Dimension reduction based on centroids and least squares for efficient processing of text data

***M. Jeon*[*], *H. Park*[†], *and J.B. Rosen*[‡]**

Abstract

Dimension reduction in today's vector space based information retrieval system is essential for improving computational efficiency in handling massive data. In our previous work we proposed a mathematical framework for lower dimensional representations of text data in vector space based information retrieval, and a couple of dimension reduction method using minimization and matrix rank reduction formula. One of our proposed methods is CentroidQR method which utilizes orthogonal transformation on centroids, and the test results showed that its classification results were exactly the same as those of classification with full dimension when a certain classification algorithm is applied. In this paper we discuss in detail the CentroidQR, and prove mathematically its classification properties with two different similarity measures of L_2 and cosine.

[*]The work of all three authors was supported in part by the National Science Foundation grant CCR-9901992. Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, U.S.A., e-mail: jeon@cs.umn.edu.

[†]Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, U.S.A., e-mail: hpark@cs.umn.edu.

[‡]Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455 and Dept. of Computer Science and Engineering, Univ. of California, San Diego, La Jolla, CA 92093, U.S.A. e-mail: jbrosen@cs.ucsd.edu.

Introduction

To handle today's massive high dimensional data efficiently, dimension or feature reduction of data is essential in a information retrieval system. Grouping similar data into one category through clustering presents more related output for user's query without much overhead [12]. Classification is the process of assigning new data to predefined proper group called class or category. On the other hand, clustering is grouping the data without any predefined categories, which is usually performed to build categories for classification task. The classification problem may be complicated by imperfect class definitions, overlapping categories, random variations in the new data [1], and nonlinearity of classifier. A common classification system is composed of data collection, feature generation, feature selection, classifier design, and finally system evaluation and feedback [6, 13, 16]. Among them feature selection is of great importance for the quality of classification and computational cost of the classifier. Several examples of available classification methods are k-nearest neighbor, perceptron, and decision tree [9, 16]. Another simple and fast method we can consider is the one based on centroids of classes which provide useful background for a couple of dimension method such as discriminant analysis, in addition to Centroid, CentroidQR methods we proposed in [14] and others [4, 11].

The dimension reduction method that we will discuss in this paper is based on the vector subspace computation in linear algebra [5]. Unlike other probability and frequency based methods where a set of representative words are chosen, the vector subspace computation will give reduction in the dimension of term space where for each dimension in the reduced space we cannot easily attach corresponding words or a meaning. The dimension reduction by the optimal lower rank approximation from the SVD has been successfully applied in numerous applications, e.g. in signal processing. In these applications, often what the dimension reduction achieves is the effect of removing noise in the data. In case of information retrieval or data mining, often the data matrix has either full rank or close-to full rank. Also the meaning of *noise* in the data collection is not well understood, unlike in other applications such as signal processing [15] or image processing. In addition, in information retrieval, the lower rank approximation is not only a tool for rephrasing a given problem into another one which is easier to solve, but the data representation in the lower dimension space itself is important [8] in further processing of data.

Several dimension reduction methods have been proposed for clustering and classification of high dimensional data, but most of them provide just approximation of original data. One attractive and simple algorithm is one based on the centroids of classes and minimization [14]. In [14] we proposed a dimension reduction method named CentroidQR and test results showed it gives exactly identical classification results in full dimensional space and reduced dimensional space when classification is determined by comparing the new data to the centroids of the clusters. In this paper, we revisit the CentroidQR method, and prove mathematically its surprisingly good classification properties with two different similarity measures of L_2 and cosine. Before CentroidQR method is investigated in detail, lower dimensional representation of term-document matrix and representation of each cluster will be discussed in the following sections.

Lower Dimensional Representation of Term-Document Matrix

To mathematically understand the problem of lower dimensional representation of the given document sets, we will first assume that the reduced dimension, which we will denote as k ($k << min(m,n)$), is given or determined in advance. Term-document matrix $A \in \mathbb{R}^{m \times n}$ is defined as the matrix whose column vector represents each document and each component of the column vector does a word of the document. Then given a term-document matrix $A \in \mathbb{R}^{m \times n}$, and an integer k, the problem is to find a linear transformation $G^T \in \mathbb{R}^{k \times m}$ that maps each column a_i of A in the m dimensional space to a vector y_i in the k dimensional space :

$$G^T : a_i \in \mathbb{R}^{m \times 1} \rightarrow y_i \in \mathbb{R}^{k \times 1}, 1 \leq i \leq n. \tag{1}$$

This can be rephrased as an approximation problem where the given matrix A has to be decomposed into two matrices B and Y as

$$A \approx BY \tag{2}$$

where both $B \in \mathbb{R}^{m \times k}$ with rank(B) = k and $Y \in \mathbb{R}^{k \times n}$ with rank(Y) = k are to be found. This lower rank approximate factorization is not unique since for any nonsingular matrix $Z \in \mathbb{R}^{k \times k}$,

$$A \approx BY = (BZ)(Z^{-1}Y),$$

and $rank(BZ) = k$ and $rank(Z^{-1}Y) = k$. The solution for problem (2) can be found by finding $B \in \mathbb{R}^{m \times k}$ with rank(B) = k and $Y \in \mathbb{R}^{k \times n}$ with rank(Y) = k in the minimization problem

$$\min_{B,Y} ||A - BY||_F. \tag{3}$$

For example, when we use centroid vectors for B, the solution vectors $Y = (B^T B)^{-1} B^T A$ will be the reduced dimensional representation of data matrix A. When the matrix B has orthonormal columns, since $B^T B = I$, we have $Y = B^T A$ which shows that $G = B$. It is well known that the best approximation is obtained from the singular value decomposition(SVD) of A. The commonly used latent semantic indexing [2] exploits the SVD of the term-document matrix. For successful rank reduction scheme, it is important to exploit a priori knowledge. The incorporation of a priori can be translated to adding a constraint in the minimization problem (3). However, mathematical formulation of the a priori knowledge as a constraint is not always easy or even possible. In this paper, we will concentrate on exploiting clustered structure for dimension reduction.

Representation of Each Cluster

First we will assume that the data set is cluster structured and already grouped into certain clusters. This assumption is not a restriction since we can cluster the data set if it is not already clustered using one of the several existing clustering algorithms such as k-means

[4, 9]. Also especially when the data set is huge, we can assume that the data has a cluster structure and it is often necessary to cluster the data first to utilize the tremendous amount of information, in an efficient way.

Suppose we are given a data matrix A whose columns are grouped into k clusters. Instead of treating each column of the matrix A equally regardless of its membership in a specific cluster, which is what is done in the SVD, we want to find the matrices B and Y with k columns and k rows, respectively, so that the k clusters are represented well in the space with reduced dimension. For this purpose, we want to choose each column of B so that it *represents* the corresponding cluster. To answer the question of which vector can represent each cluster well, we first consider an easier problem with scalar data. For any given scalar data set $\alpha_1, \alpha_2, \cdots, \alpha_n$, the *mean* value

$$m_\alpha = \frac{1}{n}\sum_{i=1}^{n} \alpha_i \tag{4}$$

is often used to represent the data set. The use of mean value is justified since it is the expected value of the data or the one that gives the minimum variance

$$\sum_{i=1}^{n}(\alpha_i - m_\alpha)^2 = \min_{\delta\in\mathbb{R}} \sum_{i=1}^{n}(\alpha_i - \delta)^2 = \min_{\delta\in\mathbb{R}} \|(\alpha_1 \cdots \alpha_n) - \delta(1\cdots 1)\|_2^2. \tag{5}$$

The mean value is often extended to the data sets in a vector space as follows. Suppose $a_1, a_2, \cdots, a_n \in \mathbb{R}^{m\times 1}$. Then its *centroid* defined as

$$c_a = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n} Ae \tag{6}$$

where $A = [a_1 a_2 \cdots a_n]$ and $e = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$, is used as a vector that represents the vector data set. The centroid is the vector which achieves the minimum variance in the following sense:

$$\sum_{i=1}^{n} \|a_i - c_a\|_2^2 = \min_{x\in\mathbb{R}^{n\times 1}} \sum_{i=1}^{n} \|a_i - x\|_2^2 = \min_{x\in\mathbb{R}^{n\times 1}} \|A - xe^T\|_F^2. \tag{7}$$

It is clear from (7) that the centroid vector gives the smallest distance in Frobenius norm between the matrix A and the rank one approximation xe^T where x is to be determined. Since one of the vectors in this rank one approximation is fixed to be e, this distance cannot be smaller than the distance obtained from rank one approximation from the SVD: the rank one approximation from the SVD would choose *two* vectors $y \in \mathbb{R}^{m\times 1}$ and $z \in \mathbb{R}^{n\times 1}$ such that $\|A - yz^T\|_F$ is minimized, and

$$\min_{y,z} ||A - yz^T||_F \le \min_{x} ||A - xe^T||_F.$$

However, the centroid vector has the advantage that for each cluster, we can find *one* vector in $\mathbb{R}^{m\times 1}$ to represent it instead of *two* vectors.
For other alternatives for representatives, such as *medoid*, see [14].

Minimization with an Orthogonal Basis of the Cluster Representatives

If the factor B has orthonormal columns in a rank k approximation $A \approx BY$, then the matrix Y by itself can give a good approximation for A in the sense that the correlation of A can be well approximated with the correlation of Y:

$$A^T A \approx Y^T B^T B Y = Y^T Y, \quad \text{where} \quad B^T B = I.$$

In addition, most of the common similarity measures can directly be inherited from the full dimensional space to the reduced dimensional space, since for any vector $y \in \mathbb{R}^{k\times 1}$,

$$\|By\|_2 = \|y\|_2,$$

where B has orthonormal columns. Accordingly, for any two vectors $a, q \in \mathbb{R}^{m\times 1}$ and their projections $\hat{a}, \hat{q} \in \mathbb{R}^{k\times 1}$ via B,

$$||a - q||_2 \approx ||B\hat{a} - B\hat{q}||_2 = ||\hat{a} - \hat{q}||_2.$$

and

$$Cos(a, q) \approx Cos(B\hat{a}, B\hat{q}) = Cos(\hat{a}, \hat{q}),$$

where for any two vectors x and y in the space of same dimension,

$$Cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}.$$

Therefore, for comparing two vectors in the reduced space, the matrix B does not need to be involved. No matter how the matrices B and Y are chosen, this can be achieved by computing the reduced QR decomposition of the matrix B if it does not already have orthonormal columns. In the following theorem, we summarize the well known QR decomposition [3, 5] to establish our notations.

Theorem 1 (QR Decomposition) *Let $B \in \mathbb{R}^{m\times k}, m \geq k$ be any given matrix. Then there is an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that*

$$B = Q \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where $R \in \mathbb{R}^{k\times k}$ is upper triangular.

Partitioning Q as

$$Q = (Q_k, Q_r), \quad Q_k \in \mathbb{R}^{m\times k}, \quad Q_r \in \mathbb{R}^{m\times(m-k)},$$

we have

$$B = (Q_k, Q_r) \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_k R. \tag{8}$$

Algorithm 0.1 CentroidQR

Given a data set $A \in \mathbb{R}^{m\times n}$ with k clusters, it computes a k dimensional representation $\hat{q}$ of a given vector $q \in \mathbb{R}^{m\times 1}$.

1. Compute the centroid b_i of the ith cluster, $1 \le i \le k$
2. Set $B = \begin{bmatrix} b_1 & b_2 & \cdots & b_k \end{bmatrix}$
3. Compute the reduced QR decomposition of B, which is $B = Q_k R$.
4. Solve $\min_{\hat{q}} \|Q_k\hat{q} - q\|_2$ (in fact, $\hat{q} = Q_k^T q$).

The right-hand side of Eqn. (8) is called the reduced QR decomposition of B, where Range(B) = Range(Q_k). Premultiplying $(Q_k, Q_r)^T$ onto both sides of Eqn. (8) gives

$$\begin{pmatrix} Q_k^T \\ Q_r^T \end{pmatrix} B = \begin{pmatrix} Q_k^T B \\ Q_r^T B \end{pmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{9}$$

where we see $Q_k^T B = R$ and $Q_r^T B = 0$. With the reduced QR decomposition of B shown in Eqn. (8), where $Q_k^T Q_k = I_k$ and R is upper triangular, the k-dimensional representation of A is the solution for

$$\min_z ||Q_k Z - A||_F. \tag{10}$$

Then $Z = Q_k^T A = RY$ where Y is the solution for

$$\min_Y ||BY - A||_F. \tag{11}$$

where B is the matrix whose columns are the centroids of classes. Eqn. (11) gives the Centroid method in our previous work [14] by which full dimensional data matrix A and centroid matrix B are transformed to Y and I_k in the reduced dimensional matrices, respectively. By the minimization problem (10) the data matrix A is transformed to Z, and the centroid matrix B is transformed to R, as

$$Z = Q_k^T A \quad \text{and} \quad R = Q_k^T B. \tag{12}$$

Above steps are summarized in Algorithm CentroidQR.

It is interesting to note that when the columns of B are the centroids of the classes in the full dimensional space, the corresponding centroids in the reduced space obtained by the CentroidQR method are the columns of the upper triangular matrix R, while those reduced by Centroid method are the columns of the identity matrix I_k [14].

There are many algorithms developed for classification [7, 14, 10]. In one of the simpler but effective algorithms we simply compare the data with each centroid [7, 14], which is summarized in Algorithm Centroid based classification.

We will show that the dimension reduction by CentroidQR algorithm has a special property when it is used in conjunction with Centroid based classification.

Algorithm 0.2 Centroid based Classification

Given a data set A with k clusters and k corresponding centroids, b_i, $1 \leq i \leq k$, it finds the index j of the cluster in which the new vector q belongs.

1. Find the index j such that $sim(q, b_j)$, $1 \leq j \leq k$, is minimum, where $sim(q, b_j)$ is the similarity measure between q and b_j.
   (For example, with L_2 norm, $sim(q, b_j) = \|q - b_j\|_2$ and the index j which gives minimum value is to be found, and with cosine, $sim(q, b_j) = cos(q, b_j) = \frac{q^T b_j}{\|q\|_2 \|b_j\|_2}$) which gives maximum value is to be found.

We now investigate the relationship between classification results from Algorithm Centroid based Classification in the full dimensional space and the reduced space obtained by CentroidQR method. It is well known that norm is invariant under orthogonal transformation. That is

$$\|Q^T(a_i - b_j)\|_2^2 = (a_i - b_j)^T Q Q^T (a_i - b_j)$$

where $Q^T Q = Q Q^T = I$. Our transformation does not hold invariance of norm, since we use Q_k, and $Q_k Q_k^T \neq I$. However we now show that the transformation by Q_k still has very interesting properties.

Definition 1 (Ordering) *Let S(q,B) denote an ordering of column indices of $B \in \mathbb{R}^{m \times k}$ which is sorted in an non increasing order of similarity between a vector $q \in \mathbb{R}^{m \times 1}$ and the k columns of $B \in \mathbb{R}^{m \times k}$*

For example, suppose the matrix $B = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \in \mathbb{R}^{m \times 3}$ and in L_2 norm similarity, $\|q - b_1\|_2 \leq \|q - b_3\|_2 \leq \|q - b_2\|_2$ the $S(q, B) = (1, 3, 2)$.

Theorem 2 (Order Preserving in L_2) *The order $S(q, B)$ with L_2 measure in the full dimensional space is completely preserved in the reduced space obtained with transformations by (12). i.e. $S(q, B) = S(\hat{q}, \hat{B})$ when $\hat{q} = Q_k^T q$ and $\hat{B} = Q_k^T B$, and the reduced QR decomposition of B is $Q_k R$.*

Proof:
Let's start with norm preserving property of orthogonal transformation (13). Since $\|Q_r^T b_j\| = 0$ from (9), $\|q - b_j\|_2^2$ can be expressed as

$$\|q - b_j\|_2^2 = \|Q^T(q - b_j)\|_2^2 \tag{13}$$
$$= \|Q_k^T(q - b_j)\|_2^2 + \|Q_r^T(q - b_j)\|_2^2 \tag{14}$$
$$= \|Q_k^T(q - b_j)\|_2^2 + \|Q_r^T q\|_2^2 \tag{15}$$

Thus if $\|q - b_j\|_2 \leq \|q - b_l\|_2$, then we have $\|Q_k^T(q - b_j)\|_2 \leq \|Q_k^T(q - b_l)\|_2$ since the term, $\|Q_r^T q\|_2^2$ of (15) does not involve b_j nor b_l and is a constant for any class. This means that our reduction method preserve the order of L_2 similarity in full dimensional space after dimension reduction. $\square$

Theorem 3 (Order Preserving in Cosine) *The order $S(q, B)$ with cosine measure in the full dimensional space is completely preserved in the reduced space obtained with transformations by (12). i.e. $S(q, B) = S(\hat{q}, \hat{B})$ when $\hat{q} = Q_k^T q$ and $\hat{B} = Q_k^T B$, and the reduced QR decomposition of B is $Q_k R$.*

Proof:
Let $cos(q, b_j)$ be cosine between vectors $q \in A$ and $b_j \in B$. Then

$$\begin{aligned} cos(q, b_j) = cos(Q^T q, Q^T b_j) &= \frac{(Q^T q)^T Q^T b_j}{\|Q^T q\|_2 \|Q^T b_j\|_2} \\ &= \frac{(q^T Q_k \quad q^T Q_r) \begin{pmatrix} Q_k^T b_j \\ Q_r^T b_j \end{pmatrix}}{(\|Q_k^T q\|_2^2 + \|Q_r^T q\|_2^2)^{\frac{1}{2}} \|Q_k^T b_j\|_2} \\ &= \frac{q^T Q_k Q_k^T b_j}{(\|Q_k^T q\|_2^2 + \|Q_r^T q\|_2^2)^{\frac{1}{2}} \|Q_k^T b_j\|_2} \end{aligned} \tag{16}$$

Thus when $cos(q, b_j) \leq cos(q, b_l)$, we have

$$\frac{q^T Q_k Q_k^T b_j}{(\|Q_k^T q\|_2^2 + \|Q_r^T q\|_2^2)^{\frac{1}{2}} \|Q_k^T b_j\|_2} \leq \frac{q^T Q_k Q_k^T b_l}{(\|Q_k^T q\|_2^2 + \|Q_r^T q\|_2^2)^{\frac{1}{2}} \|Q_k^T b_l\|_2}.$$

When we take out the second common factor $\|Q_r^T q\|$ from the denominator of the above expression, still holds

$$\frac{q^T Q_k Q_k^T b_j}{\|Q_k^T q\| \|Q_k^T b_j\|} \leq \frac{q^T Q_k Q_k^T b_l}{\|Q_k^T q\| \|Q_k^T b_l\|}. \tag{17}$$

In Eqn. (17), since the left term represents $cos(\hat{q}, \hat{b}_j)$, and the right term represents $cos(\hat{q}, \hat{b}_l)$, where $\hat{q}$ is a reduced representation of q and $\hat{b}_i = Q_k^T b_i$ which is the kth dimensional representation of b_i, $1 \leq i \leq k$, the expression (17) is equivalent to

$$cos(\hat{q}, \hat{b}_j) \leq cos(\hat{q}, \hat{b}_l).$$

Thus

$$cos(q, b_j) \leq cos(q, b_l) \quad \text{then} \quad cos(\hat{q}, \hat{b}_j) \leq cos(\hat{q}, \hat{b}_l). \quad \square$$

The above two theorems show that we can completely recover the orders of both L_2 and cosine similarities when original dimension is reduced to dimension k, the number of categories by Algorithm CentroidQR, and classification is achieved by Algorithm Centroid based Classification. In other words, we can produce exactly the same classification results with a reduced data as those with a full dimensional data, whose computational cost saving is obvious especially for high dimensional data.

Note that the order preserving property of the dimension reduction obtained by Algorithm CentroidQR holds *regardless of the quality of the clustering*. This means that no matter how the clustering in the full dimensional space is obtained, the ordering structure between any data and the centroids of the clusters is preserved after dimension reduction via Algorithm CentroidQR. Next section gives some experimental results showing the property of our algorithm numerically.

Experimental Results

In the first test we use some artificial clustered data which is generated by an algorithm which is a modified version of what is presented in [9] to examine the relationship between numerical values of similarity measures in the full dimensional and the reduced dimensional space expressed in Eqns (15) and (16). In generating data set using the program, we can optionally choose the dimension of the data, total number of data and minimum number of data for each class. For a simplicity of presentation, we first choose the data set which is composed of three classes with 20-dimensional data. Each class has 5, 5 and 3 items, and thus total 13 number of data are selected. The matrix form of the test data is a dense matrix of size 20×13 . Since it has three classes, the data vectors are reduced to dimension 3 from 20 by the CentroidQR algorithm. Then we compare classification in the full and reduced space. Detailed values are shown in the Table 1 and Table 2.

Table 1 shows classification in the full dimensional space and the reduced space with L_2 measure, and their numerical relationship. First column of the table contains the label of each data, the numerical values in the next three columns are Euclidean distances between data a_i and centroids b_j in full dimensional space, the next three columns represent those in reduced space, and the last column does L_2 norms of components which are orthogonal to $Q_k^T(a_i - b_j)$ of the full dimensional data $\|Q^T(a_i - b_j)\|$. From Eqn. (15) we know that Euclidean distance in full dimensional space is decomposed into Euclidean distance in reduced space and constant value which is independent of centroid vectors. For example, distances between a_5 and b_1, b_2 and b_3 in full dimensional space are 3.39, 4.71 and 6.05 respectively, which are decomposed into the constant 3.29 and 0.85, 3.38 and 5.08 of distances in reduced space, respectively. That is, $3.39 = \sqrt{0.85^2 + 3.29^2}, 4.71 = \sqrt{3.38^2 + 3.29^2}$ and $6.05 = \sqrt{5.08^2 + 3.29^2}$. Those classification results exactly follow the Theorem 1.

Similarly, Table 2 shows the cosine values between data vectors and centroid vectors in full dimensional space and reduced space. With the cosine measure, item 9 is misclassified in full dimensional space, and also is is misclassified in reduced space too.

Another interesting fact is that with both similarity measures, the values determining the class of data becomes more pronounced after reduction of dimension. In Table 1 for data a_1 the minimum distance is 2.90 to b_1, and next shortest distance is 4.13 to b_2. In the reduced dimensional space, they are 0.74 and 3.04, respectively. With cosine measure in Table 2 corresponding values in the full dimensional space are 0.77 and 0.43, and 0.98 and 0.55 in the reduced dimensional space. Thus the dimension reduction by CentroirQR makes class-deciding measure difference clearer.

In the next test, a bigger and higher dimensional data set is tested for classification in the full and reduced dimensional space. This data set consists of 5 categories, which are all from the MEDLINE [1] database. Each category has 500 documents, and total number of terms are 22095 after preprocessing with stopping and stemming algorithms [12]. The categories have many common words related to a cancer. By Algorithm CentroidQR the dimension 22095 is dramatically reduced to 5, the number of classes, classification of the full dimensional data is completely preserved in this 5 dimensional space. Table 3 presents as expected from Eqns. (15) and (16), the classification results are identical in the full and

[1] http://www.ncbi.nlm.nih.gov/PubMed

Table 1. *L_2 norm similarity between data and centroids*

	$\|a_i - b_j\|$			$\|Q_k^T(a_i - b_j)\|$			$\|Q_r^T a_i\|$
data	b_1	b_2	b_3	b_1	b_2	b_3	
a_1	2.90	4.13	5.45	0.74	3.04	4.68	2.80
a_2	4.25	5.46	5.68	0.83	3.54	3.87	4.16
a_3	3.61	4.85	5.93	0.49	3.28	4.74	3.57
a_4	3.42	4.66	4.93	0.85	3.28	3.65	3.31
a_5	3.39	4.71	6.05	0.85	3.38	5.08	3.29
a_6	5.10	3.72	5.78	3.84	1.61	4.70	3.36
a_7	5.26	4.10	5.39	3.60	1.43	3.77	3.84
a_8	6.48	4.88	6.14	4.66	1.90	4.18	4.50
a_9	5.57	5.01	5.13	3.72	2.82	3.02	4.15
a_{10}	4.52	3.98	6.44	2.98	2.06	5.47	3.40
a_{11}	4.55	4.49	3.29	3.33	3.25	1.10	3.10
a_{12}	5.23	4.60	2.63	4.71	4.00	1.30	2.28
a_{13}	6.87	6.33	4.50	5.50	4.81	1.83	4.11

Table 2. *cosine similarity between data and centroids*

	$cos(a_i, b_j)$			$cos(Q_k^T a_i, Q_k^T b_j)$		
data	b_1	b_2	b_3	b_1	b_2	b_3
a_1	0.77	0.43	0.23	0.98	0.55	0.29
a_2	0.64	0.24	0.34	0.97	0.36	0.52
a_3	0.66	0.23	0.14	0.99	0.35	0.21
a_4	0.70	0.29	0.40	0.97	0.41	0.56
a_5	0.69	0.23	0.07	0.97	0.33	0.10
a_6	0.34	0.75	0.25	0.45	0.99	0.33
a_7	0.01	0.26	0.15	0.03	0.88	0.49
a_8	0.16	0.66	0.34	0.24	0.98	0.50
a_9	0.01	0.01	0.29	0.03	0.02	0.86
a_{10}	0.44	0.59	0.00	0.62	0.82	0.00
a_{11}	0.39	0.31	0.73	0.52	0.42	0.98
a_{12}	0.10	0.17	0.81	0.12	0.20	0.95
a_{13}	0.31	0.48	0.80	0.38	0.59	0.98

reduced dimensional space for both measures. Classification results of each data are not shown in the table, but they are completely identical.

Table 3. *Misclassification Rate*

class	Data from MEDLINE	
	category	no. of data
1	heart attack	500
2	colon cancer	500
3	diabetes	500
4	oral cancer	500
5	tooth decay	500

	Misclassification Rate (in %)	
	Full	CentroidQR
Dimension	22095×2500	5×2500
L_2	11.76	11.76
Cosine	7.80	7.80

Concluding Remarks

In this paper we presented mathematical proof of what we observed in the Experimental results of our previous research [14] regarding Algorithm CentroidQR. For the centroid based classification, Algorithm CentroidQR gives a dramatic reduction of dimension without losing any information on the class structure.

Currently, we are studying relationship between classifications in the full dimensional and reduced space using criteria such as traces of scatter matrices . What is also remarkable is that the ordering structure between any data and the centroids based on cosine or L_2 norm similarity measures is completely preserved after dimension reduction through our CentroidQR algorithm regardless the cluster quality.

Bibliography

[1] M. R. Anderberg Cluster analysis for applications. *Academic Press, New York and London*, 1973.

[2] M.W. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573-595, 1995.

[3] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[4] I. S. Dhillon. Concept Decompositions for Large Sparse Text Data using Clustering. *IBM Research Report*, RJ 10147, 1999

[5] G.H. Golub and C.F. Van Loan. *Matrix Computations*, third edition. Johns Hopkins University Press, Baltimore, 1996.

[6] E. Gose, R. Johnsonbaugh and S. Jost. Pattern Recognition and Image Analysis. *Prentice Hall Ptr*, 1996.

[7] E. Han and G. Karypis. Centroid-Based Document Classification: Analysis & Experimental Results. *PKDD 2000*.

[8] L. Hubert, J. Meulman, and W. Heiser. Two Purposes for Matrix Factorization: A Historical Appraisal. *SIAM REVIEW*, Vol. 42, No. 1, pp.68-82, 2000.

[9] A.K. Jain, and R.C. Dubes. Algorithms for Clustering Data. *Prentice Hall*, 1988.

[10] Y. Jung, D. Du and H. Park. Advanced document classification using KNN with centroid. *preprint.*

[11] G. Karypis and E. Han. Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization. *CIKM 2000.*

[12] G. Kowalski. Information Retrieval System: Theory and Implementation, *Kluwer Academic Publishers*, 1997.

[13] M. Nadler and E.P. Smith. Pattern Recognition Engineering, *John Wiley & Sons*, 1993.

[14] H. Park, M. Jeon and J.B. Rosen. Representation of High Dimensional Text Data in a Lower Dimensional Space in Information Retrieval, *SIAM Journal on Matrix Analysis and Applications(submitted).*

[15] J.B. Rosen, H. Park, and J. Glick. Total least norm formulation and solution for structured problems. *SIAM Journal on Matrix Anal. Appl.*, 17-1:110-128, 1996.

[16] S. Theodoridis and K. Koutrousmbas. Pattern Recognition, *Academic Press*, 1999