Pairwise Constrained Clustering for Sparse and High Dimensional Feature Spaces

Su Yan[1,2], Hai Wang[1,3], Dongwon Lee[1,2], and C. Lee Giles[1,2]

[1] College of Information Sciences and Technology
The Pennsylvania State University
[2] University Park, PA 16802, USA
[3] Dumore, PA 18512, USA
{syan, dongwon, haiwang, giles}@ist.psu.edu

Abstract. Clustering high dimensional data with sparse features is challenging because pairwise distances between data items are *not* informative in high dimensional space. To address this challenge, we propose two novel semi-supervised clustering methods that incorporate prior knowledge in the form of pairwise cluster membership constraints. In particular, we project high-dimensional data onto a much reduced-dimension subspace, where rough clustering structure defined by the prior knowledge is strengthened. Metric learning is then performed on the subspace to construct more informative pairwise distances. We also propose to propagate constraints locally to improve the informativeness of pairwise distances. When the new methods are evaluated using two real benchmark data sets, they show substantial improvement using only limited prior knowledge.

1 Introduction

Clustering is one of the most important and fundamental techniques in data mining, information retrieval, and knowledge management. Most clustering techniques rely on the pairwise distances between data items. However, it is commonly believed that pairwise distances in the high-dimensional space is not informative, and the nearest neighborhood is not meaningful either [2]. As a result, many learning algorithms (including clustering methods) lose their algorithm effectiveness for high dimensional cases.

Recently, semi-supervised clustering has shown effectiveness in improving clustering accuracy by exploring "weak" supervision in the form of pairwise "must-link" or "cannot-link" constraints [1,3,5,9]. That is, if data items a and b are must-linked/cannot-linked, then a and b belong to the same/different cluster(s). Two basic semi-supervised approaches are metric learning that learns a distance measure based on constraints, and constraint enforcement that enforces constraints by modifying the objective function of a clustering method. However, most existing semi-supervised clustering methods have difficulties in handling data with high-dimensional sparse features. For example, in order for a metric-learning method [4,10] to train a distance measure, the number of independent

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 620–627, 2009.

variables to be learned is proportional to the dimension of the feature space. For data with hundreds or thousands(e.g., text data) features, metric learning is computationally expensive. Besides, due to the sparseness, only a small portion of total features are covered by constrained data items. Therefore, training a distance measure for sparse features is not effective. For another example, graph-based methods can usually handle high-dimensional data better since they work on the low-dimensional representation (affinity matrix) of the high dimensional data. However, the performance of a graph-based method partially relies on the affinity matrix, which is built upon pairwise distances. Since the pairwise distances in high dimensional space is not informative, the performance of graph-based method is impaired.

In this paper, toward these challenges, we propose two methods to tackle the high-dimensional sparse feature space problem with the help of pairwise constraints. User-provided constraints reflect user's expectation of how the data set should be clustered. Therefore, constraints define a rough clustering structure to a data set. The first method seeks a low-dimensional representation of the data through orthogonal factorizations. The clustering structure defined by prior knowledge is kept and strengthened in the subspace. Metric learning is then performed in the subspace. The second method does space-level generalization of pairwise constraints by local constraints propagation. Both methods can construct more informative pairwise distances for high-dimensional data. Our proposed schemes of exploiting constraints can be applied to any unsupervised clustering model and any high-dimensional data set. We apply the schemes to the widely used Normalized Cut method (NC), and the document data sets to demonstrate the idea and concept. Experimental results base on real data verify the effectiveness and efficiency for our proposals.

2 Main Proposal

2.1 Metric Learning in Structure-Preserving Subspace

This approach is to construct more informative distances between data items through metric learning in the reduced-dimension subspace. The motivation is obvious. In the much reduced-dimension subspace, features are not sparse. Therefore, metric learning is more effective. In addition, since the number of variables to be learned is significantly reduced, metric learning is also more efficient. Most importantly, we argue that user-provided constraints define a clustering structure that best satisfies the user's requirement. If we find a low-dimension representation of data where the clustering structure is more evident, we can expect more informative distances metric to be learned in the subspace.

We now introduce how to find such a structure-preserving subspace. Suppose we have n data items in the full space described by a matrix of column vectors $W \in \mathbb{R}^{f \times n}$, where the feature space is of f dimensions ($f \gg n$). Given pairwise constraints, we first do transitive closure to the constraints and generate d small data groups, where the i-th group with m data items represented by matrix $W_i \in \mathbb{R}^{f \times m}$. Then a matrix $C \in \mathbb{R}^{d \times n}$ is generated by using the centroids of

each groups as column vectors. The above two steps incorporates constraints information into the data representation C. We now seek a data projection by splitting the feature space of C, which is the same as the feature space of W into two parts, attributes and noise. That is, we seek a projection matrix $P = \begin{pmatrix} U & V \end{pmatrix} \in \mathbb{R}^{f \times f}$, $U \in \mathbb{R}^{f \times r}$ and $V \in \mathbb{R}^{f \times s}$, such that $P^T C = \begin{pmatrix} \widehat{C} \\ \widehat{C}^{\perp} \end{pmatrix} = \begin{pmatrix} U_r^T C \\ V_s^T C \end{pmatrix}$, where $r+s = f$. Suppose $\widehat{C}$ is the r-dim attribute part, and $\widehat{C}^{\perp}$ is the s-dim noise, a desired projection satisfies that $\widehat{C}$ is orthogonal to $\widehat{C}^{\perp}$ and $\widehat{C}^{\perp} = V_s^T C = 0$, which means that the structure-irrelevant noise that existed in the full space is now removed by the projection, and only relevant dimensions are kept in the reduced space. Since we only care about the attributes part, all we need to find is the projection U. This projection can be found by computing the orthonormal basis $U \in \mathbb{R}^{f \times r}$ for the column space of C, where $rank(C) = r$. It is easy to see that V is in fact the orthonormal basis of the left null space of C. The subspace data representation is then derived by projecting data using U, that is, $\widehat{W} = U^T W \in \mathbb{R}^{r \times n}$ is the reduced r-dim representation of data. We then do metric learning in the reduced space $\widehat{W}$ for informative distances.

Note that, the idea of using centroid to represent a group of data originates from work [7]. However, [7] solves the classification problem, where the number of data groups is fixed to the number of classes and each data group contains a large amount of training data such that the centroid of a group is the rank-1 approximation with less noise. On the contrary, we solve the clustering problem. The number of data groups generated by transitive closure is not fixed and is usually large. Due to the nature of pairwise constraints and the small amount of available constraints, most of the data groups only contain a very small amount of data items (i.e., 1 or 2). The centroids for such data groups may contain spurious information.

It is easy to see that the sparse-feature problem is solved in the subspace. This is because that U is a full rank matrix and $rank(U) = rank(C) = r$, $rank(\widehat{W}) = r < n \ll f$. The subspace thus provides a more compact data representation than the original full dimensional space. The number of variables to be learned is also greatly reduced. Now, we use the following two Lemmas[1] to show that the clustering structure defined by constraints is more evident in the subspace $\widehat{W}$. Proofs are straightforward and thus skipped.

Lemma 1 (Group Volume Shrinkage Property). *Given any data item w_i and its corresponding centroid c_i, the following holds: $\|\hat{\mathbf{w}}_i - \hat{\mathbf{c}}_i\|_2 \leq \|\mathbf{w}_i - \mathbf{c}_i\|_2$. Since data items in the subspace $\widehat{W}$ get closer to their corresponding centroid than in the full space W, the volume of any given data group shrinks to its centroid.* ∎

Lemma 2 (Constant Center-to-Center Distance). *The pairwise distance between any two given centroids $\mathbf{c}_i$ and $\mathbf{c}_j$ of the full space W is strictly preserved in the subspace $\widehat{W}$: $\|\mathbf{c}_i - \mathbf{c}_j\|_2 = \|\hat{\mathbf{c}}_i - \hat{\mathbf{c}}_j\|_2$.* ∎

[1] We only show the Lemmas in the L_2 norm measure to save space. It is easy to prove that the corresponding theorems still hold for the cosine similarity.

According to Lemma 1, data items in the subspace $\widehat{W}$ move towards their corresponding centroids. According to Lemma 2, any data group W_i keeps constant distance away from any other group W_j in the sense of constant center-to-center distance. Geometrically, the volume of a data group W_i shrinks and groups are still well separated in the subspace. Therefore, the projection in fact strengthens the clustering structure defined by constraints.

2.2 Constraint-Guided Local Propagation

Graph-based methods are well known for clustering high-dimensional data with better accuracy. For example, the representative Normalized Cut (NC) method has been successfully applied to image segmentation and document clustering problems. However, as we mentioned in section 1, the performance of a graph-based method can be impaired by noninformative pairwise distance. We propose a simple yet effective method to directly enforce and propagate constraints on the affinity matrix K. The idea is to do space-level generalization of pairwise constraints based on triangle geometry.

Our idea is the following. Given pairwise constraints that data item x and y must belong to the same cluster (i.e., must-link), we set the distance between the two items as 0, that is $dist(x, y) = 0$. For any other data item z, we set $dist(z, x) = dist(z, y) = \min(dist(z, x), dist(z, y))$. Symmetrically, given pairwise constraints that data item x and y belong to different clusters (i.e., cannot-link), we set $dist(x, y) = 1$, where 1 is the largest value for a normalized distance measure. Since constraints are propagated to at most one-hop neighbors of the constrained data items, we consider this method *local*. After local propagation, the matrix K contains more informative pairwise distances, which enable better clustering performance. The following Lemma justifies our idea. Again, proof is straightforward and thus skipped.

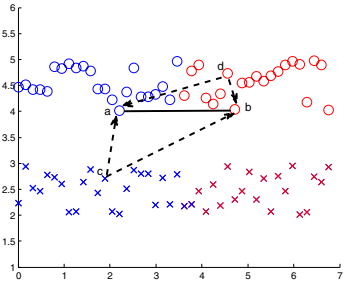


Fig. 1. The two bands data set (The true classification is indicated by marker shape, and the unsupervised K-means clustering results are indicated by marker color. Real line: must-link; Dotted line: cannot-link).

Lemma 3 (Distance propagation property). *Given three data items x, y, and z, suppose $dist(z, x) \leq dist(z, y)$. If data items x and y get closer, the 3rd item z is equally far away from x and y, That is, $dist(x, y) \rightsquigarrow 0$, then $dist(z, y) \rightsquigarrow dist(z, x)$.* ∎

The effectiveness of this method can be illustrated by Figure 1. Unsupervised clustering methods ignore the band structure of the data. If we know that data item a and b belong to the same cluster, and set $dist(d, a) = dist(d, b)$, $dist(c, b) = dist(c, a)$, data items in the upper band will get closer to each other

and the band effect is reduced. That is, constraints on data items a and b are generalized to the whole space. The local propagation method has the time complexity of $O(nq)$, where n is the number of data items and q is the number of constrained data items. It is faster than the related global propagation method [6], which is based on all-pairs-shortest-path and has the $O(n^2q)$ time complexity.

3 Experimental Validation

3.1 Set-Up

Data Sets. We have evaluated the performance of our clustering algorithms using two public available data sets: the *Reuters*-21578 document corpus and the 20-*Newsgroups* 18828 version document corpus. For the Reuters corpus, we included only documents with a *single* label to ensure unambiguous results.We pre-processed each document by tokenization, stop-words removal, and stemming. Terms that appear in only one document are removed.

12 data sets were generated from the two corpora as summarized in Table 1. Without the lose of generality, we first generated data sets of different number of clusters ranging from 2 to 6. For each given cluster number k, 5 test sets were created by first randomly picked k topics from one corpus, and then 40 documents of each of the picked topics were randomly selected and mixed together (Table 1, Reu-2∼6, News-2∼6 report the statistics of the data sets randomly chosen from the pool of 5 data sets for each cluster number k). We also created two challenging data sets from 20-*Newsgroups* corpus. News-Mediocre contains 3 related topics {*talk.politics.misc*, *talk.politics.guns*, and *talk.politics.mideast*}, and News-Difficult contains 3 very similar topics {*comp.windows.x*, *comp.os.ms-windows.misc*, and *comp.graphics*}.

Evaluation Metrics. In addition to the *running time* (RT) as a metric to evaluate the speed of algorithms, to avoid biased accuracy result using a single metric, we evaluate clustering accuracy by employing three widely-used evaluation metrics, which are (1) *Normalized Mutual Information* (NMI), (2) *Rand Index*, and (3) *F-measure*. All the three metrics take values between zero and one, with one meaning best accuracy.

We also implemented four state-of-the-art semi-supervised clustering methods: (1) *L-NC* method does metric learning in the full feature space [10], then uses NC as the unsupervised clustering model; (2) *MPCKmeans* combines metric learning and constraint-enforcement into K-means through an EM process

Table 1. Summary of data sets

20-Newsgroups				*Reuters*			
Data set name	# data items	# features	# cluster	Data set name	# of data items	# features	# cluster
News-2	80	2,422	2	Reu-2	80	1,213	2
News-3	120	3,895	3	Reu-3	120	2,028	3
News-4	160	4,520	4	Reu-4	160	2,347	4
News-5	200	6,203	5	Reu-5	200	2,343	5
News-6	240	5,991	6	Reu-6	240	2,952	6
News-Difficult	300	3,570	3	News-Mediocre	300	4,457	3

Table 2. Summary of all the experimented algorithms

Type	Algorithm	Description
Baseline	NC	Normalized Cut [8]
State-of-the-art	MPCKmeans	learning feature weights & enforce constraint through EM [3]
	L-NC	learning feature weights in the original feature space [10], followed by NC
	C-NC	Constrained Normalized Cut [5]
	Glo-NC	NC with globally adjusted affinity matrix [6]
Our proposals	**RL-NC**	learning feature weights in reduced space (Section 2.1), followed by NC
	RLC-NC	hybrid of **RL-NC** and C-NC
	Lo-NC	NC with locally-adjusted affinity matrix (Section 2.2)

[3]; (3) *C-NC* is a graph-based constraint-enforcement method that has been shown effective in clustering documents [5]; and (4) Glo-NC method *globally* propagates constraints to adjust pairwise distances [6]. Table 2 summarizes the baseline method and seven variations of semi-supervised clustering methods that we have evaluated. Our three proposals – **RL-NC**, **RLC-NC**, and **Lo-NC** – are bold faced.

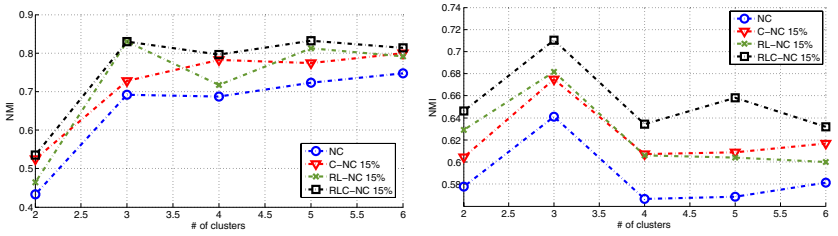
3.2 Experimental Results

Metric Learning. We reported the experimental results based on two challenging data sets News-Mediocre and News-Difficult. We controlled the experiment by varying the amount of constrained data items ranging from 2.5% to 15% of the total documents. Constraints were generated by paring constrained documents based on ground truth.The final performance score was obtained by averaging the scores from 10 test runs.

We compared our sub-space metric learning method (RL-NC) with the full-space learning method (L-NC). Performance comparisons are reported in Table 3. For both data sets and various amount of constraints, RL-NC achieves higher and more stable learning accuracy. Metric learning in the subspace is also much faster than in the full feature space. When the amount of constraints increases, the learning time of both methods increases too, with the subspace learning method scales much better than the full space learning method. Last, note that although the three accuracy metrics (i.e. NMI, RI, and F) show quite different absolute values, they show overall similar patterns for different algorithms. For simple presentation, we will only use NMI as the evaluation metric from here forward.

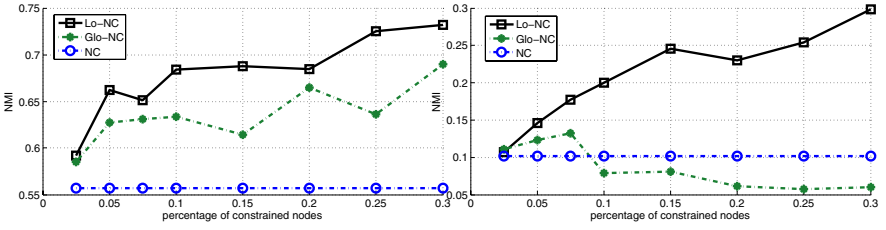
Table 3. Running time and accuracy for subspace metric learning

Algorithm	Metric	*News-Mediocre*					*News-Difficult*				
		2.5%	5%	7.5%	10%	15%	2.5%	5%	7.5%	10%	15%
NC	NMI	0.5568					0.1016				
RL-NC		**0.5865**	**0.6003**	**0.6164**	**0.6405**	**0.6374**	**0.1134**	0.104998	**0.1142**	**0.1151**	**0.1404**
L-NC		0.5220	0.5362	0.4656	0.5246	0.5756	0.1084	**0.1077**	0.1060	0.1060	0.1271
NC	RI	0.7432					0.5098				
RL-NC		**0.7693**	**0.7689**	**0.8029**	**0.8409**	**0.8387**	**0.5427**	0.5242	0.5380	**0.5624**	**0.5550**
L-NC		0.7311	0.7674	0.7112	0.6786	0.8316	0.5398	**0.5478**	**0.5397**	0.5559	0.5363
NC	F	0.6629					0.4401				
RL-NC		**0.6834**	**0.7072**	**0.7222**	**0.7666**	**0.7593**	0.4423	**0.4490**	**0.4424**	0.4436	**0.4636**
L-NC		0.6484	0.6770	0.6455	0.6191	0.7298	**0.4425**	0.4436	0.4418	**0.4443**	0.4555
RL-NC	RT	**0.6129**	**1.0875**	**5.3013**	**9.1093**	**20.4659**	**0.6209**	**2.2153**	**5.2884**	**9.2359**	**17.2411**
L-NC		2.0133	7.9069	18.9226	32.6786	77.0943	0.8736	6.4430	11.2585	26.5236	60.2636



(a) Reuters corpus (b) Newsgroups corpus

Fig. 2. Accuracy of combining *subspace metric learning* (RL-NC) and *constraint enforcement* (C-NC) (% of constraints = 15%)



(a) *News-Mediocre* (b) *News-Difficult*

Fig. 3. Local vs. global propagation

Integrating Metric Learning with Constraint Enforcement. Metric learning and constraint enforcement are two basic schemes of exploiting constraints to improve clustering performance. In this experiment, we combined the two schemes and expected to generate better clustering performance. We compared the subspace metric learning method (RL-NC), with the graph-based constraint enforcement method (C-NC), and with the hybrid approach (**RLC-NC**). Figure 2 shows that both RL-NC and C-NC individually improved upon the regular NC method, and both show similar accuracy given the same amount of constraints. However, the hybrid algorithm, RLC-NC, always significantly outperforms the individual ones. These results empirically validate our hypothesize that the hybrid method, RLC-NC, more comprehensively utilizes available constraints.

Constraint-Guided Local Propagation. We compared the effectiveness of the local propagation method (Lo-NC) with the global adjustment method (Glo-NC). The regular NC approach was adopted as the unsupervised clustering model for both methods. Figure 3 shows the clustering results on the two representative Newsgroups data sets. Lo-NC made sizable improvement over the unconstrained version of NC even with a small amount of constraints. Glo-NC is less effective and its performance for the *News-Difficult* data set is worse than the unconstrained method when the number of constraints increases. Both the local and the global methods exploit the triangle geometry. But the global method also propagates constraints based on the pairwise distances among all the data

Table 4. Lo-NC, Glo-NC on Reuters and Newsgroups corpora (C: % of constraints)

Algorithms	C	Reuters corpus					Newgroup corpus				
		NMI for different # of clusters					NMI for different # of clusters				
		2	3	4	5	6	2	3	4	5	6
Lo-NC	5%	**0.4474**	**0.7462**	**0.7159**	0.7399	**0.742**	0.6179	**0.6724**	**0.5959**	**0.58347**	**0.5972**
Glo-NC		0.4402	0.713	0.7112	**0.7421**	0.7397	**0.7068**	0.6573	0.5789	0.5819	0.5692
Lo-NC	10%	**0.4809**	**0.7202**	**0.7651**	0.7649	**0.7886**	**0.6767**	0.6845	0.6056	**0.6045**	**0.6188**
Glo-NC		0.4557	0.7109	0.7500	**0.7716**	0.7638	0.6169	**0.7031**	**0.6074**	0.5645	0.6014
Lo-NC	15%	**0.5005**	**0.7275**	0.7508	0.777	**0.7817**	**0.667**	**0.694**	**0.6433**	**0.6154**	**0.6384**
Glo-NC		0.4913	0.7038	**0.7564**	**0.7783**	0.7701	0.6371	0.66	0.6033	0.5789	0.6027

points, which may decrease the discriminative power of the constraints. Detailed results on multiple Reuters and Newsgroups data sets are shown in Table 4.

4 Conclusion

Two novel semi-supervised clustering techniques are proposed for high dimensional and sparse data. The first method projects data onto a reduced-dimension subspace such that clustering structure defined by constraints is strengthened. Metric learning is then applied to the subspace to generate informative pairwise distances. The second method exploits the triangle geometry to generalize pairwise constraints by "local" propagation. The validity of our proposals are empirically validated using extensive experiments.

References

1. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: ACM KDD, Seattle, WA, USA, pp. 59–68 (2004)
2. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbor meaningful? In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 217–235. Springer, Heidelberg (1998)
3. Bilenko, M., Basu, S., Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering. In: ICML, Banff, Alberta, Canada (2004)
4. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Technical report, Cornell University (2003)
5. Ji, X., Xu, W.: Document clustering with prior knowledge. In: ACM SIGIR, Seattle, WA, USA, pp. 405–412 (2006)
6. Klein, D., Kamvar, S.D., Manning, C.D.: From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: ICML, Sydney, Australia, pp. 307–314 (2002)
7. Park, H., Jeon, M., Rosen, J.B.: Lower dimensional representation of text data based on centroids and least squares. BIT Numerical Mathematics 43, 427–448 (2003)
8. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2000)
9. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: ICML, pp. 577–584 (2001)
10. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning, with application to clustering with side-information. In: NIPS, Vancouver, Canada (2003)