International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 9 – No.6, September 2015 –www.ijais.org

Enhanced Classification via Clustering Techniques using Decision Tree for Feature Selection

Balogun Abdullateef O.
Department of Computer Science
University of Ilorin, Ilorin

Mabayoje Modinat A.
Department of Computer Science
University of Ilorin, Ilorin

Salihu Shakirat
Department of Computer Science
University of Ilorin, Ilorin

Arinze Salvation A.
Department of Computer Science
University of Ilorin, Ilorin

ABSTRACT
Information overload has raggedly increased as a result of the advances in the aspect of storage capabilities and data collection in previous years. The growth seen in the number of observation has partly cause a collapse in analytical method but the increases in the number of variable associated with each observation has grossly collapse it. The number of variables that are measured on each observation.is referred to as the dimension of the data, and a major problem of dataset containing high dimensions is that, there exist only few "important" measured variables for understanding the fundamental occurrences of interest. Hence, dimension reduction of the original data prior to any modeling of the data is of great necessity today. In this paper, a précis of K-Means, Expectation Maximization and J48 decision tree classifier is presented with a framework on the performance measurement of base classifiers with and without feature reduction. A performance evaluation was carried out based on F-Measure, Precision, Recall, True Positive Rate, False Positive Rate, ROC Area and Time taken to build model. The experiment revealed that the reduced dataset yielded improved results than the full dataset after performing classification via clustering.

General Terms
Data mining, Intrusion detection, features reduction, and classification algorithms.

Keywords
K-Means (KM), Expectation Maximization (EM), Decision Tree, Feature Selection, Data mining

1. INTRODUCTION
In data mining, learning techniques are generally categorized in terms of unsupervised and supervised. In supervised learning methods, the training instances is characterized by pairs of input and output patterns, in contrast, the unsupervised learning networks consists of the training instances only. Classification methods basically uses a training set where all objects are already assigned to a known class labels. Classification algorithm learns and builds a model from a training dataset and uses the model to classify new objects [1]. A term commonly used in data mining to describe the techniques and tools existing for reducing the inputs of a dataset to a manageable size before processing and analysis takes place is referred to as "Feature Selection". Application of feature selection on a dataset is critical for effective analysis, because most time, datasets contain a lot of information than required to build the model [2]. Selecting the right set of features is one of the most important problems encountered in unsupervised classification because very often we do not know what the relevant features are because irrelevant features may reduce the overall mining performance.

Clustering which can be viewed as unsupervised classification [3]. It is an important aspect of data mining whose aim is to separate the training dataset based on similar characteristics [4]. It tends to group the training data based on the information found in the data describing it. Its purpose is that the objects in a certain group are of the same nature and are unrelated to the other objects in the different groups. It has been pointed out that the more the similarity within objects in a group, and the more dissimilar different groups are, the more distinct each group [5]. K-Means and Expectation Maximization are top techniques for clustering data [6]. They have been used separately and successfully for clustering data. With lower-dimensional dataset, mining techniques always perform better because higher-dimensional dataset contain much irrelevant or redundant attributes that often impair its performance. Data preprocessing in the field of data mining is fundamentally and extensively based feature Selection techniques [7]. In this research, irrelevant attributes, which could hinder these algorithms in performing optimally, are removed. Then, a performance evaluation is performed on the algorithms with and without feature selection, based on the following criteria F-Measure, Precision, Recall, True Positive Rate, False Positive Rate, ROC Area and the time taken to build model.

2. RELATED WORKS
[8] applied three standard feature selection methods, which are Gain Ratio (GR), Information Gain (IG), and Correlation-based Feature Selection (CFS), albeit they proposed a method. Comparison of feature reduction methods was carried out by computing the decision tree classifier's results to reveal that the proposed model performed efficiently for network intrusion detection. Their experiment pointed out that their method has lower false alarm rate and higher detection rate than that of full dataset and also performed as good as other methods.

[9] Carried out a research on EM and K-means using random projection and principal components analysis (PCA) to reduce the dimensionality for high dimensional data. They made an observation that PCA was only marginally better, if at all, than a random projection despite its computational intensity.

Another work done by [10] presents a hybrid model using Expectation Maximization, K-Nearest Neighbor, and Genetic

Algorithms. This work removes data that are difficult to learn in order to achieve a successful result.

This research focused on J48 decision tree as a feature selection method in the enhancement of the EM and K-Means algorithms. The new contribution of this work is the exclusion of the most informative features for improving the classification accuracy of these clustering algorithms, derived from a modern process of data acquiring.

2.1 K-Means Clustering

A popular partitioning method is K-means algorithm. It classifies objects by their membership to one of the k groups, k chosen a priori. For determining a cluster membership, the centroid for each group is being calculated and each object to the group is assigned with the closest centroid. The method discussed, iteratively reallocate the cluster members thereby decrease the overall within-cluster dispersion. For example, a dataset of m data points a_1, a_2… a_n such that each data point is in S_d, finding the minimum variance clustering of the dataset into k clusters which is the problem, is solved by finding k points $\{w_j\}$ (j=1, 2, …, k) in R_d such that is minimized, where $g(a_i, w_j)$ denotes the Euclidean distance between x_i and m_j.

$$\frac{1}{m}\sum_{i=1}^{m}[min_j\, g^2(a_i, w_j)] \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The cluster centroids are the points $\{m_j\}$ (j=1, 2… k). Finding k cluster centroids is the problem in Eq.(1). Implementation of the approximate solution to Eq.(1) is what k-means algorithm provides easily. Until convergence, the algorithm iterates between two phases. The first phase is where the assigning of each data point to its closest centroid take place, resulting to data partitioning, while the second phase involve the relocation of "mean" i.e. the moving of each cluster representative to the center (mean) of all data points that as being assigned to it. Convergence to the local minimum is what k-means algorithm does. The local minimum is always dependent of the starting cluster centroids.

2.2 Expectation Maximization

EM (an iterative algorithm) is a model based method for solving clustering problems. It is applied to problems where data is considered incomplete or contains latent variables. The basis for the concept of the EM algorithm is the Gaussian mixture model (GMM) whose method involve enhancing the density of a given set of sample data modeled as a function of the probability density of a single density estimation method with multiple Gaussian probability density function to model the distribution of the data.

Expectation Maximization algorithm function as a distance based algorithm. It assumes that the dataset can be modeled as linear combination of multivariate normal distributions. "Log like hood" - the distribution parameters that maximize a model quality measure, is what EM finds. The inputs to this algorithm are the data set (x), the accepted error to converge (e), the maximum number of iterations, and the total number of clusters (M). The algorithm can be subdivided into two stages, namely the initialization stage and the iterative stage consisting of two steps, maximization step (M-step) and expectation step (E-step) executed iteratively until some form of convergence is reached. The probability of each point belonging to each cluster is estimated by the E-Step, after which the re-estimation of the parameter vector of the probability distribution of each class is done by M-step. The algorithm ends at the convergences of the distribution parameters or when the maximum number of iterations is reached. The Expectation-Maximization (EM) algorithm is an optimization procedure which calculates the Maximum-Likelihood (ML) estimate of the unknown parameter $\theta \in \Theta$ when only incomplete (y is unknown) data T_x are presented. In other words, the EM algorithm maximizes the likelihood function.

$$\ell(\theta \mid T_x) \prod_{i=1}^{l} \quad \prod_{i=1}^{l} \sum_{y \in Y} \quad = P(T_x \mid \theta) =$$

$$P(x_i \mid \theta) = P(x_i \mid y_i\, \theta)\; P(y_i \mid \theta)$$

With respect to the parameter $\theta \in \Theta$ [11].

2.3 J48 Decision Tree

An often used data mining's classification technique is the "Decision Tree" – a predictive modeling technique. Given a predefined dataset, this classification algorithm inductively learned to construct a model. Decision tree classification technique may be seen as mapping from a set of features to a particular class where each data item is defined by values of the features and every non-terminal node in a decision tree signifies a decision or test on the considered data item and the choice of the branch depends on the outcome of the test. Classification of data items begins at the parent node, following the assertions down until a terminal node or leaf is reached. At every terminal node along the path, a decision is always made [12].Classification of a given data item by a decision tree is usually done using the values of the attributes for picking the best attribute that divides the data items into their class, thereby partitioning the data items. Deciding the attribute with which partitioning of the data into various classes can be done is one of the main problem faced. Classifying an object that is unknown begins at the root node of the tree and following the branch specified by the result of each condition until a leaf node is reached which holds the class name after classification is accomplished. Decision trees are able to process both numerical and categorical data. This algorithm is also known to be unstable and trees created from numerical datasets can be complex [12].

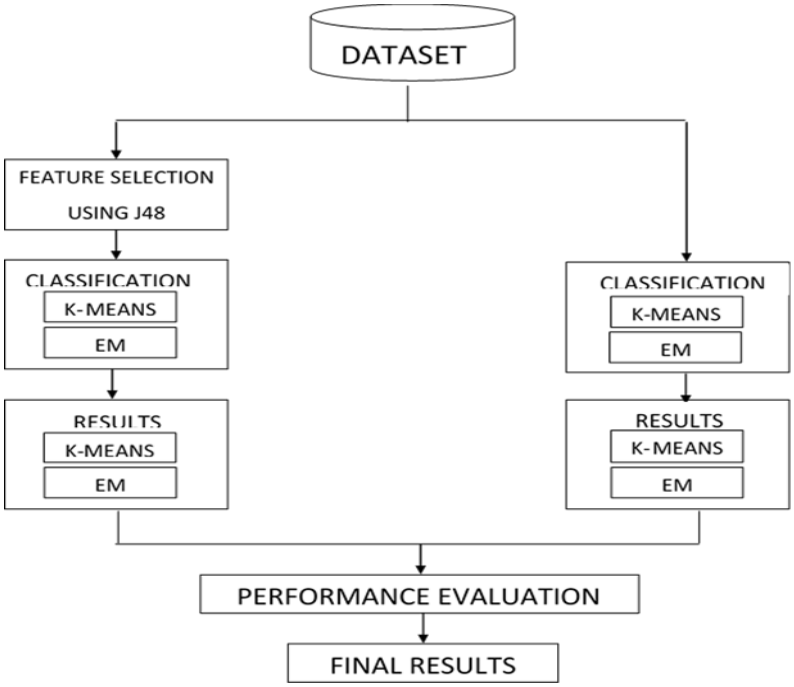

Fig 1: Developed System Framework

3. METHODOLOGY

The experiments was carried out on Acer Aspire E1-510 running on 64-bit Windows 8 Professional operating system, with 4 GB of RAM and a Pentium (R) Quad-core CPU N3520 at 2.16Hz per core using the WEKA tool. For the purpose of this research the datasets that will be used are the R2L dataset in the KDD'99 dataset and the vote dataset from the Congressional Quarterly Almanac (CQA). The Best First Search classifier is used to estimate the merits of the attributes. The attributes with higher merit value are considered as potential attributes and used for classification. It searches the space of attribute subsets by augmenting with a backtracking facility. The BFS is used for the attribute evaluator. The dataset is then processed by the J48 algorithm for the feature selection. After which the resulting data set is classified by K-means and Expectation Maximization algorithm.

4. RESULT AND DISCUSSION

The tables 1 - 8 and figures 2 - 5 shows the performance of EM and KM on the full and reduced form of the datasets mentioned earlier.

ORIGINAL R2L DATASET

Table 1: Performance Evaluation of KM and EM on the *original* dataset

PARAMETERS	K-MEANS	EM
CORRECTLY CLASSIFIED INSTANCES (%)	69.4494	69.3606
INCORRECTLY CLASSIFIED (%)	30.5506	22.2913
UNCLASSIFIED INSTANCES (%)	0.0000	8.3481
KAPPA STATISTICS	-0.004	-0.0255
MEAN ABSOLUTE ERROR	0.0266	0.0211
ROOT MEAN SQUARED ERROR	0.163	0.1454
RELATIVE ABSOLUTE ERROR(%)	157.4921	132.3471
ROOT RELATIVE SQUARED ERROR (%)	185.7582	167.1007

Table 2: Performance Measurement of KM and EM on the *Original* Dataset

PARAMETERS	K-MEANS	EM
TP RATE	0.694	0.757
FP RATE	0.779	0.848
PRECISION	0.857	0.848
RECALL	0.694	0.757
F-MEASURE	0.756	0.782
ROC AREA	0.458	0.424
TIME TAKEN TO BUILD MODEL (secs)	0.33	21.74

REDUCED R2L DATASET

Table 3: Performance Evaluation of KM and EM on the *reduced* dataset

PARAMETERS	K-MEANS	EM
CORRECTLY CLASSIFIED INSTANCES (%)	69.5382	61.9893
INCORRECTLY CLASSIFIED (%)	30.1066	26.4654
UNCLASSIFIED INSTANCES (%)	0.3552	11.5453
KAPPA STATISTICS	0.1728	-0.0631
MEAN ABSOLUTE ERROR	0.0263	0.026
ROOT MEAN SQUARED ERROR	0.1621	0.1613
RELATIVE ABSOLUTE ERROR(%)	156.0791	166.1212
ROOT RELATIVE SQUARED ERROR (%)	184.7606	185.6057

Table 4: Performance measurement of KM and EM on the *reduced* dataset

PARAMETERS	K-MEANS	EM
TP RATE	0.698	0.701
FP RATE	0.295	0.862
PRECISION	0.873	0.781
RECALL	0.698	0.701
F-MEASURE	0.757	0.739
ROC AREA	0.7	0.379
TIME TAKEN TO BUILD MODEL (secs)	0.02	1.31

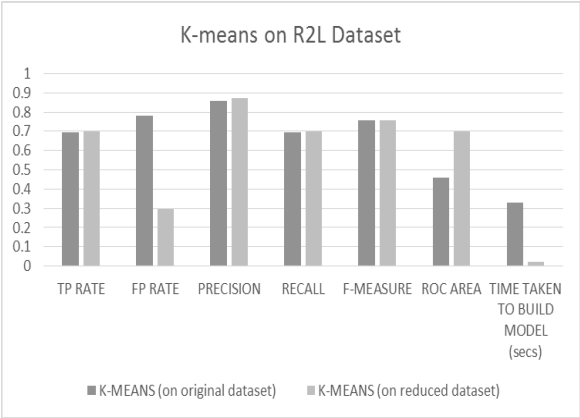

Fig 2: Performance measurement of K-means on R2L dataset

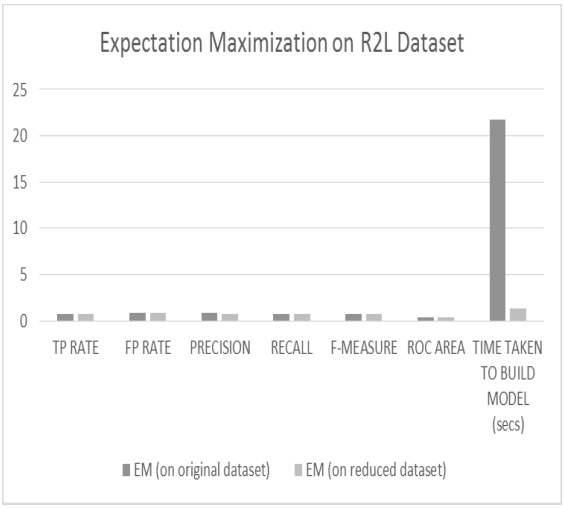

Fig 3: Performance measurement of EM on R2L dataset

Prior to feature selection, it was discovered that K-means performed slightly better than EM on this dataset though few of the instances were not classified by the EM algorithm. It was observed that EM took more time in building its classification model. Their Kappa Statistic was less than 0 which means that they performed less well than chance. Likewise, after feature selection, K-means recorded a slight improvement after feature selection and this was not same for EM. K-means also outperformed EM on this dataset though both recorded few unclassified instances. It was also observed there was significant improvement in time taken to build model for both. Moreover, the Kappa Statistic of K-means was more than 0.

ORIGINAL VOTE DATASET

Table 5: Performance Evaluation of KM and EM on the *original* dataset

PARAMETERS	K-MEANS	EM
CORRECTLY CLASSIFIED INSTANCES (%)	85.0575	60.4598
INCORRECTLY CLASSIFIED (%)	14.9425	2.7586
UNCLASSIFIED INSTANCES (%)	0.0000	36.7816
KAPPA STATISTICS	0.6978	0.9128
MEAN ABSOLUTE ERROR	0.1494	0.0436
ROOT MEAN SQUARED ERROR	0.3866	0.2089
RELATIVE ABSOLUTE ERROR(%)	31.5078	14.0713
ROOT RELATIVE SQUARED ERROR(%)	79.3921	52.1959

Table 6: Performance measurement of KM and EM on the *original* dataset

PARAMETERS	K-MEANS	EM
TP RATE	0.851	0.956
FP RATE	0.123	0.037
PRECISION	0.868	0.96
RECALL	0.851	0.956
F-MEASURE	0.852	0.956
ROC AREA	0.864	0.801
TIME TAKEN TO BUILD MODEL (secs)	0.02	27.94

REDUCED VOTE DATASET

Table 7: Performance Evaluation of KM and EM on the *reduced* dataset

PARAMETERS	K-MEANS	EM
CORRECTLY CLASSIFIED INSTANCES (%)	86.2069	72.4138
INCORRECTLY CLASSIFIED (%)	13.7931	3.908
UNCLASSIFIED INSTANCES (%)	0.0000	23.6782
KAPPA STATISTICS	0.7147	0.8963
MEAN ABSOLUTE ERROR	0.1379	0.0512
ROOT MEAN SQUARED ERROR	0.3714	0.2263
RELATIVE ABSOLUTE ERROR (%)	29.0842	13.7929
ROOT RELATIVE SQUARED ERROR (%)	76.2774	51.885

Table 8: Performance measurement of KM and EM on the *reduced* dataset

PARAMETERS	K-MEANS	EM
TP RATE	0.862	0.949
FP RATE	0.133	0.051
PRECISION	0.867	0.949
RECALL	0.862	0.949
F-MEASURE	0.863	0.949
ROC AREA	0.864	0.847
TIME TAKEN TO BUILD MODEL (secs)	0.00	2.39

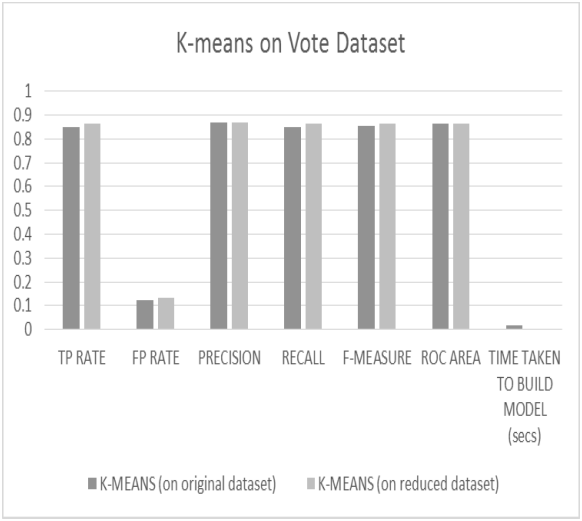

Fig 4: Performance measurement of K-means on vote dataset

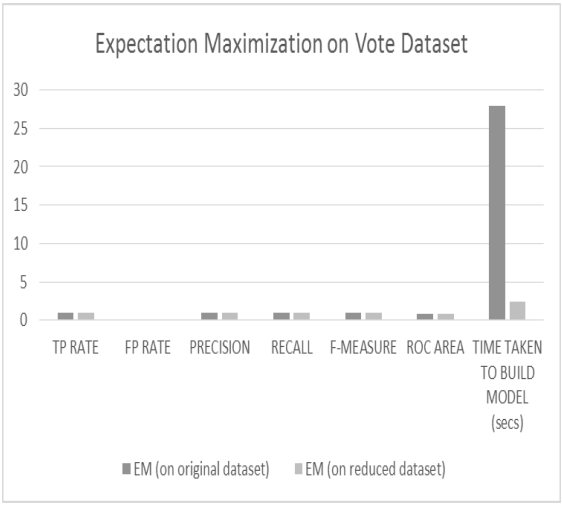

Fig 5: Performance measurement of EM on vote dataset

Prior to feature selection, it was observed that K-means classified more instances than EM though some of the instances were not classified by EM but in performance measurement, EM out perform K-means. Also, it was observed that EM took more time in building its classification model. Likewise, after feature selection, it was observed that K-means performed better though same cannot be said for EM, it was observed that there was significant improvement in time taken in building its classification model.

5. CONCLUSION

From the analysis, the experimental result revealed that feature selection improved the performance of both K-means and EM algorithm in both datasets, though EM didn't perform well on the reduced datasets. Perhaps the algorithm may have considered the selected features too little but there was significant improvement in the time taken to build the classification model in both algorithms. This experiment recorded an improvement on both algorithms. Therefore considering the general outcome of the experiment, it can be said that the optimization performed on the dataset caused an improvement on certain aspects of the algorithms such as the time taken to build the classification model, hence such data pre-processing should be carried before data analysis.

From this research, it was discovered that data pre-processing such as feature selection caused an improvement in base classifiers but researchers should exercise discretion in the choice of attribute evaluators to use in feature pre-selection before data analysis is carried out. Also, the search method used should be considered carefully because different search method produces different results on the dataset. Perhaps, EM could have performed better if a different search method that chose 10 of all ranked attributes was used. Further study should be carried out on these algorithms, applying different feature reduction techniques with varied search methods and ensemble methods in terms of the classifiers used.

6. REFERENCES

[1] Patil, T.R. and Sherekar, S.S (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Computer Science and Applications. Vol. 6, No.2.

[2] Preeti K. and Rajeswari K. (2014). Selection of Significant Features using Decision Tree Classifiers. International Journal for Engineering Research and Applications (IJERA).

[3] Osama, A.A. (2008). Comparisons between Data Clustering Algorithms. The International Arab Journal of Information Technology, Vol. 5, No. 3.

[4] Yong, G.J., Min S.K. and Jun H. (2014). Clustering Performance Comparison using K-means and Expectation Maximization Algorithms. Biotechnology & Biotechnological Equipment, 28:sup1, S44-S48.

[5] Ian H.W., Eibe F. and Mark A.H. (2011). Data Minig: Practical Machine Learning Tools and Techniques (3rd edition). Morgan Kaufmann Publishers Inc., Inc., San Francisco, CA, USA.

[6] Bezdek, J.C. (1980). A Convergence Theorem for the Fuzzy C-means Clustering Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] Mahdi, E. and Fazekas, G. (2011). Feature Selection as an Improving Step for Decision Tree Construction. 2009 International Conference on Machine Learning and Computing IPCSIT Vol. 3. p. 35.

[8] Sang-Hyun C. and Hee-Su C. (2014). Feature Selection using Attribute Ratio in NSL-KDD data. International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014), Feb 4-5, 2014 Bali (Indonesia).

[9] Neil, A., Andrew, S. and Doug, T (n.d). Clustering with EM and K-Means.

[10] Mehmet, A., I. Cigdem and A. Mutlu. (2010). "A hybrid classification method of K Nearest Neighbor, Bayesian Methods and Genetic Algorithm," Expert Systems with Applications vol. 37, p. 5061–7.

[11] Namita B., Deepti M., (2013). Comparative Study of EM and K-Means Clustering Techniques in WEKA interface. International Journal of Advanced Technology & Engineering Research (IJATER) Volume 3, Issue 4, Pp 40.

[12] Kesavalu, E., Reddy, V.N. and Rajulu, P.G. (2011). A Study of Intrusion Detection in Data Mining. Proceedings of the World Congress on Engineering 2011 Vol IIIWCE 2011, July 6-8, 2011, London, UK.