LOCALIZED FEATURE SELECTION FOR CLUSTERING AND ITS APPLICATION IN IMAGE GROUPING

Yuanhong Li, Ming Dong and Jing Hua

Department of Computer Science,
Wayne State University, Detroit, MI 48202

ABSTRACT

In clustering, global feature selection algorithms attempt to select a common feature subset that is relevant for all clusters. Consequently, they are not able to identify individual clusters that exist in different feature subspaces. In this paper, we propose a localized feature selection algorithm for clustering. The proposed algorithm computes adjusted and normalized scatter separability for individual clusters. A sequential backward search is then applied to find the optimal (maybe local) feature subsets for each cluster. Experiment results on both synthetic data clustering and content-based image grouping show the need for feature selection in clustering and the benefits of selecting features locally.

1. INTRODUCTION

Clustering is a common unsupervised learning technique used to discover the nature groups of similar objects, represented by vectors of measurements, in multidimensional spaces. It is one of the most important techniques for fast retrieval of the relevant information from the databases. Applications of clustering are widely found in data mining, information discovering, image retrieving, and image grouping. Specifically, it provides the users quick browsing environments of going over a large content-based image database by grouping images logically and predictably using clustering techniques [1].

A clustering algorithm typically considers all dimensions, or features, of the data in an attempt to learn as much as possible about the objects. With high dimensional data such as in visual recognition and document classification cases, however, many features are redundant or irrelevant. The redundant features are no help for clustering; and the irrelevant features, even worse, may hurt the clustering results by hiding clusters in noises. To alleviate this problem, one of the most extensively used methods is *feature selection*. The objective of feature selection is three-fold: improving the performance of clustering, providing fast and cost-efficient solution, and providing a better understanding of the underlying process that generates the data.

Feature selection in supervised learning has been widely studied [2, 3]. However, for unsupervised learning, the research is relatively recent [4–7]. The objective is to select important features for clustering in the absence of class labels. In [7], a maximum information compression index is used to measure feature similarity so that feature redundancy is detected. In [6], weights are assigned to different feature spaces for k-means clustering based on within-cluster and between-cluster matrix. Feature saliency is integrated in EM algorithm in [5] so that feature selection is performed simultaneously with clustering process. Dy and Brodley recently proposed a wrapper criterion for clustering [4], which evaluates the quality of clusters using normalized cluster separability (for k-means) or normalized likelihood (for EM clustering). In their approach, the bias on the feature subsets with respect to dimensionality is ameliorated by cross-projection normalization.

In the aforementioned algorithms, the candidate feature subsets are evaluated globally. Regardless of the choices of evaluation criteria, global feature selection approaches compute them over the entire dataset. Thus, they can only find one relevant feature subset for all clusters. However, it is the local intrinsic properties of data counts during clustering [8]. Such a global approach cannot identify individual clusters that exist in different feature subspaces. An algorithm that performs feature selection for each individual cluster separately is highly preferred [9, 10]. In the case of image grouping, localized feature selection is particularly important because it can reveal the natural similarity in each individual cluster, and thus provides us an efficient way to interpret the clustering results.

In this paper, we propose a localized feature selection algorithm for clustering. The proposed algorithm computes *adjusted and normalized scatter separability* for individual clusters. A sequential backward search is then applied to find the optimal (maybe local) feature subsets for each cluster. Our experiment results show that the proposed localized feature selection outperforms global approaches on both synthetic data and content-based image grouping problems.

The rest of the paper is organized as follows. The motivation and details of the proposed algorithm are described in Section 2. Our algorithm is evaluated using both a synthetic dataset and a real-world image datasets in Section 3. In Section 4, some conclusions are provided.

2. LOCALIZED FEATURE SELECTION FOR CLUSTERING

2.1. Motivation

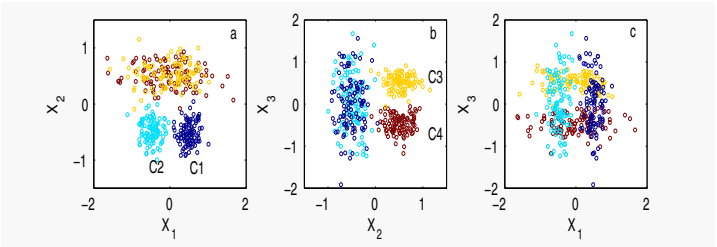

Fig. 1. Synthetic data plotted in different feature sets. Data from different clusters are marked with different colors. a: in X_1 and X_2. b: in X_2 and X_3. c: in X_1 and X_3.

Our motivation of localized feature selection can best be illus-

trated using a synthetic dataset. We generate 400 data points with 4 clusters $\{C_1, C_2, C_3, C_4\}$ in 4 dimensional space $\{X_1, X_2, X_3, X_4\}$. Each cluster contains 100 points. Clusters C_1 and C_2 are created in dimensions X_1 and X_2 based on a normal distribution. X_3 and X_4 are white noise features in these two clusters. The means and standard deviations are: $\mu_1 = [0.5, -0.5, 0, 0]$, $\mu_2 = [-0.5, -0.5, 0, 0]$, and $\sigma_1 = \sigma_2 = [0.2, 0.2, 0.6, 0.6]$, respectively. Clusters C_3 and C_4 exist in dimensions X_2 and X_3 with white noise in X_1 and X_4, and are created in the same manner. The means and standard deviations are: $\mu_3 = [0, 0.5, 0.5, 0]$, $\mu_4 = [0, 0.5, -0.5, 0]$, and $\sigma_3 = \sigma_4 = [0.6, 0.2, 0.2, 0.6]$, respectively. Figure 1 shows the data in different subspaces. A general clustering algorithm, such as k-means or EM, is unable to obtain satisfactory clustering results for this data either on all features $\{X_1, X_2, X_3, X_4\}$, or on relevant feature subset $\{X_1, X_2, X_3\}$ (may be generated by a global feature selection algorithm) because each cluster still has one irrelevant feature. For data in higher dimensional space, this problem becomes more prominent.

On the other hand, if we further remove X_3 from the feature subset $\{X_1, X_2, X_3\}$, we can completely separate C_1 and C_2, as shown in Figure $1a$. Similarly, C_3 and C_4 can be well separated by removing X_1 as shown in Figure $1b$. In addition, the clustering results of localized feature selection provide a better understanding of the underlying process that generates the data. For example $C_1 \sim \{X_1, X_2\}$ clearly indicates that cluster C_1 is mainly generated by features X_1 and X_2.

Usually, there are two major components for a feature selection algorithm: evaluation criteria and feature subset search methods. In the following, we first discuss the evaluation criterion for the localized feature selection algorithm, then the search method.

2.2. Evaluation Criteria

In this section, we first provide a brief introduction to scatter separability criterion, one of the well-known clustering criteria, and then show how this criterion could be adapted to localized feature selection.

Let S_w and S_b denote within-class scatter matrix and between-class scatter matrix, respectively, we have,

$$S_w = \sum_{i=1}^{k} \pi_i E[(X - \mu_i)(X - \mu_i)^T | C_i] = \sum_{i=1}^{k} \pi_i \Sigma_i, \quad (1)$$

$$S_b = \sum_{i=1}^{k} \pi_j (\mu_i - \mu_0)(\mu_i - \mu_0)^T \quad (2)$$

where π_i is the probability that an instance belongs to cluster C_i, X the d-dimensional input dataset, k the number of clusters, μ_i the sample mean vector of cluster C_i, μ_0 the total sample mean, Σ_i the sample covariance matrix of cluster C_i and $E[\cdot]$ the expected value operator.

Since S_w measures how scattered the samples are from their cluster mean, and S_b measures how scattered the cluster means are from the total mean, the scatter separability is defined as

$$CRIT = trace(S_w^{-1} S_b) \quad (3)$$

This measure enjoys a nice property that it is invariant under any nonsingular linear transformation [11].

Similar to the definition of S_w, we define $S_w^{(i)}$, the within-class matrix of an individual cluster C_i as,

$$S_w^{(i)} = \frac{1}{n_i} E\{(X - \mu_i)(X - \mu_i)^T | C_i\} = \frac{1}{n_i} \Sigma_i \quad (4)$$

where n_i is the number of points in cluster C_i. Now we are ready to define the scatter separability of cluster C_i.

Definition 1. *The scatter separability of cluster C_i is defined by,*

$$CRIT(C_i) = trace(S_w^{(i)-1} S_b) \quad (5)$$

Assuming that identical clustering assignments are obtained when more features are added, the scatter separability $CRIT$ prefers higher dimensionality since the criterion value monotonically increases as features are added [11]. The same conclusion could be drawn for the scatter separability for an individual cluster [1]. That is, $CRIT(C_i)$ monotonically increases with dimensions as long as the clustering assignments remain the same. To alleviate this problem, normalization of the separability criterion with respect to dimensions is necessary for feature selection [4]. Moreover, for localized feature selection strategies, each cluster is associated with a distinct feature subset. It is usually impossible to compute S_b without proper normalization.

In the proposed algorithm, the normalization is performed using cross-projection over individual clusters. Suppose we have a cluster set C,

$$C = \{(C_1, S_1), \ldots (C_i, S_i), \ldots, (C_k, S_k)\} \quad (6)$$

where S_i is the feature subset corresponding to cluster C_i. To calculate the scatter separability of (C_i, S_i) in cluster set C, we project all the clusters of C into feature subset S_i, and extend the scatter separability of cluster C_i as follows,

Definition 2. *The scatter separability of cluster C_i in cluster set C on feature subset S_i is given by,*

$$CRIT(C_i, S_i)|_C = trace(S_w^{(i)-1} S_b)|_{C,S_i} \quad (7)$$

where $|_{C,S_i}$ denotes the project of cluster set C onto feature subset S_i.

Assume an iteration of search produces a new cluster set C' on subspace S_i',

$$C' = \{(C_1', S_i'), \ldots (C_i', S_i'), \ldots, (C_k', S_i')\} \quad (8)$$

Let's also assume that cluster (C_i', S_i') corresponds to cluster (C_i, S_i), i.e., (C_i', S_i') is the cluster that has the largest overlapping with (C_i, S_i) in set C'. We then generate a new cluster set, C^*, by replacing (C_i, S_i) in C with (C_i', S_i'),

$$C^* = \{(C_1, S_1), \ldots (C_i', S_i'), \ldots, (C_k, S_k)\} \quad (9)$$

Note that $CRIT(C_i, S_i)|_C$ and $CRIT(C_i', S_i')|_{C^*}$ can not be compared directly because of the dimension bias. We have to cross-project them onto each other,

$$NV(C_i, S_i)|_C = CRIT(C_i, S_i)|_C \cdot CRIT(C_i, S_i')|_C \quad (10)$$

$$NV(C_i', S_i')|_{C^*} = CRIT(C_i', S_i')|_{C^*} \cdot CRIT(C_i', S_i)|_{C^*} \quad (11)$$

After the cross-projection, the bias is eliminated and the normalized value NV can be used to compare two clusters in different feature subspaces. A larger value of NV indicates larger separability, i.e., better cluster structures.

Localized feature selection implicitly creates overlapping and/or unassigned data points. Overlapping points are the data which belongs to more than one cluster, while unassigned points are the data

[1] we omit the straightforward proof due to space constrains

which belongs to non-cluster. Specifically, the overlapping measure O can be computed as,

$$O = \sum_{i \neq j}^{k} \frac{|C_i \cap C_j|}{mean(|C_i|, |C_j|)} \quad (12)$$

where C_i and C_j are two different clusters. And unassigned measure U can be computed as $U = \frac{n_u}{n}$ where n and n_u are the total number of data and the number of unassigned points, respectively. Overlapping and/or unassigned data are allowed in some applications, and may be forbidden by other applications. Depending on the domain knowledge, we could adjust the impact of overlapping and unassigned points by introducing a penalty and obtain the adjusted normalized value ANV.

Definition 3. *The adjusted and normalized scatter separability pair of cluster C_i in cluster set C on feature subset S_i and cluster C_i' in cluster set C^* on feature subset S_i' is given by,*

$$ANV(C_i, S_i)|_C = NV(C_i, S_i)|_C \cdot e^{(-\alpha\Delta O - \beta\Delta U)} \quad (13)$$

$$ANV(C_i', S_i')|_{C^*} = NV(C_i', S_i')|_{C^*} \cdot e^{(\alpha\Delta O + \beta\Delta U)} \quad (14)$$

where ΔO and ΔU are the changes on the overlapping and unassigned measure, respectively, if cluster (C_i, S_i) is replaced by cluster (C_i', S_i'). α and β are two constants.

In Definition 3, α and β are used to control the sensitivity with respect to overlapping points and unassigned points. Large α and β discourage the occurrence of overlapping and unassigned data. On the other hand, if α or β is zero, the corresponding affect of overlapping or unassigned will be ignored when comparing two clusters. The values for α and β depend on the given application and have to be determined empirically. For example, if a large portion of data is unassigned after clustering, β needs to be increased.

When two clusters (C_i, S_i) and (C_i', S_i') are compared, if $ANV(C_i, S_i)|_C > ANV(C_i', S_i'))|_{C^*}$, we choose (C_i, S_i). If $ANV(C_i, S_i)|_C = ANV(C_i', S_i'))|_{C^*}$, we prefer the cluster in the lower dimensional space.

2.3. Search Methods

The cross-projection normalization scheme assumes that the clusters to be compared should be consistent in the structure of the feature space [4]. Consequently, we select sequential backward search instead of the sequential forward search adopted in [4]. The trade off is the slower clustering speed.

Specifically, the data are first clustered based on all available features. Then, for each cluster, the algorithm determines if there exists a redundant or noisy feature based on the adjusted normalized value ANV defined in Equations (13) and (14). If so, it will be removed. The above process is repeated iteratively on all clusters until no change is made, at which time the clusters with the associated feature subsets will be returned.

3. EXPERIMENT AND RESULTS

In this section we evaluate the localized feature selection algorithm using both synthetic and real-world image datasets. The experiment results are obtained by choosing k-means as the clustering algorithm. However, note that the adjusted normalized value ANV is not restricted to k-means. It can be used together with any general clustering algorithm.

In general, it is difficult to evaluate the performance of a clustering algorithm on high dimensional data. Localized feature selection presents an additional layer of complexity by associating clusters to different features subsets. Therefore, we take a gradual approach for our evaluation. We first test the proposed algorithm on a small synthetic data set with known data distribution along each feature dimension. Then we apply our algorithm to an important and challenging application: content-based image grouping. On all datasets, we perform a semi-supervised learning strategy for evaluation purpose. This makes it possible for us to compute a pseudo-accuracy measure for easy comparison among different algorithms. However, one should be aware that the "true" class labels are not always consistent with the nature grouping of the underlying data set. Thus, the quality of clusters should be further analyzed in addition to the pseudo-accuracy.

On each dataset, we compare our localized feature selection algorithm (with k-means, denoted by LFS-k-means) with global feature selection algorithm (also with k-means, denoted by GFS-k-means), and k-means without feature selection. GFS-k-means is implemented in a similar fashion as [4]. The only difference is that we adopted the backward search strategy in our evaluation. How to determine the number of clusters k is out of scope of this paper. In all our experiments, the number of clusters k is fixed to the number of "true" classes. There are algorithms available to choose a suitable value for k in the case of unknown [11].

3.1. Synthetic data

The synthetic data is described in Section 2.1, and illustrated in Figure 1. Penalty of overlapping and unassigned points (α and β) is set at 1 in our experiment.

Table 1. Feature subset distribution on the synthetic data. C1 - C4 are the output cluster labels.

	Feature Subset(s)				Error
	C1	C2	C3	C4	
LFS-k-means	{1, 2}	{1, 2}	{2, 3}	{2, 3}	0.01
GFS-k-means	{4}				0.708
k-means	{1, 2, 3, 4}				0.225

Table 1 shows the selected feature subsets and error rate of k-means, GFS-k-means, and LFS-k-means, respectively. Clearly, by employing all four available features, k-means performs poorly with a error rate of 0.225, which indicates that irrelevant features greatly reduces the clustering performance. Meanwhile, GFS-k-means does a terrible job with an unacceptable error rate of 0.708. The output feature subset $\{4\}$ contains only the noisy feature X_4! This surprising result could be explained as follows. Since each feature is irrelevant to at least two clusters and each cluster has at least two irrelevant features, NO feature subset are relevant to all clusters. On the other hand, the proposed localized feature selection algorithm produces an excellent result with a error rate of 0.01. From Table 1, we can see clearly that the relevant features for each cluster are selected correctly, and the clusters are well separated in the corresponding feature subspaces. This result confirms that selecting features locally is meaningful and necessary in clustering.

3.2. Content-based Image Grouping

The data used in this experiment are chosen from Corel CDs, which contain $31,438$ general-purpose images of various contents, such as

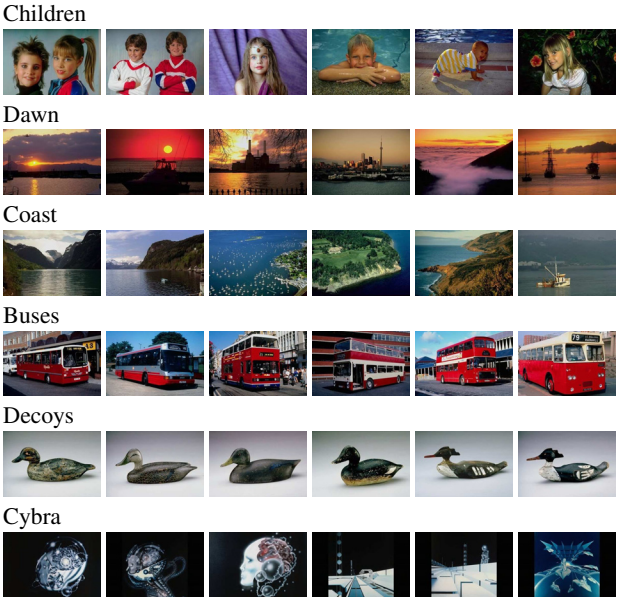

Fig. 2. Image examples

plants, animals, buildings and human society, etc. To evaluate our algorithm, we construct an image dataset with 596 images and six classes: "children", "dawn", "coast", "buses", "decoys", and "cybra". Some image examples from each class are provided in Figure 2.

A large number of visual contents are extracted from each image. They belong to 9 components: red channel (R, 3 dimensions), green channel (G, 3 dimensions), blue channel (B, 3 dimensions), color histogram (CH, 12 dimensions), color coherence vector (CCV, 24 dimensions), Gabor wavelet based texture (Gab, 24 dimensions), edge direction histogram (EDH, 9 dimensions), edge coherence histogram (ECH, 9 dimensions), and angle histogram (AH, 6 dimensions). In our experiment, we treat each component as an individual feature and perform feature selection.

Table 2. Feature subset distribution on image data.

Feature subset of GFS-k-means	
	{G B CH CCV EH AH }
Feature subsets of LFS-k-means	
buses	{ R G EH ECH }
children	{ R G CH ECH }
coast	{ R G B CH CCV Gab EH ECH AH }
cybra	{ R CH Gab EH CEH }
dawn	{ R G B CH CCV Gab EH ECH AH }
decoys	{ R G B CH Gab EH }

Table 2 shows the feature subsets chosen by GFS-k-means and LFS-k-means. The error rates for k-means, GFS-k-means and LFS-k-means are 0.307, 0.392 and 0.270, respectively. Specifically, all three algorithms successfully recognized classes "buses", "cybra" and "decoys", but failed on "coast" class. GFS-k-means selects a global feature subset consisting of Green, Blue, Color Histogram, Color Coherence Histogram, Edge Histogram, and Angle Histogram. GFS-k-means outperforms k-means on class "cybra", but gets worse results on class "children". This result indicates that globally selected feature subsets are not always suitable for individual clusters. The proposed algorithm, on the other hand, creates different feature subsets for each cluster. It outperforms both GFS-k-means and k-means on class "children ". It provides almost the same error rate on classes "buses" and "decoys", however, with less features. In addition, the result suggests that, for example, class "buses" is grouped together mainly based on feathers from red channel, green channel, edge histogram, and edge coherence histogram. This information is consistent with our visual judgments and very useful for object grouping.

4. CONCLUSIONS

In clustering, global feature selection algorithms attempt to select a common feature subset that is relevant for all clusters, which may not be feasible for many high dimensional datasets with many clusters. In order to identify individual clusters that exist in different feature subspaces, we propose a localized feature selection algorithm. We develop adjusted and normalized scatter separability (ANV) for individual clusters, based on which our algorithm is capable of reducing redundant/noisy features for each cluster separately. The proposed algorithm can also provide us better understanding of the underlying process that generates the data. Our experiment results on both synthetic and real-world image grouping problems show the need for feature selection in clustering and the benefits of selecting features locally.

5. REFERENCES

[1] J. Chen, C.A. Bouman, and J.C. Dalton, "Hierarchical browsing and search of large image databases," *IEEE Trans. Image Processing*, vol. 9, pp. 442–455, 2000.

[2] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Machine Learning Research*, vol. 5, pp. 1205–1224, Oct 2004.

[3] M. Dong and R. Kothari, "Feature subset selection using a new definition of classifiability," *Pattern Recognition Letters*, vol. 23, pp. 1215–1225, 2003.

[4] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning.," *J. Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[5] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. PAMI*, vol. 26, no. 9, pp. 1154–1166, 2004.

[6] D. Modha and S. Spangler, "Feature weighting in k-means clustering," *Machine Learning*, vol. 52, pp. 217–237, 2003.

[7] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. PAMI*, vol. 24, no. 4, 2002.

[8] Qifa Ke and Takeo Kanade, "Robust subspace clustering by combined use of knnd metric and svd algorithm," in *IEEE Conf. CVPR 2004*. June 2004, pp. 592–599, IEEE.

[9] J. Friedman and J. Meulman, "Clustering objects on subsets of attributes" *JRSS B*, vol. 66, no.4, pp. 815–849, 2004.

[10] L. Parsons, E. Haque and H. Liu, "Subspace clustering for high dimensional data: a review", *SIGKDD*, vol. 6, no. 1, pp. 90–105, 2004.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Willey & Sons, Inc, 2nd edition, 2000.