Journal of Engineering Sciences

Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

Subspace Metric Ensembles for Semi-supervised Clustering of High Dimensional Data

M. Pavithra[1], Dr.R.M.S.Parvathi[2]

Assistant Professor, Department of CSE, Jansons Institute of Technology, Coimbatore, India[1].

Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, India [2].

ABSTRACT:

A critical problem in clustering research is the definition of a proper metric to measure distances between points. Semi-supervised clustering uses the information provided by the user, usually defined in terms of constraints, to guide the search of clusters. Learning effective metrics using constraints in high dimensional spaces remains an open challenge [2]. This is because the number of parameters to be estimated is quadratic in the number of dimensions, and we seldom have enough side information to achieve accurate estimates. We address the high dimensionality problem by learning an ensemble of subspace metrics. This is achieved by projecting the data and the constraints in multiple subspaces, and by learning positive semi-definite similarity matrices therein. This methodology allows leveraging the given side-information while solving lower dimensional problems. We demonstrate experimentally using high dimensional data (e.g., microarray data) the superior accuracy achieved by our method with respect to competitive approaches [3]. It addresses the problem of supervised clustering with multi-view data of high dimensionality. We propose a new algorithm which learns discriminative subspaces in a supervised fashion based upon the assumption that a reliable clustering should assign same-class samples to the same cluster in each view. The framework combines the simplicity of k-means clustering and Linear Discriminant Analysis (LDA) within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another [4]. The effectiveness of the proposed algorithm is demonstrated empirically under scenarios where the conditional independence assumption is either fully satisfied (audio-visual speaker clustering) or only partially satisfied (handwritten digit clustering and document clustering). Significant improvements over alternative multi-view clustering approaches are reported in both cases. The new algorithm is flexible and can be readily adapted to use different distance measures, semi-supervised learning, and non-linear problems. Recently, both semi-supervised clustering and cluster ensemble have received tremendous attention due to their accurate and reliable performance [5]. There are mainly two kinds of existing semi-supervised clustering algorithms called constraint-based and metric-based. In this paper, we present a semi-supervised clustering ensemble approach which takes both pairwise constraints and metric measure into account [6]. Firstly, under the assistance of supervised information included pairwise constraints and labeled data, the approach generates different base clustering partitions respectively using constraint-based semi-supervised clustering and metric-based semi-supervised clustering, in which the latter develops a new metric function [7].

KEYWORDS:

High dimensional data semi-supervised clustering subspace metric ensemble competitive approach high dimensionality problem dimensional problem critical problem enough side information effective metric positive semi-definite similarity matrix subspace metric open challenge multiple subspace superior accuracy accurate estimate microarray data high dimensional space.

I. INTRODUCTION

Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning. The clustering problem concerns the discovery of homogeneous groups of data

according to a certain similarity measure, such that data in a cluster are more similar to each other than data assigned to different clusters. The definition of a proper similarity measure is a difficult problem that lies at the core of the field of machine learning. The structure of the groups discovered in the data by a given clustering technique strongly depends on the similarity measure used. Data mining adds to clustering the complication of large data sets with high dimensionality. Large amounts of unlabeled data are available in real-life data mining tasks, e.g., unlabeled messages in an automated email classification system, or genes of unknown functions in microarray data. This imposes unique computational requirements on clustering algorithms. Furthermore, the sparsity of the data in high dimensional spaces can severely compromise the ability of discovering meaningful clustering solutions. Recently, semi-supervised clustering has become a topic of significant research interest. While labeled data are often limited and expensive to generate, in many cases it is relatively easy for the user to provide pairs of similar or dissimilar examples. Semi-supervised clustering uses a small amount of supervised data, usually under the form of pairwise constraints on some instances, to aid unsupervised learning. The main approaches for semi-supervised clustering can be basically categorized into two general methods: constrained-based [2, 5, 3, 4] and metric-based [4, 8, 2]

However, when facing high dimensional data, learning an effective metric with limited supervision remains an open challenge. The reason is that the number of parameters to be estimated is quadratic in the number of dimensions, and we seldom have enough side-information to achieve accurate estimates. For example, in our experiments we deal with microarray data with 4026 dimensions and less those 100 samples (a typical scenario with these kinds of data). Learning a similarity metric with a limited number of constraints becomes a very hard problem under these conditions due to the large parameter space to be searched. In this paper, we address the high dimensionality problem by learning an ensemble of subspace metrics. This is achieved by projecting the data and the pairwise constraints in multiple subspaces and by learning positive semi-definite similarity matrices therein [4]. This methodology allows leveraging the given side-information while solving lower dimensional problems. The diverse clustering discovered within the subspaces are then combined by means of a graph based consensus function that leverages the common structure shared by the multiple clustering results [9]. Our experimental results show the superior accuracy achieved by our method with respect to competitive approaches, which learn the metric in the full dimensional space.

It presents our efforts to address the problems of high dimensionality and multi-modal fusion in a unified framework. We assume that each data sample is represented by two feature vectors corresponding to two independent views. We further assume significant information in each feature vector to be unrelated to the underlying class label and that there exists a lower dimensional subspace in which classes are maximally separated [3]. Inspired by the concept of co-training, we describe a new multi-view subspace clustering algorithm which reflects the intuition that a true underlying clustering should assign samples to the same cluster irrespective of the view. It seeks a discriminant subspace for each view which results in a clustering policy with maximal agreement across views [4]. Discriminant subspaces in one view are learned using cluster labels for the same samples in another view, and vice versa. The process is iterative and is repeated until a maximum agreement is achieved [5]. The proposed algorithm simultaneously outputs cluster indicators, discriminant subspaces for each view, and compact models of different clusters. As a result, the algorithm copes naturally with out-of-sample data and is readily extended to semi-supervised classification [6].

II. RELATED WORK

Semi-supervised clustering ensemble applies these two strategies simultaneously, namely semi-supervised clustering and cluster ensemble. Similarly, it combines different clustering results of various semi-supervised clustering algorithms by using ensemble function to create a single target clustering with more optimal performance than those of individual semi-supervised clustering. What's more, it has strengths of low sensitivity to noise, outliers

and variables. It [5] proposes a feature selection method based semi-supervised cluster ensemble framework for tumor clustering from bio-molecular data. It [6] proposes an incremental semi-supervised clustering ensemble framework for high dimensional data clustering. At present, two typical approaches of semi-supervised clustering called constraint-based and metric-based are researched a lot. The former revises objective function of algorithm to guide the process of clustering by using supervised information provided in advance. It [7] studies the active learning problem of selecting pairwise constraints for semi-supervised clustering. Wang et al. [8] propose a semi-supervised nonnegative matrix factorization method with pairwise constraints. The latter exploits a specific distance/similarity metric for clustering to satisfy the given pairwise constraints. It [9] proposes a semi supervised clustering method with multi-viewpoint based similarity measure. It [10] develops a semi-supervised fuzzy clustering algorithm with metric learning and entropy regularization simultaneously (SMUC). Although the two kinds of methods have their own singular focus respectively, they aren't not only separated completely, but also exists symbiotic relationship between them. Inspired by the work of. [11], many scholars have begun to turn their attention to the field of exploitation of hybrid approaches, which aims to combine the advantages of constraint-based with that of metric-based. In order to sufficiently solve the violation problem of pairwise constraints and to mitigate the problem of manually tuning the kernel parameters owing to the fact that no sufficient supervision, an adaptive semi-supervised clustering kernel method based on metric learning (SCKMM) is proposed by [12]. It [5] presents an extension of soft-constraint semi-supervised affinity propagation (E-SCSSAP) which incorporates metric learning in the optimization objective and acquires desirable clusters. Firstly, this paper improves the performance of semi supervised clustering by introducing ensemble mechanism, which unites different results respectively produced from constraint-based algorithm and metric-based algorithm. They have certain preoccupations in their fields that one concentrates on the adjustment of objective function according to pairwise constraints while the other is concerned with the introduction of metric function to measure the distance/similarity between samples more precisely. The combination method obtains benefits beyond what a single algorithm achieves.

The second one is cluster ensemble learning approach. From the generation perspective, a set of diverse ensemble members is generated and known as base clustering in various forms. Base clustering can result from different views, different initialization parameter, different methods, and so on. But then it is difficult to learn a suitable consensus function to summarize the base clustering's and search for an optimal unified clustering decision. In light of that theoretical analysis, a great amount of well-known ensemble methods emerge. Three graph-based ensemble methods are introduced in study [2], all of which partition clusters based on a constructed similarity graph. The cluster-based similarity partition algorithm (CSPA) uses METIS to partition the induced similarity graph. The hyper-graph partition algorithm (HGPA) uses HMETIS to partition the hyper-graph. The meta-clustering algorithm (MCLA) collapses related hyper-edges and assigns each object to the collapsed hyper-edge in which it participates most strongly. An iterative voting consensus (IVC) [4] is a feature-based approach, in which each base clustering provides a cluster label as a new feature describing each data point that is utilized to formulate the final solution. By exploiting the significance of attribute defined in rough set theory, it [5] applies the proposed two feature selection algorithms to a cluster ensemble selection problem. The third one is semi-supervised clustering ensemble learning approach. Due to the lack of prior knowledge about cluster labels, cluster ensemble is still a challenging problem. By leveraging limited supervision information in cluster ensemble, semi-supervised clustering ensemble offers an effective solution to overcome this limitation and obtains accurate, robust and stable results. It [6] construct semi-supervised cluster ensemble based on binary similarity matrix (BSMSCE), which takes the strengths of known information to improve the quality of clustering. It [7] analyze convergence of semi supervised clustering ensemble and proposed a new relabeling approach for semi-supervised clustering ensemble by majority voting (we called it MVSCE for short). It [5] views the expert's knowledge as constraints in the process of clustering and proposes a framework called FSSSCE, which not

only applies the feature selection technique to perform gene selection on the gene dimension.

III. CLUSTER ENSEMBLES

In an effort to achieve improved classifier accuracy, extensive research has been conducted in classifier ensembles. Recently, cluster ensembles have emerged. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature. In fact, it is well known that off-the-shelf clustering methods may discover very different structures in a given set of data. This is because each clustering algorithm has its own bias resulting from the optimization of different criteria [6]. Cluster ensembles can provide robust and stable solutions by leveraging the consensus across multiple clustering results, while averaging out spurious structures due to the various biases to which each participating algorithm is tuned. In the following we formally define the clustering ensemble problem. [7] Consider a set of data X = {x1, x2, ⋯ , xN }. A clustering ensemble is a collection of S clustering solutions: C = {C1, C2, ⋯ , CS}. Each clustering solution Cl, for l = 1, ⋯ , S, is a partition of the set X, i.e. Cl = {C1 l , C2 l , ⋯ , CKl l }, where K CK l = X. Given a collection of clustering solutions C and the desired number of clusters K, the objective is to combine the different clustering solutions and compute a new partition of X into K disjoint clusters. Different methods have been introduced in the literature to solve the clustering ensemble problem. The techniques presented in [10] compute a matrix of similarities between pairs of points, and then perform agglomerative clustering to generate a final clustering solution. In [2, 1] the authors introduce new features to describe the data, and apply K-means and EM to output the final clustering solutions. Recently, several approaches have modeled the clustering ensemble problem as a graph partitioning problem [7, 2]. In the following, we provide the necessary definitions of graph partitioning

IV. PROPOSED WORK

IV a SUBSPACE METRIC CLUSTER ENSEMBLE ALGORITHM (SMCEA)

A limited number of pairwise constraints may not be effective for learning a distance metric in high dimensional spaces due to the large parameter space to be searched. We tackle this issue by reducing the given high dimensional problem with fixed supervision into a number of smaller problems, for which the dimensionality is reduced while the amount of supervision is unchanged [8]. To achieve this goal, we utilize and leverage the paradigm of learning with ensembles. It is well known that the effectiveness of an ensemble of learners depends on both the accuracy and diversity of the individual components [2]. A good accuracy-diversity trade-off must be achieved to obtain a consensus solution that is superior to the components. Our method generates accurate learners by assigning each of them a problem of lower dimensionality, and, at the same time, by providing each of them the entire amount of constraints. Furthermore, diversity is guaranteed by providing the learners different views (or projections) of the data. Since such views are generated randomly from a (typically) large pool of dimensions, it is highly likely that each learner receives a different perspective of the data, which leads to the discovery of diverse (and complementary) structures within the data. The experimental results presented in this paper corroborate the motivation behind our approach [9]. The details of our subspace metric ensemble algorithm follow. We are given a set X of data in the D dimensional space, a set of must link constraints ML, and a set of cannot-link constraints CL. We assume that the desired number of clusters to be discovered in X is fixed to K. We reduce a D dimensional semi-supervised clustering problem into a number (S) of semi-supervised clustering problems of reduced dimensionality F [10]. To this end we draw S random samples of F features from the original D dimensional feature space. Moreover, for each must-link constraint (xi, xj) ∈ ML, we generate the projected must-link constraints (xi, xj)Fl , for l = 1, ⋯ , S. This gives new S sets of must-link constraints: MLF1 , ⋯ ,MLFS . Similarly, for each cannot link constraint (xi, xj) ∈ CL, we generate the projected cannot-link constraints (xi, xj)Fl , for l = 1,

⋯ , S. This results in S new sets of cannot-link constraints: CLF1 , ⋯ ,CLFS .

Algorithm: *Subspace metric cluster ensemble*
Input: X, ML, CL, K, number of features F, ensemble size S
Output: Partition of X into K clusters
Method:

1. Generate S subspaces $F_1, F_2, \cdots, F_S$ by random sampling F features without replacement from the D-dimensional original space.
2. For each constraint $(\mathbf{x}_i, \mathbf{x}_j) \in ML$, generate the projected constraints $(\mathbf{x}_i, \mathbf{x}_j)|_{F_l}$, for $l = 1, \cdots, S$; this gives the new S sets $ML_{F_1}, \cdots, ML_{F_S}$. Likewise, generate new S sets of cannot-link constraints $CL_{F_1}, \cdots, CL_{F_S}$.
3. Learn matrix A_l in subspace F_l, using the corresponding sets of constraints ML_{F_l} and CL_{F_l}, for $l = 1, \cdots, S$, according to the method presented in [24].
4. Cluster data X in each subspace F_l with K-means, using the metric d_{A_l} and the number of clusters fixed to K. This gives an ensemble of S clusterings.
5. Use the HBGF algorithm [9] to construct the bipartite graph $G = (V, E)$ from the resulting S clusterings.
6. Use spectral graph partitioning to obtain the final clustering result.

IV b MULTI-VIEW SUBSPACE CLUSTERING (MSC)

A co-training algorithm in this section, we apply the concept of co-training to the problem of discriminant subspace learning for multi-view clustering [3]. Since we assume unsupervised clustering, the standard semi supervised co-training algorithm cannot be applied directly. However, the goal remains the same, i.e. to learn a subspace for each view which results in a common clustering policy [4]. For clarity, samples assigned to the same cluster in the subspace of one view should be assigned to the same cluster in the subspace of the other view and, conversely, samples assigned to different clusters in the subspace of one view should be assigned to different clusters in the subspace of the other view [5].

$$CAI(H^{(1)}, H^{(2)}) = \frac{1}{n}\sum_{i=1}^{n}\delta\left(h_i^{(1)}, \mathrm{map}(h_i^{(2)})\right)$$

$$H^{(v)} = \arg\min_{H^{(v)}} \sum_{k=1}^{K}\sum_{h_i^{(v)}=k} \|P^{(v)^T}\mathbf{x}_i - P^{(v)^T}\mathbf{m}_k\|^2 \quad (v = 1, 2)$$

Algorithm 1. CoKmLDA

Input: a set of n multi-view samples $X = \{X^{(v)} | v = 1, 2\}$, where $X^{(v)} = \{x_1^{(1)}, \ldots, x_n^{(v)}\}$, and the expected number of clusters K.
Output: view dependent cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$, and projection matrices $P^{(1)}, P^{(2)}$
Initialize:
1. Center the feature vectors in each view and apply PCA if the dimensionality of the feature space is too high;
2. Perform k-means clustering in each view to estimate cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$;
3. For each view v, identify the single sample closest to each of the K clusters, $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.

for $t = 1$ **to** *iter* **do**
  for $v = 1$ **to** 2 **do**
  1. Use $X^{(v)}$ and $H^{(3-v)}$ to train LDA projections $P^{(v)}$ and project samples into the LDA subspace;
  2. Using seeds $S^{(v)}$, perform k-means clustering on projected samples to estimate new cluster indicators $H^{(v)}$;
  3. Update seeds $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.

  end for
end for

IV c DETERMINISTIC SUBSPACE ALGORITHM (DSA)

The proposed algorithm is based on the idea of creating subspaces incrementally, in a manner guided by both the quality of individual subspaces and the diversity of the whole ensemble [2]. The preference toward either quality or diversity can be adjusted by modifying the algorithms hyper parameter αα. For the approach to be computationally feasible, we had to make several simplifications. Firstly, we create the subspaces in a greedy manner based on a round-robin strategy, which may produce a non-optimal solution. Secondly, we make a strong assumption that a subspace consisting of individually strong features is itself of a high quality [3]. This assumption does not have to hold in practice; in fact, it can be easily shown that two weak features can together have a high discriminant power [4]. However, it was necessary to make training on a highly dimensional data feasible. The proposed algorithm has three parameters: the number of subspaces to be created k, the number of features selected for every subspace n, the weight coefficient αα, indicating preference toward either the feature quality or the diversity. Lower values of αα lead to creation of more diverse subspaces, with features allocated close to evenly among them. On the other hand, by choosing a higher value we force the algorithm to pick the individually strong features more often [5]. Setting αα to 0 would make the algorithm disregard feature quality completely, whereas setting it to 1 would result in

creation of a single subspace, consisting of individually strongest features.

Smaller number of features per subspace n should, in principle, result in producing weaker base learners. Additionally, the subspaces created in that case are more diverse, since there is less overlap between their features. Larger number of subspaces k leads to creation of bigger ensemble, at the same time decreasing the diversity of the subspaces [6].

Algorithm 1 Deterministic Subspace algorithm

1: **Input:** set of features $X = \{x^{(1)}, x^{(2)}, ..., x^{(d)}\}$
2: **Parameters:** number of subspaces k, number of features per subspace n, weight coefficient α, feature quality measure $qual\_m(x^{(c)})$
3: **Output:** feature subspaces S
4: **for** $i = 1$ to k **do**
5: $S_i \longleftarrow \emptyset$
6: **end for**
7: **repeat**
8: **for** $i = 1$ to k **do**
9: **for** $c = 1$ to d **do**
10: **if** $x_c \notin S_i$ **then**
11: $f_{score}(x^{(c)}) \longleftarrow \alpha \times qual\_m(x^{(c)}) + (1-\alpha) \times div\_m(S, S_i, x^{(c)})$
12: **end if**
13: **end for**
14: $x_{best} \longleftarrow \text{argmax}_{x^{(c)}} f_{score}(x^{(c)})$
15: $S_i \longleftarrow S_i \cup x_{best}$
16: **end for**
17: **until** every subspace consists of n features
18: **return** S

IV d RANDOM SUBSPACE METHOD (RSM)

The Random Subspace Method (RSM) is an ensemble classifier technique that is proposed by Ho [4]. In the RSM, the training data is modified. However, this data modification is carried out in the feature space. Hence, each training incidence Xi (i = 1, . . . , n) in the training sample set X = [X1; . . . ; Xn] is defined as a p-dimensional vector Xi = (xi1, xi2, . . . , xip), defined by p features. Then, randomly r < p features from the p-dimensional data set X are selected. Consequently, the modified training set Xfb = Xfb 1 , Xfb 2 . . . , Xfb n , is composed of r-dimensional training incidences. After this step, classifiers are built into the random subspaces Xfb and aggregated by utilizing a majority voting. Therefore, the RSM is implemented in the following way: 1. Repeat for b = 1, 2, . . . , B: 2. Choose an r-dimensional random subspace Xfb from the original p-dimensional feature space X. 3. Build a classifier C b (x) (with a decision boundary C b (x) = 0) in Xfb . 4. Aggregate classifiers C b (x), b =1, 2, . . . , B, by utilizing majority voting for the final decision. The RSM can benefit from using random subspaces for both building and combining the classifiers. When the number of training incidences is comparatively small as compared to the data dimension, by building classifiers in random subspaces, the small sample size problem can be solved [6]. The subspace dimension will be less than the original feature space, while the number of training incidence is kept the same. Thus, the relative training sample size increases [7]. Once the data have several redundant features, the better classifier can be found in random subspaces than in the original feature space. The aggregated decision of such classifiers might be better than a single classifier build on the original training set in the entire feature space [5]. There are parameters to be tuned for Random Subspace ensemble learning algorithms. After many experiments, the best results were achieved by applying the following values for parameters.

Algorithm 1 – Random subspace method
Inputs: *TrainingExamples(X), EnsembleSize(S), FeatureSet(D), SubspaceSize(d)*
Outputs: *EnsembleOfModels(H)*
while S ≠ ∅ **do**
$d_{set} = \emptyset$
while $d \neq \emptyset$ **do**
$d_{set} \leftarrow d_{set} \cup$ select $d \in D$ with replacement
$d = d - 1$
end while
$h_{temp} \leftarrow InduceTree(X, d_{set})$

V. DATASETS

The data sets used in our experiments include six UCI data sets1. Here is some basic information of those data sets. Table 5 summarizes the basic information of those data sets.

• Balance. This data set was generated to model psychological experimental results. There are totally 625 examples that can be classified as having the balance scale tip to the right, tip to the left, or be balanced.
• Iris. This data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
• Ionosphere. It is a collection of the radar signals belonging to two classes. The data set contains 351 objects in total, which are all 34-dimensional.
• Soybean. It is collected from the Michalski's famous soybean disease databases, which contains 562 instances from 19 classes.

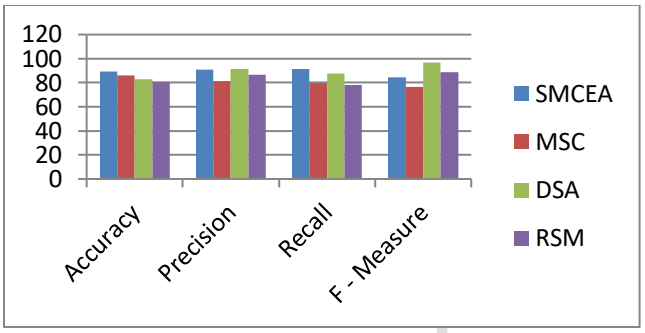
Datasets	Size	Classes	Dimensions
Balance	625	3	4
Iris	150	3	4
Ionosphere	351	2	34
Soybean	562	19	35

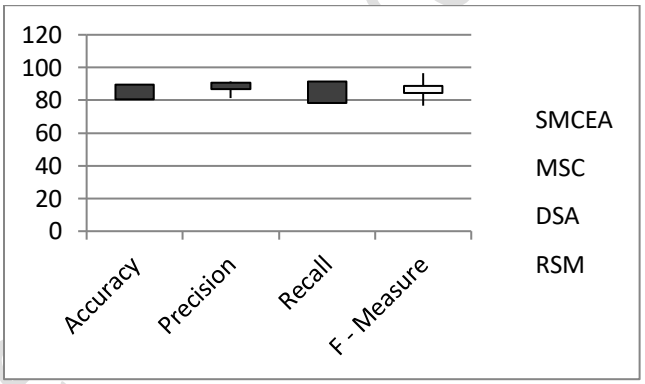
VI . EXPERIMENTAL RESULTS

VI a BALANCE DATASET RESULTS

Balance Dataset				
Algorithm	Accuracy	Precision	Recall	F - Measure
SMCEA	89.45	90.67	91.45	84.45
MSC	86.05	81.23	79.98	76.67
DSA	82.77	91.56	87.88	96.56
RSM	80.56	86.78	78.34	88.67

The above graph shows that performance of Balance dataset. The Accuracy of SMCEA algorithm is 89.45 which is higher when compare to other three (MSC, DSA, RSM) algorithms. The Precision of DSA algorithm is 91.56 which is higher when compare to other three (MSC, SMCEA, RSM) algorithms. The Recall of SMCEA algorithm is 91.45 which is higher when compare to other three (MSC, DSA, RSM) algorithms. The F-Measure of DSA algorithm is 96.56 which is higher when compare to other three (MSC, SMCEA, RSM) algorithms.

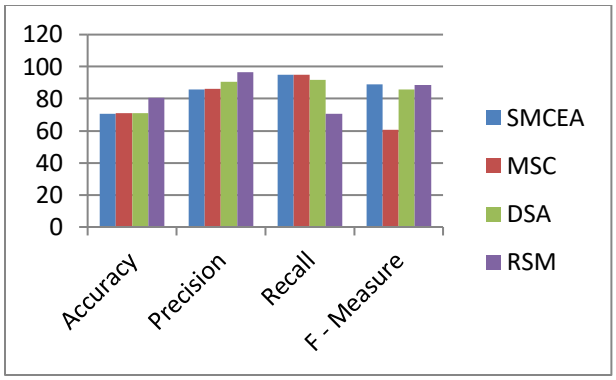


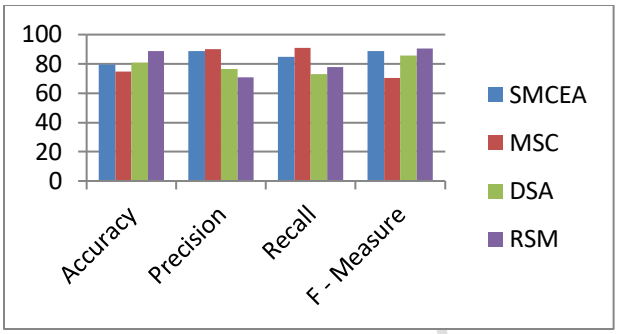


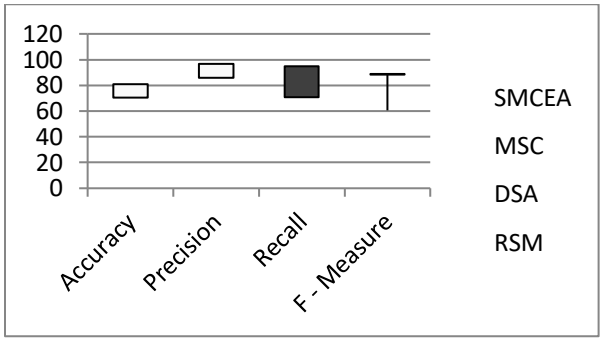
VI b IRIS DATASET RESULTS

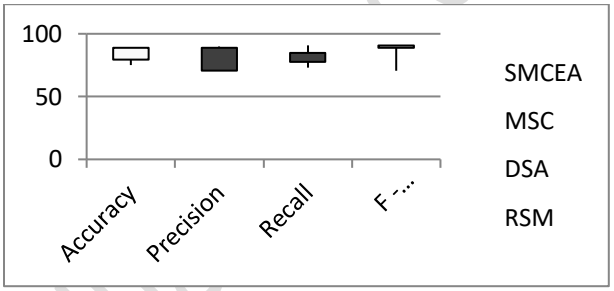
Iris Dataset				
Algorithm	Accuracy	Precision	Recall	F - Measure
SMCEA	70.45	85.91	94.77	88.89
MSC	70.91	86.08	94.78	60.56
DSA	70.92	90.67	91.89	85.78
RSM	80.67	96.67	70.78	88.67

The above graph shows that performance of Iris dataset. The Accuracy of RSM algorithm is 80.67 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms. The Precision of RSM algorithm is 96.67 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms. The Recall of MSC algorithm is 94.78 which is higher when compare to other three (SMCEA, RSM, DSA) algorithms. The F-Measure of ISSCE algorithm is 88.89 which is higher when compare to other three (DEMS, IEMS, PAM) algorithms.









VI c IONOSPHERE DATASET RESULTS

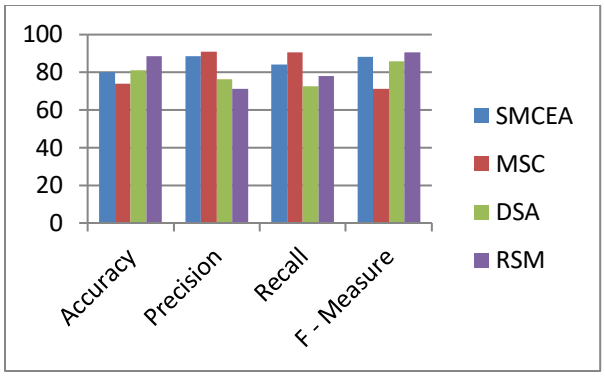
Ionosphere Dataset				
Algorithm	Accuracy	Precision	Recall	F - Measure
SMCEA	79.45	88.91	84.77	88.89
MSC	74.91	90.08	90.78	70.56
DSA	80.98	76.67	72.89	85.78
RSM	88.67	70.67	77.78	90.67

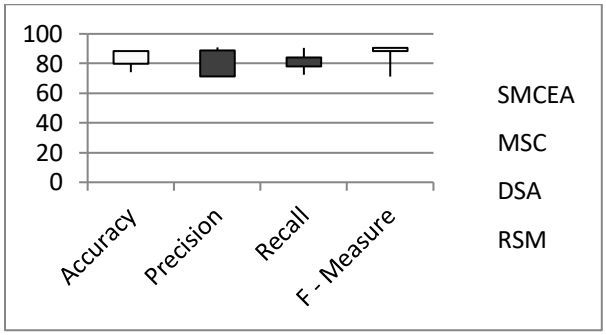
The above graph shows that performance of Ionosphere dataset. The Accuracy of RSM algorithm is 88.67 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms. The Precision of MSC algorithm is 90.08 which is higher when compare to other three (SMCEA, RSM, DSA) algorithms. The Recall of MSC algorithm is 90.78 which is higher when compare to other three (SMCEA, RSM, DSA) algorithms. The F-Measure of RSM algorithm is 90.67 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms.

VI d SOYBEAN DATASET RESULTS

Soybean Dataset				
Algorithm	Accuracy	Precision	Recall	F - Measure
SMCEA	79.89	88.65	84.23	88.34
MSC	74.03	90.89	90.67	71.23
DSA	81.08	76.32	72.45	85.9
RSM	88.54	71.32	77.89	90.56

The above graph shows that performance of Soybean dataset. The Accuracy of RSM algorithm is 88.54 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms. The Precision of MSC algorithm is 90.89 which is higher when compare to other three (SMCEA, RSM, DSA) algorithms. The Recall of MSC algorithm is 90.67 which is higher when compare to other three (SMCEA, RSM, DSA) algorithms. The F-Measure of RSM algorithm is 90.56 which is higher when compare to other three (SMCEA, MSC, DSA) algorithms.





VII. CONCLUSIONS AND FUTURE WORK

We have addressed the problem of learning effective metrics for clustering in high dimensional spaces when limited supervision is available. We have proposed an approach based on learning with ensembles that is capable of producing components which are both accurate and diverse [2]. In our future work we will investigate the sensitivity of our approach with respect to the dimensionality of subspaces, and possibly define a heuristic to automatically estimate an "optimal" value for such parameter. Furthermore, we will explore alternative mechanisms to credit weights to features by utilizing the constraints; consequently we will bias the sampling in feature space to favor the estimated most relevant features [3]. It proposes a new co-training framework for unsupervised, multi-view subspace clustering. It applies the results of unsupervised clustering in one view to learn discriminant subspaces in another. The general framework assumes conditionally independent views [4]. We show, however, that the new algorithm still performs well when the conditional independence is weak. Furthermore, the framework is straightforward and combines well known, even trivial algorithms to positive effect. The paper also presents a theoretical treatment which shows how LDA projections learned from samples with random label noise are equivalent to those learned with entirely clean labels and that the cross-view labeling, or co-training, is efficient in correcting erroneous sample labels [5]. Experiments in audio-visual speaker clustering, multi-view handwritten digit clustering and text document clustering demonstrate the effectiveness of our algorithm and superior performance to existing state-of-the-art approaches [6].

We present a novel semi-supervised clustering ensemble approach for data clustering. Our method is different from the previous studies that integrate the constraint-based method and the metric method into a semi supervised clustering ensemble approach in the hope of gaining the more optimal accuracy, robustness and stability of clustering dramatically [7]. Specifically, we construct a new metric function with two forms in our proposed metric-based semi-supervised clustering algorithm. One is for general data clustering based on the LMNC distance metric. The other is for image data clustering by combining the similarity of image pixels with the LMNC metric from the image perspective; concretely it builds collections of the inherent attributes and spatial information of pixels, which efficiently and accurately reflects the relationship between image pixels [9]. Moreover, we conduct two group comparison experiments, respectively on general data sets and image data sets. Multiple comparison results indicate that this proposed scheme can achieve better clustering performance than a number of competing clustering algorithms on the whole [8]. Empirically as well as theoretically, it confirms the feasibility and effectiveness of the proposed method with encouraging results. However, what we are done still leaves much to be desired. For instance, we should add noise factors into the experiments to test the sensitivity of clustering algorithms to noise, which can reveal the stability and robustness of clustering approaches. At the same time, there are a great many interesting directions to extend our work [10]. As we all known, clustering is often viewed as a foundation technology in image process and computer vision. To investigate further, our future work will develop clustering into the mapping process from low-level features of images to high level semantic

comprehension in conjunction with other related techniques [1].

REFERENCES

1. S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering", SIAM International conference on Data Mining, 2004.

2. S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi supervised clustering", International Conference on Knowledge Discovery and Data Mining, 2004.

3. M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and Metric learning in semi-supervised cluster", International Conference on Machine Learning, 2004.

4. C. L. Blake and C. J. Merz, "UCI repository of machine learning databases", http://www.ics.uci.edu/ mlearn/MLRepository.html, 2008.

5. D. Cohn, R. Caruana, and A. McCallum, "Semi-supervised clustering with user feedback", TR2003-1892, Cornell University, 2003.

6. X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning", International Conference on Machine Learning, 2004.

7. A. L. N. Fred and A. K. Jain, "Data clustering using evidence accumulation", International Conference on Pattern Recognition, 2002.

8. B. Kulis, S. Basu, and I. Dhillon, and R. Mooney, "Semi-supervised graph clustering: a kernel approach", International Conference on Machine Learning, 2005.

9. A. Strehl and J. Ghosh, "Cluster ensembles - knowledge reuse framework for combining multiple partitions", Machine Learning Research, 3, pages 583-417, 2002.

10. A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering", AAAI, 2005.

11. A. Topchy, A. K. Jain, and W. Punch, "Combining multiple weak clustering's", IEEE International Conference of Data Mining, 2003.

12. A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles", SIAM International Conference on Data Mining, 2004