International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9555-9560

MPSKMean Model to form Projected Attribute based Clusters to Extract Heuristic Pattern in Seasonal Data

S.Gokila[1], Dr. K. Ananda Kumar[2] and Dr. A. Bharathi[3]

[1]Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India.
[1]Orchid Id: 0000-0003-1680-8795
[2]Department of Computer Science and Engineering, Bannariamman Institute of Technology, Erode, India.
[3]Department of IT, Bannariamman Institute of Technology, Erode, India.

Abstract

The data analysis of some domain handles unequal weight of attributes that expresses importance of attribute varies. Another requirement is extracting sensitive pattern of data. These two are achieved by assigning unequal importance to the attribute and by slicing the data set into relevant group. The same is implemented in MPSKMeans (MPSKM) algorithm. The MPSKM suggested in this paper incorporate work on complete set of attributes to form set clusters on unequal attribute for each of the grouped data set. The result of proposed work notifies the extreme events patterns as a separate clusters and the same as an outline. The MPSKM works well on unequal complete set of attributes to form a clusters of different pattern by identifying rank index and forming dynamic clusters. The same rank index support to find the impact of attribute in data slice and also to predict the attribute in next slice of data. The predictive result of MPSKM clusters gives the better recall and precision. The domain of data taken for study are 11 years Weather data set which needs the season wise study for sensitive changes in weather.

Keywords: Data Mining, Time Series, Cluster, Projected Attribute,

INTRODUCTION

The dataset of some domain consists more sensitive set of values. That sensitivity extraction is possible when dataset are sliced occurring to the relevance. Another key note to be considered is the number of attributes involved in unsupervised mining. Automated selection of number of cluster are another challenge. All these are handled in proposed model. The attributes are selected based on the projected space algorithms, number of clusters are equal to the number of attributes in the data set. These method of decision helps to get the pattern based on the particular attribute.

Domain required sliced data analysis is Weather data analysis. Indian weather are serrated into four seasons and that are interdependent. The detailed analysis of each season support the accurate predictions of next season, in which the attribute of one season may not be consider in next season but there is a interdependency.

Climate plays major role in deciding future of all the sectors because it is used for many of human activates. Predicating accurate climate of future is a challenging job for all the climate scientists. Weather is

day-to-day variation in a particular region, whereas the climate is a long term fusion of the variation. The weather conditions are obtained from automated weather stations, ground observation, Doppler radar, aircraft, sensors and satellites. Weather data includes temperature, Wind Speed, Evaporation Radiation, SunShine, CloudForm, Humidity, Precipitation and Rain fall. Weather data are generally classified as synoptic data for climate data and used in weather forecast models (Mathematical calculations). Climate data are official data provided after some quality control on synoptic data. Weather varies for time to time and for each region. In a data mining work weather data can be include in spatio-temporal data sector[4]. As the nature of region varies the quality control on weather considers nature of the region to create official climate data from weather data. The nature of region are predicted based on the latitude and longitude in which it is located [1]. There the data mining techniques used to do either descriptive mining (describe general properties) or predictive mining (attempt to predict based on inference of data) on large volume of data to provide accurate forecasting even for long days and accurate prediction about climate for long term. The figure 1 expresses this work flow of weather prediction and the roll played by data mining in that.

CLUSTERING

Cluster analysis is a explore the structure of data. Core Cluster analysis is a clustering. Clustering analysis in a data is a unknown label class (unsupervised) [3][11]. So it is learned by observation not learned by example [2]. Clustering divide the data set into classes using the principle of "Maximum intra class similarity and Minimum inter class similarity". It doesn't have any assumption about the category of data. The basic

clustering techniques are Hierarchical, Partitional, Density based, Grid based and Model based clustering. Some sort of measure that can determine whether two objects are similar or dissimilar is required to add them into particular class. The distance measuring type varies for different attribute type. Clustering can also used to detect outline in data which may occur due to human error or some abnormal events occurred while creating data set [2][11]. Cluster work well on scalable, heterogeneous and high dimensional data set. In all the clustering algorithms user defined parameters are given as input to find either similarity, dissimilarity among clusters and for root attribute of cluster and for maximum or minimum number of clusters.

Projected Space Clustering

In some data set the clusters compared may vary in attributes. The projected space clustering filter the less important dimension[8][9]. Treat that as outline and eliminate those dimension. The clustering done with the remaining data set. In that the clusters varies in attributes. The quality of clusters are identified and that are optimized. Cluster formation done using any of the basic clustering method discussed above. The variation in the basic projected space clustering is important in data set like weather data. Because the dimension removed as outline may influence a predication of climate [17][18]. For example during the summer season the rain attribute may have less reading but that is also one of the value decides the climate of next season. So the attributes must be kept as such even it is less important in that cluster. MPSKM suggested in this paper ensemble the data with all level of attributes.

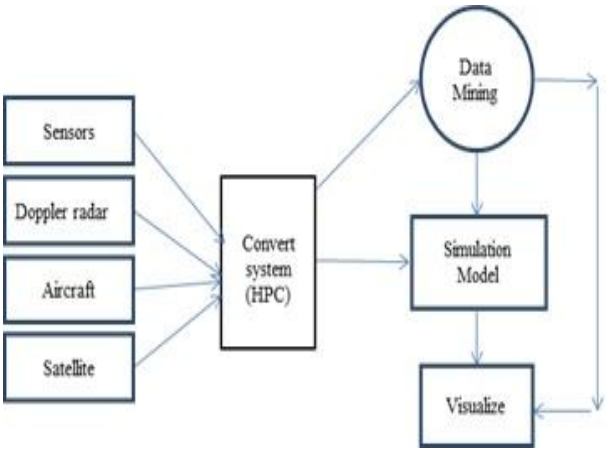

Figure 1: Data Mining in Weather Predication

Outline

The Outline is an extreme (low/ high) in data. Many of the discusses says it is error data entry in data set [15]. But it may not be 100% true in all data. For example the data taken in the MPSKM is weather data, the extreme low or extreme high data in particular season of data may occur due to some natural events occurrence in the same. This outline data also to be taken for season analysis to predict the next season which are all interdependent in each other. In MPSKM the extreme data also formed into cluster to study the same season of data.

The rest of this paper is arranged as Related work in projected space clustering, Proposed Model to compare the clusters with different attribute, Motivation of MPSKM model, Result Analysis and Conclusion.

RELATED WORK

In [4] reported Spatio-Temporal pattern in climate data using clustering. Cluster the climate data reduces the computational complexity. Climate data pre-processed to eliminate anomaly and the outline. Spatial similarity algorithm create a cluster of similar weather stations on 46/77 latitude/longitude with similar climate behaviour. In [5] used data mining techniques to find the increase and decrease in global temperature. One of the aims of climate analysis is to find the increase or decrease in global temperature. K-means clustering algorithm used to group data set with minimum temperature and also applies J48 Classification algorithm to fine the suitable attribute to split the data set. The size of K=5. The data set taken for study is 112 years long. In [6] applied K-means algorithm with the size of k=4 applied on nine years of data to form a cluster. Seasonal study is of great help for agriculturalists; The clusters used for summer, rainy, spring and autumn. The study of the clustering analysis was done for variation in rain, temperature, humidity and wind speed compare to same season of each nine years. Such study helps to predict the climate of all the four seasons in near future. All these clustering required user input to decide the number of cluster if it is KNN or K Means clustering method. when the clustering is decision tree base the initial attribute selection plays a major roll. These problem are talked in automated projected space clustering.

In [7] developed SUBCAD projected space clustering algorithm, in which the data from data set are selected in sampling technique to form a initial set of clusters. The remaining data are added into any of the initially formed cluster in which the new one have less impact on cluster quality. Later the clusters are pruned to optimize. The problem in this method is to identify the initial clusters with all attributes. In [8] found graph partition algorithm called CLICK. The weight of each attribute is calculated. The graph based partition of data by applying weighted attributes. The attributes are vertex of graph the edge between the vertex. The problem with this CLICK is handling high dimension data. This also eliminates the less weight attributes. Needs the user input to find threshold values in calculating attribute weight. In [9] developed AT-DC algorithm based on decision tree clustering. AT-DC form a single top level cluster with entire data set. Then the sub clusters are formed. The sub clusters are

accepted when the its quality higher than the original clusters formed in previous step. AT-DC handles the high dimensional data automatically that is without user input to form a projected space cluster, but it doesn't handle outliner in data set.

In [10] developed PROCAD algorithm to ensemble data in projected space. It handle high dimension data and also the outline with out user input. In the initial state the attribute weight are calculated based on which the clusters are formed. The quality of cluster is analyzed as good when it contains desired number of dimension and data point. Even in this PROCAD algorithm attribute with less important are not considered to form cluster. And also all these algorithms work only for categorical data not for numerical data. The PCKA compared with PCFA in [16][17] works on high dimension data set. The Fuzzy C Means algorithm used to identify and project the relevant attributes and all the attributes are treated equally.

The modified model of projected space cluster in this paper suggest the procedure to form a clusters of one season data pattern with required amount of dimension even few attribute are less in requirement which decides the weather of next season. Here the data is completely numerical and treated as unequally weighted. The weighting of attribute and limitation of number of clusters are done in dynamic nature to eliminate basic error like cluster study with in limited number of clusters.

MPSKM Model

The MPSKM model starts with cluster formation of each slice of data based on the patterns exist, from which the extreme event fall on separated cluster. In traditional clustering algorithm the prefixed set of clusters include the extream event of data which affect the statistical values (mean, centroid, etc) of the cluster.

The model suggested here is to work specifically on Indian weather data set. Naturally India weather segregated into four seasons. The clustering has to be done for all the four seasons (Winter December to February, Summer or Pre Monsoon – March to May, Rainy or Monsoon – June to August, Autumn or Post Monsoon – September to November) each of three months long[13] . The Model derives four sets of clusters each are for each seasons respectively. This cluster derivation expressed in figure 2 as algorithm. The number of clusters are decided based on the number of attributes taken for the study. The pattern based cluster formation including all the attributes in projected space data set instead of eliminating less important attribute. Model works for each season of separately from slicing entire set of data into G_i groups. For each group model decides K (No of attributes) clusters. The random points are assigned to each of K cluster and that are considered as initial centroids. The Distance calculation differ in this model by calculating individually for attributes. The threshold value for each attribute is decided based on the importance of attribute in that season. Base on the threshold satisfaction the data points are reassigned to other cluster and this process continues until the centroid is stagnant. The resulted cluster sets are base to study the weather information in season wise.

The outcome of the model is set of four clusters C_{1k} (Clusters of Season 1), C_{2k} (Clusters of Season 2), C_{3k} (Cluster of Season 3) and C_{4k} (Clusters of Season 4). Where K is number of attributes in each slice of data. So the number of clusters are based on number of attributes in data set, so no data point is omitted from clustering. The clusters formed using the MPSKM model identifies many different patterns in particular season and number of occurrences of extreme events in each season. This nature of model identifies the changes in weather data pattern because of minor changes in any of the projected attribute.

The extreme high and low events in each season can be identify from the cluster with less data point in it. This cluster also include for the season vice comparison even as extreme. Because these may have some interdependency either with previous season cluster or with next season cluster. And it is an alert to the application area when the same is repeated in almost all the years of same season. This identification can be takes places for $C_{xk,}$ $C_{yk,}$ C_{wk} and C_{Zk} clusters separately. The dynamic clusters formed using MPSKM model used to study the entire season and get the result like extreme events, cause of unequal rain fall, extreme temperature in rainy and summer season respectively.

The main aim of meteorological is accurate weather predication for many applications. One sub part of this prediction is analyzing of weather seasonal dependency because Indian weather is seasonal dependent[13]. The clusters of four seasons are cross compared to find the particular pattern of cluster which have more influence in climate of next season. The prediction may using Neural Network which gives accurate result when compared with traditional algorithms[12][14].

Motivation of model

Few domain like weather data consists of seasonal vice result dependency. Especially in India the climate of one season mostly depends on the previous season. This is like a chain activity. The attributes recorded in this domain is uniform in all the season but the importance of attribute in one will not have same level of importance in another season and vice versa. And the season pattern of weather will also vary. Less and more variation in pattern may also influence the climate, so no pattern or dimension to be eliminated in forming season wise patterns. In this case the clusters of different seasons to be compared have unequal weight of complete dimensional attribute.

Projected space clustering using in similar kind of domain is modified in this model to handled the clusters of projected space without removing any of the lease weighted attribute as

outline. The model clusters the similar patterns in different clusters which deviated due to any of the attribute. Number of clusters are equated to number of attributes to cover all possible patterns of weather in each season. The clustering happens on the basis of single attribute changes instead of entire data object value. The reason of weather variation and the influence of extreme event in this variation can be compared. The suggested model will perform well in handling clusters of unequal weight of attributes. Inter seasonal clusters with differently weighted attribute dimensional may back propagated using ANN to predict the weather of next season.

Algorithm

D – Data set.
G_i – Slice of data in each year of data set {G_{i1}, G_{i2}…G_{im}},
where 1 < i <4 and 1 <m < No of Data in G_i.
A_j – Attributes, 1 < j < n
K – number of clusters in each G_i,
Where 1 < k <n. Depends on number of attributes
C_{xk}, C_{yk},C_{wk},C_{zk} – clusters of each G_i,
Where x=1, y=2, w=3, z=4.

1. Slice the data into continuous four seasons of data.
// do for each season of data G_i
2. For i in 1 to 4
3. Fix threshold value of C_{ik} clusters.
4. Do for objects in G_i {d_1, d_2,…d_j}
5. Arbitrarily choose *k* data points from G_i as initial centroids of K clusters.
// Clusters C_{ik}.
6. Assign G_{im} objects to C_{ik} Clusters.
7. Repeat
8. Calculate the new mean for each cluster treat that as new centroid.
9. Calculate distance between G_{im} and Centroid of C_{ik} using

10. $Diss_k$= Centroid(C_{ik})-d_{jk}

11. Compare $Diss_1$…..$Diss_k$ with Threshold of each attribute.
12. Reassign the data point to C_{ik} clusters based on the threshold compatibility of respective attribute.
13. Evaluate the cluster quality.

14. Until No change in Centroid

Figure 2 : MPSKM Algorithm

Result Analysis

The time series based MPSKM model implemented and tested by applying on weather data set. This section gives the detail result analysis of the implementation.

Data Set

The model tested on Indian weather data different which need the four seasonal study.The Data set of weather station in 11.0^o latitude and 76.9^o longitude is taken for analysis. The data set includes ten years of data from 2003 to 2013. The attributes projected for the analysis are Temperature, Rain, Humidity, Wind Speed, Radiation, Evaporation. All the attributes are of numerical type. The algorithm applied on this data set forms six clusters each respectively for attributes.

RESULT AND DISCUSSION

As the model works to form a dynamic set of clusters it is implemented using JAVA programming language. Each year of data are sliced into 4 group and given as input to the model. So each season produces different set of cluster based on the index rank proceeded for each seasons separately. The data set taken for the study projected with 6 attributes, all are with different importance in each season, for example the temperature in summer definitely high when compare to summer season, so the cluster of such data have to be taken place with in the data of same season.

Table 1: Rank Matrix 2003 to 2006

Year	2003				2004				2005				2006			
Season	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Temp	0.3	0.2	0.1	0.3	0.2	0.4	0.2	0.3	0.2	0.1	0.2	0.3	2.0	0.1	0.2	0.2
Rain	0.5	0.4	0.5	0.6	0.3	0.4	0.4	0.5	0.2	0.3	0.5	0.4	0.4	0.3	0.3	0.5
Hum	0.6	0.5	0.4	0.4	0.6	0.5	0.4	0.4	0.6	0.4	0.4	0.5	0.6	0.4	0.4	0.5
Wind	0.2	0.4	0.3	0.1	0.5	0.3	0.3	0.2	0.2	0.3	0.5	0.1	0.3	0.4	0.4	0.2
Radiation	0.1	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.1	0.2	0.2
EVP	0.0	0.2	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1

Even with in the season the range of each attribute vary, so the distance calculation with centroid using traditional method may leady the cluster quality to low. The rank index of each season data are represented in table 1, table2 and table 3. The ranked attribute supports to find the impact of attribute in sliced set and also the prediction of another attribute in next season.

Table 2: Rank Matrix 2007 to 2010

Year	2007				2008				2009				2010			
Season	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Temp	0.1	0.2	0.3	0.2	0.2	0.3	0.2	0.2	0.2	0.4	0.2	0.3	0.2	0.2	0.3	0.1
Rain	0.2	0.4	0.4	0.3	0.5	0.3	0.4	0.4	0.3	0.4	0.4	0.5	0.3	0.3	0.3	0.4
Hum	0.3	0.4	0.5	0.4	0.6	0.5	0.4	0.4	0.6	0.5	0.4	0.4	6.0	0.4	0.3	0.3
Wind	0.1	0.3	0.5	0.3	0.2	0.4	0.3	0.3	0.5	0.3	0.3	0.2	0.5	0.3	0.4	0.2
Radiation	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.1
EVP	0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1

Table 3: Rank Matrix 2011 to 2013

Year	2011				2012				2013			
Season	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Temp	0.3	0.1	0.2	0.1	0.2	0.3	0.2	0.2	0.1	0.2	0.1	0.2
Rain	0.2	0.4	0.2	0.4	0.5	0.3	0.4	0.4	0.4	0.2	0.3	0.4
Hum	0.5	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.4	0.6	0.4	0.6
Wind	0.3	0.5	0.4	0.5	0.2	0.4	0.3	0.3	0.3	0.4	0.5	0.2
Radiation	0.2	0.1	0.2	0.2	0.2	0.1	0.2	0.2	0.1	0.2	0.1	0.1
EVP	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1

So the distance measured among relevant attribute yield considerable quality improvement in cluster. The precision and recall of the cluster for all the 10 years of are represented in the figure 3.

The visual clearly explains that the recall of each year of data is high, which shows that cluster clearly secretes the minor pattern of data in each season and there is no non cluster object in data set.

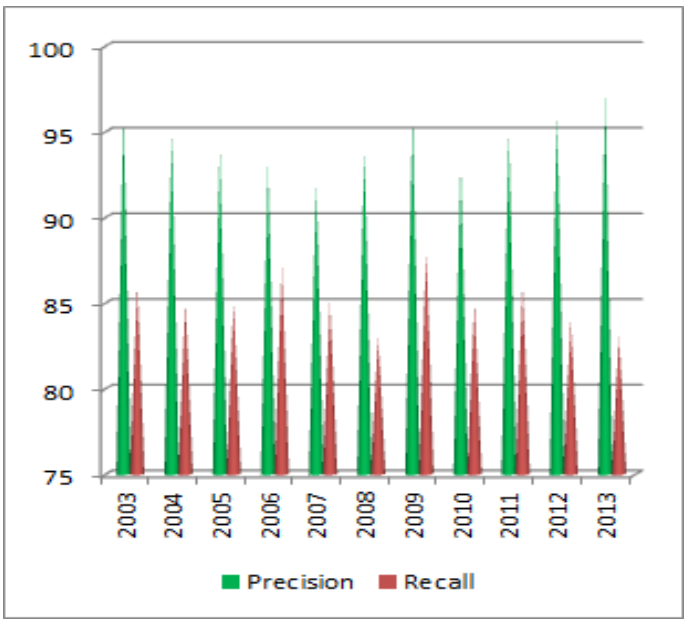

Figure 3: Precision and Recall of 10 years Weather Data

CONCLUSION

Projected Space Clustering is good on non contiguous numerical time series data. But the same needed some modification to handle all the attributes of data point without the omitting single unequal attribute. The MPSKM model will solve the problem of handling unequal attribute and to identify the extreme (low/ high) event. One more challenge in unsupervised data mining is deciding the number of labels to cluster the data. The MPSKM decides that based on the number of attributes projected in data set. One more key concept handled to get the sensitive pattern is slicing of data based on the timing sequence in which these are recorded. One of the application area where the slice of data need is climate, in which four seasons of data to be clustered separately to get different patterns. So MPSKM Clustering techniques applied on climate data helps to produce similar pattern of climate with the consideration of spatial nature. The model can be enhanced further by changing the threshold value dynamically based on the previous year actual weather report of meteorological department. This is required because of changing nature of climate.

REFERENCES

[1] Badhiye, S. S., B. V. Wakode, and P. N. Chatur. "Analysis of Temperature and Humidity Data for Future value prediction." *Intl. Journal of Computer Science and Information Technologies (IJCSIT)"*, Vol 3, pp : 3012-3014, 2012,.

[2] Jiawei, Han, and Micheline Kamber. "*Data mining: concepts and techniques*" San Francisco, CA, itd: Morgan Kaufmann 5 . ISBN : 978-1-55860-901-3, 2006.

[3] K. P. Soman, Shyam Diwakar, V. Ajay, "Insight into Data Mining Theory and Practice" – PHI Learning , Delhi, 7, 2014

[4] Daniel Levy, "*SpatioTemporal Pattern Detection in Climate Data*",ITiCSE, Vol 1,pp : 67-71 , 2013.

[5] T V Rajini kanth, V V SSS Balaram and N.Rajasekhar, "Analysis Of Indian Weather Data Sets Using Data Mining Techniques", *Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS*, Vol 1, pp. 89–94, 2012.

[6] Sarah N. Kohail, Alaa M. El-Halees "Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip)", *International Journal of Information and Communication Technology Research"*, Vol 1, pp : 96-100 , July 2011.

[7] Gan G, Wu J "Subspace clustering for high dimensional categorical data.", *ACM SIGKDD Explor ,Newsl*, Vol 6, pp : 87–94, 2004.

[8] Zaki MohammedJ, Peters M, Assent I, Seidl T "CLICKS:

an effective algorithm for mining subspace clustersin categorical datasets.", *Data & Knowledge Engineering*, Vol 60, pp :51–70 , 2007.

[9] Cesario E, Manco G, Ortale R "Top-down parameter-free clustering of high-dimensional categorical Data" , *Knowledge and Data Engineering, IEEE Transactions*, Vol 19, pp :1607–1624, 2007.

[10] M. Bouguessa, "Clustering categorical data in projected spaces", *Data Mining and Knowledge Discovery*, Vol 29 , pp :3–38 , 2015.

[11] Mohammed J Zaki, Wagner Meira JR, "Data Mining and Analysis Fundamental Concepts and Algorithms", *Cambridge University Press, New York, First Edition*, 2014.

[12] MuYeong, Kang, Jae-Do, Shin and Byungkya Kim **"**Automatic Subject Classification of Korean Journals based on KSCD*",,Indian Journal of Science and Technology*, Vol 8, pp : 452-456 , January 2015.

[13] S. D. Attri and Ajit Tyagi, "Climate profile of india", *Environment Monitoring and Research Centre, India Meteorological Department*, Lodi Road, New Delhi-110003, Met Monogram No. Environment Meteorology 01/2010.

[14] Farhad Soleimanian Gharehchopogh, Seyyed Reza Khaze, Isa Maleki, "A New Approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network Algorithms", *Indian Journal of Science and Technology*, Vol 8, pp : 237–246 , February 2015.

[15] Kavita Thawkar, Snehal Golait, Rushi Longadge, "A Framework for an Outlier Pattern Detection in Weather Forecasting", *International Journal of Computer Science and Mobile Computing*, Vol. 3, pp.348 – 358, May 2014.

[16] Ilango Murugappan, Mohan Vasudev, "PCFA: Mining of Projected Clusters in High Dimensional Data Using Modified FCM Algorithm", *The International Arab Journal of Information Technology*, Vol. 11, pp. 168-177, March 2014.

[17] Amardeep Kaur , Amitava Datta, "A novel algorithm for fast and scalable subspace clustering of high-dimensional data", *Journal of Big Data*, Vol 2, pp. 1-24 , August 2015.

[18] Sung-Soo Kim, " Variable Selection and Outlier Detection for Automated K-means Clustering", *Communications for Statistical Applications and Methods*, Vol. 22, pp. 55–67, 2015