# Convex Mixture Models for Multi-view Clustering

Grigorios Tzortzis and Aristidis Likas

Department of Computer Science, University of Ioannina,
GR 45110, Ioannina, Greece
{gtzortzi,arly}@cs.uoi.gr

**Abstract.** Data with multiple representations (views) arise naturally in many applications and multi-view algorithms can substantially improve the classification and clustering results. In this work, we study the problem of multi-view clustering and propose a multi-view convex mixture model that locates exemplars (cluster representatives) in the dataset by simultaneously considering all views. Convex mixture models are simplified mixture models that exhibit several attractive characteristics. The proposed algorithm extends the single view convex mixture models so as to handle data with any number of representations, taking into account the diversity of the views while preserving their good properties. Empirical evaluations on synthetic and real data demonstrate the effectiveness and potential of our method.

**Keywords:** clustering, mixture models, multi-view learning.

## 1 Introduction

The most common approach for the machine learning setting, is to assume that data are represented in a single vector or graph space. However, in many real-life problems multi-view data arise naturally. Multi-view data are instances that have multiple representations (views) from different feature spaces. Usually these multiple views are from different vector spaces or different graph spaces or a combination of vector and graph spaces. The most typical example are web pages. Web pages can be represented with a term vector for the words in the web page text, another term vector for the words in the anchor text and a hyper-link graph.

The natural and frequent occurrence of multi-view data has raised interest in the so called *multi-view learning*. The main challenge of multi-view learning is to develop algorithms that use multiple views simultaneously, given the diversity of the views. Most studies on this topic address the semi-supervised classification problem and multi-view classification algorithms have often proven to utilize unlabeled data effectively and improve classification accuracy (e.g. [1,2,3]).

This work focuses on multi-view unsupervised learning and particularly in *multi-view clustering*. Multi-view clustering explores and exploits multiple representations simultaneously in order to produce a more accurate and robust

C. Alippi et al. (Eds.): ICANN 2009, Part II, LNCS 5769, pp. 205–214, 2009.

partitioning of the data than single view clustering. The available literature for this topic (e.g. [4,5,6,7]) is still limited, with encouraging results though. Borrowing the terminology of [7], there exist two approaches in multi-view clustering: *centralized* and *distributed*. Centralized algorithms simultaneously use all available views to cluster the dataset, while distributed algorithms first cluster each view independently from the others, using an appropriate single view algorithm, and then combine the individual clusterings to produce a final partitioning.

Most studies in multi-view clustering follow the centralized approach and extend well-known clustering algorithms to the multi-view setting. Bickel and Scheffer [4] developed a two-view EM and a two-view $k$-means algorithm under the assumption that the views are independent. They also studied the problem of mixture model estimation with more than two views and showed that co-EM [8] is a special case of their formulation [9]. De Sa [5] proposed a two-view spectral clustering algorithm that creates a bipartite graph of the views and is based on the "minimizing-disagreement" idea. This method also assumes that the views are independent. An algorithm that generalizes the single view normalized cut to the multi-view case and can be applied to more than two views was introduced by Zhou and Burges [6]. Following the distributed approach, Long *et al.* [7] proposed a general model for multi-view unsupervised learning which handles more than two views and representations from both vector and graph spaces.

In this paper we follow the centralized approach and present a multi-view clustering algorithm based on the *convex mixture model* of Lashkari and Golland [10]. Convex mixture models are a special case of mixture models that identify exemplars in the dataset by optimizing a convex criterion and have shown promising results in [10]. One of many attractive features is their applicability when only the dataset pairwise distance matrix is available and not the data points. The proposed *multi-view convex mixture model* finds exemplars *based on all views* and handles any number of views. The experiments with our algorithm demonstrate a considerable improvement on the clustering results compared to i) single view convex mixture models applied on the individual views and ii) single view convex mixture models that use the concatenation of the views.

The rest of this paper is organized as follows. Section 2 reviews the single view convex mixture model, while the proposed multi-view algorithm is presented in section 3. The experimental evaluation on artificial data and linked documents is discussed in section 4 and section 5 concludes this work.

## 2 Convex Mixture Models

Exemplar-based mixture models [10], also called *convex mixture models* (*CMM*), result in soft assignments of data points to clusters and in the extraction of representative exemplars from the dataset. They are simplified mixture models whose components equal in number the dataset size, the components' distributions are centered at the dataset points, thus representing all data points as cluster center candidates (candidate exemplars), and the only adjustable parameters are the components' priors.

Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\mathbf{x}_i \in \Re^d$ the convex mixture model distribution is $Q(\mathbf{x}) = \sum_{j=1}^{N} q_j f_j(\mathbf{x})$, $\mathbf{x} \in \Re^d$, where $q_j$ denotes the prior probability of the $j$-th component, satisfying the constraint $\sum_{j=1}^{N} q_j = 1$, and $f_j(\mathbf{x})$ is an exponential family distribution on random variable $\mathbf{X}$ with its expectation parameter equal to the $j$-th data point. Taking into account the bijection between regular exponential families and Bregman divergences [11], we write $f_j(\mathbf{x}) = C(\mathbf{x}) \exp(-\beta d_\varphi(\mathbf{x}, \mathbf{x}_j))$ with $d_\varphi$ denoting the Bregman divergence corresponding to the components' distributions.

A clustering is produced by maximizing the log-likelihood $L\left(\{q_j\}_{j=1}^{N}; \mathcal{X}\right)$, shown in (1), over $\{q_j\}_{j=1}^{N}$ s.t. $\sum_{j=1}^{N} q_j = 1$. The constant $\beta$ controlls the sharpness of the components and also *the number of clusters* identified by the convex mixture model when the soft assignments are turned into hard ones. Higher $\beta$ values result in more clusters in the final solution.

$$L\left(\{q_j\}_{j=1}^{N}; \mathcal{X}\right) = \frac{1}{N} \sum_{i=1}^{N} \log \left[ \sum_{j=1}^{N} q_j f_j(\mathbf{x}_i) \right] = \frac{1}{N} \sum_{i=1}^{N} \log \left[ \sum_{j=1}^{N} q_j e^{-\beta d_\varphi(\mathbf{x}_i, \mathbf{x}_j)} \right] + \text{const.} \tag{1}$$

The log-likelihood function (1) can be expressed in terms of the Kullback-Leibler (KL) divergence if we define $\hat{P}(\mathbf{x}) = 1/N, \mathbf{x} \in \mathcal{X}$ to be the empirical distribution of the dataset and by noting that

$$D(\hat{P} \| Q) = -\sum_{\mathbf{x} \in \mathcal{X}} \hat{P}(\mathbf{x}) \log Q(\mathbf{x}) - \mathbb{H}(\hat{P}) = -L\left(\{q_j\}_{j=1}^{N}; \mathcal{X}\right) + \text{const.}\,, \tag{2}$$

where $\mathbb{H}(\hat{P})$ is the entropy of the empirical distribution that does not depend on the parameters of the convex mixture model. Now the maximization of (1) is equivalent to the minimization of (2). This minimization problem is *convex* and can be solved with an efficient-iterative algorithm. As proved in [12], the updates on the components' prior probabilities are given by

$$q_j^{(t+1)} = q_j^{(t)} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\hat{P}(\mathbf{x}) f_j(\mathbf{x})}{\sum_{j'=1}^{N} q_{j'}^{(t)} f_{j'}(\mathbf{x})} \tag{3}$$

and the algorithm is *guaranteed to converge to the global minimum* as long as $q_j^{(0)} > 0, \forall j$. The prior probability $q_j$ associated with data point $\mathbf{x}_j$ is a measure of *how likely this point is to be an exemplar* and will be of great importance when we present our multi-view algorithm in section 3.

Clustering with a convex mixture model requires to select a value for the parameter $\beta$ ($0 < \beta < \infty$). It is possible to identify a reasonable range of $\beta$ values by determining a reference value $\beta_0$. In [10], the following empirical value (4) has been proposed, achieving good results in their experiments.

$$\beta_0 = N^2 \log N / \sum_{i,j} d_\varphi(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

Convex mixture models showed their potential when a Gaussian convex mixture model, i.e. with Euclidean distance as the Bregman divergence, outperformed a fully parametrized Gaussian mixture model in [10]. This proved that the smaller flexibility of convex mixture models, as $\{q_j\}_{j=1}^N$ are the only parameters, is well compensated by their ability to avoid the initialization problem and always locate the global optimum. Another important feature is that only the pairwise data distances take part in the calculation of the priors, thus the values of the data points are not required. As stated in [10], the method can be extended to any proximity data as long as the distance matrix $\mathbf{D}$ is available, by simply replacing $d_\varphi(\mathbf{x}_i, \mathbf{x}_j)$ with $D_{ij}$ in (1) and the convexity is not affected.

## 3 Multi-view Convex Mixture Models

Motivated by the potential and the advantages of the convex mixture models of section 2, in this work we extend them to data with multiple representations. Following the centralized approach, exemplars are identified by defining for each view a convex mixture model distribution, *with common priors $q_j$ across all views*, as well as the corresponding empirical distribution and minimizing the KL divergence between those two distributions summed over all views.

### 3.1 Model Description

Suppose we are given a dataset with $N$ instances, $\mathfrak{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and for each instance $V$ views are available. Let us define $\mathcal{X} = \{\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^V\}$, such that $\mathcal{X}^v$ contains the representations of the instances in the $v$-th view, i.e. $\mathcal{X}^v = \{\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_N^v\}$, $\mathbf{x}_i^v \in \Re^{d^v}$. Also, assuming that no prior information for the data is available in any view, define for each view a uniform empirical dataset distribution (5), as in [10], and also a convex mixture model distribution (6). Note that all distributions $\{Q^v(\mathbf{x})\}_{v=1}^V$ *share the same prior probabilities* $\{q_j\}_{j=1}^N$, but have different component distributions $f_j^v(\mathbf{x})$.

$$\hat{P}^v(\mathbf{x}) = \begin{cases} \frac{1}{N}, \mathbf{x} \in \mathcal{X}^v \\ 0, \quad \text{otherwise} \end{cases} \tag{5}$$

$$Q^v(\mathbf{x}) = \sum_{j=1}^N q_j f_j^v(\mathbf{x}) = C^v(\mathbf{x}) \sum_{j=1}^N q_j e^{-\beta^v d_\varphi^v(\mathbf{x}, \mathbf{x}_j^v)}\ ,\ \mathbf{x} \in \Re^{d^v} \tag{6}$$

Our aim is to locate high quality exemplars (cluster centroids) in the dataset, by considering all views simultaneously, around which the remaining instances will cluster. To achieve this, the proposed *multi-view convex mixture model* minimizes the sum of the KL divergences between the empirical distribution and the convex mixture distribution of each view, given by the following equation:

$$\min_{\substack{q_1,\ldots,q_N \\ \text{s.t. } \sum_{j=1}^N q_j = 1}} \left\{ \sum_{v=1}^V D(\hat{P}^v \| Q^v) = -\sum_{v=1}^V \sum_{\mathbf{x} \in \mathcal{X}^v} \hat{P}^v(\mathbf{x}) \log Q^v(\mathbf{x}) - \sum_{v=1}^V \mathbb{H}(\hat{P}^v) \right\}, \tag{7}$$

where $\mathbb{H}(\hat{P}^v)$ is the entropy of the empirical distribution of the $v$-th view that does not depend on the parameters of the multi-view convex mixture model.

It is well known that the sum of convex functions is also a convex function. Therefore, the above optimization problem, which is a generalization of the single view case, is also *convex*, since its objective function is the sum of the single view objectives which are convex functions. To solve (7) the same efficient-iterative algorithm as in (2) can be used. It can be shown that the updates on the components' prior probabilities are given by

$$q_j^{(t+1)} = \frac{q_j^{(t)}}{V} \sum_{v=1}^{V} \sum_{\mathbf{x} \in \mathcal{X}^v} \frac{\hat{P}^v(\mathbf{x}) f_j^v(\mathbf{x})}{\sum_{j'=1}^{N} q_{j'}^{(t)} f_{j'}^v(\mathbf{x})} \tag{8}$$

and the algorithm is *guaranteed to converge to the global minimum* as long as $q_j^{(0)} > 0, \forall j$. The prior $q_j$ associated with instance $\mathbf{x}_j$ is again a measure of *how likely this instance is to be an exemplar* and takes into account all views.

In the derivation of the above multi-view convex mixture model the following facts were considered. Different views can have very different statistical properties, therefore we allow the convex mixture model distribution (6) of each view to have its own $\beta$ value and Bregman divergence, i.e. different component distributions. For example, for one view we can use a Gaussian CMM and for another a Bernoulli CMM. An important property of the single view convex mixture model is convexity and we wish to preserve this property in the multi-view setting. As a result, summing the single view objectives to construct the multi-view objective is a natural choice. Finally, since our target is to extract representative exemplars from the dataset based on all views, we require all convex mixture model distributions to have common priors $q_j$. Intuitively, this means that an instance whose corresponding prior probability has a high value, is more or less a good exemplar in all views.

### 3.2 Algorithm Implementation

We follow the same steps as in [10] to implement the algorithm that optimizes (7). Letting $s_{ij}^v = \exp(-\beta^v d_\varphi^v(\mathbf{x}_i^v, \mathbf{x}_j^v))$ and using an auxiliary matrix $\mathbf{Z}$ and an auxiliary vector $\mathbf{n}$ we update the prior probabilities $q_j$ as follows:

$$Z_{iv}^{(t)} = \sum_{j=1}^{N} s_{ij}^v q_j^{(t)} \,, \quad n_j^{(t)} = \frac{1}{V} \sum_{v=1}^{V} \sum_{i=1}^{N} \frac{\hat{P}^v(\mathbf{x}_i^v) s_{ij}^v}{Z_{iv}^{(t)}} \,, \quad q_j^{(t+1)} = n_j^{(t)} q_j^{(t)} \,, \tag{9}$$

where $q_j^{(0)} > 0$ for all instances we want to consider as possible exemplars. Obviously, our formulation requires only the pairwise distances in each view and not the instances themselves in order to calculate the priors. Thus it can be extended to use proximity values, analogously to the single view case.

Suppose we wish to partition a multi-view dataset into $M$ disjoint clusters $C_1, C_2, \ldots, C_M$ using the multi-view convex mixture model. To identify $M$ exemplars (cluster centroids), the instances with the $M$ highest $q_j$ values are determined. Specifically, we run the algorithm until the $M$ highest $q_j$ values correspond to the same instances for a number of consecutive iterations. Moreover, we

require that the order among the $M$ highest $q_j$ values remains the same during these iterations. This convergence criterion differs from that in [10]. After finding the $M$ exemplars, we assign each of the remaining $N - M$ instances to cluster $C_k$, associated with the $k$-th exemplar, that has the largest posterior probability over all views, according to (10). Note that we refer to the exemplar instances as $\mathfrak{X}^E = \{\mathbf{x}_1^E, \mathbf{x}_2^E, \ldots, \mathbf{x}_M^E\} \subset \mathfrak{X}$ and their prior probabilities and component distributions in the $v$-th view as $q_k^E$ and $f_k^{vE}(\mathbf{x})$, $k = 1, \ldots, M$ respectively.

$$C_k = \{\mathbf{x}_k^E\} \cup \left\{ \mathbf{x}_i \left| \sum_{v=1}^{V} \frac{q_k^E f_k^{vE}(\mathbf{x}_i^v)}{\sum_{j=1}^{N} q_j f_j^v(\mathbf{x}_i^v)} > \sum_{v=1}^{V} \frac{q_l^E f_l^{vE}(\mathbf{x}_i^v)}{\sum_{j=1}^{N} q_j f_j^v(\mathbf{x}_i^v)}, \forall l \neq k, \mathbf{x}_i \notin \mathfrak{X}^E \right. \right\} \tag{10}$$

A final issue on the implementation of the multi-view convex mixture model is the choice of appropriate values for the $\beta^v$ parameters. Since a separate single view convex mixture model is defined for each view, we can identify a reasonable range of $\beta^v$ values in the same way as in the single view case. Following the ideas of the single view setting the following empirical $\beta_0^v$ value is derived:

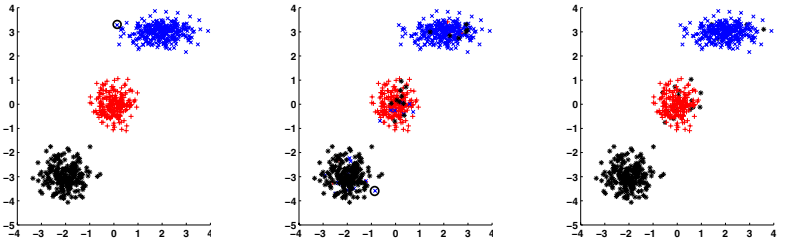
$$\beta_0^v = N^2 \log N / \sum_{i,j} d_\varphi^v(\mathbf{x}_i^v, \mathbf{x}_j^v) \; . \tag{11}$$

As for the complexity of the algorithm, calculation of the auxiliary quantities and the update of the priors costs $O(N^2V)$ scalar operations per iteration. If the distance matrices of the views are not given, computing the $s_{ij}^v$ quantities usually costs $O(N^2Vd)$, $d = \max\{d^1, d^2 \ldots, d^V\}$. Assuming $\tau$ iterations are required until convergence, the overall cost becomes $O(N^2V(\tau + d))$ scalar operations.

## 4 Experimental Evaluation

We aim to examine whether simultaneously considering all views helps to improve the clustering results obtained from the individual views, i.e. compare single view clustering to multi-view clustering. Since a very common approach to cluster multiple represented data is to concatenate all the views and then apply a single view algorithm on the concatenated view, we wish to investigate whether a multi-view algorithm provides any gains compared to a single view algorithm applied on the concatenated view. To answer these questions, we study the performance of the single view and multi-view convex mixture model on multi-view artificial data and two collections of linked documents, where multiple representations for the data occur naturally.

In all experiments we use Gaussian convex mixture models (Gaussian CMM), i.e. $d_\varphi^v(\mathbf{x}_i^v, \mathbf{x}_j^v) = \|\mathbf{x}_i^v - \mathbf{x}_j^v\|^2, \forall v$ and a uniform empirical dataset distribution (5). We report clustering results i) separately for each view, ii) for the concatenated view and iii) for the multiple views. To assess the clustering quality we use the *average entropy* metric, as in [4,5,9], which measures the impurity of the returned clusters. Average entropy is given by (12), where $N$ is the dataset size, $M$ the



(a) (b) (c)

**Fig. 1.** Examples of the artificial dataset: (a) the original dataset generated from three Gaussian distributions belonging to three classes; (b) one of the five views for $\omega = 50$ and zero translation; (c) clustering into three clusters with the three-view dataset ($\omega = 50$, $\beta^v = \beta_0^v$) using a Gaussian multi-view CMM. Only 25 instances are misplaced.

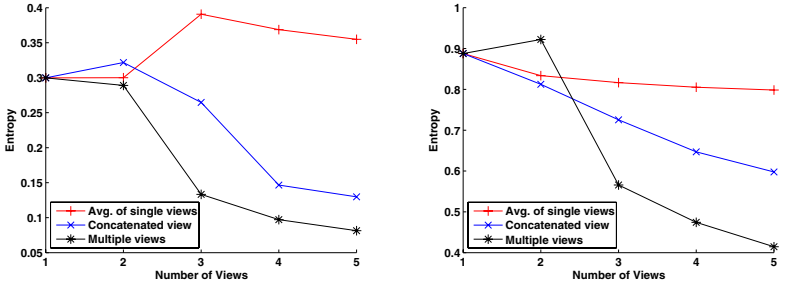
number of clusters, $c$ the number of classes, $n_i^j$ the number of points in cluster $i$ from class $j$ and $n_i$ the size of the $i$-th cluster. Lower average entropy values indicate that each cluster consists of instances belonging to the same class.

$$H = \sum_{i=1}^{M} \frac{n_i}{N} \left( -\sum_{j=1}^{c} \frac{n_i^j}{n_i} \log \frac{n_i^j}{n_i} \right) \tag{12}$$

### 4.1 Artificial Dataset

As a first step towards evaluating the performance of the multi-view convex mixture model, we generated a synthetic dataset, illustrated in Fig. 1(a) and consisting of 700 instances, from three two-dimensional Gaussian distributions. Each distribution represents a different class. Views were constructed with the following mechanism: for each view, we equally translated all instances of the original dataset and randomly selected $\omega$ of them to be moved to a different class. For example, assume that instance $\mathbf{x}_i$ had been selected, shown in circle in Fig. 1(a), that was generated by the first distribution (first class). We randomly picked one of the other two classes and generated a new point from the corresponding Gaussian distribution. This point, shown in circle in Fig. 1(b), is the representation of instance $\mathbf{x}_i$ in the view. Hence, an instance of the first class is now wrongly represented as an instance belonging to another class.

The above view generation mechanism will help us discover if simultaneously considering multiple views can correct some of the errors of the individual views and approach the optimum of $H = 0$ achieved by a convex mixture model on the well separated original dataset, since a convex mixture model on a single view will most probably misclassify all $\omega$ instances. For the experiments $\omega = 50$ and five views were generated, one of which is illustrated in Fig. 1(b). Five multi-view datasets were created, containing $1, \ldots, 5$ of the five views respectively. Results for these datasets are reported in Fig. 2(a) for three clusters and $\beta^v = \beta_0^v, \forall v$.




(a) $\omega = 50$, $\beta^v = \beta_0^v$. (b) $\omega = 200$, $\beta^v = \beta_0^v$.

**Fig. 2.** Artificial dataset results with Gaussian CMMs in terms of entropy for different number of views and three clusters

The multi-view convex mixture model constantly achieves the lowest entropy and it always outperforms the model that uses the concatenated view. Four of the five individual views have an entropy around 0.3 while one has $H = 0.57$. This view is included in the three-view dataset and explains the peak in the graph for the single views average. The corresponding clustering is illustrated in Fig. 1(c). At the same time our method achieves $H = 0.08$ with five views, confirming that it can considerably boost the clustering performance. Finally, the multi-view convex mixture model takes advantage of every available view as the entropy constantly falls as the number of views increases.

We also executed the same experiments as above, but with views for which $\omega = 200$. Fig. 2(b) depicts the results for this case. The multi-view convex mixture model is again the best algorithm and for five views it achieves $H = 0.41$, which is almost half the entropy of the individual views average and 33% less than the entropy of the concatenated view.

### 4.2 Document Archives

We selected two archives of linked documents. The *WebKB* dataset is a collection of academic web pages from computer science departments of various universities, while the *Citeseer* dataset is a collection of scientific publications. Both are very popular datasets for evaluating multi-view clustering algorithms [4,5,9] and multi-view classifiers [1,2]. We used the Bickel and Scheffer [9] version in which both collections have six classes and two or three views respectively. The first view of web pages is their text and the second the anchor text of the inbound links. Publications are represented in terms of a text view, consisting of the title and abstract of each paper, and two link views, made up of the inbound and outbound references. For some of the web pages no inbound links with anchor text exist, while some papers do not have inbound or outbound references. Such instances were removed, resulting in 2076 web pages and 742 papers.

For each view we generated tfidf vectors and normalized them to unit length (normalized tfidf), so that square Euclidean distances reflect the commonly used

**Table 1.** *WebKB* results with Gaussian CMMs in terms of entropy and six clusters

| Method-View | ***WebKB* Entropy** | |
|---|---|---|
| | $\beta^v = \beta_0^v$ | $\beta^v = \alpha\beta_0^v$ |
| **Single view CMM-text** | 1.54 | 1.49 ($\alpha = 1.5$) |
| **Single view CMM-anchor text** | 1.55 | 1.44 ($\alpha = 3.5$) |
| **Single view CMM-concat. text & anchor text** | 1.56 | 1.48 ($\alpha = 1.7$) |
| **Multi-view CMM-text & anchor text** | **1.5** | **1.4** ($\alpha = 1.5$) |

**Table 2.** *Citeseer* results with Gaussian CMMs in terms of entropy and six clusters

| Method-View | ***Citeseer* Entropy** | |
|---|---|---|
| | $\beta^v = \beta_0^v$ | $\beta^v = \alpha\beta_0^v$ |
| **Single view CMM-text** | 1.61 | 1.56 ($\alpha = 0.5$) |
| **Single view CMM-inbound references** | 1.65 | 1.65 ($\alpha = 1$) |
| **Single view CMM-outbound references** | 1.57 | 1.56 ($\alpha = 1.5$) |
| **Single view CMM-concat. text & two link views** | 1.6 | 1.54 ($\alpha = 1.5$) |
| **Multi-view CMM-text & two link views** | **1.5** | **1.5** ($\alpha = 1$) |

cosine similarity. Both datasets were partitioned into six clusters. Tables 1 and 2 report results for the *WebKB* and *Citeseer* collections respectively, where the multi-view convex mixture model is compared to its single view counterpart.

In a first series of experiments we set $\beta^v = \beta_0^v, \forall v$. As can be seen, the multi-view algorithm improves the clustering of the individual and concatenated views, making once again apparent the potential of our method and the advantages of using simultaneously multiple views. Remarkably, for both datasets the concatenated view's performance is even worse than that of some of the single views. This result explains the need to develop multi-view algorithms and not resort to tricks that allow single view algorithms to handle multiple represented data.

We also investigated the impact of the $\beta^v$ parameter by searching around the range of values defined by $\beta_0^v$ and selecting the fraction $\alpha$ of $\beta_0^v$ that yields the smallest entropy (shown in parentheses in Tables 1, 2). A decrease in entropy for the two collections can be observed and again the multi-view convex mixture model is the best performer. Note that setting $\beta^v = \beta_0^v$ is the best choice for the inbound references view and the multi-view setting ($\alpha = 1$) of the *Citeseer* dataset, indicating that $\beta_0^v$ provides a good range of values for the $\beta^v$ parameter.

## 5 Conclusions and Future Work

We have proposed the multi-view convex mixture model, a method that extends convex mixture models [10] to the multi-view case and identifies exemplars in the dataset by simultaneously considering all available views. The main advantages of our method are the convexity of the optimized objective, the ability to handle

views with different statistical properties and its applicability when only pairwise distances are available and not the data points. Our empirical evaluation with multi-view artificial data and two popular document collections, showed that the presented algorithm can considerably improve the results of a single view convex mixture model based either on the individual views or the concatenated view.

As for future work, we plan to compare our algorithm to other multi-view methods and experiment using additional datasets so as to thoroughly investigate the potential of the multi-view convex mixture model. We also aim to use multi-view convex mixture models in conjunction with other clustering algorithms which will treat the exemplars as a good initialization. Finally, another interesting research direction is the assignment of different weights to different views and the ability to learn those weights automatically under our framework.

**Acknowledgments.** We would like to thank Steffen Bickel and Tobias Scheffer for kindly providing their processed datasets.

## References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, pp. 92–100 (1998)
2. Muslea, I., Minton, S., Knoblock, C.A.: Active + semi-supervised learning = robust multi-view learning. In: Proceedings of the 19th International Conference on Machine Learning, pp. 435–442 (2002)
3. Brefeld, U., Scheffer, T.: Co-em support vector learning. In: Proceedings of the 21st International Conference on Machine Learning (2004)
4. Bickel, S., Scheffer, T.: Multi-view clustering. In: Proceedings of the 4th IEEE International Conference on Data Mining, pp. 19–26 (2004)
5. de Sa, V.R.: Spectral clustering with two views. In: Proceedings of the 22nd International Conference on Machine Learning Workshop on Learning with Multiple Views, pp. 20–27 (2005)
6. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1159–1166 (2007)
7. Long, B., Yu, P.S., Zhang, Z.M.: A general model for multiple view unsupervised learning. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 822–833 (2008)
8. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Proceedings of the 9th International Conference on Information and Knowledge Management, pp. 86–93 (2000)
9. Bickel, S., Scheffer, T.: Estimation of mixture models using co-em. In: Proceedings of the 16th European Conference on Machine Learning, pp. 35–46 (2005)
10. Lashkari, D., Golland, P.: Convex clustering with exemplar-based models. In: Advances in Neural Information Processing Systems, vol. 20, pp. 825–832 (2008)
11. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Machine Learning Research 6, 1705–1749 (2005)
12. Csiszár, I., Shields, P.C.: Information theory and statistics: A tutorial. Communications and Information Theory 1(4), 417–528 (2004)