AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Amalgamation of Opportunistic Subspace & Estimated Clustering on High Dimensional Data

[1]M. Ravichandran and [2]Dr.A.Shanmugam

[1]*Assistant Professor(Senior Grade) of Department of IT, Bannari Amman Institute of Technology, Sathyamangalam, Erode District, Tamilnadu, India*
[2]*Dean, Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode District, Tamilnadu, India.*

ARTICLE INFO

Article history:
Received 12 January 2014
Received in revised form 20 March 2014
Accepted 25 March 2014
Available online 2 April 2014

Key words:
High Dimensional Data, Opportunistic Subspace, Search retrieval, Cluster rationale, estimated clustering.

ABSTRACT

Problem statement: Clustering contains the numerous techniques to develop the various fields such as statistics, pattern recognition, and data mining. Subspace clustering developed from the group of cluster objects in all subspaces of a dataset. When clustering high dimensional objects, the accuracy and efficiency of traditional clustering algorithms are very poor, because data objects may belong to diverse clusters in different subspaces comprised of different combinations of dimensions. **Approach:** To overcome the above issue, we are going to implement a new technique termed Opportunistic Subspace and Estimated Clustering (OSEC) model on high Dimensional Data to improve the accuracy in the search retrieval. **Results:** Performance of Opportunistic Subspace and Estimated Clustering technique to discover the efficient cluster validation is evaluated in terms of computational cost and energy usage based on attribute size. **Conclusion:** It considers the problem of accuracy in clustering the high dimensional data. An analytical and empirical result shows the better cluster rationale with the efficient estimated clustering of our proposed scheme.

To Cite This Article: M. Ravichandran and Dr.A.Shanmugam., Amalgamation of Opportunistic Subspace & Estimated Clustering on High Dimensional Data. ***Aust. J. Basic & Appl. Sci.,*** *8(3): 88-97, 2014*

INTRODUCTION

Clustering is an accepted data mining technique for a diversity of applications. One of the incentives for its recognition is the capability to work on datasets with minimum or no a prior knowledge. This builds clustering realistic for real world applications. In recent times, high dimensional data has awakened the interest of database researchers due to its innovative challenges brought to the community. In high dimensional space, the space from a report to its adjacent neighbor can approach its space to the outermost reports. In the circumstance of clustering, the problem causes the space between two reports of the same cluster to move toward the space between two reports of different clusters. Traditional clustering methods may not succeed to distinguish the precise clusters and providing the accuracy in retrieval of data.

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. Cluster is used to group items that seem to fall naturally together. Various types of clustering: hierarchical (nested) versus partitioned (un-nested), exclusive versus overlapping versus opportunistic, and complete versus partial. Clustering is an unverified learning process that partitions data such that similar data items grouped together in sets referred to as clusters. This activity is important for condensing and identifying patterns in data.

Clustering is supposed as an unverified process in which the process of organizing objects into groups whose members are similar in some way. The authority of clustering results has to be evaluated by discovery the optimal number of clusters that best fits the given data set. Clustering objects in high dimensional spaces may detain to clustering the objects in subspaces which may be of diverse dimensions. The trial-and-error approach may not succeed because of the following difficulties:

- Predefining the numeral of clusters primarily is not easy.
- Re-initialization at every phase increases the computational cost.
- The sparsity is called as “curse of dimensionality”.

Corresponding Author: M. Ravichandran, Assistant Professor(Senior Grade) of Department of IT,Bannari Amman Institute of Technology, Sathyamangalam, Erode District , Tamilnadu, India
E-mail: mraveen2007@gmail.com

In view of the above, we have offered a new fuzzy subspace clustering algorithm for clustering high-dimensional datasets, and an algorithm for detecting the attacks based on Mahalanobis distance.

Fuzzy techniques have been used for handling vague boundaries of arbitrarily oriented clusters. However, traditional clustering algorithms tend to break down in high dimensional spaces due to inherent sparsity of data. Charu Puri., and Naveen Kumar., 2011 propose a modification in the function of Gustafson-Kessel clustering algorithm for projected clustering and prove the convergence of the resulting algorithm. It present the results of applying the proposed projected Gustafson-Kessel clustering algorithm to synthetic and UCI data sets, and also suggest a way of extending it to a rough set based algorithm.

As in the case of traditional clustering, the purpose of discrete dimensional projected clustering algorithms is to form clusters with most encouraging quality. However, the traditional functions used in estimate the cluster quality may not be appropriate in the predictable case. The algorithms will consequently be likely to select few attributes values for each cluster, which might be inadequate for clustering the reports correctly. In some previous works on projected clustering (Mohamed Bouguessa., and Shengrui Wang., 2009), the clusters are evaluated.

Spontaneously, a small standard space between attribute values in a cluster indicates that the associate reports agree on a small range of values, which can make the reports easily restricted. A large number of selected attributes value towards the reports are analogous at a high dimensional, so they are very credible to belong to the same real cluster. Finally, a large number of reports in the cluster point out there are a high support for the selected attributes value, and it is improbable that the small distances are merely by chance.

All these are indicators for a high-quality multiple clusters, but here is essentially a tradeoff between them. Suppose a given set of reports, it selects only attributes value that tends to generate the ordinary space among reports, fewer attributes will be selected. Similarly for a space obligation, locating more reports into a cluster will most likely amplify the average number of attribute value chosen.

It's important to point out that in this work; we focus on Opportunistic Subspace and Estimated Clustering (OSEC) model on high Dimensional Data. Estimated clustering method focused to find clusters in small estimated subspaces for data of high dimensionality. It presents an effectual method for finding regions of superior density in high dimensional data in a way which has high-quality scalability and accuracy. Opportunistic subspace uses the difference subspace clustering method as the initialization technique. It combines opportunistic logic for influential the clusters in subspaces and complete space. The ability to detect attacks can be improved using a mutual perception of attack detection and cluster identification.

We provide here an overview of Multi cluster Dimensional Projection on Quantum Distribution. The rest of this paper is arranged as follows: Section 3 introduces architecture diagram of the proposed scheme. Section 3.1 and 3.2 describes about proposed method; Section 4 shows the evolution and experimental evaluation; Section 5 evaluated the results and discuss about it. Section 6 describes conclusion and prospect.

MATERIALS AND METHODS

Most existing clustering algorithms become substantially inefficient if the required similarity measure is computed between data points in the full-dimensional space. To address this problem, Bouguessa, M., and Shengrui Wang., 2009 a number of projected clustering algorithms have been proposed. However, most of them encounter difficulties when clusters hide in subspaces with very low dimensionality.

Enrico Bertini.,*et al*,, 2011 present a systematization of techniques that use quality metrics to help in the visual exploration of meaningful patterns in high-dimensional data. Satish Gajawada., and Durga Toshniwal., 2012 propose VINAYAKA, a semi-supervised projected clustering method based on DE. In this method DE optimizes a hybrid cluster validation index. Subspace Clustering Quality Estimate index (SCQE index) is used for internal cluster validation and Gini index gain is used for external cluster validation in the proposed hybrid cluster validation index. Proposed method is applied on Wisconsin breast cancer dataset.

Hierarchical clustering is one of the most important tasks in data mining. However, the existing hierarchical clustering algorithms are time-consuming, and have low clustering quality because of ignoring the constraints. In this paper, GuoYan Hang., *et al*., 2009, a Hierarchical Clustering Algorithm based on K-means with Constraints (HCAKC) is proposed.

B Shanmugapriya and M Punithavalli., 2012., an algorithm called Modified Projected K-Means Clustering Algorithm with Effective Distance Measure is designed to generalize K-Means algorithm with the objective of managing the high dimensional data. The experimental results confirm that the proposed algorithm is an efficient algorithm with better clustering accuracy and very less execution time than the Standard K-Means and General K-Means algorithms.

L.Jegatha Deborah., *et al*., 2010 present a detailed description of the mathematical working of few cluster validity indices and not all, to classify these indices and to explore the ideas for the future promotion of the work in the domain of cluster validation. Survey by Hans peter kriegel., *et al*., 2009, tries to clarify: (i) the different problem definitions related to subspace clustering in general; (ii) the specific difficulties encountered in this

field of research; (iii) the varying assumptions, heuristics, and intuitions forming the basis of different approaches; and (iv) how several prominent solutions tackle different problems.

Rahmat Widia Sembiring., *et al*., 2010, PROCLUS performs better in terms of time of calculation and produced the least number of un-clustered data while STATPC outperforms PROCLUS and P3C in the accuracy of both cluster points and relevant attributes found.

Inspired from the recent developments on manifold learning and L1-regularized models for subset selection, Deng Cai Chiyuan., and Zhang Xiaofei He., 2010 propose in a new approach, called Multi-Cluster Feature Selection (MCFS), for unsupervised feature selection. Specifically, we select those features such that the multi-cluster structure of the data can be best preserved. The corresponding optimization problem can be efficiently solved, since it only involves a sparse eigen-problem and a L1-regularized least squares problem.

Existing clustering techniques normally merge small cluster with big ones results in removing the identity of those small clusters. The proposed algorithms work on split and merge technique to overcome this limitation. Clustering ensemble method based on a novel two-staged clustering algorithm is proposed by B.A Tidke., *et al*., 2012.

Yun Yang., and Ke Chen., 2011., proposed weighted clustering ensemble algorithm provides an effective enabling technique for the joint use of different representations, which cuts the information loss in a single representation and exploits various information sources underlying temporal data but does not contain the extracted feature. Jung-Yi Jiang., *et al*., 2011., have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data.

To evolve a high dimensional data, a new technique named Opportunistic Subspace and Estimated Clustering (OSEC) model is presented.

PROPOSED OPPORTUNISTIC SUBSPACE AND ESTIMATED CLUSTERING ON HIGH DIMENSIONAL DATA MODEL

The proposed work is efficiently designed for estimating the clusters in high dimensional by adapting the Opportunistic Subspace and Estimated Clustering (OSEC) model.

The architecture diagram of the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model is shown in Fig. The proposed opportunistic subspace clustering is processed under different input, intermediate and output processes. The input process takes the high dimensional data. We select the Difference Subspace Clustering algorithm as the basic clustering algorithm for initialization, which take over the advantages of opportunistic type clustering algorithms such as easiness of calculation, effortlessness, and can covenant with noise and overlap clusters. Our proposed system consists of two modules namely:

- ✓ Clustering module
- ✓ Attack Detection module.

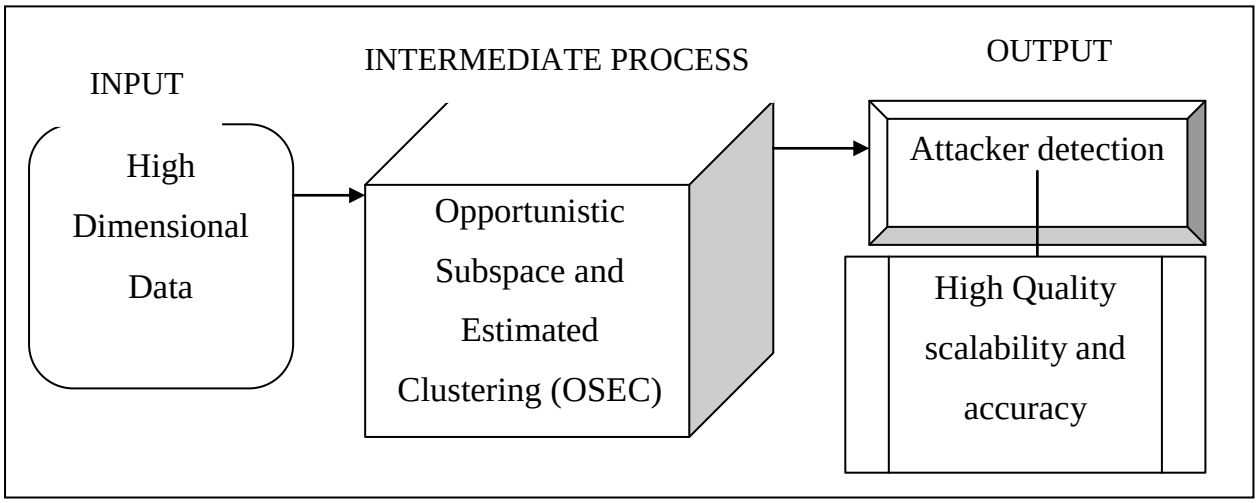


Fig. 1: Architecture Diagram of Opportunistic Subspace and Estimated Clustering (OSEC) model

The clustering module is supplementary separated into two sub modules, which are

- Initialization using Difference Subspace clustering
- Clustering using opportunistic logic

Australian Journal of Basic and Applied Sciences, 8(3) March 2014, Pages: 88-97

Difference Subspace Clustering:

Subspace clustering try to find clusters in different subspaces within a real dataset. This means that a data summit might fit in to multiple clusters, each accessible in a different subspace. Subspace algorithm determines the each cluster center and then determines all their centroid points. Frequently in high dimensional data, several dimensions may be inappropriate and can cover accessible clusters in noisy data. Subspace clustering algorithms usually restrict the investigation for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces.

Steps to perform the Difference Subspace Clustering:

Step 1: Highest point Data object selected as initial cluster center

Step 2: Neighborhood data objects are removed from the initial cluster center

Step 3: Goto Step 1 and Step 2 until the data points are within the radii of cluster

Algorithm for initialization using Difference Subspace clustering:

Input: Dataset Y= {y1, y2, …, yn} ∈ Real $_d$ (Initialize the centre E (0) and set E $_{max}$ =k)

Step 1: For each $y_i \in Y$, compute the mass index

$$\mathbf{F_i} = \sum_{j=1}^{n} \exp o[-\left\| yi-yj \right\|^2 /(.5q_a)^2$$

Step 2: Let F_{c1}=max {Fi,i=1,2,…,n},then select y_{c1} as the initial cluster center;

Step 3: Repeat the step if y_{ck} be the k_{th} cluster center, and the mass index be F_{ck}.

Step 4: For each yi ∈ Y, update the mass index,

$$\mathbf{F_i = F_i - F\,ck} \sum_{j=1}^{n} \exp o[-\left\| yi-yck \right\|^2 /(.5q_b)^2$$

Step 5: For each yi ∈ Y, until F_{ck} +1 / Fc1 < δ, where qa, qb and δ need pre assignment.

$$q_a = q_b = 1/2^{\min_k \{ \max_i \{ \| y1-yk \| \} \}}$$

In general, each phase tests the number of clusters ‘c’ between C_{min} and C_{max}. The cluster centers should be initialized at the establishment of running the difference algorithm. In difference clustering algorithm, the production order of the cluster centers is determined by the mass index. The superior of mass index is that the earlier of cluster center generated. Thus, at each step the top ‘c’ cluster centers can be selected as the new initialization cluster centers, and there is no need to re-initialize the cluster centers. After initialization of the centroid, the next step is to concern the opportunistic algorithm to get hold of the association degree for every data point with deference to each cluster.

Australian Journal of Basic and Applied Sciences, 8(3) March 2014, Pages: 88-97

Opportunistic Subspace & Estimated Clustering method:

Opportunistic subspace method of clustering is a data clustering system in which a real dataset is grouped into 'n' clusters. Each data point in the real dataset belongs to every cluster of a confident degree. For illustration, a certain data point that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and a different data point that lies far away from the middle of a cluster will have a low degree of belonging to that cluster.

With opportunistic method, the centroid point of a cluster is computed as the mean of all points. It is weighted by their degree of belonging to the cluster. The performance depends on initial centroid points. In our approach, Opportunistic Subspace clustering algorithm is used to conclude the possible of each cluster center and then conclude all of the centroid points.

For a dynamic approach there are two ways.

- Using an opportunistic subspace algorithm we can determine all of the centroid points.
- Run opportunistic algorithm several times each starting with different initial centroid points.

The opportunistic steps include

Step1: Modernize association matrix (A)

Step 2: Determine association for each point

Step 3: Repeat step (1) and step (2) until the centroid points are stabilized.

Algorithm for Opportunistic Subspace and Estimated Clustering (OSEC) model after centroid point initialization:

The below describes the steps to be performed

Step 1: Initialize the K-step

Step 2: Calculate the centers vectors $C^{(k)} = [cj]$ with $A^{(k)}$.

$$C_j = \sum_{i=1}^{N} a_{ij}^{m} . y_i / \sum_{i=1}^{N} a_{ij}^{m}$$

Step 3: Update the points $A^{(k)}$,$A^{(k+1)}$.

$$C\,ij = \frac{1}{\sum_{k=1}^{c} (\|yi - yj\| / \|yi - yk\|)^{2/m-1}}$$

Step 4: If $\| A^{(k+1)} - A^{(k)} \| < \delta$ then goto Step 6.

Step 5: Otherwise return to Step 3.

Step 6: End

Here, δ is a predefined value which is specified as input. Generally δ is taken as 0.0001.

By using the difference clustering as a part of OSEC algorithm, the problem of initialization and the maximal number of clusters is determined. Difference Subspace clustering is used as the basic clustering algorithm, and joint with the proposed indices which are particularly distinct for subspace clusters, to decide the optimal number of clusters in high dimensional spaces.

Thus, in our proposed system, every dimension donates to the detection of clusters, but the dimensions with superior weights form the subsets of dimensions of cluster. The difference subspace clustering of the multiple data points based on their attributes using opportunistic logic is used. After that clusters with their centroid points are obtained. Since the figure of the clusters as specified is circular, there may be abnormal data points which are called as attackers. Detection of attackers is very important to obtain good quality clustering results.

Australian Journal of Basic and Applied Sciences, 8(3) March 2014, Pages: 88-97

Attacker Detection:

When analyzing data in real datasets, sometimes outlying observations cause problems. Peter J. Rousseeuw and Mia Hubert, 2011 aims at detecting the attackers by searching for the model by the majority of the data. In real data sets, often some observations are different from the majority. Such observations are called attackers.

Outlying observations may be different from the majority of that points that have been recorded under outstanding circumstances. Consequently, they do not fit the model well. To avoid slough effects, vigorous statistics finds a fit that is close to the fit found without the outliers. The attackers are recognized by their big deviation from that robust fit.

Multiple location and estimation points:

Assume that the real dataset contained 'n' data objects, which are 'r' dimensional and stored in an n × r data matrix, Y= {y1, y2, …, yn} ^{S} with xi = (yi1, ….. yir)S the i th observation.

Experiential mean, $\overline{y}$ is

$$\overline{y} = 1/n \sum_{i=1}^{n} y_i$$

Experiential Estimation matrix T_x is obtained using,

$$T_x = \sum_{i=1}^{n} (y_i - \overline{y})\,(y_i - \overline{y})^{s} /(n-1)$$

Experimental Evaluation :

The proposed Opportunistic Subspace and Estimated Clustering (OSEC) is implemented in JAVA. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms are functional honestly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also further appropriate for increasing new machine learning schemes. Attribute Relationship File Format (ARFF) is the text format file used by Weka to accumulate data in a database.

Our team has incorporated numerous standard Machine Learning (ML) techniques into a software "workbench" called Waikato Environment for Knowledge Analysis (WEKA). With it, a proficient in a particular field is able to use ML to gain helpful knowledge from databases that are far too large to be analyzed by hand. WEKA's users are ML researchers and industrial scientists, but it is also extensively used for teaching. CORTINA Dataset 10 contains Million images using image content, text and annotations

Another dataset named the Ski Resort Data Set is given from Data Mining Course by Yong Bakos. It uses weka to weight the data file and save the dataset as final.arff file in default ARFF file format for future processing. It contains 989 data objects and each object has 16 attributes. All the data attribute are nominal. The attribute set rating, Survey, Prize, Punishment represents the overall assessment from a subject. The other attribute set Aspen, Snowmass, ..., Eldora communicate to the different ski resorts a subject rates.

In this section, we develop a progression of experiments considered to estimate the correctness of the proposed algorithm in terms of

i) Energy usage,
ii) Execution Time,
iii) Computational cost.

RESULTS AND DISCUSSION

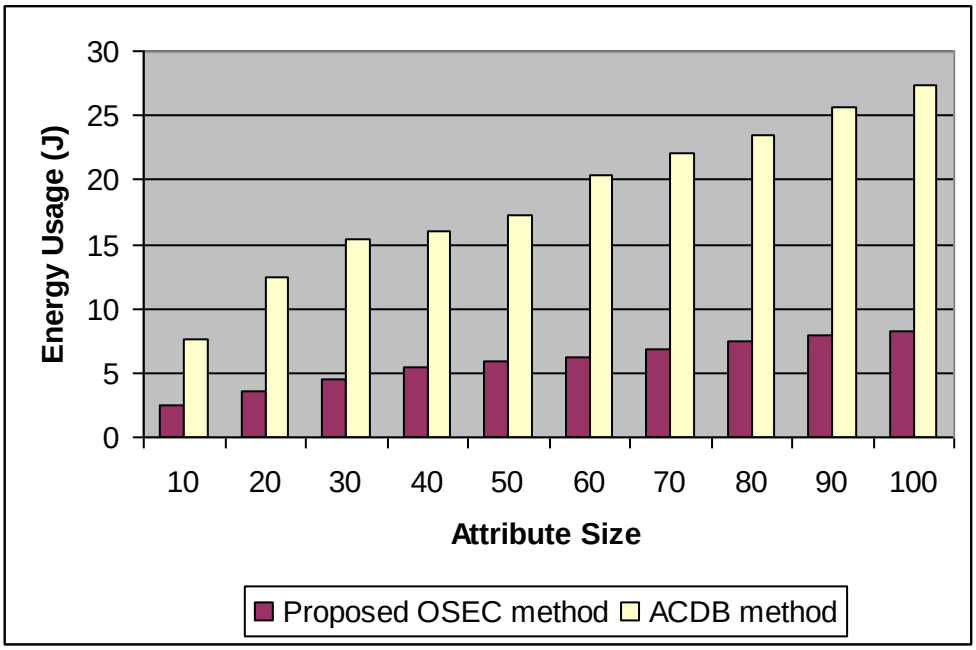
In this work we have seen how the clusters have been estimated in high dimensional spaces. The below table and graph describes the performance of the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model. In the consequence, we compared Adaptive Cluster Distance Bounding for High-Dimensional Indexing (ACDB) and Opportunistic Subspace and Estimated Clustering (OSEC) model, in terms of energy usage.

Attribute Size	Energy Usage (Joules)	
	Proposed OSEC method	ACDB method
10	2.5	7.6
20	3.6	12.5
30	4.5	15.4
40	5.5	16.0

50	5.9	17.2
60	6.2	20.3
70	6.9	22.1
80	7.5	23.5
90	7.9	25.6
100	8.2	27.4

Attribute Size vs. Energy Usage:

The above table describes the energy usage based on the attribute size formed with respect to the CORTINA dataset. The dimensionality of the cluster of the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model, in terms of energy usage is compared with an existing Adaptive Cluster Distance Bounding for High-Dimensional Indexing (ACDB).



Attribute Size vs. Energy Usage:

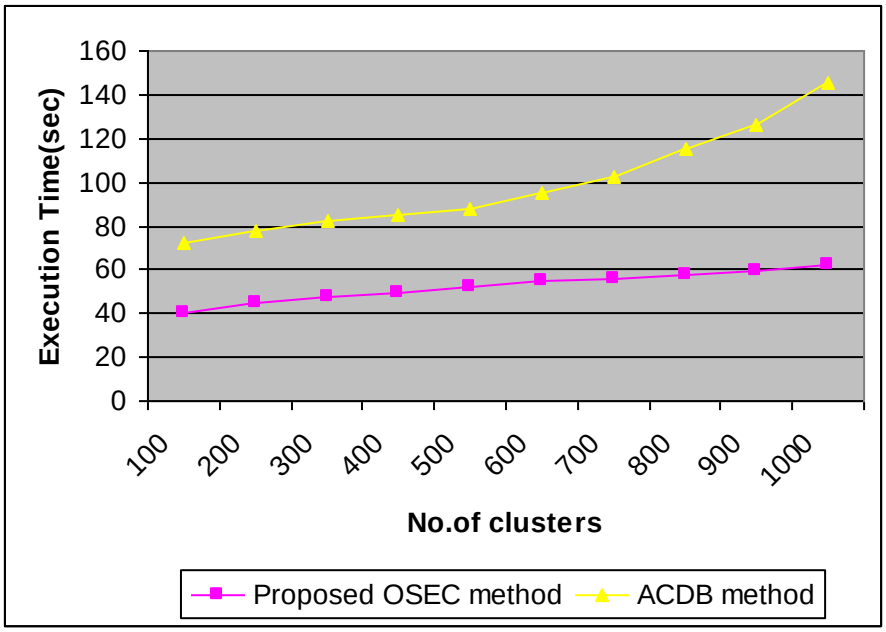
Fig describes the average energy consumption based on the attribute size in the real CORTINA dataset. The set of experiments was used here to examine the impact of energy consumed in the Opportunistic Subspace and Estimated Clustering (OSEC) model. OSEC is capable to complete vastly precise results and its performance is normally reliable. As we can see from Fig., OSEC is more scalable and accuracy in getting the radii of the clusters as input and obtains the optimal number of clusters than the existing ACDB algorithm. If the average attributes size is very low, higher energy usage by providing unsatisfactory results in ACDB. Experiments showed that the proposed OSEC algorithm efficiently identifies the clusters using the difference subspace and its dimensions precisely in a variety of situations.

OSEC eradicates the choice of inappropriate dimensions in all the data sets used for experiments. This can be achieved by the fact that OSEC initiates its process by detecting all the attackers who using the attributes and consuming the energy. Compared to an existing ACDB, the proposed OSEC achieved lesser energy usage and the variance is approximately 30-40% low.

No. of clusters	Execution Time (sec)	
	Proposed OSEC method	ACDB method
100	40	72
200	45	78
300	48	82
400	49	85
500	52	88
600	55	95
700	56	102
800	58	115
900	59	126
1000	62	145

No. of clusters vs. Execution Time:

The above table describes the presence of time taken to execute based on the number of cluster partitioned with respect to the Ski Resort Data Set. The execution time of the cluster of the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model, in terms is compared with an existing Adaptive Cluster Distance Bounding for High-Dimensional Indexing (ACDB).



No. of clusters vs. Execution Time:

Fig describes the presence of time to execute based on the number of clusters with respect to the Ski Resort Data Set. As observed from the figure , OSEC exhibit reliable performance from the first set of experiments on data sets with lesser execution time taken. In tricky cases, OSEC presents much improved results than the existing ACDB method. The results stated in Fig. recommend that the proposed OSEC is more interested to the proportion of data sets in execution time parameter.

Fig describes the consumption of time to perform the opportunistic subspace of estimated clustered based on the clusters described in the dataset. The proposed OSEC balances linearly with the increase attributes in the dataset of the Ski Resort. As specified in the scalability experiments with respect to the data set size, the execution time of OSEC is generally provides improved results than that of ACDB when the time required to develop the clusters in high dimensionality employed for regular runs is also included.

The time consumption is measured in terms of seconds. Compared to the existing ACDB the proposed OSEC consumes less time since it gives better cluster dimensionality result and the variance in time consumption is approximately 20-30% low in the proposed OSEC.

Cluster Object Size	Computational Cost	
	Proposed OSEC method	ACDB method
50	52	90
100	53	91
150	55	92
200	57	93
250	59	95
300	60	96
350	62	97
400	63	98
450	65	95
500	67	96

Table Cluster Object Size vs. Computational Cost:

The above table describes the computational cost with respect to the Ski Resort dataset dimensionality. The computational cost with respect to the cluster object size for the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model, in terms of computational cost is compared with an existing Adaptive Cluster Distance Bounding for High-Dimensional Indexing (ACDB).

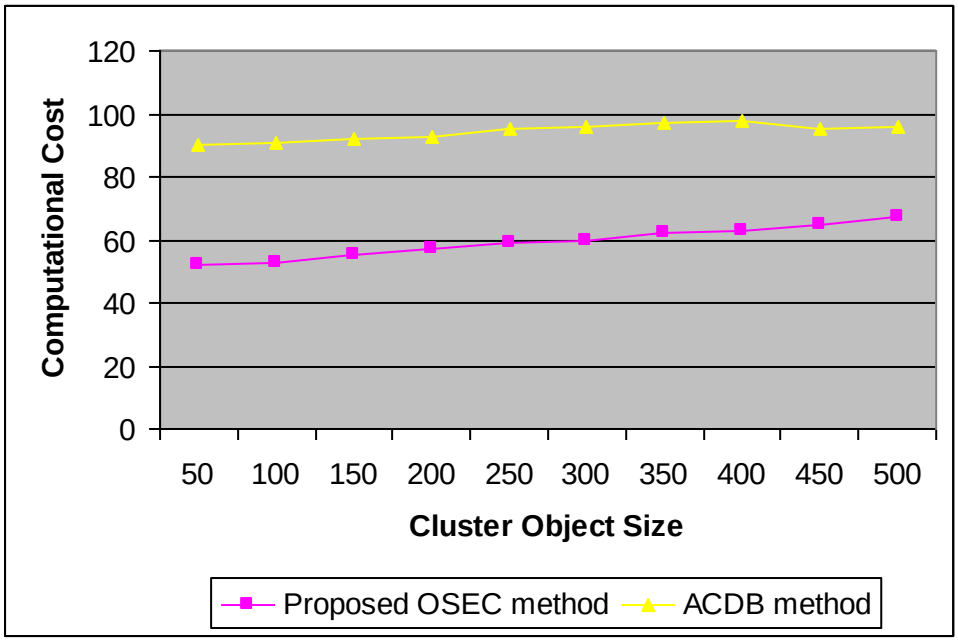


Fig. : Cluster Object Size vs. Computational Cost

Fig describes the computational cost with the help of the cluster object size. In the ACDB method, the CORTINA dataset used to find the computational cost but in the proposed system it is calculated using the Ski resort dataset. The ski resort dataset produces the lesser cost consumption in the proposed OSEC method. The main objective of OSEC is to detect the attackers and to improve the accuracy. Compared to an existing ACDB, the proposed OSEC provides the efficient cluster formation and the variance in is approximately 20-25% lesser cost in the proposed OSEC method.

Conclusion:

In this work, we efficiently achieve the high dimensional data clustering concept in Ski Resort Data Set, CORTINA Dataset and real data set by professionally introducing the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model. The proposed scheme describes the difference subspace model by analyzing the data; rectify the redundancy and improving the accuracy occurs on the attribute value in the dataset. We compared OSEC with Adaptive Cluster Distance Bounding for High-Dimensional Indexing, in terms of accuracy, computational cost and execution time. Our experimental evaluations showed that dimensional estimation clusters considerably outperforms difference subspace clustering algorithm especially on high dimensional data. The experimental results showed that the proposed OSEC scheme for the data attributes worked efficiently by improving 30 – 35 % accuracy and less execution time. We show that the authority of clusters on data dimension indicates the influence of the opportunistic subspace. The proposed method provides a high quality clusters by detecting the attackers. In addition, still to improve the quality of the clustering results heuristic approach can be introduced.

REFERENCES

Enrico Bertini., Andrada Tatu., and Daniel Keim., "Quality Metrics in High-Dimensional Data Visualization: An

Overview and Systematization," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2011

HANS-PETER KRIEGEL., PEER KRO¨ GER., and ARTHUR ZIMEK., "Clustering High-Dimensional Data: A

Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering, 2009. " ACM Transactions on Knowledge Discovery from Data, 3(1) 1.

Mohamed Bouguessa., and Shengrui Wang., 2009. "Mining Projected Clusters in High-Dimensional Spaces," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 21(4).

Charu Puri., Naveen Kumar., 2011. "Projected Gustafson-Kessel Clustering Algorithm and Its Convergence," Transactions on Rough Sets in Computer Science, 6600: 159-182

Rahmat Widia Sembiring., Jasni Mohamad Zain., Abdullah Embong., 2010."Clustering High Dimensional Data using subspace and projected clustering algorithms" International Journal of Computer Science & Information Technology (IJCSIT) Vol.2, No.4, DOI : 10.5121/ijcsit.2010.2414 162.

Satish Gajawada., and Durga Toshniwal., 2012."VINAYAKA: A Semi-Supervised Projected Clustering Method Using Differential Evolution", International Journal of Software Engineering & Applications (IJSEA), 3(4), DOI : 10.5121/ijsea.2012.3406 77.

Deng Cai Chiyuan., Zhang Xiaofei He., 2010."Unsupervised Feature Selection for Multi-Cluster Data," ACM Transactions on Knowledge Discovery from Data.

Bouguessa, M., Shengrui Wang., 2009. "Mining Projected Clusters in High-Dimensional Spaces," IEEE Transactions on Knowledge and Data Engineering, 21: 4.

GuoYan Hang., Dongmei Zhang., Jiadong Ren., Changzhen Hu., 2009. "A Hierarchical Clustering Algorithm Based on K-Means with Constraints International Conference on Innovative Computing, Information and Control (ICICIC), 2009 Fourth Date of Conference: 7-9.

Tidke., B.A., R.G Mehta, D.P Rana, 2012."A Novel Approach for High Dimensional Data Clustering," International Journal Of Engineering Science & Advanced Technology, 2(3): 645-651.

Shanmugapriya, B. and M. Punithavalli., 2012."A Modified Projected K-Means Clustering Algorithm with Effective Distance Measure," International Journal of Computer Applications., 44(8): 32-36.

Jung-Yi Jiang., Ren-Jia Liou., and Shie-Jue Lee., 2011. "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 23: 3.

Jegatha Deborah, L., R. Baskara, A. Kannan, 2010. "A Survey on Internal Validity Measure for Cluster Validation," International Journal of Computer Science & Engineering Survey (IJCSES) 1(2).

Yun Yang., and Ke Chen., 2011. "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 23: 2.

Peter, J., Rousseeuw and Mia Hubert. Robust statistics for outlier Detection, 2011. John Wiley & Sons, Inc. WIREs Data Mining Knowledge Discovery, pp: 73-79.