ISSN: 2321-7782 (Online)
Volume 2, Issue 10, October 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Study on Prediction Performance of Some Data Mining Algorithms

M.Jayakameswaraiah[1]
Research Scholar, Dept.of Computer Science, Sri Venkateswra University, Tirupati,
Andhra Pradesh, India

S.Ramakrishna[2]
Professor, Dept.of Computer Science, Sri Venkateswara University, Tirupati,
Andhra Pradesh, India

Abstract: ***Data mining algorithms create how the cases for a data mining model are calculate. Data mining illustration algorithms present the decision-making capabilities necessary to categorize, segment, associate and examine data for the processing of data mining columns that present predictive, variance, or possibility information about the case set. With a huge quantity of data stored in databases and data warehouses, it is more and more important to develop dominant tools for analysis of such data and mining interesting information from it. Data mining is a process of inferring data from such huge data. Data Mining has three most important components Clustering/Classification, Association Rules and Sequence Analysis.***

Keywords: Data mining, K-Means, ID3, NEA, J48, SVM, FL, PART, Naïve Bayes, RFC, Apriori.

I. INTRODUCTION

Data mining functionalities are used to specify the kind of patterns to be found in data mining responsibilities. In common, data mining tasks can be classified into two categories: descriptive and predictive. The descriptive mining tasks characterize the general properties of the data in the database. Analytical mining tasks execute inference on the current data in order to create predictions. Data mining functionalities consist of the discovery of concept/class descriptions, associations and correlations, classification. Data mining is an iterative process within which progress is defined by detection, during either regular or manual methods.

Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. The selection and implementation of the appropriate data - mining technique is the main task in this segment. This method is not straightforward; generally, in practice, the implementation is based on some models, and select the most excellent one is a further task. By simple definition, in classification/clustering we analyze a set of data and generate a set of grouping rules which can be used to classify future data. One of the important problems in data mining is the Classification-rule learning which involves finding rules that partition given data into predefined classes. An association rule is a rule which implies certain association relationships among a set of objects in a database. In sequential Analysis, we seek to discover patterns that happen in order. This deal among statistics that appear in separate transactions [1].

The KDnuggets conducted a survey poll on data mining methods/algorithms in the year of 2011-2012 is as shown below.

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Figure 1: KDnugget survey poll on data mining algorithms

In this survey will gives the priority to the algorithm is first Decision Tree/Rules, Second Regression and third one is clustering as shown in the figure 1.

II. Algorithms For Data Analysis In Data Mining System

A. Sample Survey on Applied Data Mining Algorithms:

In this paper, all the algorithms are explained individually. Performance and outcome are compared of all algorithms and evaluation is done by already obtainable datasets. All the algorithms have a acceptable performance but a few of them give accuracy more. This section has revealed a survey of data mining techniques that have been applied to some datasets from UCI machine learning repository by various research groups [2].

1. K-Means Clustering

In K-Means clustering, (Kusum K. B.,2010) task of the data points to clusters is depend upon the distance between cluster centroid and data point. Accuracy of k-means clustering depends upon the value of k. determining the appropriate number of clusters is challenging area for researchers.

2. Nearest Cluster Algorithm (NEA).

Nearest cluster algorithm is a condensed version of K-nearest neighbor clustering algorithm. Input to this algorithm is a set of cluster centers generated from the training data set using standard clustering algorithms like K-Means, E&M binary split, and leader algorithm.

3. ID3 Algorithm

ID3, (Amanpreet C.,2011) is considered to be a very useful Inductive Logic Programming technique developed. ID3 is a characteristic based machine-learning algorithm that constructs a decision tree which is said to be based on a given training data set.

4. J48 Algorithm

The J48 algorithm (Huy A. N.,2008), is based on the algorithm intended with features which simply address the loopholes that are there in ID3. This algorithm was mainly considered as the enhanced version of C4.5 as the major drawback of C4.5 was the capacity of CPU time it took and the system memory it essential.

5. Partial Decision Tree

The PART algorithm (Mohammed M. M.,2009), was developed by Frank and Witten. This name was chosen because this algorithm generates rules by repeatedly producing partial decision trees. This algorithm is derived from C4.5 and RIPPER algorithms. Both C4.5 and RIPPER use decision trees to generate the rule set Unlike those rules, PART does not need to perform global optimization. decision tree is been generated, then transformed it into a rule set and finally it simplifies the rules.

6. Support Vector Machines (SVM)

First maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature (Huy A. N.,2008). A Support Vector Machines classifier is considered for binary classification. The simplification in this approach generally depends on the geometrical characteristics of the certain training data, and not on the terms of the put in space. This technique transforms the training data into a feature space of a vast dimension. That is, to split a set of exercise vectors which belong to two different classes (Xu X., 2006).

7. Fuzzy Logic (FL)

It processes the input data from the network and describes measures that are significant to the anomaly detection (Amanpreet C.,2011). Fuzzy logic (or fuzzy set theory) is based on the concept of the fuzzy phenomenon to happen regularly in the actual world. Fuzzy set hypothesis considers the set organization values for reasoning and the values series between 0 and 1. That is, in fuzzy logic the level of truth of a statement can range between 0 and 1 and it is not constrained to the two accuracy values (i.e. true, false).

8. Naïve Bayes

It classifier provides a simple approach based on the inferences of probabilistic graphic models which specify the probabilistic dependencies underlying a particular model using a graph structure (Huy A. N.,2008). In its simplest type, a probabilistic graphical form is a graph in which nodes represent accidental variables, and the arcs symbolize restricted confidence assumptions. Therefore it provides a compact representation of combined probability distributions. An undirected graphical form is called as a Markov network, even though a directed graphical form is called as a Bayesian network or a Belief network (Mrutyunjaya P., 2009).

9. Random Forest Classifier (RFC)

Random Forest was formulated in 1995 (Yeung D. Y., 2002). These processes combine bagging and the random selection of features to construct a group of decision trees with restricted dissimilarity. The choice of a random subset of features is a method of random subspace approach, which is a way to execute stochastic bias proposed by Eugene Kleinberg. The presentation of Decision Table and Random Forest classifiers are used to predict the classification accuracy. Based on this, Random Forest outperforms on the most techniques.

10. Apriori

It is a confidence-based Association Rule Mining algorithm. The fundamental idea of the Apriori algorithm is to produce frequent item sets for a certain dataset and then scan those frequent item set to distinguish most frequent items in this dataset. The process is iterative. Because generated frequent item sets from a step can construct another item sets by joining with previous frequent item sets (Mohammed M. M., 2009) [3].

III. Performance Comparison of Some Algorithms

S.No	Algorithm	Percentage of Successful Prediction (PSP)%	Training Time(TT) Sec.
1	K-Means	89.90	50.1
2	Nearest Cluster	92.22	10.63
3	ID3	93.42	20.51
4	J48	92.06	15.85
5	Partial Decision Tree	45.67	169
6	Support Vector Machines	81.38	222.28
7	Fuzzy Logic	94.80	873.9
8	Naïve Bayes	78.32	5.57
9	Random Forest Classifier	92.01	491
10	Apriori	87.50	18

Table 1: Top 10 Data mining Algorithms

Table 1 shows the accuracy of some data mining classification and clustering algorithms applied on some data sets using 10-fold cross validation is observed. It shows that Nearest Cluster, ID3 and J48 technique has highest accuracy compared to other methods. K-means and Fuzzy Logic algorithms also showed an acceptable level of accuracy. The table also shows the time complexity in seconds of various classifiers and clusters to build the model for training data [4, 5].

IV. Conclusion

After studying through the vast resources of technological papers written on data mining, here are a few of conclusions that I have made regarding the algorithms used for data mining. In this paper, ten existing decision tree algorithms have been applied on some data datasets for predicting the performance. All the algorithms are applied for the efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to draw from the tree. The predictions obtain from the classification have helped and improve their performance. The best algorithms among all the ten are nearest cluster, ID3, J48 and K-means because these provide better accuracy and efficiency than the further algorithms. At rest, efficient algorithms for decision tree have to be developed.

References

1. Dr.V.Karthikeyani, I.Parvin Begum," Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease", International Journal on Computer Science and Engineering (IJCSE), Vol. 5 No.03, ISSN: 0975-3397, Mar 2013.
2. P. IndiraPriya, Dr. D.K.Ghosh ," A Survey on Different Clustering Algorithms in Data Mining Technique", International Journal of Modern Engineering Research (IJMER), Vol.3, Issue.1, pp-267-274, ISSN: 2249-6645, Jan-Feb. 2013.
3. Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan," A SURVEY ON DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, ISSN (Online): 2278-1021, January 2014.
4. Nikita Jain, Vishal Srivastava," DATA MINING TECHNIQUES: A SURVEY PAPER", IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11, eISSN: 2319-1163, Nov-2013
5. Ms. Ruth D, Mrs. Lovelin Ponn Felciah M," A Survey on Intrusion Detection System with Data Mining Techniques", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, ISSN 2348 – 7968, May 2014.

