University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 2, Issue 2, April 2018
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Studying the Reduction Techniques for Mining Engineering Datasets

Mustafa Ali Abuzaraida

Computer Science Department, Faculty of Information Technology, Misurata University, Libya
abuzaraida@it.misuratau.edu.ly

Abstract: Over the world, companies often have huge datasets as data warehouses collection. The enormous size could make difficulty to analyze the data. The main reason, the complexity of data in terms of number of attributes and number of cases. To overcome this problem could be done by using a sufficient number of attributes and cases before mining this dataset. In Data Mining field, several methods could be used to reduce the attributes number and similar cases. This paper presents a study to test three reduction methods on engineering domain using five datasets. The three methods are: Genetic Algorithm (GA), Principal Component Analysis (PCA), and Johnson technique. The five datasets where obtained from UCI machine learning archive. The study examines which reduction method can be proper for datasets in Engineering field. It can be done by identifying the three reduction methods ranking based on percentage accuracy and number of selected attributes.

Keywords: Data mining, Data reduction, Engineering dataset.

I. INTRODUCTION

Data Mining field is one of the most motivating research field which used in different categories. The objective of applying Data Mining algorithms is to find significant information from enormous data. lately, Data Mining is used to analyze data in many fields like engineering field. There is a necessity of detecting unknown and valuable information by using an efficient analytical methodology in engineering category. In engineering category, Data Mining could provide many benefits like identification of good manufacturing methods, finding the technology solution to the factory at lower cost, detection of causes of losing, and detection of the fall of some products in the markets [1].

However, dealing with huge data in engineering category could make analyzing the data more difficulty. Data could be not simple in terms of attributes number and cases number. To minimize this matter, sufficient attributes number and cases should be used as a good solution[2]. Here, a lot of methods which can be used to reduce data in the field of Data Mining. Holtes' IR, Genetic Algorithm, Principal Component Analysis (PCA), Classification and Regression Tree (CART), and Johnson algorithm are examples of these methods. Most reduction techniques perform differently when applied to various problems [3]. To date there is no research that can identify which reduction technique is the best. This is because one reduction technique may be suitable to be used on one problem domain but unsuitable when applied on another problem domain.

In this paper, three reduction techniques namely PCA, GA, and Johnson, have been tested on five datasets of engineering domain. Used data was chosen because they cover important areas in our society. Discovering knowledge from these data could assist people in these fields to solve many real problems.

PCA, GA, and Johnson reduction techniques were chosen because these techniques were found to be well known techniques in many research areas such as engineering, marketing, and education. PCA can be a useful statistical tool for reducing large number of attributes into less number of attributes to capture important information. It has been used in some researches like as pattern recognition and image processing. PCA can be considered as a regular method to find the best patterns of data of huge dimension [4]. GA is another reduction technique that applies Darwinian principles to obtain an optimized solution. John Holland was introduced this technique at University of Michigan in 1975 [5]. Johnson Algorithm is an algorithm, which was developed by Donald B. Johnson (1977) in The Pennsylvania State University. The algorithm uses the all pairs shortest path method to find relationships [6].

The experiment tests the reduction methods to find the most suitable for engineering datasets. In addition, the research identifies the ranking of the three techniques.

II. PROBLEM STATEMENT

In Data Mining, the use of a large number of data leads to several problems. First, the use of a large number of attributes causes inefficiencies in data analysis in terms of cost, time, and process. More people are required to collect data for the attributes. This adds up operating costs and in turn, lowers profits. In terms of time, more hours or days are needed to gather all the data. Again this causes an increase in operating costs. Lastly, in terms of process, irrelevant attributes may be included in the Data Mining process. This causes complexity in the analysis and may affect the accuracy of results.

Second, in order to obtain good results, data to be processed need to be selected or reduced. This can be done by the use of a proper reduction technique in Data Mining

process. There are many reduction techniques, however, not all techniques are suitable to be used for a particular data set. One technique may be suitable to be used for one type of data and may not be suitable for others. Thus, the identification of the best technique is required.

III. RELATED WORKS

Data Mining techniques have been used by industrial companies and research groups like clustering and classification. They used these techniques to increase the capabilities of making decisions as regards of industry and technology issues. Data Mining could offer new helpful information about engineering category which in turn helpful for making administrative as well as engineering decisions like estimating the companies staff performances, decisions about the policy of industry insurances, selecting of the best products, prediction of costs, and other subjects [7,8]. Many studies have focused on different Data Mining issues and challenges in engineering field [8-10] . Data Mining is used also for analyzing and predicting solution for different industry issues [7,9,11]. A number of studies proposed an improvement in available methodology of Data Mining in order to get better the results. On the other hand, some other researches implement new frameworks and methodologies for companies systems [12].

For data reduction and features selection, several studies have been conducted such as Nemes and Lantos who used the Genetic Algorithm to reduce the data set in a robot manipulator dynamic model. The samples number in the representative training data set were reduced from over 10000 sample into just 90 sample. They obtained accuracies for more than 75% [13].

Godınez et al, compared Johnson Algorithm with GA in terms of speed of training. Results showed that the two algorithms took the same total elapsed time and they record that, Johnson's algorithm showed a better result [15]. Sun et al, used GA to reduce frontal facial images data set. Their result showed that the average Neural Network error of classification decreased from error average rate of 17.7% into 11.3% [14].

Banthia et al, conducted a research on electronic manufacturing to reduce the computation time for accurate binning of integrated chips (ICs). In their study, PCA was used to reduce the dimensionality of the raw data. Their result showed that PCA gave a higher accuracy than hierarchical clustering (HC) [14]. Cohen et al, introduced a new method for reducing the dimensionality of feature sets. They choose the original features subset that contained the most of the essential information. To do that, they use the same criterion of the PCA. The proposed method known as Principal Feature Analysis (PFA) was used for two applications which are content based image retrieval and face tracking. Results from the study demonstrated that PFA did not produce comparable results to PCA [15].

Based on above, it can be seen that GA, PCA, and Johnson Algorithm are popular reduction techniques and produced good results in various researches. However, which technique is the best and for engineering domain are still unknown. Since most reduction techniques are usually suitable for certain kinds of data, there is thus, a need for a study to determine which technique is suitable for which set of data. Thus, the purpose of this study is to provide this information gap.

IV. EXPERIMENTAL PROCEDURE

Rosetta software is used in this experimental study. This software is toolkit which used to analyze datasets base on rough set theory framework. It is developed to support the knowledge discovery and overall Data Mining process. Starting from browsing and data preprocessing, computing the minimal attributes and generating the if-then-rules, and finally validating and analyzing the induced patterns or rules. This software is proposed as a general purpose instrument for discernibility based modeling. Figure.1 shows the main interface of ROSETTA software.

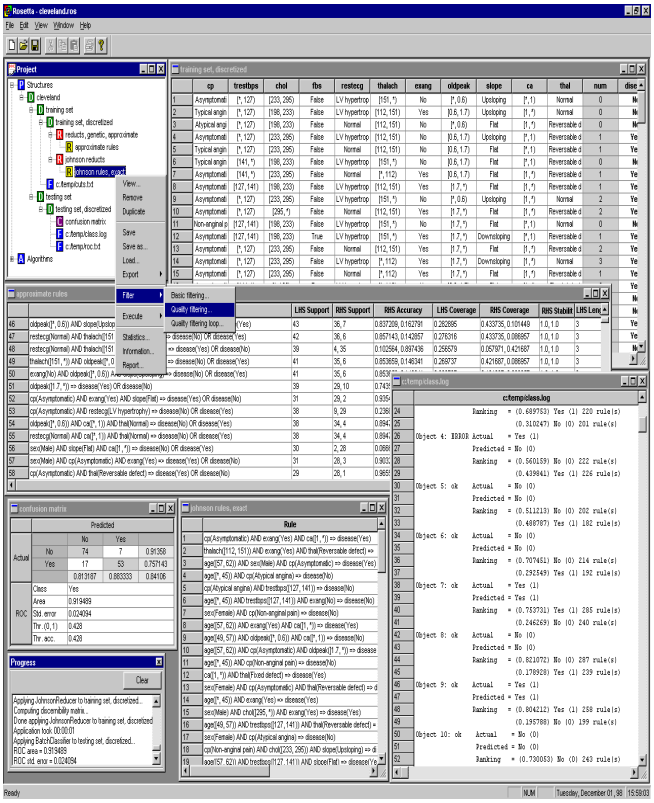

Figure 1. Main Interface of ROSETTA Software

V. METHODOLOGY

This study adopts the knowledge discovery in database (KDD) process as shown in Figure.2 . The KDD process has been applied by several researchers such as [16], and [17].

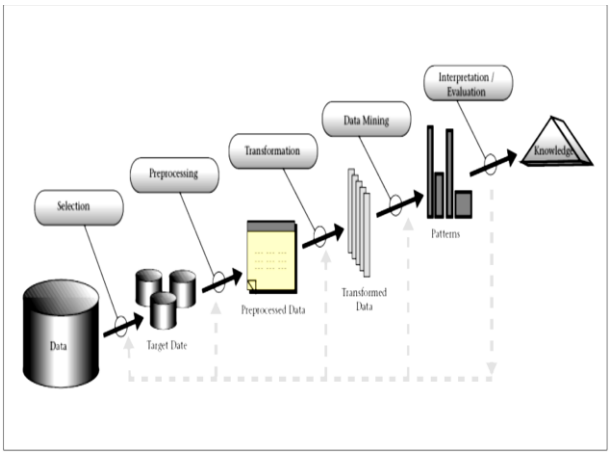

Figure 2. KDD Process [17]

According to Fayyad et al, [19] "Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process". [19] the knowledge discovery process are structured in various stages which are listed as following:

A. .Data Collection

Five different engineering datasets are used in the study. All the datasets were taken from UCI datasets Machine Learning Archive. The datasets are Auto -Mpg, Image Segmentation, Ionosphere, Letter Image Recognition, and CPU Performance datasets. All the datasets have an average of 400 instances.

B. Data Preprocessing

The preprocessing phase, two steps are performed which are transformation of data and handling noisy data and the missing values.

Data Transformation: Some of the attributes in these datasets come in character format. The attribute values are in non -numeric form. Thus, these values have to be converted to numeric form before performing Data Mining task.

Handling missing values and noisy data: Incomplete and noisy data are common problem in real database and warehouses. Missing data can happen due to many reasons. Important attributes might not often be existing, like patient age and could happen to other reasons. Data noisy could occur with incorrect attribute values. Several problems can cause perplexity of the procedure of the mining. For example, there might be computer errors or human mistakes occurring at data collection or entry.

In this study some techniques were considered to handle these matters. These techniques were used depends on the type of the matter. Here, SPSS software was used to handle missing values. There are several techniques that can be used. For example, series mean, nearby points mean and median, and linear interpolation or linear trend at point. For the presented study, nearby points mean was used to handle missing values. No noisy data was found in all used datasets. After this step, all datasets are in numeric format, complete and have no noisy data.

C. Data Discretization

This step is necessary for any Data Mining tasks to be performed. It is performed for reducing the values number for continuous attributes. This was done by separating the attribute range into intervals ranges. The labels of intervals could then be used for replacing actual values of data.

Some Data Mining algorithms only accept categorical attributes and cannot handle a range of continuous attribute value. For example, there is an attribute whose values are from 1 to 10. Interval 0 will take the range from 1 to 3. Interval 1 will take the range from 4 to 6. Interval 3 will take the range from 7 to 10. In this work, discretization step has been done using ROSETTA software. There are several techniques to discretize the data in ROSETTA software. Not all the techniques are suitable to discretize the chosen datasets. However, for this work Equal Frequency Binning has been used to discretize the datasets. From pervious experiment, "Equal Frequency Binning" was found to give the best results when it was used for discretization.

D. Data Reduction

This step aims to find the best less number of features for representing the data. Finding the core attributes is reached of using a dimensionality lessening technique to reduce the attributes number. PCA, GA, and Johnson Algorithm have been used in this study. The reduction process was done using ROSETTA and SPSS software.

E. Data Mining

This step includes choosing the appropriate Data Mining task. It is performed by deciding the purpose of study. Examples of Data Mining tasks are summarization, classification, regression, and clustering. In our study, classification is used to analyze the reduced data. The results of performing Data Mining classification are measured using percentage of accuracy and the percentage of the select attribute number.

Several classification techniques can be used to test the reduced datasets. These are Naive Bayes (NB), Standard Voting (SV), Voting with Object Tracking (VT), and Standard Tuned Voting (RSES). A small experiment was conducted to choose which technique is suitable for this study. Split Factor of 0.1 to 0.5 and Random seed from 100 to 1000 were tested. As a result, the best classification method was Standard Tuned voting (RSES) and the best split factor was 0.1. Thus, these methods are used to conduct tests for the reduced datasets obtained before.

VI. EXPERIMENTAL RESULTS

Five engineering datasets which have been reduced were tested. The measurements used in the tests were percentage of accuracy and percentage of selected attributes. The two experiments were as following:

A. *Accuracy Rate*

After conducting the first experiment, the results are shown summarized in Table I.

TABLE I. Accuracy Averages of Reduction Techniques

Dataset	PCA	GA	JA
Auto -Mpg	64.77%	42.28%	38.23%
Image Segmentation	65.22%	71.96%	45.05%
Ionosphere	54.01%	80.80%	50.38%
Letter Image Recognition	40.67%	49.61%	14.87%
CPU Performance	71.83%	58.87%	43.01%
Average	59.30%	60.70%	38.31%

From Table I, the results show that GA has the highest rate of accuracy with an average of 60%. The second algorithm is PCA with an average of 59%. The last algorithm is Johnson with an average of 38%. Figure.3 illustrates the result in graphical form.

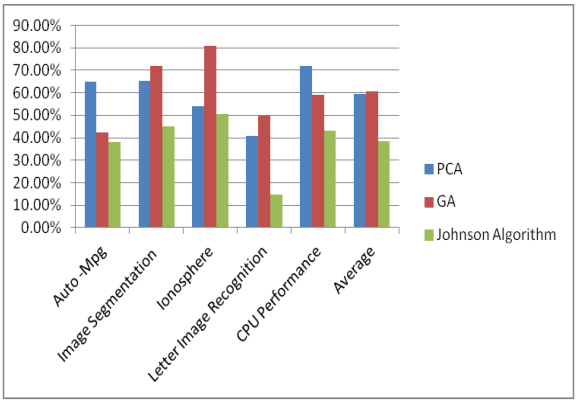

Figure 3. Accuracies Average of Engineering Datasets

B. *Selected Attributes*

The percentage of selected attributes was taken by dividing the number of selected attributes on the number of attributes in the reduced datasets. Table II shows the average percentages of attributes selected from Engineering datasets using the three reduction techniques.

TABLE II. Percentage of Selected Engineering Attributes Datasets

Dataset	PCA	GA	JA
Auto -Mpg	75%	67%	66%
Image Segmentation	40%	24%	17%
Ionosphere	14%	12%	9%
Letter Image Recognition	29%	34%	29%
CPU Performance	33%	48%	39%
Average	38%	37%	32%

From Table II, results show that Johnson selected the least number of attributes with an average of 32%. The second algorithm is GA with an average of 37%. The last algorithm is PCA with an average of 38%. Figure.4 illustrates the result in graphical form.

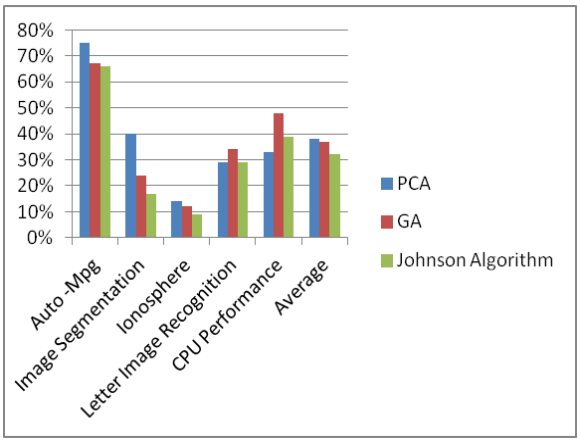

Figure 4. Percentage of Selected Engineering Attributes Datasets

VII. CONCLUSION

In this paper, three data reduction techniques known as PCA, GA and Johnson have been tested on engineering category. Five datasets were used for this study. The experiments used ROSETTA software to train and test the datasets. For this study Standard Tuned voting (RSES) has been chosen the suitable method for classification. Split factor of 0.1 has been chosen to partition the datasets for training and testing. The experiments conducted have produced significant results for all the study objectives. The results have shown that GA gave the highest rate of accuracy with an average of 60%. The second algorithm was PCA with an average of 59%. The last algorithm was Johnson with an average of 38%.

The experimental results to select the attributes that contributes to the study, the results have shown that Johnson algorithm selected the least number of the attributes with an average of 29%. The second algorithm was PCA with an average of 37%. The last algorithm was GA with an average of 38%.

From above, Johnson Algorithm gave the worst accuracy rate, but it had the least number of the attributes. It can be seen that for this part of the study, results produced were that satisfactory. This may be due to data not properly cleaned or discretization method used was not suitable.

The results, however, did not produce high accuracies due to several reasons. If more reduction techniques were used, better results can be obtained. Datasets were limited to the ones in UCI Machine Learning Archive. If more datasets were tested, there is a high chance of getting better results.

REFERENCES

[1] Jia Li, Yimin Zhang, Dongyun Du and Zhengyu Liu, "Improvements in the decision making for Cleaner Production by data mining: Case study of vanadium extraction industry using weak acid leaching process," *Journal of Cleaner Production*, vol. 143, pp. 582-597, 2017.

[2] Mustafa Ali Abuzaraida and Amel Faraj Elramalli, "Identifying the Suitable Reduction Technique for Mining Medical Data," In Proceeding of the The 8th International Conference on Information Technology (ICIT 2017), Amman, Jordan, 2017.

[3] Jiawei Han, Jian Pei and Micheline Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.

[4] Lindsay I Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, p. 52, 2002.

[5] John H Holland, "Genetic algorithms," *Scientific american*, vol. 267, pp. 66-72, 1992.

[6] Donald B Johnson, "Efficient algorithms for shortest paths in sparse networks," *Journal of the ACM (JACM)*, vol. 24, pp. 1-13, 1977.

[7] Wilko Henecka and Matthew Roughan, "Privacy-Preserving Fraud Detection Across Multiple Phone Record Databases," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, pp. 640-651, 2015.

[8] Elusade O Moses and Osuolale A Festus, "Multidimensional Analysis and Mining of Call Detail Records Using Pattern Cube Algorithm," *Computer Engineering & Information Technology*, vol. 2017, 2017.

[9] Yi Lou, Juqin Shen and Shiye Yuan, "The development and application of hydraulic engineering migration risk early warning system based on data mining," In Proceeding of the IEEE International Conference on Computer Communication and the Internet (ICCCI), 2016, pp. 346-349, 2016.

[10] PENG Chen, ZHAO Rong-Cai, Shan ZHENG, XUN Jia and YAN Li-Jing, "Android Malware of Static Analysis Technology Based on Data Mining," *DEStech Transactions on Computer Science and Engineering*, 2016.

[11] Hao Wang and Jinhai Sun, "Quantitative Analysis of Data Mining Application and Sports Industry Financing Mechanism based on Cloud Computing," *International Journal of Grid and Distributed Computing*, vol. 9, pp. 233-244, 2016.

[12] Leandro L Minku, Emilia Mendes and Burak Turhan, "Data mining for software engineering and humans in the loop," *Progress in Artificial Intelligence*, vol. 5, pp. 307-314, 2016.

[13] Attila Nemes and Bela Lantos, "Training data reduction for optimisation of fuzzy logic systems for dynamic modeling of robot manipulators by genetic algorithms," In Proceeding of the Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference, 2001., pp. 1418-1423, 2001.

[14] Ashish S Banthia, Anura P Jayasumana and Yashwant K Malaiya, "Data size reduction for clustering-based binning of ICs using principal component analysis (PCA)," In Proceeding of the IEEE International Workshop on Current and Defect Based Testing, 2005., pp. 24-30, 2005.

[15] Ira Cohen, Qi Tian, Xiang Sean Zhou and Thomas S Huang, "Feature selection using principal feature analysis," *Univ. of Illinois at Urbana-Champaign*, 2002.

[16] Pang-Ning Tan, *Introduction to data mining*: Pearson Education India, 2006.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27-34, 1996.