International Journal of Database Theory and Application
Vol.9, No.11 (2016), pp.29-36
http://dx.doi.org/10.14257/ijdta.2016.9.11.03

An Active Learning Based LDA Algorithm for Large-Scale Data Classification

Xu Yu[1] *, Yan-ping Zhou[1] and Chun-nian Ren[1]

[1]*School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China*
Email: yuxu0532@163.com

Abstract

As traditional Linear Discriminant Analysis algorithm runs slowly in large data set, this paper proposed a fast LDA algorithm based on active learning. In the proposed algorithm, the original training set is divided into three parts, i.e. initial training set, correction set and testing set. Secondly, LDA algorithm is running on the initial training set, and the projection vector can be obtained. Thirdly, we select from correction set the samples whose projection is farthest from the mean vector, add them into the initial training set and compute the projection vector again. Repeat this step until the classification precision attains the expected target or the correction set is empty. The simulation experiments on the UCI data set and the MNIST data set show that the proposed algorithm running fast on large data set, and has a good classification precision.

Keywords*: Large scale data set; Linear Discriminant Analysis; Active learning; the MNIST data set*

1. Introduction

Linear Discriminant Analysis is an analytical method in statistics, and it use the idea of dimensionality reduction to do classification task. The research on LDA can date back to the Fisher's typical paper entitled 'The use of multiple measurements in taxonomic problems' [1].

Currently, LDA is applied widely in many aspects of human life, such as face recognition [2-3], speech recognition [4-5], fault diagnosis [6-7] and network intrusion detection [8-9], and achieves a good performance. However, a serious drawback of LDA is that it runs slowly on large-scale classification data set. The computational complexity of finding the optimal line for the Fisher linear discriminant is dominated by the calculation of the within-category total scatter and its inverse, which is an $O(d^2n)$ calculation.

For this shortcoming, in this paper we proposed a fast LDA algorithm based on active learning, and gives the detailed implementation. The proposed algorithm chooses the most significant samples to do classification based on active learning, which can get a higher performance with a small number of samples and achieve a fast classification.

Our paper is organized as follows. In Section 2, the active learning method and the LDA algorithm are reviewed. In Section 3, a fast LDA algorithm based on active learning is proposed. The simulation experiments on the UCI data set and the MNIST data set are conducted in Section 4, and the experimental results and a detailed analysis are also given in this part. Section 5 concludes the whole paper.

ISSN: 2005-4270 IJDTA

2. Review on Active Learning and the LDA Algorithm

2.1. Review on Active Learning

Active learning is an iterative type of supervised learning that is suitable for situations where data are abundant, yet the class labels are scarce or expensive to obtain. The learning algorithm is active in that it can purposefully query a user (*e.g.*, a human oracle) for labels. The number of tuples used to learn a concept this way is often much smaller than the number required in typical supervised learning [10].

To keep costs down, the active learner aims to achieve high accuracy using as few labeled instances as possible. Let D be all of data under consideration. Suppose that a small subset of D is class-labeled. This set is denoted L. U is the set of unlabeled data in D. Various strategies exist for active learning on D. A pool-based approach to active learning is given in the following [11].

(1) An active learner trains a classifier on L which is the initial training set.

(2) It then uses a querying function to carefully select one or more data samples from U (also referred to as a pool of unlabeled data) and requests labels for them from an oracle (*e.g.*, a human annotator).

(3) The newly labeled samples are added to L, and the active learner trains a classifier again on the expanded L with supervised way.

(4) Repeat the third step until the classifier trained by the active learner achieve a good performance.

An illustration about this process is given in Figure 1.

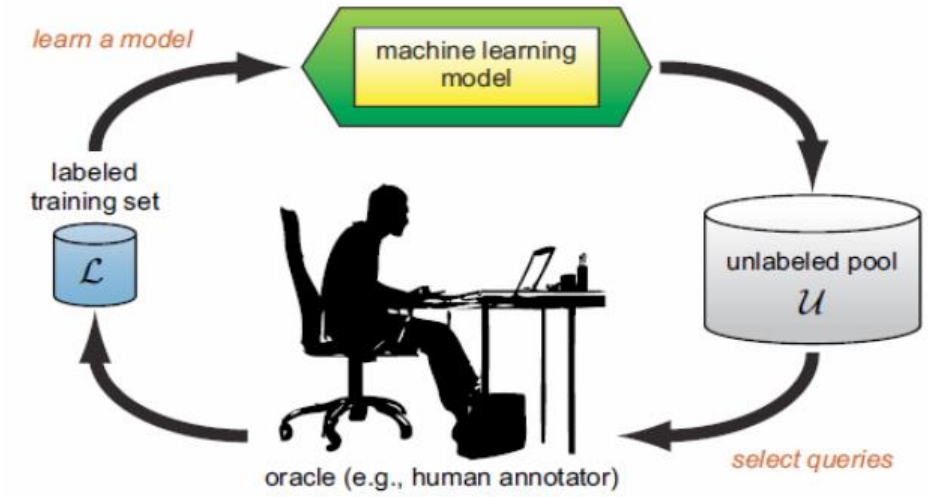

Figure 1. The Pool-Based Active Learning Cycle

2.2 Review on the LDA algorithm

LDA algorithm is a typical pattern classification algorithm based on dimensionality reduction. For a binary classification problem, the LDA algorithm converts the problem from *n*-dimensional space to 1-dimensional space by projecting the two classes of samples on a certain line. Thus the complex classification problem can be solved in an easier way. The LDA algorithm requires that the projected samples are well separated. Figure 2 shows a comparison between projection of samples onto two different lines. The figure on the right shows greater separation between the red and black projected points.

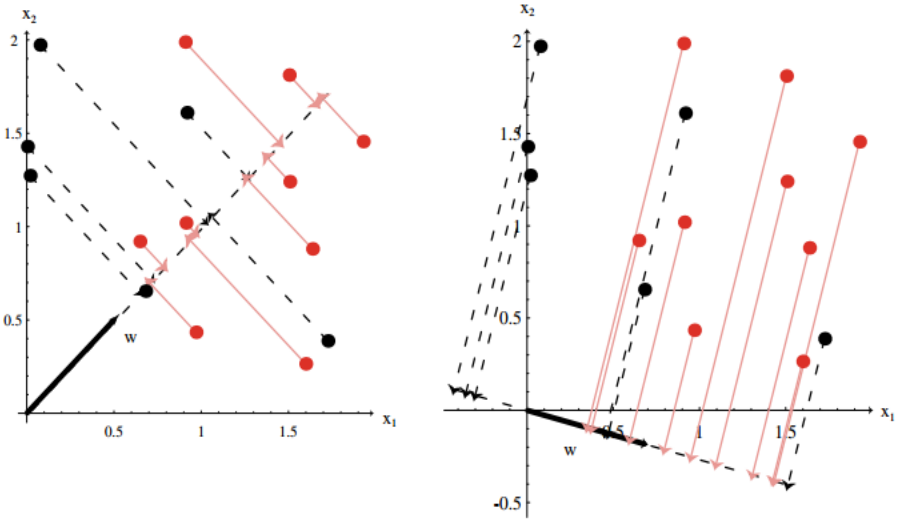

Figure 2. Projection of Samples Onto Two Different Lines

The idea proposed by Fisher is to maximize a function that will give a large separation between the projected class means while also giving a small variance within each class, thereby minimizing the class overlap [12].

The detailed computation on LDA is given in the following. Let two classes of samples are denoted by $\chi_1 = \{x_1^1, ..., x_{l_1}^1\}$ and $\chi_2 = \{x_1^2, ..., x_{l_2}^2\}$. Let $\chi = \chi_1 \bigcup \chi_2 = \{x_1, ..., x_l\}$, where $l = l_1 + l_2$.

The LDA algorithm achieves the direction vector w of the best projection line by maximize the following function

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

where

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$
$$S_W = \sum_{i=1,2} \sum_{x \in \chi_i} (x - m_i)(x - m_i)^T$$

denote the between-class scatter matrix and the within-class scatter matrix respectively, and $m_i = \frac{1}{l_i} \sum_{j=1}^{l_i} x_j^i$ ($i = 1, 2$) denotes the center of the i-th class of original samples.

For similarity, in this paper we just give the result of the direction vector w by the following. Detailed analysis on maximizing the function $J(w)$ can refer to [13].

$$w = S_w^{-1}(m_1 - m_2)$$

3. Algorithm Design

In this Section, we will design a fast LDA algorithm based on active learning, which is abbreviated as FLDBAL. Firstly, the original training set T is divided into three parts, namely A, B, C, where A is the initial training set, B is the correction set and C is the testing set. Our algorithm firstly uses A to train an initial LDA classifier. Then select proper samples from B, add them into A and retrain the LDA classifier. Repeat this step until the classification precision attains the expected target or the correction set is empty. Strict algorithm is shown in ALGORITHM 1.

ALGORITHM 1

Input: the original training set T
Output: the projection line
method:

1. Divide the original training set into 3 different part, *i.e*. the initial training set A, the correction set B and the testing set C. The proportion of the sizes of A, B, and C is 2:6:2.

2. Train a LDA learner on the initial training set A, and the direction vector of the best projection line is *w*

3. Compute the precision on the testing set C

4. If the precision of the LDA learner attained the expected target or the correction set B is empty, output the direction vector *w*, else repeat Step 5-8.

5. Project the samples in the initial training set A onto *w*, and compute the projection center of the two classes of samples, denoted as m_1 and m_2.

6. Select x_1 and x_2 from each class of B, where x_1 denotes the farthest sample from the projection center m_1 and x_2 denotes the farthest sample from the projection center m_2. That is to say,

$$d(x_1, m_1) = \max_{x \in one\ class\ of\ B} d(x, m_1)$$

$$d(x_2, m_2) = \max_{x \in the\ other\ class\ of\ B} d(x, m_2)$$

7. Add x_1 and x_2 into A, and remove them from B, *i.e*., $A = A \bigcup \{x_1, x_2\}$, $B = B / \{x_1, x_2\}$

8. Train a new LDA learner on A, and compute the classification precision on the testing set C.

The FLDBAL algorithm is illustrated in Figure 3. As shown in Figure 3, the red samples consists of the initial training set A, and the yellow samples consists of the correction set B, where the square samples denote one class and the circle samples denote the other class. The FLDBAL algorithm firstly trains a LDA learner on A and obtains the best line *l*, the projections and the projection centers m_1 and m_2. Then compute x_1 and x_2, remove them from B and add them into *A*. Finally, train a new LDA learner on *A* until the classification precision on the testing set *C* attains the expected goal or the correction set *B* is empty.

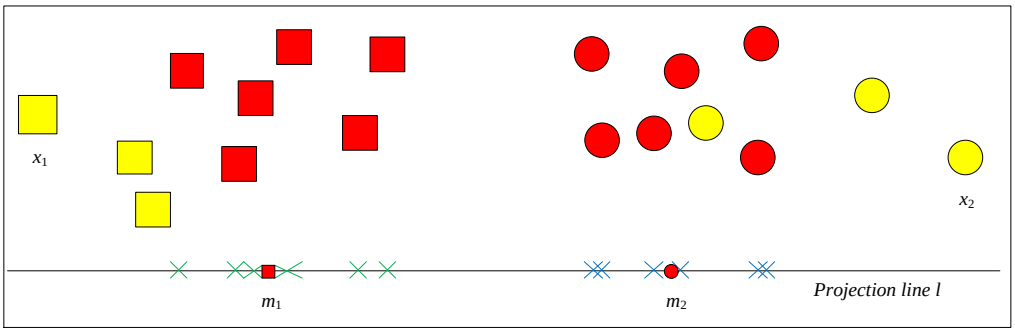

Figure 3. An Illustration of the FLDBAL Algorithm

The proposed algorithm employs active learning to train a fast LDA classifier. It firstly uses 20% of the samples to train, and then choose the most important samples from the correct set to correct. On one side, it improves the efficiency by largely reducing the number of samples needed for training. On the other side, it guarantees the classification precision by selecting the most important samples. In order to test the performance of the algorithm furtherly, we conduct two experiments on the UCI data set and the MNIST data set in the next section.

4. Experiments

In this paper, we perform experiments on the UCI standard data sets and the MNIST data set to test the performance of the proposed FLDBAL algorithm and the traditional LDA algorithm. All the algorithms are implemented with Matlab 2015 and libsvm-mat-2.83. All experiments are run on 2.00 GHz, Intel (R) Core (TM) 2 CPU with 2GB main memory under window 7.

4.1. Experiments on the UCI Data Sets

Firstly, we use 3 different UCI data sets to test the performance of the proposed FLDBAL algorithm and the traditional LDA algorithm. For the traditional LDA algorithm, we select 80% of samples from each UCI data set for training and the remaining 20% is used for testing. For the FLDBAL algorithm, we randomly select 20% of samples from the original training sample set to form *A*, randomly select 60% of samples to form B, and the rest as *C*. The selected data information is shown in Table 1.

Table 1. The Experimental Data Information of the FLDBAL Algorithm

Data set	Size of A	Size of B
mashroom	1600	4800
credit	120	360
glass	40	120

Five runs of 10-fold cross-validation are performed for each algorithm on the three UCI data sets, and the average classification precision and running time are reported in Table 2 and Table 3.

Table 2. Comparison of Classification Precision on 3 UCI Data Set

Data set	Classification precision (%)	
	LDA	FLDBAL
mashroom	96.6	96.3
credit	93.1	92.9
glass	92.2	92.1

Table 3. Comparison of Running Time on 3 UCI Data Set

Data	Running time (ms)	
	LDA	FLDBAL
mashroom	10233.2	7215.3
credit	205.6	162.3
glass	35.5	13.6

4.2. Experiments on the MNIST Data Set

In this part, we conduct the experiment on the MNIST data set. This data set consists of 70000 handwritten digits, including 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image. The MNIST data set consists of 10 classes of images, including ′0′ to ′9′. Each image is with a gray level of 8, and can be represented by a 784-dimensonal vectors. Figure 4 gives some samples in the MNIST data set. For similarity, we just choose the ′0′ and ′1′ samples for classification. For the traditional LDA algorithm, we randomly select 800 ′0′ and 800 ′1′ as training samples, and randomly select 200 ′0′ and 200 ′1′ as testing samples. For the FLDBAL algorithm, we randomly select 200 ′0′ and 200 ′1′ as training samples, and randomly select 600 ′0′ and 600 ′1′ as correction samples.

The preprocessing was done by Le Cun *et al.* [14] and a linear transform was performed such that all patterns were centered at 28 × 28 window while keeping the

aspect ratio. The pixel values of resulting gray-scale images were scaled to fall in the range from –1.0 to 1.0.

Since patterns on MNIST are not truly located at the center, first the preprocessing was performed by enclosing the pattern with a rectangle, and then translating this rectangle into the center of a 28 × 28 box. Then patterns were blurred using the following mask:

$$\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

Finally, DeCoste and Schölkopf's [15] idea was used to normalize each pattern by its Euclidean-norm scalar value such that the dot product was always within [−1, 1].

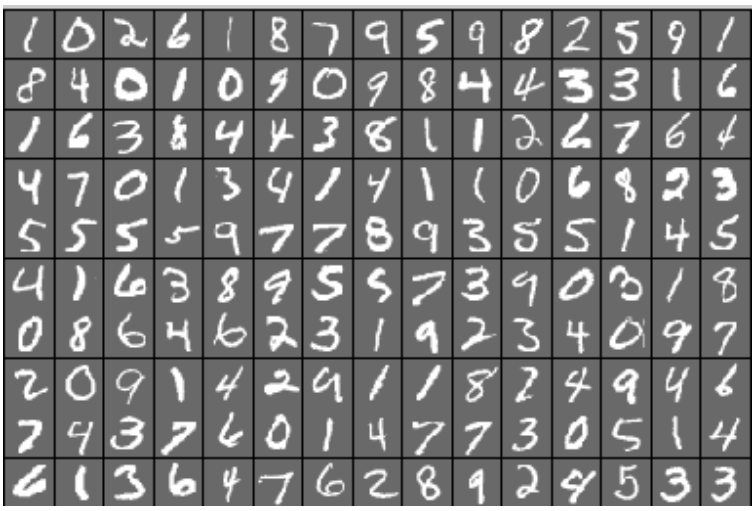

Figure 4. Some Samples in the MNIST Data Set

Five runs of 10-fold cross-validation are performed for each algorithm, and the average classification precision and running time are reported in Table 4.

Table 4. The Result Comparison between LDA and FLDBAL on the MNIST Data Set

Algorithm	Classification precision (%)	Running time (s)
LDA	95.2	722.1
FLDBAL	94.1	425.6

4.3. The Experimental Result and Analysis

As shown in Table 2, Table 3 and Table 4, the classification precision of the FLDBAL algorithm is slightly lower than that of the traditional LDA algorithm, the FLDBAL algorithm is much more efficient. The main reason is that the traditional LDA algorithm require to train on the whole data set, but the FLDBAL algorithm just need a few important training samples for training based on active learning, thus its efficiency is higher and the classification precision is always acceptable.

5. Conclusion

In this paper, we proposed a fast LDA algorithm based on active learning. The proposed algorithm uses active learning for training, and firstly trains an initial LDA learner on the initial training set, then selects the most important samples for correction. The proposed algorithm is with a higher efficiency than the traditional LDA algorithm,

but its classification precision can almost match that of LDA. The simulation experiments on the UCI data set and the MNIST data set show the effective of the proposed algorithm.

Acknowledgments

This work is sponsored by the National Natural Science Foundation of China (Nos. 61402246, 61402126, 61370083, 61370086, 61303193, and 61572268), a Project of Shandong Province Higher Educational Science and Technology Program (No. J15LN38, J14LN31), Qingdao indigenous innovation program (No. 15-9-1-47-jch), the Project of Shandong Provincial Natural Science Foundation of China (No. ZR2014FL019), the Open Project of Collaborative Innovation Center of Green Tyres & Rubber (No.2014GTR0020), the National Research Foundation for the Doctoral Program of Higher Education of China (No. 20122304110012), the Science and Technology Research Project Foundation of Heilongjiang Province Education Department (No. 12531105), Heilongjiang Province Postdoctoral Research Start Foundation (No. LBH-Q13092), and the National Key Technology R&D Program of the Ministry of Science and Technology under Grant No. 2012BAH81F02.

References

[1] R. A. Fisher D. F. R. S. Sc, "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, vol. 7, no. 2, **(1936)**, pp. 179-188.
[2] Y. Hua and Y. Jie, "A direct LDA algorithm for high-dimensional data with application to face recognition", Pattern Recognition, vol. 34, no. 00, **(2001)**, pp. 2067-2070.
[3] L. F. Chen, H. Y. M. Liao and M. T. Ko, "A new LDA-based face recognition system which can solve the small sample size problem", Pattern Recognition, vol. 33, no. 10, **(2000)**, pp. 1713-1726.
[4] A. Martin and L. Mauuary, "Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments", Speech Communication, vol. 48, no. 2, **(2006)**, pp. 191-206.
[5] D. Kolossa, S. Zeiler amd R. Saeidi, "Noise-Adaptive LDA: A New Approach for Speech Recognition Under Observation Uncertainty", IEEE Signal Processing Letters, vol. 20, no. 11, **(2013)**, pp. 1018 - 1021.
[6] P. M. Frank, "Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results", Automatica, vol. 26, no. 3, **(1990)**, pp. 459-474.
[7] K. Guo, Y. Zhu and Y. San, "Analog Circuit Fault Diagnosis Using LDA and OAOSVM Approach", Advanced Materials Research, **(2012)**, pp. 490-495:1130-1134.
[8] B. Mukherjee, L. T. Heberlein and K. N. Levitt, "Network intrusion detection", IEEE Network the Magazine of Global Internetworking, vol. 8, no. 3, **(1994)**, pp. 26-41.
[9] R. X. Zhang and Y. Wang, "Fusion of PCA and LDA for Intrusion Detection", Computer Technology & Development, **(2009)**.
[10] J. Han, M. Kamber and J. Pei, "Data mining concept and technology (the third edition)", Beijing: China machine press, **(2012)**.
[11] A. Mccallum and K. Nigam, "Employing EM and Pool-Based Active Learning for Text Classification", Icml, **(1998)**, pp. 350-358.
[12] C. M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer-Verlag New York, Inc., **(2006)**.
[13] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification (2nd Edition)", En Broeck the Statistical Mechanics of Learning Rsity, **(2000)**.
[14] Y. LeCun, L. D. Jackel, L. Bottou, J. S. Denker, H. Drucker, I. Guyon, U. A. Müller, E. Sackinger, P. Simard and V. N. Vapnik, "Comparison of learning algorithms for handwritten digit recognition", In: Proceedings of International Conference on Artificial Neural Network, **(1995)**, pp. 53-60.
[15] D. DeCoste and B. Schölkopf, "Training invariant support vector machines", Machine Learning, vol. 46, no. 1-3, **(2002)**, pp. 161-190.

