International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23583-23597
© Research India Publications
http://www.ripublication.com

Data Analysis and Visualization Technique: An Empirical Framework

Manohara M, Dinesh R, Sowmya M S
[1]*kotamanu@gmail.com*
[2]*dr.dineshr@gmail.com*
[3]*mssowmya.sbmjce@gmail.com*
[1,3]*Research Scholar,*[2] *Research Supervisor*
Department of Information Science & Engineering,
School of Engineering and Technology, Jain University, Bangalore, Karnataka State,(India).

Abstract

Choosing suitable Data Analysis and visualization techniques for a given data set are an important task in data mining applications. Large varieties of classifiers and visualization techniques are proposed in the literature, however, it is very difficult for non-experts to choose the appropriate data analysis and visualization technique to suite their problem. One way of choosing the appropriate data analysis and visualization technique is through brute force method which is cumbersome and practically not viable. Hence there is a need for selecting appropriate Data analysis and visualization technique for given data set and objectivity in more effective and efficient way. In this paper, we have developed a method for suggesting the appropriate Data analysis and visualization technique for a given large dataset. The proposed method automatically selects an appropriate data analysis technique by considering both supervised and unsupervised classifiers (six different classifiers, four clustering, and three association rule mining algorithm). Also the proposed method suggests the appropriate visualization technique for a given dataset. The method has been tested on various datasets to establish the appropriateness in selecting the models. The proposed system has also been compared with the manually selected classifier, association rule, and cluster to establish its superiority.

Key words: Classifiers, Clusters, Data Mining, Knowledge Discovery from Data (KDD), Machine Learning Tools, Visualization Techniques.

Introduction

Social Networks, Satellites, Emails, Chats, Text messages, Power Grid, Share market, and Online trading generates large quantity of data (SOURCE) daily, making a drastic increase in the amount of the data stored in electronic format, which are useful for the analysis of various applications and to make valid business decisions. Analysis and visualization of such data will add benefits for the industry growth, by the use of the data mining and visualization techniques. Generally, the experts or data miners chose the data mining technique by using two main parameters irrespective of technical characteristics expected. The final choice of best data analysis depends basically on:

- Goal of the problem
- Structure of the available data

The results of data mining technique always depend on various parameters, viz., the number of instances, number of attributes, & types of the attributes in the dataset, and each technique is based on statistical, probabilistic, regression & other mathematical concepts. This creates difficulty in deciding which data analysis tool would give the efficient result for the given data sets.

Major problem in selecting the best data analysis tools is to decide on what basis user will select the best classifier/ clustering or association rule. Always the best data analysis tool has to satisfy the following three parameters:

- Accuracy: the percentage of correctness in the classification using data analysis.
- Scalability: even for the dynamic changes in the number of instance of the given data set, the proposed model must give the same classifiers/ cluster/ association rule as best data analysis tools.
- Execution time: best data analysis should take less execution time for analysing the data set when compared with other data analysis tools.

In this paper, we have proposed a solution, providing a method for automatically selecting the data mining and visualization techniques for the given data set. Section II provides the brief literature review on the related topic, followed by the broad view on classifiers and Weka machine learning tool in Section III. Section IV describes the proposed approach, and its implementation details are presented in section V, Finally, the conclusions and result analysis are presented in Section VI.

Related Work

In literature, industries, organizations and many other users of SOURCE, first have to pre-process the data set. Once the pre-processing is done, the user has variety of options in selecting the data analysis tools namely, classifiers, clustering, association rules, regression analyses, and various visualization techniques for analysing and representing the result in the format specified as per the end user requirement. A visualization technique depends on the purpose for which the data set is used or analysed. The technique involved in data mining for automatically selecting visualization tool is complex as per the literature. According to the classical scheme

of the Data mining/ Knowledge discovery from data (KDD) [1], the process is shown in the fig 1.

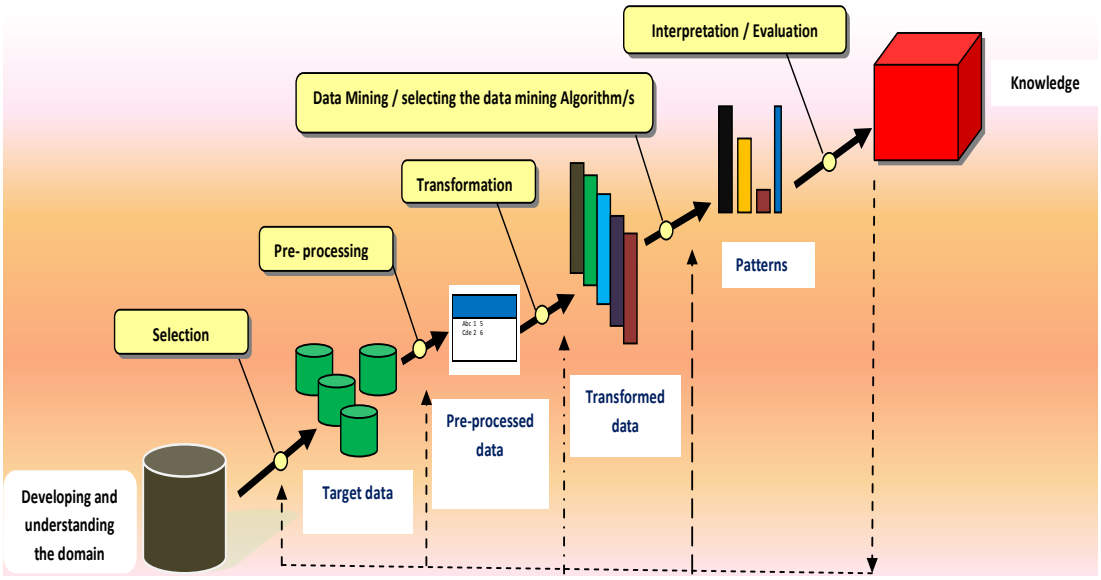


Figure 1: Data Mining: a KDD Process

Our research mainly concentrates on selecting the data mining algorithms for the pre-processed datasets. The data mining algorithm selection is mainly based on the Meta features of the datasets and its important attributes. Finding the important attributes, plays a major role in the selection of data mining algorithm, such attributes are identified using the principle component analysis (PCA). The bird eye view on Meta features and PCA are given in the subsequent sections.

A. Meta Learning:

The experts in the Meta learning model designing & Meta feature selection domain suggests which classifiers can be used for the user objectives, otherwise user have to test accuracy of the dataset using all available classifiers in the market, practically, this is time consuming process and sometimes it is impossible approach. The designing of the Meta learning model improves the generalization ability & stability of the learned models and support data mining automation in the issues related to algorithm and parameter selection [2]. Algorithm proposed in the literature runs in two phases; primarily the acquisition phase, here, all classifiers results are stored in the system with various possible parameter values of the classifiers, on different samples of dataset. Secondly, the advisory phase. In these approaches, all the solutions are displayed using Euclidian distance in the descending order of the priority, for each sampling dataset. The Algorithm uses the same dataset with varying samples for building the training dataset, and parameter considered for selecting the best classifiers are only the performance/accuracy of the classifiers, and not scalability & execution time.

Various Meta learners and the various Meta features [3], are derived from different concepts, and therefore can be categorized into five groups: simple, statistical, information-theoretic, model-based, and land marking meta-features. These Meta features play a major role in deciding classifiers performance. Based on the extracted Meta features and classifier's accuracy, it is possible to predict the best classifiers using linear regression techniques. However, a possible drawback of the regression approach is that for each target algorithm; a separate regression model has to be trained, which is impractical. Initially user inputs the data, then its Meta features will be selected and ask user to select the required classifier for test. [4] Meta learning and accuracy are used as the target value for the regression, to predict & display the best solution accuracy and root mean square error (RMSE) to the user. For each classifier regression model will be built and found works well when similar kind of Meta featured data sets are used. The research mainly concentrates on support vector machine (SVM) parameter selection, which plays an important role in the classification accuracy of that algorithm. The best bayes classifiers (considering six varieties of bayes classifiers) are selected for the given datasets [5]. The training set is based on the number of attributes & instances, the decision tree is constructed, in which each leaf represents the best solution. Once the data set is inputted, based on the already constructed decision tree, the best solution will be provided. Authors in [6] has proposed ranking of the best classifiers based on success rate ratio (SRR) for the given datasets. Meta learning database will be created using this test sample evaluated by SRR, based on the priority order all possible solution will be displayed.

B. Principle Component Analysis (PCA):

One of the parameter which degrades the classifiers performance is dimensionality. To overcome this problem most of the researcher uses PCA [7, 8] as one of the tool to reduce the dimension of the feature vector. principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each subsequent components in turn has the highest variance possible under the constraint that are mutually orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

According to these literature surveys, the Meta feature selection varies from one researcher to another and the Meta feature is selected to decide the particular classification algorithm & its parameter to achieve better result. In the proposed method we have considered various Meta features as the literature along with few of our Meta features that are derived from our research work. For the proposed method, all of these Meta features are considered before suggesting the best data mining & visualization techniques for the user datasets.

Overview of Classifiers and Weka Machine Learning Tool

Classification is an important area of data mining problems [9]. A classifier is a global model which is used to predict the class label for data objects that are unlabeled. Various types of classifiers are available in literature few of them are listed below,

1. Rule based classifiers:

A rule-based classier consists of a set of rules, used in a given order during the prediction process, to classify unlabeled objects. For further details readers are advised to refer [10].

2. Bayes Classifiers:

Bayesian classifiers [10] use Bayes theorem, which says:

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Where,

- $p(c_j \mid d)$ = probability of instance d being in class c_j (This is what we are trying to calculate).
- $p(d \mid c_j)$ = probability of generating instance d given class c_j (One can imagine that being in class c_j, causes to have feature d with some probability).
- $p(c_j)$ = probability of occurrence of class c_j (This is just how frequent the class c_j,, in database).
- p(d) = probability of instance d occurring,(This can actually be ignored, since it is the same for all classes).

3. Decision tree classifiers:

Decision Tree Classifier is a simple and widely used classification technique [11]. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

4. Lazy classifiers:

Lazy classifiers store all of the training samples and do not build a classifier/ a general model until a new sample needs to be classified [12]. It differs from eager classifiers, such as decision tree induction, which build a general model (such as a decision tree) before receiving new samples.

5. Ensemble classifiers:

The idea of ensemble methodology is to build a predictive model by integrating the multiple models [13]. It is well known that ensemble methods improve the prediction performance.

6. KNN classifiers:

KNN is a simple machine learning method and considered as lazy learning algorithm [14]. The choice of KNN is motivated by its simplicity and flexibility to incorporate different data types. The main idea of KNN is to base estimation on a fixed number of observations, say k, which are closest to the desired output. KNN can be used both in discrete and continuous decision making known as classification and regression respectively. Even though, KNN suffer from the curse of dimensionality and over fitting problems. For classification, they select most frequent neighbour, and for regression, they calculate the average of k neighbour. KNN is a supervised learning algorithm i.e. a training set is given consisting of n pair (xi , yi) and the problem is to estimate y(x) from a new input x. In order to apply this technique, it is necessary to have a training set and a test set, to know the value of k (how many neighbours are used for prediction) and the mathematical formula of the distance calculated between the instances. The three famous distance functions used with KNN are [15] (i) Euclidean distance, (ii) Manhattan distance iii) Minkowski distance. In our research KNN classifier is used because of its non parametric in nature, simple to implement, robust with small error ratio, instance based learning, local learner, has uniform feature weighting and it can work with relatively simple information. We have considered the K value preferred to be selected is odd, to break the ties.

WEKA Machine Learning Tool:

Today there are many machine learning tools available in the market. WEKA is simple to use, and at the same time various data mining algorithm can be applied to the loaded dataset. WEKA has several graphical user interfaces that enable easy access to the various features of it. The main graphical user interface is explorer; it has a panel-based interface, where different panels correspond to different data mining tasks. The first panel is pre-processing, loaded dataset can be pre processed using the various filter options. Second panel provides all types of the classifiers, using which user can test their dataset performance. Third panel is for clusters, enabling the user to run various cluster algorithms. Forth panel is for association algorithms and last one is for visualization. Using visualization option user can view their datasets and the result using various visualization techniques. Other graphical user interfaces of WEKA are experimenter, knowledge flow and simple CLI.

Proposed Solution

In the proposed research, we have considered the most important attributes suggested by PCA, which decides the classification performance, and user can view all such PCA values of their dataset. We have also considered the Meta features, execution time, accuracy and scalability, to decide the best analysis tool for the given dataset. Proposed system suggests the best visualization techniques suitable for user applications considered. We have considered Weka (Waikato environment for knowledge analysis) machine learning tools [16] to find classifiers performance on each inputted dataset. Proposed system has taken both supervised and unsupervised

classifiers into account; it also suggests the appropriate visualization technique for given dataset.

A. System Architecture

We have proposed a system for the automatic data analysis tool and visualization techniques for the given data set to aid non expert end users in selecting appropriate classifier/clustering and visualization model for his task. The empirical framework for the automatic suggestion of data analysis tools & visualization techniques for given large dataset for expert and non-expert end users is depicted in the fig 2. Major parts of the proposed system act as black box. Black box consists of data base; all the training records characteristics (Meta features) are stored into oracle database. Oracle uses B-tree for the record retrieval so it will take less access time and also supports for large datasets. Once the user inputs the test dataset, its characteristics are fetched, and then it finds the best possible solution using the KNN classifier and K-mean clustering approach. The best matching solution from the database will be retrieved and displayed to the user. Output of the black box is shown in the fig 2, which gives the best three suggested classifiers/ clusters or association rules and the best visualization techniques. After user validates the result and if feels, the system result is not that efficient compared to actual result, then, the proposed system can update its database with the user's latest result. In this way, the proposed system will ensure more accurate results in future cases through the incorporation of active learning model. Hence, the system used in the black box becomes self learning because the latest result will be used as a training set, later it will be used for the evaluation of the next test data set. For each dataset it's all characteristics are stored as one record into the data base. So the data base of the black box supports large number of data sets.

The proposed system works in three phases they are training, assessment and reprocess.

1. Training:

Initially the system must contain few training record to evaluate the test record which is inputted by the end user. In this phase, the training records features, like the number of instance, number of attributes, data types of each attributes, class of applications of the dataset & number of the classes, standard deviation, variance and mean of the each data set are extracted and stored into the oracle database. Along with these features, the best three data analysis tools selected by our system, its execution time, performance in percentage, the three best visualization techniques and the application for which this data set is used are all stored in database. Reason to select the oracle database is that it has already built-in file structure to retrieve the record in a minimum and acceptable access time.

2. Assessment:

In this phase, end user input the data set along with its application. Once the data set is inputted, it's all features as mentioned in the training phase are extracted and using KNN classifier, the best nearest solution from the training data set is estimated. After

execution, the best solutions & visualization techniques for the given dataset will be displayed to the user.

3. Reprocess:

Assessment phase result will be evaluated by the end user. Once the assessment is done, the system expects user to input the three best solutions for the given data set, and this will be stored along with all other features of the data set as training data set into the oracle data base. Hence the number of training record increases hence the chances of getting the best solution is more.

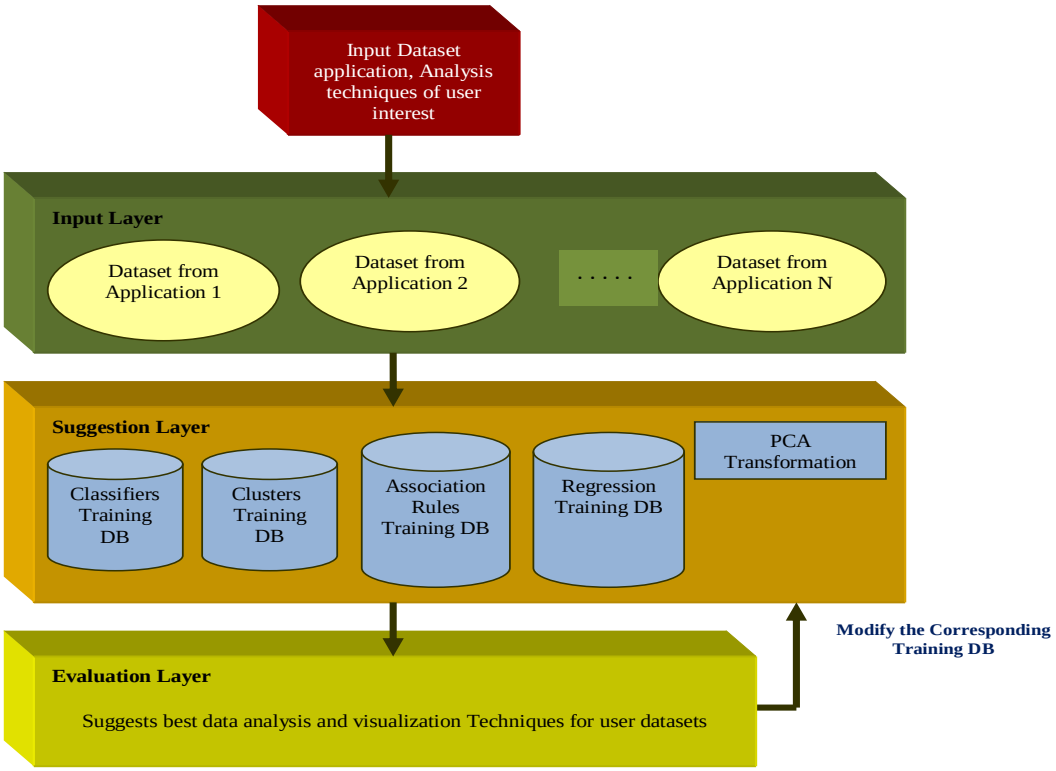


Figure 2: System Architecture

B. Flowchart for the automatic classifiers and visualization Techniques of the given dataset using KNN classifiers & K -mean cluster

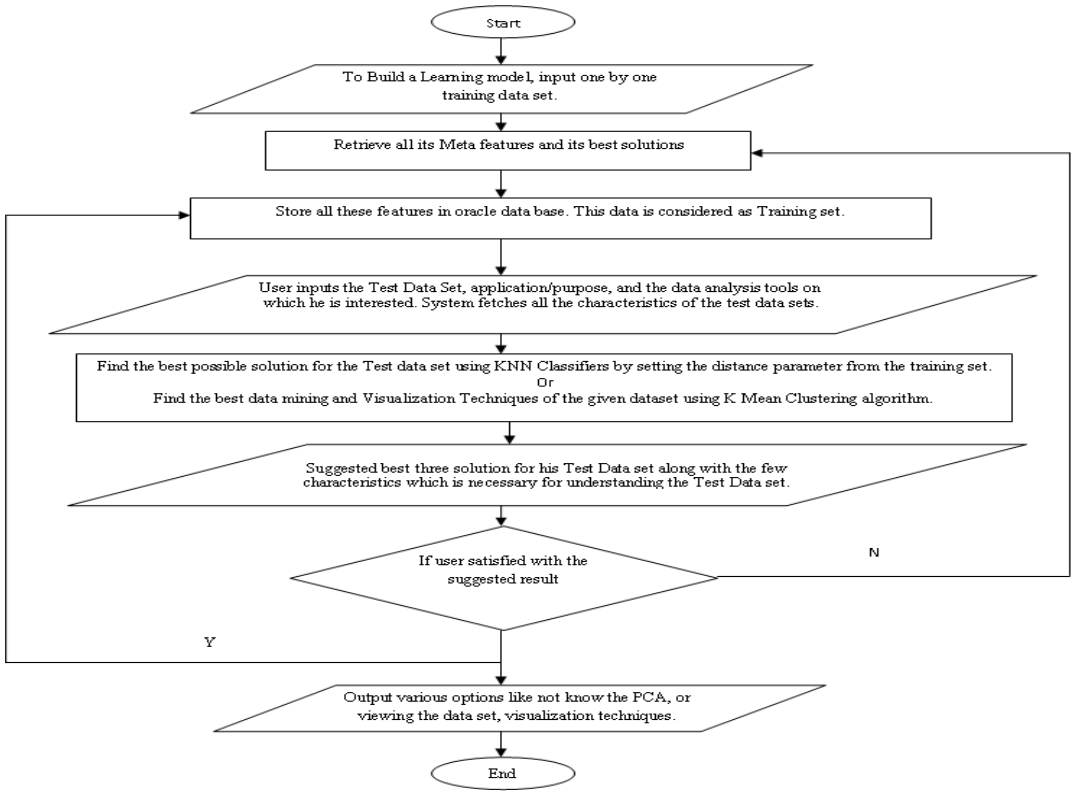


Figure 3: Flow chart of the proposed system

Fig 3, explains working of the proposed system which includes the tested datasets, Meta features (Meta features are selected based on the literature) & execution time of the training data sets. When user inputs the datasets, its Meta features are extracted and, using Euclidean distance of KNN classifiers we estimate the nearest possible Meta features in training datasets, which contains its corresponding data analysis tools and its execution time. This is displayed to the user along with its execution time. Similarly, we also developed k-mean Custer Technique for the same problem in which the result is different clusters which will include inputted data set Meta feature and suggested data sets Meta feature along with suggested solution. We provided the option to change the K value of the K-mean clusters to get the efficient result. For the testing purpose user can validate the suggested result by manually running their dataset on the available data analysis tools. Once the result displayed to the user, his inputted data sets Meta features, best three data analysis tools and, its visualization technique used to represent that data set will be stored as training datasets to our model, which internally helps in evaluation for the subsequent inputted dataset and also makes our training model stronger.

Implementation

Datasets vary depending on the source of their origin and hence come in different category of attributes like continuous, binary, Boolean, discrete, or nominal. To understand datum with the above attributes, we need statistical and mathematical measures which will give the clear idea about the distribution of the datasets. It is a natural phenomena and statically proved that more than 95 % of the data in the continuous attributes are normally distributed. When the data is normally distributed, mean, standard deviations & skewness values are the good measures to find the central tendency of the data and its distributions. For continuous attributes these are the best measures to represent the data characteristics. Nominal and binary attributes are analyzed using the maximum range, minimum range, mean, and standard deviation of the distinct values. We measure the Mean Entropy of the discrete attributes because it is used to measure disorder/impurity in the data set. If the data evenly distributed in each category of the discrete attributes then entropy value will be large. Along with these characteristics, we have also considered total number of attributes of dataset, number of nominal attributes, Boolean attributes, continuous attributes, instances, classes and its minimum and maximum imbalance ratio. These attributes & its measures will give the clear idea about the nature of data in the dataset, and so we used these attributes & measures to find the similarity between the datasets.

Proposed system is implemented using Matlab version R2010a in Intel Atom 455 single core, 1.66 GHz, 1GB RAM Window 7 operating system. The data set is stored in .arff (attribute relation file format) format. Once user executes the system, the system expects user to enter the data set name then, the system asks the user to enter the application for which the data set is used, and his interested data analysis technique. The system reads this input, to find the best possible solution for given data set, using KNN classifiers. Option will be provided for the user to view the features of the inputted dataset at any point of time. So the system is interactive and user can see the detailed view of his data set & its properties. To evaluate the system performance, we have considered Ten Training Data set of varying features and six test dataset used to test the performance of the said system. Once the user inputs a data set, its data characteristics are fetched, and subsequently checks for the similar data characteristics in the database using KNN classifiers/ K-Mean Cluster algorithm, and displays its best three solution to the user by extracting it from the database. The proposed system uses only pre-processed training and test datasets which are free from outliers, and missing values. All the training dataset is executed using the ZeroR, OneR, IBK, Naïve Bayes, J48 and AdaBoost classifiers of Weka machine leaning tools. From the literature it is proved that 10-fold stratified cross validation reduces the variance of the estimate and gives 95% of the time best result in most of the datasets. So, all proposed classifiers performance are evaluated using 10 – fold stratified cross validation. The system output can be verified by the end user, using the WEKA machine learning tools for their given dataset. Detailed description about the training and test datasets used for the evaluation of the proposed system performance is given in the Table 1. Iris data set with varying instances is considered to prove that the system will give the accurate result even if the data set is of varying size.

Table 1: Selected Meta Features of The Training And Test Datasets

Training Dataset

SI. No	Data set	ZeroR		OneR		IBk(k=1)		NB		J48		AdaBoost		Mean	Std dev	No. Attr	No. Inst	Attr.type	No. Classes
		Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)						
1	Weather	64.18	0	42.85	0	78.57	0	64.28	0	64.2	0	57.14	0.01	31.42	3.748	5	14	21122	2
2	Iris	33.33	0	94	0	95.33	0	96	0	96	0.01	95.33	0.07	2.9	0.92	5	150	11112	3
3	weather.nominal	64.28	0	42.85	0	57.14	0	57.14	0	50	0	64.28	0.01	0.685	0.638	5	14	22222	2
4	Sick	93.892	0.01	96.57	0.1	95.64	0.01	92.85	0.1	98.9	0.96	97.35	1	9.108	4.381	30	2800	1.222E+29	2
5	Labor	64.912	0.02	75.43	0	82.45	0	89.47	0	73.68	0.02	87.71	0.04	4.236	1.442	17	57	1.111E+16	2
6	Diabetes	65.104	0.01	72.78	0.03	70.12	0.01	76.3	0.08	73.8	0.2	74.34	0.66	40.02	22.92	9	768	11111111 2	2
7	Anneal	76	0.02	83.63	0.08	99.10	0	86.30	0.05	98.4	0.7	83.63	0.17	149.6		39	896	1.112E+38	6
8	audiology	25	0	46.46	0.02	77.87	0	73.45	0.02	77.8	0.06	46.46	0.02	9.75		70	226	2.22E+69	24
9	Autos	33	0	61.46	0	76.09	0.02	56.09	0	81.9	0.09	44.87	0.02	122	35.42	26	205	1.22E+25	7
10	breasr-cancer	70.279	0	65.73	0.02	72.37	0	71.67	0	75.5	0.03	70.27	0.06	31.33		10	286	2.222E+09	2

Test Dataset

SI.No	Dataset	ZeroR		OneR		IBk(k=1)		NB		J48		AdaBoost		Mean	Std dev	No. Attr	No. Inst	Attr.type	No. Classes
		Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)	Per	E.T(sec.)						
1	Bank	54	0	52.66	0	59.33	0.01	58.33	0	68.667	0.05	62	0.07	0.3078	0.144	9	300	12212222 2	2
2	mushroom-test	51.92	0.01	98.41	0.04	100	0.03	95.01	0.08	100	0.12	96.47	0.664	0.873	0.750	23	6124	2.22E+22	2
3	contact-lenses	62.5	0	70.83	0	79.16	0	70.833	0	83.3	0.43	70.83	0.02	0.7833	0.639	5	24	22222	3
4	Iris(189 Instance)	66.667	0	94	0	95.33	0	96	0	96	0.01	95.33	0.07	2.7777	0.944	5	189	11112	3
5	Iris(326 Instance)	66.667	0	94	0	95.33	0	96	0	96	0.01	95.33	0.07	2.6683	0.968	5	326	11112	3
6	Iris(1504 Instance)	66.667	0	94	0	95.33	0	96	0	96	0.01	95.33	0.07	5.627	0.819	5	1504	11112	3

A. Experimental Result

The system is evaluated for its superiority by inputting the bank dataset, mushroom dataset, devanagari handwritten Image datasets and iris dataset with varying instances. Devanagari hand written dataset contains hand written devanagari script image features, each record in the dataset represent one hand written image devanagari script features. There are around 20208 such records in one dataset with 81 features, and another dataset contains 24056 with 108 features. The class label of these datasets represents nine different devanagari letters or the symbols. These data sets are used to identify the devaganagri script letters or symbols and researcher needs to select the best classifiers for this problem. So we conducted experiment with these dataset with our proposed system and able achieve good results for these datasets. The output of the system with devanagari hand written datasets and other datasets are summarised in

the Table 2, 3, 4, 5, 6 and 7. In the output, the decimal number 1 represent the ZeroR classifier which comes under Rule based classifiers, 2 represents OneR classifiers which will also comes under Rule based classifiers, 3 for IBK which is Lazy classifier, 4 denotes Naïve Bayes classifiers belong to probabilistic model, 5 for J48 – tree based classifier and 6 is AdaBoost which is an ensemble classifiers. For the bank dataset, system suggests 6, 5 and 4 classifiers and when executed with Weka machine learning tool we have achieved maximum performance in 5 and 6 classifiers. For mushroom dataset 4,6 and 3 is suggested by the system and when we validated using Weka it was observed that 3, 2 and 6 gives the best performance. With iris dataset with various instances are tested using the system and system suggests 4,5 and 3 classifiers performance is best, and it shows the iris dataset gives same result with different number of instances, which is a good example to show our system perform well when the new samples are added into the datasets and hence the proposed method support the scalability. Fig. 4, shows our system output versus Weka machine learning tool on various test datasets, and also shows that our system performance is more than 95% accurate in all the cases.

Table 2: Bank Dataset Result

Data set		Bank
Number of attributes		9
Number of instances		300
Output of the system		
Classifiers	%	Exe.Time
6	74.34	0.66
5	73.83	0.22
4	72.79	0.3

Table 3: Mushroom Dataset Result

Data set		mushroom
Number of attributes		23
Number of instances		6124
Output of the system		
Classifiers	%	Exe.Time
4	89.47	0
6	87.72	0.04
3	82.46	0

Table 4: Iris dataset result(Instance:189)

Data set		iris
Number of attributes		5
Number of instances		189
Output of the system		
Classifiers	%	Exe.Time
4	96	0
5	96	0.01
3	95.33	0

Table 5: Iris dataset result(Instance:326)

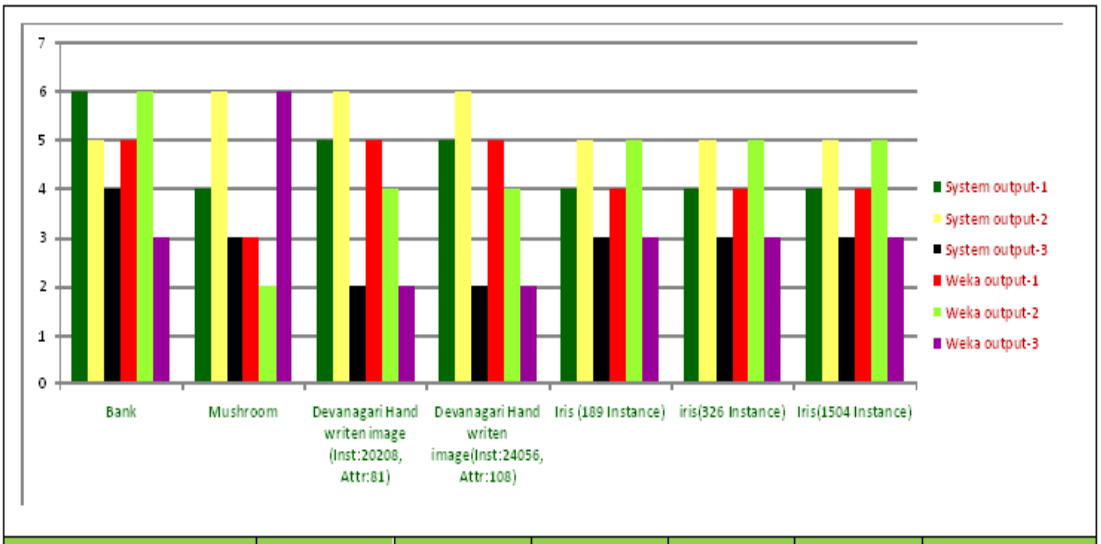
Data set		Iris
Number of attributes		5
Number of instances		326
Output of the system		
Classifiers	%	Exe.Time
4	96	0
5	96	0.01
3	95.33	0

Table 6: Iris dataset result (Instance: 1504)

Data set		Iris
Number of attributes		5
Number of instances		1504
Output of the system		
Classifiers	%	Exe.Time
4	96	0
5	96	0.01
3	95.33	0

Table 7: Devanagari handwritten image result

Data set		Devanagari
Number of attributes		81
Number of instances		20208
Output of the system		
Classifiers	%	Exe.Time
5	91	1
6	97	1
2	93	1



DataSet	System output-1	System output-2	System output-2	Weka output-1	Weka output-2	Weka output-3
Bank	6	5	4	5	6	3
Mushroom	4	6	3	3	2	6
Devanagari Handwritten image-1	5	6	2	5	4	2
Devanagari Handwritten image-2	5	6	2	5	4	2
Iris (189 Instance)	4	5	3	4	5	3
iris(326 Instance)	4	5	3	4	5	3
Iris(1504 Instance)	4	5	3	4	5	3

Figure 4: Comparison of System Output With Weka Output

Conclusion

In the proposed work, we have developed a method for suggesting appropriate data analysis and visualization technique for a given large dataset. One of the most difficult tasks of the whole KDD or data mining process was to choose the right data mining technique for the given data set and, the visualization techniques to represent the results. Though there are many data mining technologies available to classify the given data sets. It was very difficult for the user to decide the best analysis technique for their input data set. We have developed an efficient system to suggest the best

three data analysis tools and the visualization techniques for given datasets. Before suggesting the best data analysis technique for the user dataset, our system takes care of accuracy, scalability and the execution time. User can view the PCA values of the given dataset, which is helpful in understanding which attributes plays major role in the classification. The method has been tested on ten Data sets of the Weka and two devanagri handwritten image datasets. It is compared with the manually selected classifier, association rule, and cluster to establish its superiority. We have experimentally proved the superiority of our system for the user datasets by suggesting best data mining techniques. Proposed system can be utilized for stream data, time series data, social networks, World Wide Web, graph mining and biological sequence data classification.

References

[1] Karina Gibert, Miquel Sànchez-Marrè, Víctor Codina, "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation", *International Environmental Modelling and Software Society (iEMSs) International Congress on Environmental Modelling and Software Modelling for Environment's Sake*, Fifth Biennial Meeting, Ottawa, Canada,2010, pp-1-9.

[2] Silviu Cacoveanu , Camelia Vidrighin , Rodica Potolea, "Evolutional Meta-Learning Framework for Automatic Classifier Selection", *IEEE International Conference on Intelligent Computer Communication and Processing - ICCP*, 2009, Pages 27-30.

[3] Matthias reif, Faisal Shafait, Markus Goldstein, Thomas m. Breuel, Andreas Dengel. "Automatic Classifier Selection for Non-Experts", *Pattern Analysis and Applications*, May 2012, Pages 1-20.

[4] Taciana a. F. Gomes, Ricardo b. C. Prudˆencio, Carlos Soares, Andr´e l. D. Rossi, Andr´e Carvalho, "Combining Meta-Learning and Search Techniques to Select Parameters for Support Vector Machines, Neurocomputing"**,** Volume 75, Issue 1, Elsevier, 2012, Pages 3–13.

[5] Stuart Moran, Yulan Hey, Kecheng Liu, "An Empirical Framework for Automatically Selecting the Best Bayesian Classifiers", *Proceedings of the World Congress on Engineering* 2009, Vol 1WCE 2009, July 1 - 3, 2009, London, U.K. Pages 1-6.

[6] Nikita Bhatt, Amit Thakkar, Amit Ganatra, Nirav Bhatt, "Ranking of Classifiers based on Dataset Characteristics using Active Meta Learning", *International Journal of Computer Applications* (0975 – 8887) Volume 69– No.20, 2013, Pages 31-36.

[7] Pearson K., On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2: (1901), Pages 559-572.

[8] Herve Abdi, Lynne J. Williams, Principal component analysisWIREs Computational Statistics, , Volume 2, July/August 2010, (2010), Page 433-459.

[9] Arnaud Giacometti, Eynollah Khanjari Miyaneh Patrick Marcel, Arnaud Soulet, "A Generic Framework for Rule-Based classification", L.I. University Francois Rabelais de Tours, 41000 Blois, France.,(2008), Pages 37-54.

[10] Jiawei Han and Micheline Kamber, " Data Mining: Concept and Techniques, Morgan Kaufmann, 2nd ed., 2006.

[11] Tan Pang- Ning, "An Introduction to Data Mining". Pearson Education , 2007.

[12] S.Deepajothi1, Dr.S.Selvarajan, "A Comparative Study of Classification Techniques On Adult Data Set", *International Journal of Engineering Research & Technology* (IJERT) Vol. 1 Issue 8, 2012, Pages 1-8.

[13] Lior Rokach , "Ensemble-based Classifiers", *Artif Intell* Rev,(2010) 33:pp-1–39.

[14] Minakshi Sharma, Suresh Kumar Sharma, "Generalized K-Nearest Neighbour Algorithm- A Predicting Tool", *International Journal of Advanced Research in Computer Science and Software Engineering(*IJARCSSE),volume 3, issue 11, 2013, Pages 1-4.

[15] K. Q. Weinberger and l. K. Saul. "Distance metric learning for large margin nearest neighbour classification", *Journal of Machine Learning Research* 10: 2009, Pages 207-244.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard P Fahringer, "The WEKA Data Mining Software: An Update, ACM SIGKDD Explorations SIGKDD Explorations", Volume 11, Issue 1, 2009, Pages 10-18.