensland University, Australia.

[7] http://www.statsoft.com/Textbook/Data-Mining-Techniques#mining

[8] Margaret H. Danham, S. Sridhar, "Data mining, introductory and Advanced Topics", Person education, 1st ed., 2006.

[9] Sonia Singh, Priyanka Gupta "Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey" International Journal of Advanced Information Science and Technology (IJAIST) Vol.27, No.27, July 2014, pg. 97-103.

[10] Aman Kumar Sharma, Suruchi Sahni, "AComparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3 No. 5, May 2011, pg. 1890-1895.

[11] http://www.wikipedia.com/