Research Paper Computer Science

Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA

* Kaushik H. Raviya ** Biren Gajjar

* Assistant Professor, B.H. Gardi Vidyapith, Anandpar, Gujarat, India

** B.E. (CSE) Student, B.H. Gardi Vidyapith, Anandpar, Gujarat, India

ABSTRACT

Data mining is a powerful new field of computer science with great potential , is the process of discovering or extracting new patterns from large datasets or databases. It is commonly used in marketing, surveillance, fraud detection, artificial intelligence, scientific discovery and now gaining a broad way in other fields also. Classification is a data mining technique used to predict the class of objects whose class label is unknown. There are many classification mechanisms used in data mining such as K-Nearest Neighbor (KNN), Bayesian network, Neural networks, Decision trees, Fuzzy logic, Support vector machines, etc. This paper presents comparison on three classification techniques which are K-nearest neighbor, Bayesian network & Decision tree respectively. The goal of this research is to enumerate the best technique from above three analyzed under a given dataset and provide a fruitful comparison result which can be used for further analysis or future development.

Keywords: Classification, Decision tree, Bayesian network and K-Nearest Neighbor

1. Introduction

In today's world large quantities of data is being accumulated and seeking knowledge from massive data is one of the most fundamental attribute of Data Mining. It consist of more than just collecting and managing data but to analyze and predict also. Data could be large in two senses : in terms of size & in terms of dimensionality. Also there is a huge gap from the stored data to the knowledge that could be construed from the data. Here comes the classification technique and its sub-mechanisms to arrange or place the data at it's appropriate class for ease of identification and searching. Thus classification can be outlined as inevitable part of data mining and is gaining more popularity. In the present paper comparison on three classification technique is done with its fundamental knowledge.

2. Decision Tree :

Decision trees are powerful classification algorithms. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and Breiman et al.'s CART. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. Most decision tree classifiers perform classification in two phases: tree-growing (or building) and tree-pruning. The tree building is done in top-down manner. During this phase the tree is recursively partitioned till all the data items belong to the same class label. In the tree pruning phase the full grown tree is cut back to prevent over fitting and improve the accuracy of the tree in bottom up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing the over-fitting. Compared to other data mining techniques, it is widely applied in various areas since it is robust to data scales or distributions.

3. Bayesian Network :

Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. Bayesian algorithms predict the class depending on the probability of belonging to that class. A Bayesian network is a graphical model for probability relationships among a set of variables features. This BN consist of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables. Second component is a set of parameters that describe the conditional probability of each variable given its parents. The conditional dependencies in the graph are estimated by statistical and computational methods. Thus the BN combine the properties of computer science and statistics.

4. K-Nearest Neighbor :

K-Nearest Neighbor is one of the best known distance based algorithms, in the literature it has different version such as closest point, single link, complete link, K-Most Similar Neighbor etc. Nearest neighbors algorithm is considered as statistical learning algorithms and it is extremely simple to implement and leaves itself open to a wide variety of variations. The k-nearest neighbors' algorithm is amongest the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. K is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. KNN has got a wide variety of applications in various fields such as Pattern recognition, Image databases, Internet marketing, Cluster analysis etc. The Table 1 below gives the theoretical comparison on classification techniques.

Parameters	Decision Tree	Bayesian Network	K-Nearest Neighbor
Fundamentals	Decision trees are powerful classification algorithms. Decision tree algorithms build a tree which also yields a series of if- else-then rules to classify the data items.	Bayesian networks are a powerful probabilistic representation where they do graphical representation of probability distribution. It is also called belief networks.	K-Nearest Neighbor is one of the best known distance based algorithms and is considered as statistical learning algorithms

Work of classifier	It recursively partitions a data set of records using depth-fist greedy approach or breadth-first approach, until all the data items belong to a particular class are identified. It is a structure made of root, internal and leaf nodes.	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.	When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample
Phases of work	- Tree building - Tree Pruning	- Directed Acyclic Graph (DAG) - Conditional Probabilities	No Phases
Pros	- Construction does not require any domain knowledge - Can handle high dimensional data	- Naive Bayesian classifier simplifies the computations - Exhibit high accuracy and speed when applied to large databases	- Analytically tractable - Simple in implementation - Uses local information, which can yield highly adaptive behavior - Lends itself very easily to parallel implementations
Cons	- Output attribute must be categorical - Limited to one output attribute	- The assumptions made in class conditional independence - Lack of available probability data	- Large storage requirements - Highly susceptible to the curse of dimensionality - Slow in classifying test tuples
Applications	In Decision making systems, Teaching, Research area, etc.	In computational biology and bioinformatics medicine, document classification, information retrieval, semantic search, image processing, data fusion, etc.	In Pattern recognition, Image databases, Internet marketing, Cluster analysis etc.

Table 1 : Comparison of Classification Techniques

5. Experimental Work :

For the practical scenario comparison on these techniques is done through WEKA. WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. Here we have used a common database (which is a supermarket database) for all the three mechanisms, so that we can differentiate their parameters on a single instance. This super market database consist of 217 attributes (attributes like bread, tea, biscuits, department, etc.) and 4627 instances.

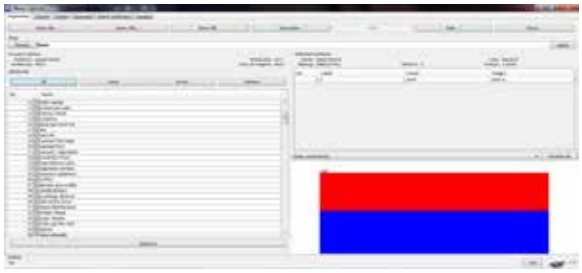

Figure 1 : WEKA 3.7.7 - Explorer window

The above is the explorer window in WEKA tool with the supermarket dataset loaded. Here we can also analyze the data in the form of graph as shown above in visualization section with red and blue code. In WEKA, all data is considered as instances features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned in numeric and percentage value and subsequently Kappa statistic, mean absolute error and root mean squared error will be in numeric value only.

Figure 2 : Classifier Result

This dataset is measured and analysed with 10 folds cross validation under specified classifier as shown in figure 2. Here it computes all required parameters on given instances with the classifiers respective accuracy and prediction rate. Based on Table 2 given below we can clearly see that the highest accuracy is 63.71% for Bayesian and Decision tree and lowest is 37.12% for KNN. For Bayesian and Decision tree there is huge difference in time taken to build the model. In fact by this experimental comparison we can say that Bayesian is best among three as it is more accurate and less time consuming.

Classifier (Total Instances, 4627)	Algorithm Implemented	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (Seconds)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Decision Tree	trees.Random Forest	63.713 (2948)	36.287 (1679)	0.78	0	0.4623	0.4808	99.9738	100.0012
Bayesian Network	bayes.Naive Bayes	63.713 (2948)	36.287 (1679)	0.05	0	0.4624	0.4808	100	100
K-Nearest Neighbor	lazy.IBK	37.1299 (1718)	62.8701 (2909)	0.01	0.0083	0.6218	0.7806	134.4739	162.3358

Table 2 : Simulation Result of each Algorithm

6. Conclusion :

Classification is one of the most popular techniques in data mining. In this paper we compared algorithms based on their accuracy, learning time and error rate. We observed that, there is a direct relationship between execution time in building the tree model and the volume of data records and also there is an indirect relationship between execution time in building the model and attribute size of the data sets. Through our experiment we conclude that Bayesian algorithms have good classification accuracy over above compared algorithms.

7. Future work :

In future we will work increasing the accuracy of the Bayesian algorithm where it is lacking behind other recently developed mechanisms.

REFERENCES

[1] Sunita B Aher, "Data Mining in Educational System using WEKA," International Conference on Emerging Technology Trends (ICETT) 2011 Proceedings published by International Journal of Computer Applications® (IJCA) | [2] Han,J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. | [3] P.Nancy & Dr.R.Geetha Ramani, "A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data," International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011 | [4] Thair Nu Phyu "Survey of Classification Techniques in Data Mining",Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, March 18 - 20, 2009, Hong Kong | [5] Matthew N. Anyanwu & Sajjan G. Shiva "Comparative Analysis of Serial Decision Tree Classification Algorithms". | [6] Gökhan Silahtaroğlu ,"An Attribute-Centre Based Decision Tree Classification Algorithm", World Academy of Science, Engineering and Technology 32 2009 | [7] Raj Kumar & Dr. Rajesh Verma "Classification Algorithms for Data Mining:A Survey," International Journal of Innovations in Engineering and Technology (IJIET) | [8] 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 | [9] Mohd Fauzi bin Othman &Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," Biomed 06, IFMBE Proceedings 15, pp. 520-523, 2007, www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007 |