ISSN 2320 - 2602

Volume 2, No.3, March 2013
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2013/ijacst03232013.pdf

COMPARATIVE INVESTIGATION OF DECISION TREE ALGORITHMS ON IRIS DATA

Mahendra Tiwari
(Asstt. Professor,UCER,Naini Allahabad)
Tiwarimahendra29@gmail.com
Vivek Srivastava
(Asstt. Professor,UCER,Naini Allahabad)
viveksrivastava@united.ac.in
Vivek Pandey
(Sr.Lecturer,UCER,Naini Allahabad)

Abstract: Data mining is a computerized technology that uses complicated algorithms to find relationships and trends in large data bases, real or perceived, previously unknown to the retailer, to promote decision support.., data mining is touted to be one of the widespread recognition of the potential for analysis of past transaction data to improve the quality of future business decisions.The purpose is to organize a collection of data items and classify them, In this paper, we use J48(c4.5) , and CART algorithm and compare the performance evaluation of both with IRIS data.

Keywords: Data Mining, Decision Tree, J48, CART

1. INTRODUCTION

People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

2. RELATED WORK

Abdullah compared various classifiers with different types of data set on WEKA, we presented their result as well as about tool and data set which are used in performing evaluation.

The article "performance evaluation and characterization of scalable data mining algorithms" [1] investigated data mining applications to identify We studied various journals and articles regarding performance evaluation of Data Mining algorithms on various different tools, some of them are described here, Ying Liu et all worked on Classification algorithms while Osama abu abbas worked on clustering algorithm, and their characteristics in a sequential as well as parallel execution environment .They first establish Mine bench, a benchmarking suite containing data mining applications. The selection principle is to include categories & applications that are commonly used in industry and are likely to be used in the future, thereby achieving a realistic representation of the existing applications. Minebench can be used by both programmers & processor designers for efficient system design.

They conduct their evaluation on an Intel IA-32 multiprocessor platform, which consist of an Intel Xeon 8-way shared memory parallel(SMP) machine running Linux OS, a 4 GB shared memory & 1024 KB L2 cache for each processor. Each processor has 16 KB non-blocking integrated L1 instructions and data caches. The number of processors is varied to study the scalability.

In all the experiments, they use VTune performance analyzer for profiling the functions within their applications, & for measuring their breakdown execution times. VTune counter monitor provides a wide assortment of metrics. They look at different characteristics of the applications: execution time, fraction of time spent in the OS space, communication/synchronization complexity& I/O complexity. The Data comprising 250,000 records. This notion denotes the dataset contains 2,00,000 transactions,the average transaction size is 20, and the average size of the maximal potentially large itemset is 6. The number of items is 1000 and the number of maximal potentially large itemset is 2000. The algorithms for comparison are ScalParc, Bayesian, K-means, Fuzzy K-means, BIRCH,HOP,Apriori, & ECLAT.

Osama compared four different clustering algorithm in his article "comparison between data clustering algorithms" [2] (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters ,type of S/W. The general reasons for selecting these 4 algorithms are:

- Popularity
- Flexibility
- Applicability
- Handling High dimensionality

Osama tested all the algorithms in LNKnet S/W- it is public domain S/W made available from MIT Lincoln lab www.li.mit.edu/ist/lnknet.
For analyzing data from different data set, located at www.rana.lbl.gov/Eisensoftware.htm

The dataset that is used to test the clustering algorithms and compare among them is obtained from the site www.kdnuggets.com/dataset .This dataset is stored in an ASCII file 600 rows,60 columns with a single chart per line "A comparison study between data mining tools over some classification methods" conducted a comparison study between a number of open source data mining S/W and tools depending on their ability for classifying data correctly and accurately.

The methodology of the study constitute of collecting a set of free data mining & knowledge discovery tools to be tested, specify the datasets to be used, and selecting a set of classification algorithm to test the tool‟s performance. For testing, each dataset is described by the data type being used, the types of attributes, whether they are categorical ,real, or integer, the number of instances stored within the data set, the number of attributes that describes each dataset, and the year the data set was created. After selecting the dataset , a 1-100 normal
101-200 cyclic
201-300 increasing trend
301-400 decreasing trend
401-500 upward shift
501-600 downward shift

Abdullah in his article number of classification algorithm are chosen that are Naïve Bayes, K-nearest, SVM,C4.5 as well as some classifiers are used that are Zero R, One R, & Decision Tree classifier. For evaluating purpose two test level modes were used; the K-fold cross validation mode and the percentage split mode. After running the four tools ,they have obtained some results regarding the ability to run the selected algorithm on the selected tools. All algorithms ran successfully on WEKA, the 6 selected classifiers used the 9 selected data sets.

The performance of two algorithm are implemented and analyzed in [3] in research paper "performance evaluation of K-means & Fuzzy C-means clustering algorithm for statistical distribution of input data points" studied the performance of K-means & Fuzzy Cmeans algorithms. These two algorithm are implemented and the performance is analyzed based on their clustering result quality. The behavior of both the algorithms depended on the number of data points as well as on the number of clusters. The input data points are generated by two ways, one by using normal distribution and another by applying uniform distribution (by Box-muller formula). The performance of the algorithm was investigated during different execution of the program on the input data points. The execution time for each algorithm was also analyzed and the results were compared with one another, both unsupervised clustering methods were examined to analyze based on the distance between the various input data points. The clusters were formed according to the distance between data points and clusters centers were formed for each cluster. The implementation plan would be in two parts, one in normal distribution and other in uniform distribution of input data points. The data points in each cluster were displayed by different colors and the execution time was calculated in milliseconds.

Velmurugan and Santhanam chose 10 (k=10) clusters and 500 data points for experiment. The algorithm was repeated 500 times (for one data point one iteration) to get efficient output. The cluster centers (centroid) were calculated for each clusters by its mean value and clusters were formed depending upon the distance between data points

Minebench applications are evaluated by Jayaprakash [4] in their paper "performance characterization of Data Mining applications using Minebench" presented a set of representative data mining applications call Minebench. They evaluated the Minebench application on an 8 way shared memory machine and analyze some important performance characteristics. Minebench encompasses many algorithms commonly formed in data mining. They analyzed the architectural properties of these applications to investigate the performance bottleneck associated with them. For performance characterization, they chose an Intel IA-32 multiprocessor platform, Intel Xeon 8-way shared memory parallel (SMP) machine running Red Hat advanced server 2.1. The system had 4 GB of shared memory. Each processor had a 16 KB non-blocking integrated L1 cache and a 1024 KB L2 cache. For evaluation they used VTune performance analyzer. Each application was compiled with version 7.1 of the Intel C++ compiler for Linux.

The data used in experiment were either real-world data obtained from various fields or widely accepted synthetic data generated using existing tools that are used in scientific and statistical simulations. During evaluation, multiple data sizes were used to investigate the characteristics of the Minebench applications, For non-bioinformatics applications, the input datasets were classified in to 3 different sizes: small, medium, & large. IBM Quest data generator, ENZO, & real image database by Corel corporation.

Association rule mining algorithms are evaluated with sort and unsort data by Pramod [5] in their research paper "performance evaluation of some online association rule mining algorithms for sorted & unsorted data sets" evaluated association rule mining algorithm for sorted and unsorted data sets. They worked on Continuous Association Rule Mining Algorithm (CARMA) and Data Stream Combinatorial approximation Algorithm (DSCA) , & estDec method. The 3 algorithms are implemented in JAVA and the results were plotted, all 3 algorithms were tested with 5 data sets and all of them are available in Frequent Itemset Mining data set (FIM) repository. The transactions of each data set were looked up one by one in sequence to simulate the environment of an online data stream. The DSCA algorithm used sorted transaction items while other 2 algorithms used unsorted transaction items.

3. ANALYSIS OF DATA MINING ALGORITHM

Data Mining is the extraction of hidden predictive information from large databases. It is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses

.Data mining tools predicts future trends and behaviors, helps organizations to make proactive knowledge-driven decisions.

3.1 Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Algorithm: Decision_Tree:

Generate a decision tree from the training tuples of data partition D.

Input:

Data partition, D: A set of training tuples and their associated class labels;
attribute_list : The set of candidate attributes;
Attribute.selection_method : A procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes. This criterion consists of a splitting, attribute and, possibly, either a split point or splitting subset**.**

Method:

1. create a node N;
2. if tuples in D are all of the same class, C then
3. return N as a leaf node labeled with the class C;
4. if attribute_list is empty then
5. return N as a leaf node labeled with the majority class in D;
6. apply Attribute_selection_method(D, attribute_list) to find the “best” splittmg_criterion;
7. label node N with splitting_criterion;
8. if splitting_attribute is discrete-valued and multiway splits allowed then 1/ not restricted to binary trees
9. attribute_list „<— splitting_attribute;
10. for each outcome j of splitting_criterion
11. let Dj be the set of data tuples in D satisfying outcome j;
12. if Dj is empty then
13. attach a leaf labeled with the majority class in D to node N;
14. else attach the node returned by Generate_decision_tree(D, attribute_list) to node N; endfor
15. returnN;

3.2 Evaluation Strategy/Methodology

H/W tools:

We conduct our evaluation on Pentium 4 Processor platform which consist of 512 MB memory, windows xp operating system, a 40GB memory, & 1024kbL1 cache.

S/W tool:

In all the experiments, we used Weka 3-6-8, we looked at different characteristics of the applications-using classifiers to measure the accuracy in different data sets, using classifier to build models etc. Weka toolkit is a widely used toolkit for machine learning and data mining that was originally developed at the university of Waikato in New Zealand . It contains large collection of state-of-the-art machine learning and data mining algorithms written in Java. Weka contains tools for regression, classification, clustering, association rules, visualization, and data processing.

Input Data set:

Input data is an integral part of data mining applications. The data used in our experiment is real world data obtained from UCI data repository and widely accepted dataset available in Weka toolkit, during evaluation, the dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset, This dataset was chosen because they have different characteristics. This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.

Number of Instances: 150 (50 in each of three classes)
Number of Attributes: 4 numeric, predictive attributes and the class Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

missing Attribute Values: None

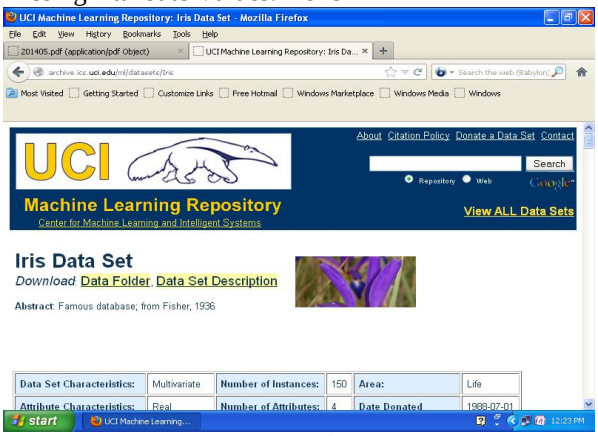

Figure 1: Iris data from UCI repository

4. EXPERIMENTAL RESULT & DISCUSSION

To evaluate the selected tool using the given dataset, several experiments are conducted. For evaluation purpose, test mode used, the k-fold cross-validation(k-fold cv) mode,. The k-fold cv refers to a widely used experimental testing procedure where the database is randomly divided in to k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process is repeated k times. At the end, the recorded measures are averaged. It is common to choose k=10 or any other size depending mainly on the size of the original dataset. Once the tests is carried out using the selected dataset, then using the available classification and test mode ,results are collected and an overall comparison is conducted. Weka has four panels to perform operations on it,Simple CLI, Explorer, Experimenter, Knowledge Flow. We used Experimenter to compare 2 Decision tree algorithms that are J48, & Simple CART.

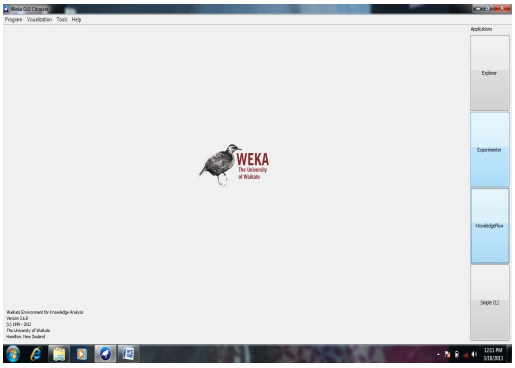

Figure 2: Weka application

The result of experiments are stored in arff file format

ie experiment2.arff.

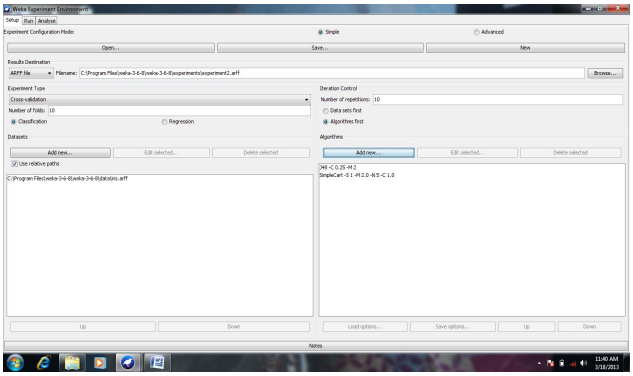

Figure 3: Experiment2.arff

For comparison of both algorithms we chose comparison fields from Analyse tab, select the *percent_correct* attribute and then perform test to generate a comparison of 2 schemes.

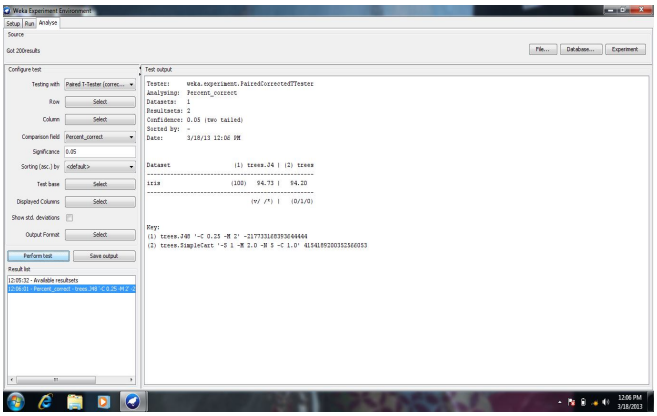

Figure 4: Percent_correct

The percentage correct for each of the 2 schemes is described in each dataset row: 94.73% for J48, and 94.20% for CART. The annotation v or * indicates that a specific result is statistically better(v) or worse(*) than the baseline scheme at the significance level specified (currently 0.05). Selecting *number_correct* as the comparison field and clicked perform test generate the average number correct (out of 50 test patterns-14% of 150 patterns in the Iris dataset).

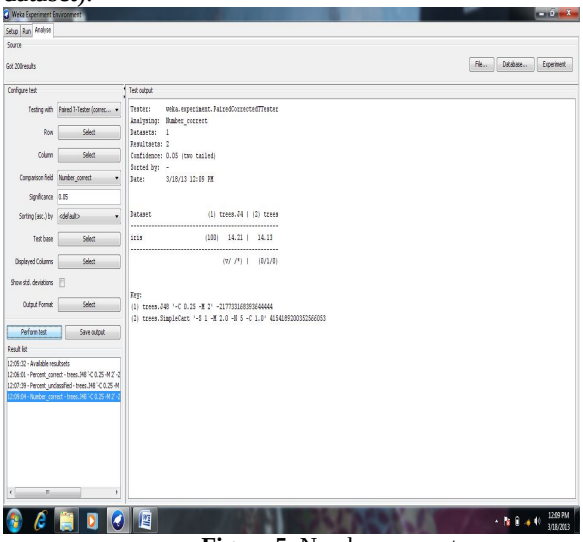

Figure 5: Number_correct

5. RESULT

The result of experiment is saved in arff format ,total comparative analysis result is large, we present some part of it over here.

relation InstanceResultListener

@attribute Key_Dataset {iris}
@attribute Key_Run {1,2,3,4,5,6,7,8,9,10}
@attribute Key_Fold {1,2,3,4,5,6,7,8,9,10}
@attribute Key_Scheme
{weka.classifiers.trees.J48,weka.classifiers.trees.SimpleCart}

@attribute Key_Scheme_options {'-C 0.25 -M 2','-S 1 -M 2.0 -N 5 -C 1.0'}
@attribute Key_Scheme_version_ID {-2177331683936444444,4154189200352566053}
@attribute Date_time numeric
@attribute Number_of_training_instances numeric
@attribute Number_of_testing_instances numeric
@attribute Number_correct numeric
@attribute Number_incorrect numeric
@attribute Number_unclassified numeric
@attribute Percent_correct numeric
@attribute Percent_incorrect numeric
@attribute Percent_unclassified numeric
@attribute Kappa_statistic numeric
@attribute Mean_absolute_error numeric
@attribute Root_mean_squared_error numeric
@attribute Relative_absolute_error numeric
@attribute Root_relative_squared_error numeric
@attribute SF_prior_entropy numeric
@attribute SF_scheme_entropy numeric
@attribute SF_entropy_gain numeric
@attribute SF_mean_prior_entropy numeric
@attribute SF_mean_scheme_entropy numeric
@attribute SF_mean_entropy_gain numeric
@attribute KB_information numeric
@attribute KB_mean_information numeric
@attribute KB_relative_information numeric
@attribute True_positive_rate numeric
@attribute Num_true_positives numeric
@attribute False_positive_rate numeric
@attribute Num_false_positives numeric
@attribute True_negative_rate numeric
@attribute Num_true_negatives numeric
@attribute False_negative_rate numeric
@attribute Num_false_negatives numeric
@attribute IR_precision numeric
@attribute IR_recall numeric
@attribute F_measure numeric
@attribute Area_under_ROC numeric
@attribute Weighted_avg_true_positive_rate numeric
@attribute Weighted_avg_false_positive_rate numeric
@attribute Weighted_avg_true_negative_rate numeric
@attribute Weighted_avg_false_negative_rate numeric
@attribute Weighted_avg_IR_precision numeric
@attribute Weighted_avg_IR_recall numeric
@attribute Weighted_avg_F_measure numeric
@attribute Weighted_avg_area_under_ROC numeric
@attribute Elapsed_Time_training numeric
@attribute Elapsed_Time_testing numeric
@attribute UserCPU_Time_training numeric
@attribute UserCPU_Time_testing numeric
@attribute Serialized_Model_Size numeric
@attribute Serialized_Train_Set_Size numeric
@attribute Serialized_Test_Set_Size numeric
@attribute Summary {'Number of leaves: 4\nSize of the tree: 7\n','Number of leaves: 5\nSize of the tree: 9\n','Number of leaves: 3\nSize of the tree: 5\n','Number of leaves: 6\nSize of the tree: 11\n'}
@attribute measureTreeSize numeric
@attribute measureNumLeaves numeric
@attribute measureNumRules numeric

@data iris,1,1,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,14,1,0,93.333333,6.666667,0,0.9,0.045016,0.169318,10.128603,35.917699,23.774438,2.632715,21.141722,1.584963,0.175514,1.409448,21.615654,1.441044,1363.7959,1,5,0,0,1,10,0,0,1,1,1,1,0.933333,0.033333,0.966667,0.066667,0.944444,0.933333,0.93266,1,0.031,0,0.0156,0,4449,11071,2791,'Number of leaves: 4\nSize of the tree: 7\n',7,4,4
iris,1,2,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,15,0,0,100,0,0,1,0.010588,0.015887,2.382303,3.37004,23.774438,0.347856,23.426581,1.584963,0.02319,1.561772,23.426581,1.561772,1478.052717,1,5,0,0,1,10,0,0,1,1,1,1,1,0,1,0,1,1,1,1,0,0,0,0,4850,11071,2791,'Number of leaves: 5\nSize of the tree: 9\n',9,5,5
iris,1,3,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,15,0,0,100,0,0,1,0.010588,0.015887,2.382303,3.37004,23.774438,0.347856,23.426581,1.584963,0.02319,1.561772,23.426581,1.561772,1478.052717,1,5,0,0,1,10,0,0,1,1,1,1,1,0,1,0,1,1,1,1,0,0,0,0,4850,11071,2791,'Number of leaves: 5\nSize of the tree: 9\n',9,5,5
iris,1,4,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,15,0,0,100,0,0,1,0.010588,0.015887,2.382303,3.37004,23.774438,0.347856,23.426581,1.584963,0.02319,1.561772,23.426581,1.561772,1478.052717,1,5,0,0,1,10,0,0,1,1,1,1,1,0,1,0,1,1,1,1,0,0,0,0,4850,11071,2791,'Number of leaves: 5\nSize of the tree: 9\n',9,5,5
iris,1,5,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,14,1,0,93.333333,6.666667,0,0.9,0.058656,0.21278,13.197674,45.137402,23.774438,5.977092,17.797346,1.584963,0.398473,1.18649,21.087633,1.405842,1330.48151,1,5,0,0,1,10,0,0,1,1,1,1,0.933333,0.033333,0.966667,0.066667,0.944444,0.933333,0.93266,0.96,0,0,0,4449,11071,2791,'Number of leaves: 4\nSize of the tree: 7\n',7,4,4
iris,1,6,weka.classifiers.trees.J48,'-C 0.25 -M 2',-2177331683936444444,20130318.0614,135,15,15,0,0,100,0,0,1,0.009401,0.014865,2.115172,3.153328,23.774438,0.308798,23.465639,1.584963,0.020587,1.564376,23.465639,1.564376,1480.516994,1,5,0,0,1,10,0,0,1,1,1,1,1,0,1,0,1,1,1,1,0,0,0,0,4850,11071,2791,'Number of leaves: 5\nSize of the tree: 9\n',9,5,5

CONCLUSION

From the above investigation, it can be said that J48 (C4.5) classification method is better than CART in small to medium size data set. The advantage of J48 is its low computation cost,and drawback is sensitive to nosy data .

REFERENCES

1. *www.ics.**uci**.edu/~mlearn/*
2. www.eecs.northwestern.ed/~yingliu/papers/pdcs.
3. Velmurugan T., T. Santhanam(2010), performance evaluation of k-means & fuzzy c-means clustering algorithm for statistical distribution of input data points., European Journal of Scientific Research, vol 46 no. 3
4. Osama A. Abbas(2008),Comparison between data clustering algorithm, The International Arab journal of Information Technology, vol 5, N0. 3
5. Pramod S., O. Vyas(2010), Performance evaluation of some online association rule mining algorithms for sorted & unsorted datasets, International Journal of Computer Applications, vol 6

6. John F. Elder et al, (1998) A Comparison of Leading Data Mining Tools, Fourth International Conference on Knowledge Discovery

7. Agrawal R, Mehta M., Shafer J., Srikant R., Aming (1996) ,the Quest on Knowledge discovery and Data Mining, pp. 244-249

8. Hen L., S. Lee(2008), performance analysis of data mining tools cumulating with a proposed data mining middleware, Journal of Computer Science

9. Giraud, C., Povel, O.,(2003), characterizing data mining software, Intell Data anal 7:181-192

10. X.Hu, (2002) "Comparison of classification methods for customer attrition analysis" in Proc, of the Third International Conference on Rough Sets and Current Trends in Computing, Springer, pp. 4897-492.

11. Kleissner, C.(1998),, data mining for the enterprise, Proceeding of the 31st annual Hawaii International conference on system science

12. Dhond A. et all (2000), data mining techniques for optimizing inventories for electronic commerce. Data Mining & Knowledge Discovery 480-486

13. A. Kusiak, (2002) Data Mining and Decision making, in B.V.Dasarathy (Ed.). Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology TV, ol. 4730, SPIE, Orlando, FL, pp. 155-165.