ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)
On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Enhancement of Subspace Clustering For High Dimensional Data

V.R.Saraswathy[#1], Dr.N.Kasturi[*2], V.Lathika[#3]

[#1]Assistant Professor(Sl.), Department of ECE, Kongu Engineering College(Autonomous), Tamilnadu, India.
[*2]Professor, Department of ECE, Kongu Engineering College(Autonomous), Tamilnadu,India.
[#3]PG Student, Department of CSE, Kongu Engineering College(Autonomous), Tamilnadu,India.

ABSTRACT— Clustering is a technique for grouping of similar objects among different objects from the dataset. Traditional clustering algorithms are more suitable for fewer dimensions. Feature selection can be used to select the relevant features from high dimensional dataset to improve the formation of clusters. In these methods the relationship between objects is not preserved since an object cannot be a member of more than one cluster. Subspace clustering is the solution to this problem, which preserves the relationship between objects. Semi-supervised learning along with Subspace clustering uses the partial background knowledge which improves the cluster formation results. Constrained Laplacian Score is semi supervised based feature selection method for selecting the relevance feature. This method preserves the local and constraints ability among data objects. The combination of semi supervised clustering and feature selection enhances the subspace clustering process. Experimental results show that the semi supervised subspace clusters formed with semi supervised feature selection have good accuracy than the semi supervised subspace clusters formed only with the relevant dimensions.

KEYWORDS— Semi-supervised learning, Feature selection, Subspace clustering, Constrained Laplacian Score.

I. INTRODUCTION

Data mining is a process which is used to extract the information from larger databases.In our world, large volume of data is available and data mining is highly important to transform this data into meaningful patterns. Many data mining tools are available. Data mining tools can be used to predict the behaviors and future trends from large amount of data. The extracted trends and behaviors are used in business for important decision making process. Data mining tools can also be used to answer the business queries which take long time to solve manually. Some specific applications of data mining includes market basket analysis, trend analysis and fraud detection.

The important tasks in data mining are classification, clustering, and association rule mining. Among these tasks, clustering is the process of grouping similar data points into different groups such that the resulting cluster will have high intra-class similarity and low inter-class similarity. In clustering process, the similarity between data points is defined in terms of distance measure. The similar data points will be near to each other and the dissimilar data points will be far apart. Clustering is an unsupervised learning method since it does not make use of any domain knowledge. Many conventional clustering algorithms [2] are available.

All of the available clustering algorithms produce clusters of good quality when the input dataset has fewer dimensions. These clustering algorithms consider all the dimensions in the input dataset to form the clusters. The increase in volume of space increases and the data points are appears to be nearly equi-distant from each other when the number of dimensions increases. So the distance measure (used for measuring the similarity between data points) becomes meaningless. The input dataset will also contain some irrelevant dimensions.

Dimensionality reduction techniques are used to solve this problem. Dimensionality reduction is the process of removing redundant and irrelevant dimensions. Many dimensionality reduction techniques are available for unsupervised and supervised learning methods. These techniques first find the relevant dimensions to form the cluster. The relevance of a dimension can be found by using methods like dispersion measures so that the irrelevant dimensions will be removed. In the semi supervised context, Constrained Laplacian score method is used for feature selection [3].

Subspace clustering can be used to overcome the drawbacks of the dimensionality reduction techniques.

Subspace clustering is the process of detecting all clusters in all subspaces and thus it tries to preserve the relationship among objects. The resulting subspace clusters will reflect the original cluster properties. Different types of subspace clustering algorithms are available[4]. The quality of the subspace clusters can be further improved by using learning methods. There are three types of learning methods: supervised, semi-supervised and unsupervised. So in order to acquire the domain knowledge semi supervised learning methodology[6] is used. Semi-supervised learning method guides the subspace clustering process by providing slight domain knowledge either in the form of class instances or constraints. There are several methods which makes use of semi-supervised learning method to the process of dimensionality reduction[5].

When the semi-supervised subspace clustering algorithms searches for subspaces in a high dimensional dataset, the quality of resulting subspace clusters will be reduced significantly. This paper discusses a method in which the features are selected based on semi supervised method. A little amount of domain knowledge will be given in the form of constraints [9] to guide the subspace clustering process. Traditional clustering algorithm is applied for the resulting subspaces. This approach improves the quality of the resulting subspace clusters.

The rest of the paper is organized as follows. In section 2, a survey of the existing works is done. Section 3 describes the methodology for semi supervised subspace clustering with semi supervised feature selection. Experimental results are discussed in section 4. Section 5 concludes the paper work.

II. LITERATURE SURVEY

A. Semi-Supervised Clustering

Semi-supervised learning method guides the subspace clustering process by providing little amount of domain knowledge either in the form of class labels or constraints. Generally, two types of semi-supervised clustering algorithms are available [2]. They are: 1) similarity-adapting based algorithms and 2) Search-based algorithms. In [12], it was suggested that the similarity-adapting based algorithms will make an assumption that the target classification will not be reflected correctly by the similarity measures taken initially. Thus the initial similarity measures should be adapted so that the traditional clustering algorithms can be applied and the constraints will be satisfied. In [11],they proposed a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields (HMRF). This model combines constraints and Euclidean distance learning and allows the use of a broad range of clustering distortion measures, including Bregman divergences and directional similarity measures. This method performs partitional semi-supervised clustering of data by minimizing an objective function derived from the posterior energy of the HMRF model.

Search-based algorithms [2] are based on the assumption that the similarity measures used between the data objects provide useful information. This type of algorithms modifies the clustering process to achieve the clustering performance. In [13], they proposed a method which explores the use of labeled data to generate initial seed clusters, and also it uses constraints generated from labeled data to guide the clustering process. This method introduces two semi-supervised constraints such as must-link and cannot-link constraints are used for K-Means clustering algorithm that can be viewed as instances of the Expectation Maximization (EM) algorithm, where labeled data provides prior information about the conditional distributions of hidden labels.

B. Semi-Supervised Subspace Clustering

In [16], the proposed algorithm can accurately identify projected clusters with relevant dimensions. The algorithm makes use of a robust objective function that combines object clustering and dimension selection into a single optimization problem. This method can also make use of domain knowledge in the form of labeled objects and labeled dimensions to improve its clustering accuracy. In [14],they proposed a semi-supervised impurity based clustering method in conjunction with k-Nearest Neighbor approach. This algorithm is based on semi-supervised subspace clustering that considers the high dimensionality as well as the sparse nature in text data. This algorithm finds clusters in the subspaces of the high dimensional text data where each text document has fuzzy cluster membership. This type of clustering approach exploits two factors, namely, chi square statistic of the dimensions and the impurity measure within each cluster. In [16], they proposed a methodwhich uses pair-wise constraints for the problem of subspace clustering. This method extends the framework of bottom-up subspace clustering algorithms by integrating background knowledge which is given in the form of instance-level constraints. The algorithm can be applied to both density and distance-based bottom-up subspace clustering techniques.

In [1], the proposed method exploits constraint inconsistence for selecting correct dimensions for subspaces. The algorithm first computes the correlation of each constraint to each dimension. Based on the correlation values, the algorithm computes the support degree of constraints to each other. The constraints are combined with other constraints which have maximum support degree to it. For each constraint union, correlated dimensions are found and corresponding subspaces are formed.

C. Feature Selection for high dimensional data

Feature selection methods are classified into two methods such as filter and wrapper. A wrapper method selects the feature search around the learning algorithms that will ultimately be applied and utilizes the learned the results to select the features. A filter method uses the intrinsic property of data. Filter methods are commonly used for feature selection process. The proposed filter in [17] is based feature selection methods for high

dimensional data which exploits the idea that feature relevance is proportional to its dispersion.

Constrained LaplacianScore proposed in [10], enhances the semi supervised constraint context. Laplacian score [7] investigates the variance of data in order to assess the locality preserving ability of features. In a semi supervised context, the unlabeled part is normally larger than the labeled one. The non−treatment of unlabeled data in such a case may mislead the learning process. In addition, choosing constraint subset is still a problematic issue, which could degrade the performance of the feature selection process. So to negotiate this degradation constraint score [8] is used which based on constraint preserving ability. In [15], the proposed algorithm includes the feature selection for high dimensional data and semi supervised subspace clustering. They proposing the mean-median based technique for selecting the features and semi supervised subspace clustering for cluster formation.

III. SEMI SUPERVISED SUBSPACE CLUSTERING WITH CONSTRAINED LAPLACIAN SCORE

The semi supervised subspace clustering with constrained Laplacian score is proposed here and the explanation is as follows:

A. Basic Framework

The basic framework of the proposed approach is given in Fig 1. The dataset with highest number of dimensions is reduced to dataset with few dimensions by using semi supervised feature selection. The dataset with reduced number of dimensions is used for forming the subspace clusters.

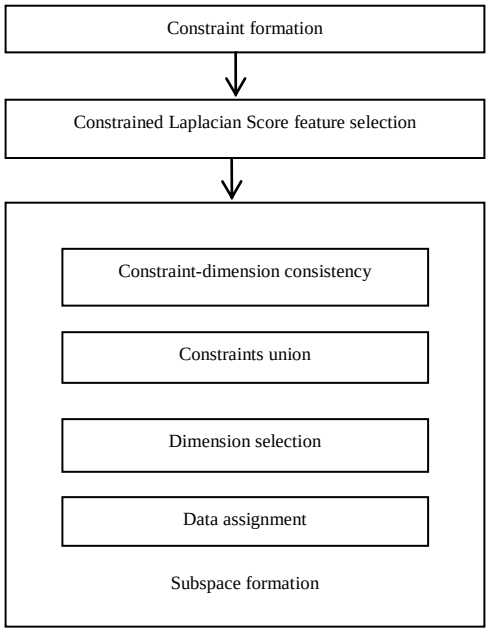

Fig. 1 Subspace formation based on semi supervised feature selection

B. Constraint Generation

The labeled part of data is used for taking the constraints formation and the pairwise constraints are as follows:

- Must-link constraints[1] are formed by taking two points from the same cluster and making it as a single pair of points.
- Cannot-link constraints[1] are formed by taking two points from different clusters and making it as a single pair of points. This process of forming the cluster is repeated until the predetermined numbers of constraints are found.

$\bar{M} = \{(x_i, x_j) | \ i,j \in (1,2,\dots,n_m, i \neq j\}$ denotes the set of must-links, where (x_i, x_j) represents a must-link between x_i and x_j. $\bar{C} = \{\langle x_i, x_j \rangle | \ i,j \in (1,2,\dots,n_c, i \neq j\}$ denotes the set of cannot-links, where $\langle x_i, x_j \rangle$ represents a cannot-link between x_i and x_j. T is used to denote a constraint when there is no need to distinguish whether it is a must-link or cannot-link constraint. Both must-link and cannot-link constraints are referred as pair-wise constraints. These constraints are given as the input to the constraint-dimension consistency module.

C. Semi supervised feature selection

In semi supervised learning, a data set of N data points $X = \{x_1, \dots, x_n\}$ consists of two subsets depending on label availability: $X_L = (x_1, \dots, x_l)$ for which the labels $Y_L = (y_1, \dots, y_l)$ are provided, and $X_U = (xl+_1, \dots, x_{l+u})$ whose labels are not given. Here,a data point x_i is a vector with m dimensions (features), and $y_i \in \{1,2,\dots,C\}$ (C is the number of different labels) and $l + u = N$ (N is the number of instances). Let $F_1, F_2, \dots, F_m$ denote the m features of X and $f_1, f_2, \dots, f_m$ be the corresponding feature vectors that record the feature value on each instance.Semi supervised feature selection and both X_L and X_U to identify the set of most relevant features $f_{j1}, f_{j2}, \dots, f_{jk}$ of the target concept, where $k \leq m$ and $j_r \in \{1,2,\dots,m\}$ for $r \in \{1,2,\dots,k\}$.

The original dataset with large number of dimensions is given as input to the feature selection module. The original dataset may have some irrelevant and redundant dimensions. These irrelevant dimensions may produce some noise in the resulting clusters. Feature selection is the task of removing irrelevant and redundant dimensions. Feature selection chooses a subset of dimensions and it helps to improve the speed of learning and predictive accuracy. The output of this module will be the dataset with reduced dimensions.

For pairwise constraint feature selection, the combination of Laplacian and Constrained Score is used which dramatically bias the selection for the features having labeled part of data and the unlabeled part.The defined constrained Laplacian score φ_r as follows:

$$\varphi_r = \frac{\sum_{i,j}(f_{ri}-f_{rj})^2(S_{ij}+N_{ij})}{\sum_{i,j}(f_{ri}-\alpha_{ij}^i)^2 D_{ii}} \quad (1)$$

where

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i-x_j\|^2}{\lambda}} & if\ x_i\ and\ x_j \\ & are\ neighbors \\ 0 & otherwise \end{cases} \quad (2)$$

and

$$\alpha_{rj}^i = \begin{cases} f_{rj} & if\ (x_i, x_j) \in \bar{M} \\ \mu_r & if\ i = j\ and\ x_i \in X_U \\ f_{ri} & otherwise \end{cases} (3)$$

and

$$N_{ij} = \begin{cases} e^{-\frac{\|x_i-x_j\|^2}{\lambda}} & if\ x_i and\ x_j\ are\ neighbors \\ & and\ (x_i, x_j) \in \bar{M} \\ (e^{-\frac{\|x_i-x_j\|^2}{\lambda}})^2 & if\ [x_i and\ x_j\ are\ neighbors \\ & and\ (x_i, x_j) \in \bar{C}]\ or \\ & [x_i and\ x_j\ are\ not\ neighbors \\ & and\ (x_i, x_j) \in \bar{M}] \\ 0 & otherwise \end{cases} \quad (4)$$

If there are no labels $L = 0, X = X_U$ then $\varphi_r = L_r$ and if $U = 0, X = X_L$ then φ_r represents an adjusted L_r, where the $\bar{M}$ and $\bar{C}$ information would be weighted by $(S_{ij} + N_{ij})$ and D_{ii} respectively in the formula, this weighting focuses on features where the neighborhood is inconsistent with the constraints. With φscore, on the one hand, a relevant feature should be the one in which those two samples, neighbors or related by a $\bar{M}$ constraint, are close to each other. On the other hand, the relevant feature should be the one with a larger variance or in which those two samples, related by a $\bar{C}$ constraint, are well separated.

D. Subspace Formation

a)Constraint-Dimension Consistency

In this module, the dimensions (features) in which the constraints are consistent need to be determined. The dimensions in which the constraints become consistent is determined based on the methodology that in a consistent dimension[1] , the two data points that are placed on a must-link constraint should have a large number of nearest neighbors in common. Also the two data points that are placed on a cannot-link constraint should not have any nearest neighbor in common or the two data points can have only a few nearest neighbors in common. The consistency between the constraints and the dimensions can be determined by calculating the number of nearest neighbors that the two data points (placed on the must-link or cannot-link constraint) have in common. Eq. (5) shows the formula[1] to calculate the number of nearest neighbors that the two data points (placed on the must-link or cannot-link constraint) have in common. (N_K^D represents the number of k nearest neighbors of x_i in dimension D and also x_i and x_j denotes the two data points that are placed on a must-link or cannot-link constraint).

$$Sim_{i,j}^D = \frac{|N_k^D(x_i) \cap N_k^D(x_j)|}{k} \quad (5)$$

Based on the number of common nearest neighbors that are shared between the two data points , the value of correlation between each of the constraint in each of the dimension can be calculated. The values are stored in a correlation matrix R.

- The value of the correlation matrix is 0, if following condition ,
  $(T_i \in \bar{M}) and\ \left(Sim_{T_i}^{D_j} < \propto\right)\ or (T_i \in \bar{C})\ and\ \left(Sim_{T_i}^{D_j} < \propto\right) is\ satisfied$[1].
- The value of correlation matrix is 1 if the following condition,
  $(T_i \in \bar{M}) and\ \left(Sim_{T_i}^{D_j} \geq \propto\right)\ or (T_i \in \bar{C})\ and\ \left(Sim_{T_i}^{D_j} < \propto\right) is\ satisfied$[1].

where, T_i represents a constraint which can be either must-link or cannot-link, $\bar{M}$ represents the must-link constraints, $\bar{C}$ represents the cannot-link constraints, $0 < \propto < 1$ is the threshold parameter.

b)Constraints Union

In this module, the constraints (must-link or cannot-link) that are sharing large number of consistent dimensions in common are united. Each one of the constraint union will denote a subspace. An effective method known as support degree [1] can be used for finding the constraints which share a large number of consistent dimensions in common. The following equation gives the formula for calculating the support of constraint T_i by constraint T_j. The support degree [1] is given by,

$$Sup_{i,j} = \frac{|D_{T_i} \cap D_{T_j}|}{|D_{T_i}|} \quad (6)$$

where D_T is used to denote the set of dimensions in which the constraint T is consistent. Every constraint is combined with every other constraint that has the maximum support degrees to it. This method is performed to find the constraint unions.

c) Dimension Selection

For each of the constraint union formed from above step, the consistent dimensions[1] should be found. Based upon the constraint union and its corresponding dimensions, the subspaces can be formed. By analyzing the correlation between the dimension and constraints, three types of dimensions can be found. They are:

- Backbone dimension: The correlation value which can be computed between the backbone dimension and each of the constraint in the constraint union will be 1. This backbone dimension must be added to the subspace that is used to denote the corresponding constraint union.
- Unrelated dimension: The correlation value which can be computed between the unrelated

dimension and each of the constraint in the constraint union will be 0. This unrelated dimension should not be used to form the subspace that corresponds to the constraint union.

- Uncertain dimension: The correlation value which can be computed between the backbone dimension and each of the constraint in the constraint union is either 0 or 1. We should decide upon adding this dimension to a constraint union. It is decided by calculating the difference between the mean distance of cannot-link constraints and the mean distance of must-link constraints[1].

d) Data Assignment

The subspaces corresponding to each constraint union will be the output from the above module. The point which lies on the must-link or cannot-link constraints will be assigned to the corresponding clusters and they form the initial cluster members. These initial cluster members can be used to form the cluster centroids. The points other than those lying on the constraints will be assigned to the cluster whose centroid is nearer to the data object. The result of this module is the subspace cluster with its members.

IV. EXPERIMENTAL RESULTS

A. Dataset

The dataset for experiments are taken from the UCI repository. The sonar dataset in the UCI repository is taken for analysis. The sonar dataset has 60 attributes and 208 instances. The instances in the dataset belong to two groups. The dataset contains patterns obtained by bouncing the sonar signal from the mine or rock. Based on the values of the attributes, the instances should be grouped as rock (the signal which bounces from rock) or mine (the signal which bounces from mine).

The ionosphere dataset from UCI repository is also taken for analysis. The dataset contains 351 instances and 34 attributes. It contains two types of class labels such as good and bad.

B. Parameters for Evaluation

F1-value is used to access the quality of resulting semi supervised subspace clusters. F1-value is the popular evaluation criteria for the top-down semi supervised subspace clustering algorithms. Precision and recall are used to calculate the F1-value as follows:

$$F1 - value = \frac{2 \times precision \times recall}{precision + recall} \quad (7)$$

Precision is the ratio of the data points correctly predicted in subspace $cluster_i$to the total data points predicted in subspace $cluster_i$ and recall is the ratio of the data points correctly predicted in subspace $cluster_i$ to the total data points in original $cluster_i$.

C. Experimental Settings

MATLAB tool is used to implement the concept. MATLAB is a numerical computing environment and it allows for matrix manipulations. Excel is used to store the output obtained from the coding of MATLAB. The excel files can be imported in MATLAB.

D.Experimental Analysis

The F1−value is calculated for Sonar dataset by varying the number of constraints. Sonar and Ionosphere datasets are taken from the UCI machine learning repository. Table 4.1 shows the semi supervised subspace clustering performances. The Table 4.2 is shows the mean−median (MM) based semi supervised subspace clustering. Table 4.3 represents the constrained Laplacian score (CLS) based semi supervised subspace clustering. By comparing both the results the Constrained Laplacian score gives the better result. Table 4.4 represents the performance analysis of semi supervised subspace clustering with mean-median and Constrained Laplacian score based feature selection for Ionosphere dataset.

The following figures show the result comparison of both Ionosphere dataset and Sonar dataset. Fig. 4.2 represents the performance analysis for Sonar dataset and fig. 4.3 shows the performance comparison for Ionosphere dataset. The figures show the enhancement of semi supervised subspace clustering by using the Constrained Laplacian score.

Number of constraints	Precision S3C	Recall S3C	F1 value S3C
10	0.44	0.60	0.51
20	0.47	0.66	0.55
30	0.49	0.67	0.56
40	0.51	0.70	0.58
50	0.54	0.75	0.63

Table 4.1 Performance Analysis of Semi Supervised Subspace Clustering for Sonar Dataset

Number of constraints	Precision MM based S3C	Recall MM based S3C	F1 value MM based S3C
10	0.55	0.76	0.58
20	0.59	0.81	0.68
30	0.60	0.83	0.70
40	0.62	0.85	0.72
50	0.70	0.96	0.81

Table 4.2 Performance Analysis of Mean Median based Semi Supervised Subspace Clustering for Sonar Dataset

Number of constraints	Precision CLS based S3C	Recall CLS based S3C	F1 value CLS based S3C
10	0.65	0.72	0.70
20	0.66	0.77	0.72
30	0.71	0.82	0.76
40	0.73	0.85	0.79
50	0.79	0.95	0.86

Table 4.3 Performance Analysis of Constrained Laplacian Score Semi Supervised Subspace Clustering for Sonar Dataset

Number of constraints	F1-value of S3C	F1-value of MM based S3C	F1-value of CLS based S3C
10	0.51	0.53	0.55
20	0.55	0.58	0.57
30	0.58	0.60	0.63
40	0.62	0.63	0.65
50	0.64	0.67	0.68

Table 4.4 Performance Analysis of Semi Supervised Subspace Clustering with MM and CLS based feature selection for Ionosphere Dataset

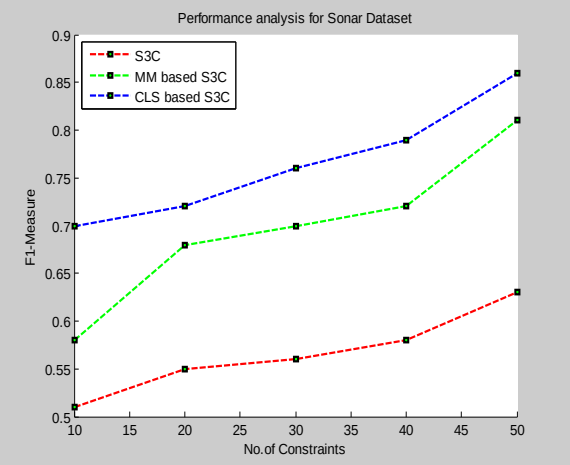

Figure 4.2 Performance Analysis of Semi Supervised Subspace Clustering for Sonar Dataset

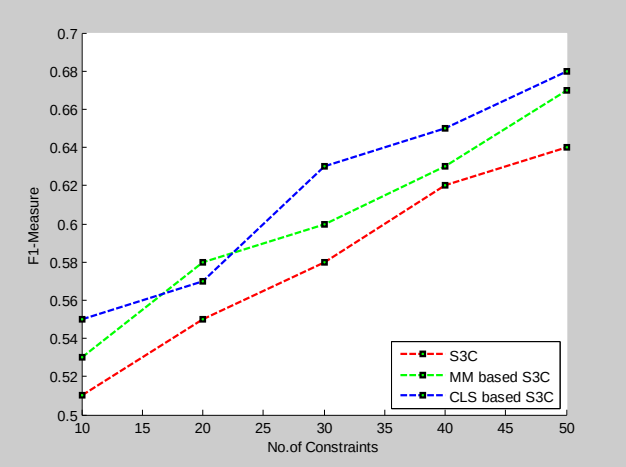

Fig. 4.3 Performance Analysis for Ionosphere Dataset

Two types of feature selection are taken for comparing the results such as ratio of mean median (MM) and constrained Laplacian score with semi supervised subspace clustering (S3C). By analyzing the F1−values, it can be found that the proposed algorithms produce more accurate results with the Sonar and Ionosphere datasets. It can be inferred that the proposed algorithms are suitable for the dataset with high number of dimensions.

V.CONCLUSION

In high dimensional data, traditional clustering algorithms produce poor quality clusters due to the curse of dimensionality. Dimensionality reduction techniques tend to remove the redundant and irrelevant dimensions, but they fail to preserve the relationship among objects. Subspace clustering finds clusters in subspaces and they preserve the relationship between objects. But searching for subspaces by considering all the input dimensions will decrease the accuracy of the resulting clusters. Semi supervised approach uses pairwise constraints to form the subspace.

The feature selection based on mean median measure and constrained Laplacian score are done for Sonar and Ionosphere data sets. The advantage of Laplacian score is that it selects the features with respect to intrinsic data behavior.Furthermore, for the constraint score, the principle is mainly based on the constraint preserving ability. This small supervision information is certainly necessary for feature selection, but not sufficient when ignoring the unlabeled data party especially if it is very large. Hence the constrained Laplacian score method is used for efficient feature selection compared to mean median measure. The relevant features are reduced and are clustered based on semi supervised subspace clustering approach.

REFERENCES

[1] X.Zhang, Y.Qiu, Y.Wu, et al., "*Exploiting constraint inconsistence for dimension selection in subspace clustering: A semi-supervised approach"*,2011, pp.3598-3608.

[2] P.Kerkhin, "*Survey of clustering data mining techniques"*, 2006, pp.25-71.

[3] Khalid B, Hindawi M (2013), "*Efficient Semi supervised Feature Selection: Constraint, Relevance and Redundancy"*, IEEE Transactions on Knowledge and Engineering.

[4] L. Parsons, E. Haque, H. Liu, "*Subspace clustering for high dimensional data: a review"*, SIGKDD Explorations (2004) 90–105.

[5] H. Cevilkalp, J. Verbeek, F. Jurie, et al., "*Semi-supervised dimensionality reduction using pairwise equivalence constraints"*, Proceedings of CVTA, 2008.

[6] E. Fromont, A. Prado, C. Robardet, "*Constraint based subspace clustering"*, Proceedings of SDM, 2009.

[7] X. He, D.Cai, N. Partha, "*Laplacian score for feature selection"', in* Advances in Neural Information Processing Systems, 17, 2005.

[8] D. Zhang,'Constraint score: "*A new filter method for feature selection with pairwise constraints"* Pattern Recognition, vol. 41(5), pp. 1440–1451, 2008.

[9] K. Wagstaff, C. Cardie, "*Clustering with instance-level constraints",* Proceedings of ICML, pp. 1103–1110, 2000.

[10] K. Benabdeslem, M. Hindawi, "*Constrained laplacian score for semi-supervised feature selection" in* Proceedings of ECML-PKDD conference, pp. 204–218, 2011.

[11] S.Basu, A.Banerjee, R.Mooney, "*Aprobabilistic framework for semi-supervised clustering",* proceedings of KDD, 2004, pp. 59-68.

[12] S.Basu, A.Banerjee, R.Mooney, "*Semi-supervised by seeding",* proceedings of ICML, 2002, pp. 19-26.

[13] Y.Kevin, W.David, K.Michael, et al., "*On discovery of extremely low dimensional clusters using semi-supervised projected clustering",* proceedings of IEEE ICDE, 2005, pp. 329-340.

[14] M.Ahmed, L.Khan, *SISC: "A text classification approach using semi supervised subspace clustering",* proceedings of IEEE ICDM, 2009, pp.1-6.

[15] V. R. Saraswathy, N. Kasthuri and M. Revathi., "*Feature Selection based Semi-Supervised Subspace Clustering', IJCA Proceedings"* on Amrita International Conference of Women in Computing, AICWIC(4):10-14, January 2013.

[16] E.Fromont, A.Prado, C.Robardet, "*Constraint based subspace clustering, in proceedings of SDM*", 2009.

[17] Ferreira, A.J., & Figueiredo, M.A, *"Efficient feature selection filters for high-dimensional data",* 2012.