2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)
August 29 – September 1, 2010, Kittilä, Finland

LOCAL DIMENSIONALITY REDUCTION FOR MULTIPLE INSTANCE LEARNING

Saehoon Kim [1] *and Seungjin Choi* [1,2]

[1] Department of Computer Science
[2] Division of IT Convergence Engineering
Pohang University of Science and Technology, Korea
{kshkawa,seungjin}@postech.ac.kr

ABSTRACT

Multiple instance learning involves labeling bags (sets of instances) rather than individual instances. Positive bags contain both true positive and false positive instances, leading to *label ambiguity*, while negative bags consist of only true negative instances. Since labels for individual instances are not known, a direct application of existing discriminant analysis or dimensionality reduction methods often yields an undesirable projection direction due to this label ambiguity in positive bags. In this paper we present a *citation local Fisher discriminant analysis* (CLFDA) where we incorporate both citation and reference information into local Fisher discriminant analysis, in order to detect false positive instances whose corresponding labels are corrected to be negative. To our best knowledge, CLFDA is the first attempt in supervised dimensionality reduction for multiple instance learning. Numerical experiments on several benchmark datasets confirm that CLFDA outperforms existing methods in the task of multiple instance learning.

Index Terms— Dimensionality reduction, Fisher discriminant analysis, multiple instance learning

1. INTRODUCTION

Fisher linear discriminant analysis (FLD), which is also known as linear discriminant analysis (LDA), is a popular supervised dimensionality reduction method, the goal of which is to find a linear combination of attributes that best separates two or more classes [6]. In FDA, linear projections are sought such that the between-class scatter is maximized and the within-class scatter is simultaneously minimized. In general, FDA works well, however, it yields undesirable results when instances in some class form several separate clusters (i.e., multimodality) [7].

In order to cope with multimodality, local Fisher discriminant analysis (LFDA) was developed [11], where the local structure of data was incorporated into FDA. Laplacian eigenmap [2] and locality preserving projection (LPP) (which is a linear version of Laplacian eigenmap) [8] are exemplary graph embedding methods in unsupervised setting where the local geometry of data is approximated by a neighborhood graph. LFDA combines the idea of this unsupervised graph embedding and FDA, seeking projections that maximize between-class separability and simultaneously minimize within-class local structure [11].

In the standard supervised learning, each instance in training set is labeled by the teacher and the task is to predict proper labels of instances in test set. Multiple instance learning involves labeling bags (sets of instances) rather than individual instances [5]. In training, labels are provided for the bags so that a bag which contains at least one positive instance is labeled as a positive bag, while negative bags consist of only negative instances. In other words, positive bags contain both true positive and false positive instances, without knowing labels for individual instances, which leads to *label ambiguity* in instances. Various methods for multiple instance learning have been developed, including diversity density (DD) algorithm [9], citation-kNN and Bayesian-kNN [12], SVM for multiple instance learning [1], etc.

In this paper we address a method of dimensionality reduction suited to multiple instance learning, which was not studied yet to our best knowledge. Due to the label ambiguity in positive bags, FDA and LFDA yield undesirable projections since both true positive and false positive instances in positive bags are treated as positive examples. We present a simple modification of LFDA where detects a false positive instances, whose corresponding labels are changed to be negative. Our method is referred to as *citation LFDA* (CLFDA). Although our idea is very simple, our proposed method is the first attempt to develop a dimensional reduction method in the framework of multiple instance learning, to our best knowledge. Numerical experiments on several benchmark datasets confirm that CLFDA outperforms LFDA in the task of multiple instance learning.

2. LOCAL FISHER DISCRIMINANT ANALYSIS

We briefly review LFDA which we base our method on. More details on LFDA can be found in [11]. Let $\boldsymbol{x}_i \in \mathbb{R}^D$ $(i =$

$1, 2, \ldots, n)$ be the D-dimensional training examples and $y_i \in \{1, 2, \ldots, l\}$ be the associated the class label of $\boldsymbol{x}_i$, where l is the number of classes and n is the number of training samples. Let n_i be the number of training samples in class i and $n = \sum_{i=1}^{l} n_i$.

LFDA seeks $d \leq l-1$ discriminant functions $\boldsymbol{W} \in \mathbb{R}^{D \times d}$ which is determined by the following optimization:

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{\arg\max}\, \mathrm{tr}\left\{\left(\boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W}\right)\right\}, \quad (1)$$

where $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ are *local* within-class and between-class scatter matrices, which are defined by

$$\boldsymbol{S}_W = \frac{1}{2}\sum_{i,j=1}^{n} \overline{A}_{ij}^{W}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (2)$$

$$\boldsymbol{S}_B = \frac{1}{2}\sum_{i,j=1}^{n} \overline{A}_{ij}^{B}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (3)$$

where

$$\overline{A}_{ij}^{W} = \begin{cases} A_{ij}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \quad (4)$$

$$\overline{A}_{ij}^{B} = \begin{cases} A_{ij}(1/n - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \quad (5)$$

where $\boldsymbol{A} = [A_{ij}]$ denotes the *affinity matrix* (adjacency matrix on a neighborhood graph). The simplest choice of the affinity matrix is to construct k-nearest neighbor (k-NN) graph. That is, $A_{ij} = 1$ if $\boldsymbol{x}_j$ belongs to the k-nearest neighbors of $\boldsymbol{x}_i$, otherwise $A_{ij} = 0$. Note that in the case of a fully-connected graph ($A_{ij} = 1$ for $i, j = 1, \ldots, n$), *local* within-class and between-class scatter matrices are equal to the standard scatter matrices used in FDA.

3. CITATION LFDA

3.1. Problem setting

In multiple instance setting, a training set is composed of $\{X_i, y_i\}_{i=1}^{N}$, where $y_i = \{+1, -1\}$ and $X_i = \{\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}\}$. X_i is called a bag and an element ($\boldsymbol{x}_{ij}$) of X_i is called an instance. In this paper, we address only the two-class multiple instance learning problem. +1 and -1 represent positive and negative class, respectively.

In order to use a supervised dimensionality reduction method (such as FLD, LFDA, and etc.) in the multiple instance setting, we assume each instance has the label of the corresponding bag. Hence, the training data is reformulated as $\{(\boldsymbol{x}_{11}, y_i), ..., (\boldsymbol{x}_{1n_1}, y_i), ..., (\boldsymbol{x}_{Ni}, y_N), ..., (\boldsymbol{x}_{Nn_N}, y_N)\}$. Let $n(= \sum_i^N n_i)$ be the total number of instances.

Similar to this approach, converting multiple instance data into supervised data is proposed in [10]. In the paper, the authors compare the performance between multiple instance algorithms and their supervised counterpart on several different domains. The experimental results remark that multiple instance algorithms are not much superior to their supervised counterpart on test domains. Hence, it is possible to attempt to apply supervised dimensionality reduction method to multiple instance data.

Table 1. The 4-nearest neighbors of 5 examples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$

	$k=1$	$k=2$	$k=3$	$k=4$
$\boldsymbol{x}_1$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	$\boldsymbol{x}_4$	$\boldsymbol{x}_2$
$\boldsymbol{x}_2$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_3$
$\boldsymbol{x}_3$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_2$
$\boldsymbol{x}_4$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	$\boldsymbol{x}_2$	$\boldsymbol{x}_1$
$\boldsymbol{x}_5$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	$\boldsymbol{x}_3$	$\boldsymbol{x}_2$

3.2. Motivation

Fig. 1 explains an example where FDA and LFDA fail to determine a proper discriminant function. There are two bags, where instances in the positive bag are filled with blue and instances in the negative bag are colored red. The positive bag contains true positive instances (marked by squares) and false positive instances (marked by circles). The negative bag contains only true negative instances (marked by triangles). In the example, the proper discriminant function (discriminative projection direction) tends to be vertical to maximize the distance between true positive instances and negative instances. If the discriminant function is horizontal, the true positive instances are highly overlapped with the false positive instances and the negative instances in the embedding space.

Since FDA do not consider multimodality, FDA determines the discriminant function to correspond to the dotted line which is almost horizontal. Although LFDA determines a better discriminant function corresponding to the dashed line due to the consideration of local structure, LFDA still has the problem that the true positive instances are overlapped with the false positive instances and the negative instances in the embedding space. If those false positive instances could be detected somehow, then LFDA finds a more desirable discriminant function associated with the solid line. Adopting the idea of citation and reference used in citation-kNN [12], we develop citation LFDA (CLFDA) which successfully determines a more desirable discriminant function.

3.3. Algorithm

We first explain the terms *citer* and *reference* used in citation k-NN [12]. C-nearest citers of $\boldsymbol{x}_i$ is a set of instances such that $\boldsymbol{x}_i$ belongs to their C-nearest neighbors (CNN), i.e.,

$$Citers(\boldsymbol{x}_i) = \{\boldsymbol{x}_j | \boldsymbol{x}_i \in CNN(\boldsymbol{x}_j), j = 1, \ldots, n\},$$

where $CNN(\boldsymbol{x}_j)$ represents the C-nearest neighbors of $\boldsymbol{x}_j$. The R-nearest references of $\boldsymbol{x}_i$ are simply the R-nearest

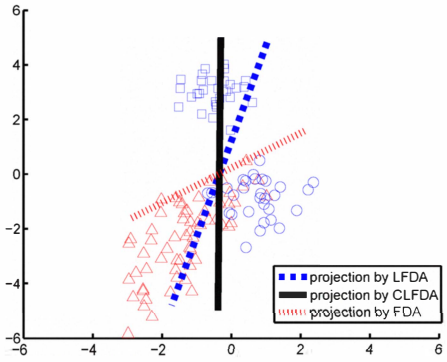

Fig. 1. Comparison of three Fisher discriminant analysis methods on a synthetic toy example: (1) the original Fisher discriminant analysis (FDA); (2) local Fisher discriminant analysis (LFDA); (3) citation LFDA (CLFDA). The positive bag contains true positive instances (squares) and false positive instances (circles), while the negative bag has all true negative instances (triangles). Three different projections are shown, and the direction determined by CLFDA well preserves the separability between positive and negative instances.

neighbors of $\boldsymbol{x}_i$. For example, the 4-nearest neighbors of 5 examples are represented in Table 1. If we assume R and C are set to be two, the 2-nearest references of $\boldsymbol{x}_1$ are $\{\boldsymbol{x}_3, \boldsymbol{x}_5\}$ and the 2-nearest citers of $\boldsymbol{x}_1$ are $\{\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_5\}$.

The reason why we introduce *citer* and *reference* is that false positive instances are detected using these concepts. Let p_i be the number of positive instances in the R-nearest references and C-nearest citers of example $\boldsymbol{x}_i$. Similarly, let n_i be the number of negative instances in R-nearest references and C-nearest citers of example $\boldsymbol{x}_i$. In both p_i and n_i, we allow some instances to be counted several times.

It is very natural to consider that a false positive instance contains at least one negative instance as one of its neighbors. Hence, if $n_i/p_i \geq t$ and $y_i = 1$ (i.e. the class of the example $\boldsymbol{x}_i$ is positive), we think of the example $\boldsymbol{x}_i$ as a false-positive instance, where t is pre-defined threshold.

Therefore, given the training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, the $\overline{A}_{ij}^W$ and $\overline{A}_{ij}^B$ matrices in LFDA are modified as follows.

$$\overline{A}_{ij}^W = \begin{cases} A_{ij}/n_c & \text{if } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j) = c \\ 0 & \text{if } l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_j) \end{cases} \quad (6)$$

$$\overline{A}_{ij}^B = \begin{cases} A_{ij}(1/n - 1/n_c) & \text{if } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j) = c \\ 1/n & \text{if } l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_j), \end{cases} \quad (7)$$

where

$$l(\boldsymbol{x}_i) = \begin{cases} -1 & \text{if } n_i/p_i \geq t \text{ and } y_i = 1 \\ y_i & \text{otherwise} \end{cases} \quad (8)$$

With these modified matrices, we can compute $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$. Therefore, we obtain the discriminant functions by solving the generalized eigenvalue problem:

$$\boldsymbol{W} = \arg\max_{\boldsymbol{W}} \operatorname{tr} \left\{ \left(\boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W} \right)^{-1} \left(\boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W} \right) \right\} \quad (9)$$

One may be curious why we need C-nearest citers, because R-nearest references look like being enough. Obviously, R-nearest references provide a good measure for detecting a false positive instance. Usually, we selected a small value for R. If we select a quite huge value for R, true positive instances could be detected as a false positive instances, because the number of true positive instances is very small compared to the one of negative instances. Therefore, we should choose a value for R as small as possible.

However, if R is very small, it is not sufficient to detect a false positive instance. In that case, the citers of a negative instance are useful information for detecting a false positive instance. In the region where the density of false positive instances is higher than the one of negative instances, the R-nearest references of a false positive instance often contain only false positive instances. On the other hand, the C-nearest citers of a negative instance contain some false positive instances, because the density of negative instances is low in the region. Therefore, it is useful to take into account both the R-nearest references and C-nearest citers.

In the implementation of CLFDA, we first construct a $\max(R, C)$-NN graph, where $\max(R, C)$-NN graph is a k-NN graph with $k = \max(R, C)$. Using the $\max(R, C)$-NN graph, we compute p_i and n_i of each example $\boldsymbol{x}_i$. Using the decision rule (8), we detect false positive instances and relabel them as negative ones. Then, we apply these modified training data into the existing LFDA algorithm. Algorithm 1 illustrates the pseudo code for CLFDA.
There are three inputs in the Algorithm 1: $\left(\{\boldsymbol{x}_i, y_i\}_{i=1}^n, C, R, t\right)$. $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is a training data set. C and R represent the number of citers and references, respectively. t is a predefined threshold. On the line 20, the LFDA represents the pseudocode for LFDA, which needs two inputs: a training data and a nearest-neighbor graph.

3.4. Time Complexity for CLFDA

In this section, we want to show that the time complexity of CLFDA is the same with the one of LFDA. Compared to LFDA, CLFDA requires three additional computation parts: $[\overline{R}_{ij}]$, $[\overline{C}_{ij}]$, and $\max(R, C)$-NN. First, we need to check the time complexity of $[\overline{R}_{ij}]$ and $[\overline{C}_{ij}]$.

Given a $\max(R, C)$-NN graph, the time complexity of gathering them is just $O(n \max(R, C))$. To explain the time complexity, we assume there exist five examples and the 4-NN graph provided in Table 1. We set R and C to be two. For the example $\boldsymbol{x}_1$, its R-nearest references is easily obtained to be $\{\boldsymbol{x}_3, \boldsymbol{x}_5\}$ with the time complexity $O(R)$. Hence, obtaining R-nearest references of all the example requires the time complexity $O(nR)$. During scanning of the $\max(R, C)$-NN

Algorithm 1 CLFDA $(\{\boldsymbol{x}_i, y_i\}_{i=1}^n, C, R, t)$

Input : $(\{\boldsymbol{x}_i, y_i\}_{i=1}^n, C, R, t)$
Output : W

$G \leftarrow \max(C, R)$-NN graph
for $i = 1$ to n **do**
  for $j = 1$ to C **do**
    $[\overline{C}_{ij}] \leftarrow j$th nearest citers of $\boldsymbol{x}_i$
  end for
  for $j = 1$ to R **do**
    $[\overline{R}_{ij}] \leftarrow j$th nearest references of $\boldsymbol{x}_i$
  end for
end for
for $i = 1$ to n **do**
  $n_i \leftarrow$ # of negative instances in $[\overline{C}_{i\cdot}]$ and $[\overline{R}_{i\cdot}]$
  $p_i \leftarrow$ # of positive instances in $[\overline{C}_{i\cdot}]$ and $[\overline{R}_{i\cdot}]$
  if $n_i/p_i \geq t$ and $y_i = 1$ **then**
    $y_i \leftarrow -1$
  end if
end for
$W = \text{LFDA}(\{\boldsymbol{x}_i, y_i\}_{i=1}^n, G)$

of $\boldsymbol{x}_1$, we store $\boldsymbol{x}_1$ as one of the C-nearest citers of the $\boldsymbol{x}_3$ and $\boldsymbol{x}_2$ with the complexity $O(C)$. Hence, the time complexity of obtaining C-nearest citers of all examples is $O(nC)$.

Second, we have to investigate the additional time complexity for constructing the $\max(R, C)$-NN graph. Efficient implementation for LFDA requires the time complexity for constructing $\max(R, C)$-NN graph in a classwise manner, which means that we only construct $\max(R, C)$-NN graph within the same class [11]. A naive implementation for constructing $\max(R, C)$-NN graph in a classwise manner requires $O(n_p^2 \log n_p + n_n^2 \log n_n)$ time complexity, where n_p and n_n represent the number of examples in positive class and negative class, respectively. In order to detect a false positive instance, $\max(R, C)$-NN graph should be generated over all examples, leading to $O(n^2 \log n) = O(n_p^2 \log n + n_n^2 \log n + 2n_p n_n \log n)$. Note that $n = n_p + n_n$. One can easily see that $O(n^2 \log n)$ is the same as $O(n_p^2 \log n_p + n_n^2 \log n_n)$. Therefore, the total additional time complexity for CLFDA is $O(n^2 \log n) + O(n \max(R, C)) = O(n^2 \log n)$. Hence, the time complexity of CLFDA is approximately the same as LFDA.

4. NUMERICAL EXPERIMENTS

4.1. Image Categorization

In this experiment, we followed the region-based image categorization approach which was conducted in [4]. To adapt multiple instance learning, each image which consists of several regions (i.e. clusters) is regarded as a bag, and the several regions are interpreted as the instances of the bag.

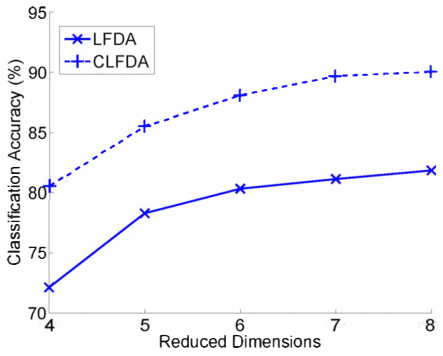

Fig. 2. The average classification accuracy of CLFDA and LFDA on Corel images.

Table 2. Comparison of classification accuracy of 5 different methods on Corel images with 95% confidence interval, where three methods (MILES, DD-SVM, and MI-SVM) are applied without dimensionality reduction.

Algorithm	accuracy (%)
CLFDA + MI-SVM ($d = 8$)	**90.6** : [89.3, 91.8]
LFDA + MI-SVM ($d = 8$)	81.8 : [80.1, 83.6]
MILES	82.6 : [81.4,83.7]
DD-SVM	81.5 : [78.5, 84.5]
MI-SVM	74.7 : [74.1, 75.3]

In this experiment, we used the Corel image data set composed of 20 categories. Each category contains 100 images. Fig. 3 represents sample images in the data set. A region of each image is obtained by k-means algorithm, and is described by a 9-dimensional feature vector. A detailed explanation about the feature extraction can be found in [4]. The preprocessed Corel data is available online[1] and we used the same version of the preprocessed Corel data.

In this experiment, we used first 10 categories among 20 categories, and then the total number of images is 1000. We used 50 images in each category for training and the other 50 images are reserved for testing. Hence, the training data consists of 500 images with 10 different classes. To reduce variance of results, we constructed 5 random splits of data into training and testing sets, and took the average over the 5 different test results.

This categorization problem is multi-class classification problem. We adapted the simple "one-versus-rest" approach. We constructed 10 different classifiers, and used the maximum value of the 10 different classifiers for the final decision of an test example.

In this experiment, we first applied CLFDA and LFDA to reduce the dimension of a instance, then we used the MI-SVM [1] to classify the test sets. The gaussian kernel, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) =$

[1] http://john.cs.olemiss.edu/ ychen/ddsvm.html

Table 3. The confusion matrix is summarized when CLFDA is applied to Corel images. Diagonal entries represent classification accuracy, while off-diagonal entries correspond to classification error.

	Cat. 0	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8	Cat. 9
Cat. 0	**77.6**	0.0	0.0	15.6	0.0	0.4	6	0.0	0.4	0
Cat. 1	0.0	**84.4**	0.0	12.4	0.0	0.0	0	0.4	1.6	1.2
Cat. 2	0.0	0.0	**82.4**	11.2	0.0	0.0	0.0	0.0	0.8	5.6
Cat. 3	0.0	0.0	0.0	**95.2**	0.0	0.0	0.0	0.0	0.0	4.8
Cat. 4	0.0	0.0	0.0	1.2	**98.8**	0.0	0.0	0.0	0.0	0.0
Cat. 5	0.0	0.0	0.4	2.8	0.0	**90.4**	0.0	4.4	0.8	1.2
Cat. 6	0.0	0.0	0.0	6.4	0.0	0.0	**93.6**	0.0	0.0	0.0
Cat. 7	0.0	0.0	0.0	0.4	0.0	0.0	0.0	**98.0**	1.2	0.4
Cat. 8	0.8	2.0	0.0	4.8	0.0	0.0	0.0	0.0	**91.6**	0.8
Cat. 9	0.0	0.0	0.4	3.6	0.0	0.0	0.0	2.0	0.4	**93.6**

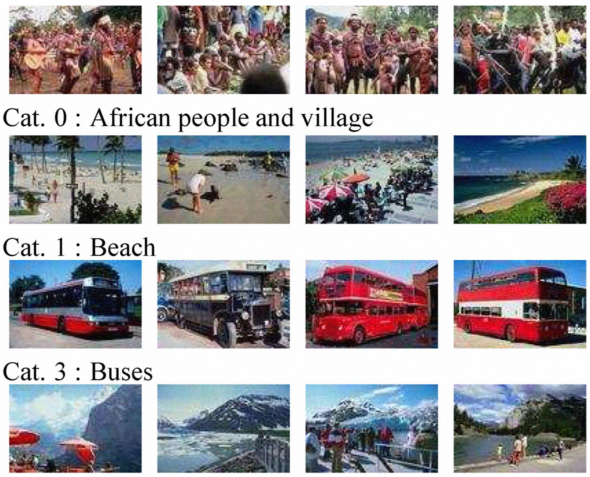

Fig. 3. Sample images from four categories in Corel data set

$e^{-||\boldsymbol{x}_i-\boldsymbol{x}_j||^2/\sigma^2}$ with $\sigma^2 = 0.05$, is used for the MI-SVM. We used the multiple instance learning library (MILL)[2] for MI-SVM.

In LFDA, there is one parameter k which is selected to be seven in this experiment. In CLFDA, there are three more parameters to be tuned (C,R, and t). We select t to be one manually. By the definition of t, this choice is very natural. For the other two parameters (C and R), we performed 5-fold cross validation on training data with $C, R \in \{1, 2, 3, 4, 5\}$.

Fig. 2 shows the classification accuracy using CLFDA and LFDA. We confirmed CLFDA generates better results than LFDA in every reduced dimension. We obtained the best classification accuracy when the dimension of the instance (d) is reduced to eight. We compared the classification accuracy of CLFDA + MI-SVM with that of MILES [3], DD-SVM [4] and MI-SVM.

Table 2 shows that CLFDA + MI-SVM reports the best performance when the reduced dimension is eight. Table 3 presents the confusion matrix of CLFDA + MI-SVM when the reduced dimension is eight. In the confusion matrix, diagonal entries show the classification accuracy and off-diagonal entries show the classification error. In the previous studies [3] and [4], the error between the category 1 and category 8 is reported relatively larger than any other error. In this experiment, the most largest error part is observed between the category 0 and category 3.

[2]http://www.cs.cmu.edu/ juny/MILL/index.html

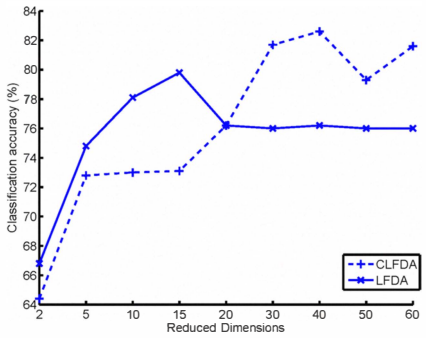

Fig. 4. Classification accuracy on Musk 1 data set with different reduced dimensions.

4.2. Drug Activity Prediction

Multiple instance learning has been applied to the drug activity prediction problem by [5]. The goal of the drug activity prediction is to find out whether the unknown drug molecule tightly binds to the protein or not. The most important factor for binding to the protein is the shape of the molecule, which varies over time. Hence, even for the same molecule, only a specific shape of the molecule binds to the protein. As the molecule is interpreted as a bag and a shape of the molecule is considered as an instance, the drug activity prediction fits well to multiple instance learning.

In this experiment, we used the Musk data set which

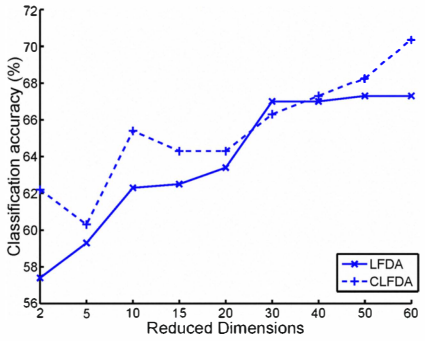

Fig. 5. Classification accuracy on Musk 2 data set with different reduced dimensions.

has been generated for the drug activity prediction and is provided by [5]. The Musk data set are available at UCI Machine Learning data repository, which distributes two versions: Musk 1 and Musk 2. Musk 1 and Musk 2 are basically same except the number of bags. An instance in Musk data is described by a 166-dimensional feature vector. We used citation-kNN as a classifier and performed 10-fold cross validation. We set $C = 4$ and $R = 2$ for citation-kNN, because these are known the best performance parameter in MILL. To select the hyperparameters, we take the same strategy in the previous section.

The classification accuracy over reduced dimensions is shown in Fig. 4 and 5. For Musk 1 and Musk 2, CLFDA is slightly superior to LFDA. Without dimensionality reduction, citation-kNN reports 92.4% classification accuracy for Musk 1 and 86.3% classification accuracy for Musk 2.

5. CONCLUSIONS

We have presented citation LFDA (CLFDA) which is local dimensionality reduction for multiple instance learning. This method is based on LFDA adapting the citation and reference concept to detect false positive instances. Experimental results of Corel data set and Musk data set confirmed that CLFDA is better than LFDA. According to [13], multiple instance learning can be considered as a semi-supervised learning imposed by a positive constraint. Therefore, for future work, we are working on incorporating semi-supervised dimensionality reduction method to multiple instance setting.

Acknowledgments: This work was supported by Korea NIPA ITRC support program (NIPA-2010-C1090-1031-0009), National Research Foundation (NRF) of Korea (No. 2010-0014306), Converging Research Center Program (No. 2009-0093714), and WCU Program (Project No. R31-2008-000-10100-0).

6. REFERENCES

[1] S. Andrew, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 15. MIT Press, 2003.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 14. MIT Press, 2002.

[3] Y. Chen, J. Bi, and J. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1931–1947, 2005.

[4] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[6] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[7] K. Fukunaga, *An Introduction to Statistical Pattern Recognition*. New York, NY: Academic Press, 1990.

[8] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 16. MIT Press, 2004.

[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 10. MIT Press, 1998.

[10] S. Ray and M. Craven, "Supervised versus multiple instance learning: An empirical comparison," in *Proceedings of the International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005.

[11] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 5, pp. 1027–1061, 2007.

[12] J. Wang and J. D. Zucker, "Solving the multiple-instance problem: A lazy learning approach," in *Proceedings of the International Conference on Machine Learning (ICML)*, San Francisco, CA, 2000.

[13] Z. H. Zhou and J. M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, Corvallis, OR, 2007, pp. 1167–1174.