International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN: 2349-9745
p-ISSN: 2393-8161

CLASSIFICATION ALGORITHM USING RANDOM CONCEPT ON A VERY LARGE DATA SET: A SURVEY

Pooja Sharma[1], Annu Mishra[2]
[1] Computer Science & engineering,BU-UIT
[2] Computer Science & engineering, BU-UIT

Abstract— Data mining environment produces a large amount of data, that need to be analyses, pattern have to be extracted from that to gain knowledge. In this new period with rumble of data both ordered and unordered, by using traditional databases and architectures, it has become difficult to process, manage and analyses 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 provides an inclusive survey of different classification algorithms and put a light on various classification algorithms including j48, C4.5, k-nearest neighbor classifier, Naive Bayes, SVM etc., using random concept.

Keywords - Classification, Classifier, Large Data, Random concept

---

I. INTRODUCTION

Data Mining is the technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data. Data Mining has also been termed as data dredging, data archaeology, information discovery or information harvesting depending upon the area where it is being used [3]. Data mining is growing in various applications widely like analysis of organic compounds, medicals diagnosis, product design, targeted marketing, credit card fraud detection, financial forecasting, automatic abstraction, predicting shares of television audiences etc. 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 [6]. These tools can include statistical models, mathematical algorithm and machine learning methods. Classification techniques in data mining are capable of processing a large amount of data. It classifies data based on training set and class labels and can predict categorical class labels and hence can be used for classifying newly available data [4]. Thus it can be outlined as an inevitable part of data mining and is gaining more popularity. The benefit of analyzing the pattern and association in the data is to set the trend in the market, to understand customers, analyse demands, and predict future possibilities in every aspect [3].

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

II. LITERATURE SURVEY

2.1 Data Mining

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories. Data mining consists of five major elements:

- Data warehouse system load, extract, transforms transaction data.
- In a multidimensional Database system Store and manage the data.
- For the business analysts and Information technology professionals provide data access.
- By application software analysis of data done.
- Data is presented in a useful format, such as a graph or Table [5].

2.2 Large Data

Large Data is usually defined in terms of the 3Vs: volume, velocity, variety. The large amount of data being processed by the Data Mining environment. In other words, it is the collection of data sets large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications, so data mining tools were used. Large Data are about turning unstructured, invaluable, imperfect, complex data into usable information [7] [10]. Data have hidden information in them and to extract this new information; interrelationship among the data has to be achieved. Information may be retrieved from a hidden or a complex data set [8]. Browsing through a large data set would be difficult and time consuming, we have to follow certain protocols, a proper algorithm and method is needed to classify the data, find a suitable pattern among them. The standard data analysis method such as classification, clustering, factorial, analysis need to be extended to get the information and extract new knowledge treasure.

2.3 Classification

Classification is the most frequently used data mining task. Classification maps the data in to predefined targets. Classification is a supervised learning because the targets are predefined [16]. 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 [9]. 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 C4.5 Algorithm

To generate a decision trees C4.5 is a well-known algorithm, is an extension of the ID3 algorithm used to overcome its drawbacks. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier. The C4.5 algorithm made a number of changes to improve ID3 algorithm [2]. Some of these are:

- For missing values of attributes handling training data.

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

- Differing cost attributes is handled.
- The decision tree is pruned after creation.
- Discrete and continuous values, attributes are handled.

Let the training data be a set S = s1, s2 ... of already classified samples. Each sample Si = xl, x2... is a vector where xl, x2 ... represent attributes or features of the sample. The training data is a vector C = c1, c2..., where c1, c2... represent the class to which each sample belongs to [8]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits data set of samples S into subsets that can be one class or the other. It is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute factor with the highest normalized information gain is considered to make the decision [13]. The C4.5 algorithm then continues on the smaller sub-lists having next highest normalized information gain. C4.5 uses a metric called "information gain," which is defined by subtracting conditional entropy from the base entropy; that is, Gain (P|X) =E (P)-E (P|X). This computation does not, in itself, produce anything new. However, it allows you to measure a gain ratio. Gain ratio, defined as Gain Ratio (P|X) =Gain (P|X)/E(X), where(X) is the entropy of the examples relative only to the attribute. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too much detail in the training data set. It uses gain ratio impurity method to evaluate the splitting attribute. Decision trees are built in C4.5 by using a set of training data or data sets as in ID3. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other [1]. Its criterion is the normalized information gain (difference in entropy) [11] that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision.

a. 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 [3].

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 [12]. 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

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

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.

b. Support Vector Machine Algorithm

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [28].

Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function $K(x, y)$ selected to suit the problem. It produce very accurate classifier, robust to noise especially popular in text classification problem where very high dimensional space are the norms but computationally SVM is expensive thus runs slow.

3.4 K-nearest neighbors Algorithm

In pattern recognition, the ***k*-Nearest Neighbors algorithm** (or ***k*-NN** for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space [19]. The output depends on whether *k*-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

A shortcoming of the *k*-NN algorithm is that it is sensitive to the local structure of the data. K-NN requires an integer k, a training data set and a metric to measure closeness.

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

3.5 J48 Algorithm

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [17] [18].
While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them.

IV. COMPARATIVE STUDY OF SOME CLASSIFICATION ALGORITHM

S.NO	ALGORITHM	ADVANTAGES	DISADVANTAGES
1.	Random J48	1) It produces the accurate result. 2) This algorithm also handles the missing values in the training data 3) Both the discrete and continuous attributes are handled [20] by this algorithm. 4) No empty branches. 5) Use split info and gain ratio.	1) The depth of the tree is linked to tree size and the run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the attributes. 2) Space complexity is very large as we have to store the values repeatedly in arrays [14] [15].
2.	Random C4.5	1) Produces the accurate result. 2)Distribution of trees is uniform [22] 3) Less Memory for large program execution. 4) Model build time is less. 5) Searching time is short.	1) Performance of randomized tree with classication noise is not as good [21]. 2) Problem of Over fitting. 3) Can't deal with missing values.
3.	Random Forest	1) Almost always have lower classification error. 2)Far easier for humans to understand 3) Deal really well with uneven data sets that have missing variables. Train faster	1)The high classification error rate while training set is small in comparison with the number of classes.[27] 2) Not do well with imbalance data. 3) Need to discrete data for some particular construction algorithm.

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

4.	Random k-Nearest Neighbour	1) It is a more effective and more efficient model for high-dimensional data 2) Applied to both qualitative and quantitative responses. 3) It is significantly more stab & & more robust for feature selection when the input data are noisy and unbalanced. 4)For multimodal classes it is well suited.[23]	1) For local structure of the data it is sensitive. 2) Memory limitation. 3) Being a supervised, it is lazy learning algorithm i.e., runs slowly.
5.	Random Naïve Byes	1) Suitable for applications where computational power and memory are limited 2) Remove irrelevant feature for improving the performance.	1) Require very large number of records to obtain good results. 2) All the variables are uncorrelated to each other. 3) It is really fragile to over fitting.[25][26]
6.	Random SVM	1) Prediction accuracy is generally high 2) Robust works when training examples contain errors.[24] 3) Especially popular in text classification problems. 4) Fast evaluation of the learned target function	1) Difficult to incorporate in domain knowledge 2) SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used. 3) Long training time so computationally expensive.

V. CONCLUSION

At present data mining is an important area of research and due to Increase in the amount of data it becomes difficult to handle the large data, methodologies associated with different algorithms used to handle such large data sets. Classification is a very suitable for solving the problems of data mining because the classification techniques show how a data can be determined and grouped when a new set of data is available. Each classification technique has got its own advantages and disadvantages as given in the paper. We select the required classification algorithm as our requirement. According to our theoretical study, the performances of the algorithms are strongly depends on the entropy, information gain, gini index and the features of the data sets. The big advantage of a decision tree classifier is that it doesn't require a lot of information about the data to create a tree that could be very accurate and very informative and, the knowledge in decision tree represented in form of [IF-THEN] rules which is easier for humans understand.

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

Decision Tree's algorithms (C4.5, j48) have less error rate and it is easier algorithm as compared to other classification algorithms.

REFERENCES

[1] S.Archana, Dr.K.Elangovan "Survey of Classification Techniques in Data Mining" International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February 2014, pg. 65-71.

[2] RanshulChaudhary, Prabhdeep Singh, Rajiv Mahajan "A Survey on Data Mining Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3 Issue 1, January 2014, pg. 5002-5003.

[3] 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.

[4] Ms.Aparna Raj, Mrs.Bincy G, Mrs.T.Mathu"Survey on Common Data Mining Classification Techniques" International Journal of Wisdom Based Computing, Vol. 2(1), April 2012, pg. 12-15.

[5] Nikita Jain, Vishal Srivastava "Data Mining Techniques: A Survey Paper" IJRET: International Journal of Research in Engineering and Technology Vol. 2 Issue 11, November 2013, pg. 116-119.

[6] 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.

[7] UN Global Pulse "Big Data for Development: Challenges and Opportunities", May 2012.

[8] DelveenLuqmanAbdAl.Nabi, ShereenShukri Ahmed, "Survey on Classification Algorithms for Data Mining :(Comparison and Evaluation)" Vol.4, March 2013, pg. 18-25.

[9] 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.

[10] ChanchalYadav, Shuliang Wang, Manoj Kumar "Algorithm and approaches to handle large Data-A Survey" IJCSN International Journal of Computer Science and Network, Vol. 2 Issue 3, April 2013, pg. 56-60.

[11] Mohd. Mahmood Ali1, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan "Extracting Useful Rules through Improved Decision Tree Induction Using Information Entropy" International Journal of Information Sciences and Techniques (IJIST) Vol.3 No.1, January 2013,pg. 27-41.

[12] 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.

[13] Mr. Brijain R Patel, Mr. Kushik K Rana "A Survey on Decision Tree Algorithm for Classification" International Journal of Engineering Development and Research, Vol. 2 Issue 1, March 2014, pg.20-24.

[14] Juneja, Deepti "A novel approach to construct decision tree using quick C4.5 algorithm" Oriental Journal of Computer Science & Technology Vol. 3(2), February 2013, pg. 305-310.

[15] Dr. Neeraj Bhargava, Girja Sharma, Manish Mathuria "Decision Tree Analysis on J48 Algorithm for Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3 Issues 6, June 2013, pg. 1114-1119.

[16] Tulips Angel Thankachan1, Dr. Kumudha Raimond "A Survey on Classification and Rule Extraction Techniques for Datamining" IOSR Journal of Computer Engineering (IOSR-JCE) Vol. 8 Issue 5, February 2013, pg. 75-78.

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

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

[19] Bhaskar N. Patel, Satish G. Prajapati and Dr.Kamaljit I. Lakhtaria "Efficient Classification of Data Using Decision Tree" Bonfring International Journal of Data Mining, Vol. 2 No. 1, March 2012, pg. 6-12.

[20] Anshul Goyal, Rajni Mehta "Performance Comparison of Naïve Bayes and J48 Classification Algorithms" International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7, November 2012, pg. 1-5.

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

[21] Thomas G.Dietterich "An Experimental Comparison Of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, And Randomization" Department Of Computer Science, Oregon State University, Corvallis, 1999, pg. 1-22.

[22] Suban Ravichandran, Vijay Bhanu Srinivasan and Chandrasekaran Ramasamy "Comparative Study on Decision Tree Techniques for Mobile Call Detail Record" Journal of communication and computer 9 ,December 2012, pg. 1331-1335.

[23] http://search.proquest.com/docview/30503150

[24] Bjorn Waske, Jon Atli Benediktsson "Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data" IEEE transaction on geoscience and remote Sensing, Vol. 48, No. 7, July 2010, pg. 2880-2889.

[25] Juan J. Rodr´ıguez and Ludmila I. Kuncheva "Naive Bayes Ensembles with a Random Oracle" Springer-Verlag Berlin Heidelberg 2007, pg. 450-458.

[26] Martin Godec, Christian Leistner, Amir Saffari, Horst Bischof "On-line Random Naive Bayes for Tracking" IEEE (ICPR), August 2010, pg. 3545-3548.

[27] http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

[28] http://cs229.stanford.edu/notes/cs229notes3.pdf

