A Probability Model for Projective Clustering on High Dimensional Data

Lifei Chen
Department of Computer Science
Fujian Normal University
Fuzhou, 350108, China
clifei@fjnu.edu.cn

Qingshan Jiang*
Software School
Xiamen University
Xiamen, 361005, China
qjiang@xmu.edu.cn

Shengrui Wang
Department of Computer Science
University of Sherbrooke
Quebec, J1K 2R1, Canada
Shengrui.Wang@usherbrooke.ca

Abstract

Clustering high dimensional data is a big challenge in data mining due to the curse of dimensionality. To solve this problem, projective clustering has been defined as an extension of traditional clustering that seeks to find projected clusters in subsets of dimensions of a data space. In this paper, the problem of modeling projected clusters is first discussed, and an extended Gaussian model is proposed. Second, a general objective criterion used with K-means type projective clustering is presented based on the model. Finally, the expressions to learn model parameters are derived and then used in a new algorithm named FPC to perform fuzzy clustering on high dimensional data. The experimental results on document clustering show the effectiveness of the proposed clustering model.

1. Introduction

The problem of clustering data has been studied extensively and a lot of work has been done in the area of clustering. However, many real-world data sets consist of very high dimensional features. In high dimensional spaces, data points may cluster differently in varying subspaces comprised of different combinations of dimensions [9, 7]. To address these challenges, projective clustering has been defined to find clusters in different subspaces of the same data set [9, 7, 1].

A projected cluster is a subset of points together with a subset of attributes. A number of algorithms were proposed to find such projected clusters in the literature and they falls into two categories. The first category includes PROCLUS [1], ORCLUS [2], and P3C [4] etc., that aims to find out the exact subspaces of different clusters. The second category is to cluster data points in the entire data space but assign different weighting values to different dimensions of clusters. The methods in this latter category, alternatively known as the soft subspace clustering, include FWKM [6], LAC [3] and EWKM [8] etc.

In this paper, we are interested in the projected clusters modeling, that is, to model the population of points that may form the structure discovered by projective clustering algorithms. Describing the population by a model based on the sample on which the clustering is applied allow us gain insights and to learn the most important aspects of the population [12]. However, most of the existing clustering methods only focus on the grouping aspect of clustering and lack the ability to provide such model [12].

Due to the empty space phenomenon [10] and the property of subspace clustering as mentioned above, the cluster modeling on high dimensional data become a difficult problem. There are a few related work in literature. Hoff [11] proposed a model of “clustering shifts in mean and variance” based on a nonparametric mixture of sequences of independent normal random variables. The model is learned by Markov Chain Monte Carlo [11], however, the computational cost may be prohibitive. Harpaz et al. [12] presented a non-parametric density estimation modeling technique, where the data are described as a mixture of linear manifolds. A Bayesian approach is used to identify groups of points that fit or are embedded in lower dimensional linear manifolds. Like Hoff [11], the objective criterion for clustering based on the model was not able to be expressed, resulting in the somewhat inefficient clustering process.

The first contribution of this paper is the introduction of a probability model for projective clustering. We present an extended Gaussian model, which by analysis is compatible with the general views on projective clustering [7, 4]. Moreover, we derive an objective criterion to be optimized by the learning algorithm. The second contribution is the proposal of a fuzzy algorithm FPC (Fuzzy Projective Clustering) based on the model for fuzzy projective clustering. FPC generates “soft” partitions of the data set and discovers clusters in the “soft” subspaces, which are identified by assigning different weighting values between 0 and 1 to all the dimensions. These two advantages make FPC more practicable in many applications, e.g., document clustering in text
mining. The performances of FPC are finally evaluated in clustering high dimensional document data, and the experimental results shown its effectiveness and stability.

The remainder of this paper is organized as follows. Section 2 presents notations and formalization of the projected clusters. Our probability model and the general clustering criterion are introduced in Section 3. Section 4 describes the new algorithm FPC. Section 5 shows experimental results. Finally, Section 6 presents our conclusions.

2. Projected Clusters

We formalize the projected clusters in this section. The notations used in the paper are summarized in Table 1.

<table>
<thead>
<tr>
<th>A = (A1,A2, ...,Ad)</th>
<th>Set of attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_i = (x_{i1},x_{i2}, ...,x_{id})</td>
<td>i-th data point in ( \mathbb{R}^d ), i=1,2, ... ,n</td>
</tr>
<tr>
<td>DB = (x_1,x_2, ...,x_n)</td>
<td>The data set</td>
</tr>
<tr>
<td>K</td>
<td>Number of clusters</td>
</tr>
<tr>
<td>C = (C1,C2, ...,C_K)</td>
<td>Partition of DB into K clusters</td>
</tr>
<tr>
<td>u_{ki}</td>
<td>Membership degree of x_i to C_k, ( k=1,2, ...,K )</td>
</tr>
<tr>
<td>v_k = (v_{k1},v_{k2}, ...,v_{kd})</td>
<td>Cluster center of C_k</td>
</tr>
<tr>
<td>w_k = (w_{k1},w_{k2}, ...,w_{kd})</td>
<td>A weight vector associated with C_k</td>
</tr>
<tr>
<td>W_k = \text{diag}(\sqrt{w_{k1}},\sqrt{w_{k2}}, ..., \sqrt{w_{kd}})</td>
<td>A d \times d diagonal matrix associated with C_k.</td>
</tr>
<tr>
<td>e_{k1}, e_{k2}, ..., e_{kd}</td>
<td>d linear independent vectors associated with C_k.</td>
</tr>
<tr>
<td>E_k = {e_{kj}}_{d \times d}</td>
<td>Matrix designating the space in which C_k exists</td>
</tr>
<tr>
<td>I_d</td>
<td>A d \times d identical matrix</td>
</tr>
<tr>
<td>x \in C_k</td>
<td>A point in C_k</td>
</tr>
<tr>
<td>y</td>
<td>Projection of x into the subspace in which C_k exists</td>
</tr>
</tbody>
</table>

Generally, a projected cluster must represent not only the data points group, but also the subspace in which it exists. Consider a partition \( C_k(=1,2, ...,K) \) of DB, since the coordinates of each point \( x \in C_k \) are determined uniquely when given a set of linear independent vectors \( e_{k1}, e_{k2}, ..., e_{kd} \) which form a basis of the space, we adopt \( E_k \) to represent the space in which \( C_k \) exists. In particular, if \( E_k \) is an identical matrix, i.e., \( E_k = I_d \), it degenerates to denote the original data space \( \mathbb{R}^d \). If \( E_k \) is a diagonal matrix, the space will be axis-aligned. Otherwise, the space is non-axis-aligned or called of arbitrarily oriented, as described in ORCLUS [2], where the new attributes are typically combined from the original dimensions of the data set.

Additionally, each partition \( C_k \) is associated with a weight vector \( w_k \) satisfying with:

\[
0 \leq w_{kj} \leq 1, \quad 1 \leq j \leq d
\]

and, the trace of the matrix \( W_k \) is fixed to a constant, i.e.,

\[
\text{Trace}(W_k) = \sum_{j=1}^{d} \sqrt{w_{kj}} = 1 \text{(any constant)} \quad (1)
\]

Here, the value of \( w_{kj} \) represents how much the \( j \)-th dimension is relevant to the partition \( C_k \).

**Definition 1.** (subspace) Let \( S_k = W_k E_k \) be the subspace of \( \mathbb{R}^d, k=1,2, ...,K \).

The matrix \( W_k \) here plays a role of feature selection from \( E_k \). According to the constraints of Eqn.(1), if all but one value of \( w_{kj} \) is 0 for \( j=1,2, ...,d \), \( S_k \) is actually a hard subspace; otherwise a soft subspace, as described in the soft subspace clustering algorithms [6, 3, 8]. Based on Def.1, we can introduce the definition of projected clusters.

**Definition 2.** (projected cluster) Let \( SC_k = (S_k, C_k) \) be the \( k \)-th projected cluster of DB, \( k=1,2, ...,K \).

In order to examine the distribution of data points in projected clusters, we need project them into their subspaces.

**Definition 3.** (projection) Let \( x \) be a data point of \( SC_k \) and \( y \) the projection in its subspace,

\[
y = \pi_{S_k}(x) = x(W_k E_k)^T
\]

Def.3 can be validated in the following derivation. Consider the Euclidean distance of two points \( y_1 \) and \( y_2 \), we have

\[
D_{S_k}(y_1, y_2) = \sqrt{(\pi_{S_k}(x_1) - \pi_{S_k}(x_2))^T (\pi_{S_k}(x_1) - \pi_{S_k}(x_2))}^T
\]

\[
= \sqrt{(x_1 - x_2)^T E_k^T W_k^T E_k (x_1 - x_2)^T}
\]

If \( E_k = I_d \), the above equation can be simplified into

\[
D_{S_k}(y_1, y_2) = \sqrt{\sum_{j=1}^{d} w_{kj} (x_{1j} - x_{2j})^2}
\]

Eqn.(3) is the very weighted Euclidean distance, which is commonly used in existing algorithms [6, 3, 8].

3. A Projective Clustering Model

In this section, we first model the projected clusters, then derive the objective criterion to be optimized in the procedure of discovering projected clusters.
3.1. Projected Clusters Modeling

It is important to note that the Gaussian mixture is a fundamental hypothesis that many partition-based and model-based clustering algorithms make regarding the data distribution model. In this case, data points are thought of as originating from various possible sources, and the data from each particular source is modeled by a Gaussian [13]. However, in high dimensional space, Gaussian functions are no more appropriate. Verleysen [10] state that, when the dimension increases, the percentage of the samples of a normalized scalar Gaussian distribution falling around its center would rapidly decrease to 0. We then are motivated to examine the marginal distribution on each dimension at first. Figure 1 gives an example.

The left figure of Figure 1 shows a data points group in 2-dimensional space, and the right one illustrates the histograms of the projections on the axis $e_{kj}$. It is reasonable to describe the projections using a 1-dimensional Gaussian function. Formally, on the $j$-th projected dimension, the probability density function is

$$G(y_j | \mu_{kj}, \sigma_k) = \frac{1}{\sqrt{2\pi} \sigma_k} \exp \left( -\frac{1}{2\sigma_k^2} (y_j - \mu_{kj})^2 \right)$$

where $\mu_{kj}$ and $\sigma_k$ denote mean and covariance of the Gaussian. Using Def.3 and Eqn.(2), we can transform the density function into the original data space and obtain

$$G(x_j | v_{kj}, \sigma_k) = \frac{1}{\sqrt{2\pi} \sigma_k} \exp \left( -\frac{w_{kj}}{2\sigma_k^2} \left( (x - v_k) \cdot e_{kj}^T \right)^2 \right)$$

(4)

The major difference between Eqn.(4) and the standard Gaussian is the introduction of weighting value $w_{kj}$, indicating the contribution of the j-th dimension to $C_k$. According to Eqn.(4), with a larger weighting value, the data points would distribute within a smaller range. The new characteristic of this extended Gaussian has met with the general requirements of projective clustering. Virtually all existing projective algorithms are based on the following general views [7, 4]: the points project along a significant dimension onto a smaller range of values than that of the other dimensions; the points are more likely uniformly distributed along each irrelevant dimension.

Based on Eqn.(4), we then suppose the $j$-th dimension of $n$ inputs $x_1, x_2, \ldots, x_n$ are independently and identically distributed from the following mixture density population:

$$F(x_j; \Theta_j) = \sum_{k=1}^{K} \alpha_k G(x_j | v_{kj}, \sigma_k)$$

with

$$\sum_{k=1}^{K} \alpha_k^d = 1, \alpha_k \geq 0, k = 1, 2, \ldots, K$$

(5)

where $\Theta_j = \{ (\alpha_k, v_{kj}, S_k, \sigma_k) | 1 \leq k \leq K \}$ is the set of parameters, and $\alpha_k$ denotes the mixing weight of the $k$-th component of the model. Our goal is to estimate the parameters $\Theta_j (j=1, 2, \ldots, d)$ for the $d$ marginal distributions, subjected to the constrains of Eqn.(1). Next, we will discuss an objective criterion to be optimized for this goal.

3.2. Objective Criterion for Clustering

Suppose $\hat{\Theta}_j = \{ (\hat{\alpha}_k, \hat{v}_{kj}, \hat{S}_k, \hat{\sigma}_k) | 1 \leq k \leq K \}$ is an estimator of $\Theta_j$. The distance between $F(x_j; \Theta_j)$ and $\hat{F}(x_j; \hat{\Theta}_j)$ can be measured by the following Kullback-Leibler divergence function [14]:

$$R(\hat{\Theta}_j) = \int \frac{F(x_j; \Theta_j) \ln F(x_j; \Theta_j)}{\hat{F}(x_j; \hat{\Theta}_j)} dx_j$$

The above equation can be decomposed into two components and the one $\int F(x_j; \Theta_j) \ln F(x_j; \Theta_j) dx_j$ is a constant irrelevant to $\Theta_j$, therefore, the following objective criterion need to be maximized (the symbol ↑ and ↓ in front of the function indicate maximizing and minimizing the function, respectively).

$$\uparrow Q_1(\hat{\Theta}_j) = \int F(x_j; \Theta_j) \ln \hat{F}(x_j; \hat{\Theta}_j) dx_j$$

$$= \sum_{k=1}^{K} \int p(k | x_j) F(x_j; \Theta_j) \ln \hat{F}(x_j; \hat{\Theta}_j) dx_j$$

with

$$p(k | x_j) = \frac{\hat{\alpha}_k G(x_j | \hat{v}_{kj}, \hat{\sigma}_k)}{\hat{F}(x_j; \hat{\Theta}_j)}, 1 \leq k \leq K$$

(6)

where $p(k | x_j)$ is the posterior probability of an input $x_j$ from the $k$-th probability density function as given $x$. Substituting $\hat{F}(x_j; \hat{\Theta}_j)$ with Eqn.(6) for $Q_1(\hat{\Theta}_j)$, and adding up the functions of $j=1, 2, \ldots, d$, we have

$$\uparrow Q_2(\Theta) = \sum_{k=1}^{K} \sum_{j=1}^{d} \int p(k | x_j) F(x_j; \Theta_j) \ln \frac{\hat{\alpha}_k G(x_j | \hat{v}_{kj}, \hat{\sigma}_k)}{p(k | x_j)} dx_j$$
with $\Theta = \{\Theta_j|1 \leq j \leq d\}$ as the estimators of $\Theta = \{\Theta_j|1 \leq j \leq d\}$. It can be seen that, given a data set $DB$, maximizing $Q_2(\hat{\Theta})$ is equivalent to the maximum likelihood (ML) learning of $\Theta$ from all the inputs $x_1, x_2, \ldots, x_n$. By the law of large number and replace $G(x_j|\hat{v}_{kj}, \hat{\sigma}_k)$ with Eqn.(4), the objective criterion can be further simplified into

$$\downarrow Q_2(\hat{\Theta}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=1}^{n} p(k|x_{ij}) \times \left( \frac{\hat{v}_k}{\sqrt{2\pi}\hat{\sigma}_k} - (x_i - \hat{v}_k) \cdot \hat{\epsilon}_{kj} \right)^2 - \ln \frac{\hat{v}_k}{\sqrt{2\pi}\hat{\sigma}_k} + \ln p(k|x_{ij})$$

By applying the objective criterion to clustering, the posterior probability $p(k|x_{ij})$ can be thought of as the fuzzy assignment measure $u_{ki}$, i.e., $\forall j \in 1, 2, \ldots, d : p(k|x_{ij}) = u_{ki}$ with the constrains of

$$0 \leq u_{ki} \leq 1; \sum_{k=1}^{K} u_{ki} = 1, i = 1, 2, \ldots, n \quad (7)$$

Upon these facts, the resulting objective criterion can be obtained as follows.

$$\downarrow J(C, V, W, E) = \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=1}^{n} u_{kj} w_{ij} \left( \frac{x_i - V_{k}}{\sqrt{2\pi}\sigma_k} \right)^2 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=1}^{n} u_{kj} \ln u_{ki}}{\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=1}^{n} u_{kj} \ln u_{ki}} \quad (8)$$

If all the $\alpha_k$s and $\sigma_k$s are forced to be constants, and only the “hard” clustering(i.e., $u_{ki} \in \{0,1\}$) and axis-aligned subspaces are considered, Eqn.(8) would degenerate to

$$\downarrow J_1(C, V, W) = \sum_{k=1}^{K} \sum_{j=1}^{d} w_{kj} \sum_{x_i \in C_k} (x_{ij} - v_{kj})^2$$

which is exactly the error measure that popularly used in the existing clustering algorithms [3, 8].

Since all inputs $x_1, x_2, \ldots, x_n$ are available, the learning of $\Theta$ via minimizing Eqn.(8) can be implemented by the Expectation Maximization (EM) algorithm in a batch way. Next, we will prefer to perform clustering and parameter learning adaptively in analog with the previous $k$-means.

### 4. A Fuzzy Algorithm for Projective Clustering

In this section, an algorithm named FPC(Fuzzy Projective Clustering) is proposed, based on the general clustering criterion of Section 3.2. For ease of description, and more importantly, to make the resulting projected clusters easily to be interpreted, the algorithm will only consider the axis-aligned subspaces, i.e., we always suppose $E_k = I_d$ in the following.

### 4.1. Learning Parameters

For the purpose, we construct the following objective function for clustering algorithm:

$$J_2(C, V, W, E) = J(C, V, W, E) + \frac{\sum_{k=1}^{K} \sum_{j=1}^{d} \sqrt{w_{kj}} - 1}{\eta} + \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ki} - 1}{\zeta_i}$$

where $\lambda_k(k=1, 2, \ldots, K)$, $\eta$ and $\zeta_i(i=1, 2, \ldots, n)$ are the Lagrange multipliers to introduce the constrains defined in Eqn.(1), Eqn.(5) and Eqn.(7), respectively. By setting the gradient of $J_2$ with respect to $u_{ki}$ and $\zeta_i$ to zero, a novel definition of the “soft” assignment measure for projective clustering can be derived, as follows.

$$u_{ki} = \left( \sum_{i=1}^{n} \frac{\alpha_k}{\sigma_k} \exp \left( -\frac{1}{2\sigma_k^2} \sum_{j=1}^{d} w_{ij} (x_{ij} - v_{ij})^2 \right) \right)^{-1} \times \frac{\alpha_k}{\sigma_k} \exp \left( -\frac{1}{2\sigma_k^2} \sum_{j=1}^{d} w_{kj} (x_{ij} - v_{kj})^2 \right) \quad (9)$$

Note that the definition of Eqn.(9) is in exponential type and based on the weighted Euclidean distance, which is quite different from the one used in the traditional algorithm, like FCM [5]. To learn the optimal values of dimension weights, we now set the gradient of $J_2$ with respect to $w_{kj}$ to zero and obtain

$$w_{kj} = \left( \frac{1}{X_{kj}} \right)^2 \quad (10)$$

with

$$X_{kj} = \sum_{i=1}^{n} u_{ki} (x_{ij} - v_{kj})^2$$

Following FWKM [6], to ensure that the denominator of Eqn.(10) is always larger than 0, we then adjust the denominator by adding an additional factor $\delta = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - o_j)^2$ [6], where $o_j$ is the mean feature value of the entire data set. The numerator of Eqn.(10) can be determined by setting $\frac{\partial J_2}{\partial \alpha_k} = 0$. In conclusion, we compute the weighting values for each dimension using the following Eqn.(11).

$$w_{kj} = \left( \frac{1}{X_{kj} + \delta} \right)^2 \left( \sum_{j=1}^{d} \frac{1}{X_{kj} + \delta} \right)^2 \quad (11)$$

Also, let $\frac{\partial J_2}{\partial \eta} = 0$ for $k=1, 2, \ldots, K$ and $\frac{\partial J_2}{\partial \eta} = 0$, we have
for each document cluster. Furthermore, FPC can make use of the feature weights to identify the candidates of few keywords for each document cluster.

\[
\alpha_k = \left( \frac{u_{k+}}{n} \right)^{\frac{1}{2}}
\]

with \( u_{k+} = \sum_{i=1}^{n} u_{ki} \). Setting \( \frac{\partial J}{\partial \alpha_k} = 0 \) for \( k = 1,2,\ldots,K \) results in:

\[
\sigma_k^2 = \frac{1}{d u_{k+}} \sum_{j=1}^{d} \sum_{i=1}^{n} u_{ki} (x_{ij} - v_{kj})^2
\]

To learn the cluster centers, we set \( \frac{\partial J}{\partial v_{kj}} = 0 \) for \( k=1,2,\ldots,K, j=1,2,\ldots,d \) and give

\[
v_{kj} = \frac{1}{u_{k+}} \sum_{i=1}^{n} u_{ki} x_{ij}
\]

The above equation is the same as the one defined in FCM [5](in case of the fuzzifier equals to 1).

4.2. The FPC Algorithm

Our algorithm FPC is summarized in Algorithm 1.

**Algorithm 1 FPC algorithm**

Input: \( DB, K \) and a termination criterion \( \varepsilon \)

Output: \( U = \{ u_{ki} | k=1,2,\ldots,K; i=1,2,\ldots,n \} \) and the associated weights of \( K \) clusters \( w_1,w_2,\ldots,w_K \)

begin

1. Initialization

   1.1 Randomly choose \( K \) cluster centers.

   Let \( V = \{ v_{kj} | k=1,2,\ldots,K; j=1,2,\ldots,d \} \)

   1.2 Let \( p \) be the number of iteration, \( p = 0 \);

   Denote \( V \) as \( V(p) \)

   1.3 For \( k=1,2,\ldots,K, j=1,2,\ldots,d \) and \( i=1,2,\ldots,n \),

   set \( \tilde{w}_{kj} = \frac{1}{d}, u_{ki} = \frac{1}{K} \); Set \( \alpha_k \) and \( \sigma_k \) to an constant

2. Repeat

   2.1 Update \( U \) using Eqn.(9)

   2.2 Update \( w_k(k=1,2,\ldots,K) \) using Eqn.(11)

   2.3 Update \( \alpha_k(k=1,2,\ldots,K) \) using Eqn.(12)

   2.4 Update \( \sigma_k(k=1,2,\ldots,K) \) using Eqn.(13)

   2.5 Update \( v_k(k=1,2,\ldots,K) \) using Eqn.(14)

   2.6 Set \( p=p+1 \) and denote \( V \) as \( V(p) \)

until \( \| \| V(p) \|_\infty - \| V(p-1) \|_\infty \| < \varepsilon \)

3. Output \( U \) and \( w_1,w_2,\ldots,w_K \)

end

From Algorithm 1, FPC not only identifies clusters in the “soft” subspaces, but also generates “soft” partitions of the data set. These two advantages make FPC practicable in document clustering of text mining. Due to the diversity, a document may include multiple topics and thus may relate to multiple categories at the same time. Fuzzy clustering is useful in this case. Furthermore, FPC can make use of the feature weights to identify the candidates of few keywords for each document cluster.

5. Experimental Evaluation

In this section, we evaluate the performance of our algorithm FPC to clustering document and compare its performance with FWKM [6], PROCLUS [1] and the classical fuzzy algorithm, FCM [5]. Both FWKM [6] and PROCLUS [1] are “hard” clustering algorithms.

We use two corpus named Email-1431(available at http://www2.imm.dtu.dk/~rem/data/) and Ling-Spam(available at http://www.aueb.gr/users/ion/data/), and one data set extracted from TanCorp(available at http://lcc.software.ict.ac.cn/~tansongbo/corpus1.php). The detailed descriptions are summarized in Table 2.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Dimension</th>
<th># categories</th>
<th># documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email-1431</td>
<td>2002</td>
<td>2</td>
<td>1431</td>
</tr>
<tr>
<td>Ling-Spam</td>
<td>4435</td>
<td>2</td>
<td>2894</td>
</tr>
<tr>
<td>Tan-Science</td>
<td>1174</td>
<td>4</td>
<td>1040</td>
</tr>
</tbody>
</table>

In our experiments, documents are represented in vector-space model(VSM) where each document is a vector in the word space and each element of the vector indicates the frequency of the corresponding word in the document. All the data sets are normalized such that each entry \( x_{ij} \) of the document satisfies \( x_{ij} \in [0,1] \).

To measure the goodness of fuzzy clusterings, we employ two well-known cluster validity indices(CVIs) \( V_{PC} \) and \( V_{PE} \) [5]. The larger the values of \( V_{PC} \) and the smaller the values of \( V_{PE} \), the better the clustering solution is. We also used the Micro-F1 and Macro-F1 [8] to measure the clustering accuracy on document categories. Since both the two measurements are adopted for the “hard” clusterings, we will try to transform the fuzzy assignments into hard ones by extracting the cluster with the highest degree of membership for each data point. Formally, each data point \( x_i \) is considered to be of cluster \( C_l \) according to the rule:

\[
l = \arg \max_{k=1,2,\ldots,K} u_{ki}
\]

In the following, we first report the best performance from ten executions of clustering on the data sets. Table 3 illustrates the values of \( V_{PC} \) and \( V_{PE} \) yielded by our algorithm FPC and competing one FCM [5].

<table>
<thead>
<tr>
<th>Data set</th>
<th>FPC</th>
<th>FCM</th>
<th>( V_{PC} )</th>
<th>( V_{PE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email-1431</td>
<td>0.90087</td>
<td>0.14483</td>
<td>0.50001</td>
<td>0.69315</td>
</tr>
<tr>
<td>Ling-Spam</td>
<td>0.80960</td>
<td>0.26912</td>
<td>0.50001</td>
<td>0.69315</td>
</tr>
<tr>
<td>Tan-Science</td>
<td>0.69151</td>
<td>0.49511</td>
<td>0.25000</td>
<td>1.38630</td>
</tr>
</tbody>
</table>
It can be seen from Table 3 that FPC is able to achieve results with high quality. On the contrary, the classical algorithm FCM approximately failed in clustering such high dimensional data. Note that the number of clusters ($K$) is 2, 2 and 4 in the three data sets, respectively. The values of $V_{PC}$ reported by FCM are actually closed to $1/K$. To examine the effectiveness of algorithms more intuitively, the clustering results (of FPC and FCM) were transformed using Eqn.(15) and then compared with the “hard” clustering algorithms (FWKM [6] and PROCLUS [1]). Due to the randomness in initialization of all the four algorithms, the clustering results may somewhat be of unstable. Table 4 shows the average accuracy over ten executions in Micro-F1 and Macro-F1.

<table>
<thead>
<tr>
<th>Data set</th>
<th>FPC</th>
<th>FCM</th>
<th>PROCLUS</th>
<th>FWKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email-1431</td>
<td>0.9224</td>
<td>0.5563</td>
<td>0.6604</td>
<td>0.6974</td>
</tr>
<tr>
<td>Ling-Span</td>
<td>0.9218</td>
<td>0.5256</td>
<td>0.6567</td>
<td>0.6567</td>
</tr>
<tr>
<td>Tan-Science</td>
<td>0.8383</td>
<td>0.6146</td>
<td>0.8400</td>
<td>0.8410</td>
</tr>
</tbody>
</table>

From Table 4, FPC have achieved the highest average accuracy in most situations. For example, on Email-1431 both the Micro-F1 and Macro-F1 are improved more than 20% comparing with three competing algorithms. We also observed that some algorithms, such as FWKM [6], have obtained clustering results with considerable high quality in the experiments, however, the algorithms tended to fail in most of the clustering processes. It is because such algorithms are strong dependent on the initial clustering which is hard to determine since it is performed in full-dimensional space. The performance achieved by FPC is largely due to the fuzzy projective clustering on high dimensional data that helps to discover the exact subspaces for each partition in the entire space.

6. Conclusions

In this paper, we first discussed the problem of providing a probability model to describe the projected clusters in high dimensional data. Based on the extended Gaussian model we proposed, an objective criterion was derived, it is because such algorithm FCM approximately failed in clustering such high dimensional data. Note that the number of clusters ($K$) is 2, 2 and 4 in the three data sets, respectively. The values of $V_{PC}$ reported by FCM are actually closed to $1/K$. To examine the effectiveness of algorithms more intuitively, the clustering results (of FPC and FCM) were transformed using Eqn.(15) and then compared with the “hard” clustering algorithms (FWKM [6] and PROCLUS [1]). Due to the randomness in initialization of all the four algorithms, the clustering results may somewhat be of unstable. Table 4 shows the average accuracy over ten executions in Micro-F1 and Macro-F1.

7. Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.10771176; the Scientific Research Foundation of Xiamen University under Grant No.0630-X01117 and the National 985 Project of China under Grant No.0000-X07204.

References