Uncorrelated Local Maximum Margin Criterion: An Efficient Dimensionality reduction Method for Text Classification

Koushik Mallick* and Siddhartha Bhattacharyya

*Department of Computer Science & Engineering, RCC Institute of Information Technology, Kolkata, India

**Department of Information Technology, RCC Institute of Information Technology, Kolkata, India

Abstract

The pragmatic realism of the high dimensionality incurs limitations in many pattern recognition arena such as text classification, data mining, information retrieval and face recognition. The unsupervised PCA ante up no attention to the class labels of the existing training data. LDA is not stable due to the small sample size problem and but corroborate best directions when each class has a Gaussian density with a common covariance matrix. But it can flop if the class densities are more general and interpreted class separability in between-class-matrices are inadequate. Maximum Margin Criterion (MMC) having lower computational cost, is more efficient than LDA for calculating the discriminant vectors barring the computation for inverse of within-class-scatter matrix. But traditional MMC disregards the discriminative information within the local structure of samples and performance is depended on choosing of a coefficient. In this paper we delineate the locality of data points by counting a distances among data points considering the supervised knowledge. We have computed the entire scatter matrix in Laplacian graph embedded space and finally the produced stastically uncorrelated discriminant vectors reduces redundancy among the extracted features and there is no constant to be chosen. Our experiment with Reauter dataset recommends this algorithm is more efficient than LDA, MMC and it manifests similar or sometime s better result than other locality based algorithm like LPP and LSDA.

1. INTRODUCTION

Dimensionality reduction has been a main problem in many fields in stastical pattern classification, clustering and retrieval. To solve the problem of excessive dimensionality one of the best methods is to reduce dimensionality by combining features. Linear combinations are one of them because they are simple to compute and analytically tractable. Principal Component Analysis (PCA)[1] and Linear
Discriminant Analysis (LDA)[2] are two popular algorithms. In the information retrieval community this PCA method has been named Latent Semantic Indexing (LSI)[3]. LSI, PCA are completely unsupervised, i.e. they pay no attention to the class labels of the existing training data. LSI aims at optimal representation of the original data in the lower dimensional space in the mean squared error sense. LDA is a supervised learning algorithm. LDA searches for the project axes on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other[4]. However, the scatter matrices are dense, and the eigen decomposition could be very expensive in both time and memory. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are required to guarantee the non-singularity of scatter matrices [26]. LDA is optimal only when the data of each class is approximately Gaussian distributed. If the data are more general (i.e. different classes share approximately same mean) then discrimination power is degraded. To avoid the small sample size problem based on a new feature extraction criterion, the maximum margin criterion (MMC)[5]geometrically, MMC maximizes the average margin between classes. It can be shown that MMC represents class separability better than PCA. As a connection to Fisher’s criterion, we can also derive LDA from MMC by incorporating some constraint. Recent studies show that the text documents possibly reside on a nonlinear manifold [6]. In the past few years have seen many manifold-based learning algorithms for discovering intrinsic low-dimensional embedding of data have proposed. Among them most well-known are isometric feature mapping ISOMAP, local linear embedding (LLE)[7][8] and Laplacian Eigenmap[9]. He et. al. proposed Locality Preserving Projections (LPP)[10], which is a linear subspace learning method derived from Laplacian Eigenmap. These methods have been shown to be effective in discovering the geometrical structure of the underlying manifold. However, they are unsupervised in nature and fail to discover the discriminant structure in the data. The projection direction determined by LPP ensures that, if samples \( x_i \) and \( x_j \) are close then their projections \( y_i \) and \( y_j \) are close as well. It possibly happens that two near samples belonging to different classes may also result in close texts after the projection of LPP. Locality Discriminating Indexing(LDI) proposed by J. Hu et al. [6] which follows the supervised approach. They assume that the documents reside on diverse class-specific manifold structures that overlap one another, and aims to linearly represent the document in lower dimensions which discount the inter-class overlap. LDI explicitly considers both the local structure and the prior class information. LDI is more discriminative than LSI,LPP and LDA/GSVD for text classification.

In this paper first we have analyzed traditional MMC in Laplacian Graph embedded space. Performance of MMC is dependant on a constant value selection that can be done manually by cross validation checking. So this will consume a bulk time. We have experimented with a little different criterion of MMC, we take the criterion function as minimization problem with reverse subtraction of original MMC criterion. Although this criterion performs good but this method is still dependent on choosing of a constant value. Then we propose a new local similarity based supervised algorithm which doesn’t need any constant value selection. As MMC criterion is reasonably principled and simple in formulation so we extended the traditional MMC which is aims to preserve the global structure of data into a local manifold based model. The basis vectors of all the methods above are statistically correlated, and so the extracted features contain redundancy, which may distort the distribution of the features and even dramatically degrade recognition performance. In this paper we introduce a supervised dimensionality reduction method called Uncorrelated Local Maximum Margin Criterion(ULMMC). First we have constructed similarity among the all data points by using some suitable function, cosine similarity among documents is considered in this paper. Then we have constructed the formulae of between class scatter and within class scatter into a Laplacian graph embedded space. Then we incorporated locality information by similarity matrix value as weight of the Laplacian graph edge. Here all the connected
2. Uncorrelated Local Maximum Margin Criterion

Suppose we have a set of m samples \([x_1, x_2, \ldots, x_m]\) belonging to c classes. So term-document matrix is \(X \in n \times m\) matrix, where \(n\) is number of terms and \(m\) is no of documents. Each vector is represented by traditional TF-idf score. The similarity between two samples is defined as: \(W_{ij} = x_i^T x_j \ldots (1)\).

if \(x_i\) and \(x_j\) are normalized vector then \(W_{ij}\) equals to the cosine similarity between the document vectors, is a strictly monotone decreasing function with respect to the distance between two samples documents.

\[
S_t = \sum_{j=1}^{m} (x_j - \mu)(x_j - \mu)^T = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (x_i - x_j)(x_i - x_j)^T \quad \ldots \ldots (2)
\]

Where \(m\) is the no. of sample data. \(\mu\) is the total data mean. Now we can consider relation of locality among each data \(x_i\) and \(x_j\). So the total scatter can be characterized by the term \(\alpha_t\)

\[
\alpha_t = tr(LS_t) = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} W_{ij} (x_i - x_j)(x_i - x_j)^T = \frac{1}{m} \sum_{i=1}^{m} (x_i D_{ii} x_i^T) - \frac{1}{m} \sum_{i=1}^{m} (x_i W_{ij} x_j^T) \ldots \ldots (3)
\]

Where \(D_{ii} = \sum_{j=1}^{m} W_{ij}\), \(W_{ij} = \frac{W_{ij}}{m}\) and \(D_{ii} = \frac{D_{ii}}{m}\) where \(D_{ii}\) is a diagonal matrix.

\[
S_w = \sum_{i=1}^{c} \sum_{j=1}^{m} (x_j^i - \mu_i)(x_j^i - \mu_i)^T = \sum_{i=1}^{c} \sum_{k=1}^{m_i} \sum_{j=1}^{m_j} (x_j^k - x_j^i)(x_j^k - x_j^i)^T \quad \ldots \ldots (4)
\]

Where \(\mu_i\) is mean of \(i\) th class. By considering the local information we can intra class scatter is characterized by the term \(\alpha_w\)

\[
\alpha_w = tr(LS_w) = \sum_{i=1}^{c} \frac{1}{m_i} \sum_{k=1}^{m_i} \sum_{j=1}^{m_j} H_{kj} (x_j^k - x_j^i)(x_j^k - x_j^i)^T = \sum_{i=1}^{c} \frac{1}{m_i} (XDW_{ij}X^T - XHX^T) \quad \ldots \ldots (5)
\]

\[
H_{kj} = \begin{cases} W_{kj} & \text{if } x_k \text{ and } x_j \text{ belong to the same class, otherwise } 0 \end{cases}
\]

Let us define a row vector \(a = \left[ \frac{1}{m_1}, \ldots, \frac{1}{m_c}, \ldots, \frac{1}{m_1}, \ldots, \frac{1}{m_c}, \ldots \right]^T\) and take another matrix \(A_{m \times m}\) with this row vector \(a\). From eq(5) we get \(\alpha_w = \left( XDW_{ij}X^T - XHX^T \right) \ldots \ldots (6)\)

where \(D_w = D_{ij} a a^T\), \(H_w = H_{ij} A\), this operation \(\ast\) defines dot product. Now defining the \(\alpha_b = tr(S_b) = LS_t - LS_w = X((D_t - D_w)-(W_t - H_w))X^T = X(D_b - W_b)X^T \ldots \ldots (7)\)
local Maximum Margin Criterion (LocalMMC) is defined as follows:
\[ J(C) = \text{tr}(C^T(LS_d - LS_w)C) = \text{tr}(C^T(LS_d - LS_w)) \]  
(8)
Where \( C \) is the optimal projection discriminant matrix. The optimal projection vectors \( c_1, c_2, ..., c_k \) can be selected corresponding to first \( k \) largest eigen vectors \( \lambda_1, \lambda_2, ..., \lambda_k \). 
\[
(LS_d - LS_w)c_i = \lambda_i c_i, \quad \lambda_1 > \lambda_2 > ... > \lambda_k \quad ....... (9)
\]
Now consider the stastically uncorrelated constraint. Assume any two different components \( y_k \) and \( y_I \) are extracted feature \( y = C^TX_l \), are uncorrelated. Then,
\[
E \left[ (y_k - E(y_k))(y_I - E(y_I))^T \right] = c_k^T LS_c c_l = 0 \quad ....... (10)
\]
(10) where \( c_l \) and \( c_l \) are two different components of matrix \( C \). If \( c_l \) is normalized then \( c_l^T LS_c c_l = 1 \) or \( C^T SC = I \) ........... (11), \( I \) is identity matrix. For details and proof see Li et al.[17]. As a result LMMC criterion function turns into Uncorrelated LMMC by adding another constraints as in eq (9). We can write \( (LS_d - LS_w)c_i = \lambda_i LS_c c_i \) \( \quad ....... (12) \). Which is a maximization problem and matrix \( C \) is uncorrelated. Let us see how to solve the above equation (12). We can write the eq (12) as:
\[
X((D_b - W_b) - (D_w - W_w))X^T C = \lambda X(D_c - W_c)X^T C \quad \text{let} \quad U = X((D_b - W_b) - (D_w - W_w))X^T \quad \text{and} \quad S = X(D_c - W_c)X^T \quad \text{so the above eq can be written as:} \quad \Rightarrow UC = \lambda SC \quad ............ (13)
\]
Let \( \text{svd}(S) = VD^T \) the dimension of \( V \) is \( m \times r \) where \( r \) is the rank of \( S \). This \( V \) can be calculated by the similar method described at[13]. Where \( S = X(D_c - W_c)X^T \) can be formulated as eigen value problem:
\[
\gamma(D_c - W_c)X^TV = DV \quad \text{here} \quad V \quad \text{is the required eigen vector matrix of} \quad S; \quad D \quad \text{is diagonal matrix of the corresponding eigen values. From eq (13) } \Rightarrow UC = \lambda VD^T V^T C = \gamma UVV^T C = \gamma VD^T C \quad \text{as} \quad VV^T = I .
\]
\[
= D^{\frac{1}{2}}V^T UV = \lambda D^{\frac{1}{2}} V^T \gamma D^{\frac{1}{2}}V^T C = \lambda VD^{\frac{1}{2}} C \quad = (VD^{\frac{1}{2}})^T U(MD^{\frac{1}{2}}) \quad \text{where} \quad M = D^{\frac{1}{2}} V^T C \quad \text{Once we compute} \quad M^* \text{ (optimal vector)} \text{ we can easily get the} \quad C^* \quad \text{ by} \quad C = M \quad MD^{\frac{1}{2}} V , \quad ........ (14). \quad \text{from where we can select} \quad k \quad \text{ largest uncorrelated eigen vector corresponding to} \quad \lambda_1 > \lambda_2 > ... > \lambda_k \quad \text{ largest eigen values. Note that we don’t need any PCA or SVD stage to preprocess the data and also note that dimension of} \quad (VD^{\frac{1}{2}})^T U(MD^{\frac{1}{2}})I \quad \text{is} \quad r = m, \quad \text{so this computationally efficient also. We can pre multiply} \quad X \quad \text{to} \quad VD^{\frac{1}{2}} \quad \text{instead of forming} \quad U \quad \text{which is} \quad n \times n \quad \text{large matrix. Then the equation (14) becomes:}
\]
\[
(X^T VD^{\frac{1}{2}})^T ((D_b - W_b) - (D_w - W_w))(X^T VD^{\frac{1}{2}}) M = \lambda M \quad ............ (15)
\]
3. Experimental Results

We have used a subset of reauter[18] data set. Multiple category documents are also included. Initially all the documents are filtered with stop word removal program ,and stemmer program. We have used normalized term frequency as feature value. We have selected 3,5 and 6 class data among some most frequent categories (earn ,acq, crude, money-fx) and some less frequent(money-supply, trade, grain) categories. Which have no of documents 333 ,788,1243 respectively and have a dimension 2700, 3525, 5719 respectively. we have used a 4 fold data set. We have compared the performance of ULMMC with ULDA[20],MMC ,LPP[10], Local MMC[16] algorithms,
Table 1: shows the micro average recognition rate of different algorithm.

<table>
<thead>
<tr>
<th>No of classes</th>
<th>ULDA</th>
<th>MMC</th>
<th>LPP</th>
<th>Local MMC</th>
<th>LSDA</th>
<th>Proposed ULMMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97.1</td>
<td>97.8</td>
<td>96.1</td>
<td>97.1</td>
<td>97.8</td>
<td>97.9</td>
</tr>
<tr>
<td>5</td>
<td>97.2</td>
<td>97.6</td>
<td>96.8</td>
<td>97.0</td>
<td>97.4</td>
<td>98.0</td>
</tr>
<tr>
<td>6</td>
<td>97.1</td>
<td>96.9</td>
<td>96.8</td>
<td>96.9</td>
<td>97.1</td>
<td>98.3</td>
</tr>
</tbody>
</table>

Reference:

11. Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang, Face Recognition Using Laplacianfaces:IEEE Transaction on Pattern Analysis And Machine Intelligence, VOL. 27, NO. 3, MARCH 2005
12. Deng Cai :SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis IEEE.
16. Wankou Yang, Jingyu Yang, Feature Extraction Base on Local Maximum Margin Criterion ,2008 IEEE
17. Yong-zhi Li, at alA improved method of face recognition by kernel maximum margin criterion, Third International Conference on Natural Computation (ICNC 2007)