Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 557–562
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

The effect of imbalanced data sets on LDA: A theoretical and empirical analysis

Jigang Xie*, Zhengding Qiu

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, PR China

Received 13 August 2005; accepted 17 January 2006

Abstract

This paper demonstrates that the imbalanced data sets have a negative effect on the performance of LDA theoretically. This theoretical analysis is confirmed by the experimental results: using several sampling methods to rebalance the imbalanced data sets, it is found that the performances of LDA on balanced data sets are superior to those of LDA on imbalanced data sets.

Keywords: Imbalanced data sets; Linear discriminant analysis (LDA); Random sampling; Tomek links; Smote

1. Introduction

Supervised learning methods have advanced to the point where they might be applied to real world problems, such as in data mining and knowledge discovery, document categorization, and financial forecasting [1]. By being applied in such domains, the problem of imbalanced data sets, a huge disproportion in the number of examples belonging to each class, is common. A lot of works [2–4] have been focused on learning from imbalanced data sets using standard supervised learning methods, including decision trees, support vector machines, nearest neighbor rule, etc. Nevertheless, none of them have explored the effect of imbalanced data sets on the performance of linear discriminant analysis (LDA). In this work, we demonstrate that the imbalanced data sets have a negative effect on LDA theoretically. We use some sampling methods to obtain the balanced data sets from the original imbalanced data sets, and use LDA to learn from both data sets. The experimental results verify the correctness of theoretical analysis.

2. Linear discriminant analysis

The classical LDA can be developed by Gaussian distribution based Bayes plug-in rule [5]. In terms of a set of discriminant functions $g_j(\mathbf{x})$, $i = 1, \ldots, c$, the classifier is said to assign an example $\mathbf{x}$ to class ω_j if

$$g_j(\mathbf{x}) > g_i(\mathbf{x}) \quad \text{for all } i \neq j. \tag{1}$$

When the underlying classes are Gaussian distributed, and the parameters of the distribution are also known, $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, the discriminant functions can be derived from the Bayes decision rule:

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{d}{2}\ln 2\pi + \ln P(\omega_j), \tag{2}$$

where d is the dimensional number of $\mathbf{x}$. When all the covariance matrices are assumed to be equal, the discriminant

* Corresponding author. Tel.: +86 10 51688636; fax: 5168 61688616.
E-mail addresses: xie_jigang@263.net (J. Xie), zdqiu@center.njtu.edu.cn (Z. Qiu).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.01.009

functions can be simplified to

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{d}{2}\ln 2\pi + \ln P(\omega_j), \tag{3}$$

where $\boldsymbol{\Sigma}$ is the common covariance matrix. The resulting functions (3) are linear in $\mathbf{x}$, hence the Bayes decision rule belongs to the class of LDA. And we also refer this class of LDA as the Gaussian-based LDA.

Following the common practice [2,3], we consider only the two-category problems and therefore, the examples are either from the minority class or the majority class, respectively. It is more common to define a single discriminant function

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}), \tag{4}$$

and to use the following decision rule: decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2. Thus, the linear discriminant functions can be written as

$$g^L(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \ln\frac{P(\omega_1)}{P(\omega_2)}. \tag{5}$$

Expansion of the quadratic form $(\mathbf{x}-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)$ results in a sum involving a quadratic term $\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}$ which here is independent of j. After this term is dropped from (5), the resulting discriminant functions are again linear:

$$g^L(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + \mathbf{w}_0, \tag{6}$$

where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{7}$$

and

$$\mathbf{w}_0 = -\frac{1}{2}(\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \ln\frac{P(\omega_1)}{P(\omega_2)}. \tag{8}$$

As $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ and $P(\omega_j)$ are in practice unknown, they have to be estimated from the training data. For the estimates of the prior probabilities $P(\omega_j)$, the relative frequencies of the examples in each class are usually used in practice. To estimate $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, one usually uses the sample mean $\hat{\boldsymbol{\mu}}_j$ and the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_j$, which are the maximum likelihood estimates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. We denote the training data by $\{\mathbf{x}_{1_1}, \mathbf{x}_{1_2}, \ldots, \mathbf{x}_{1_{n_1}}\}$ and $\{\mathbf{x}_{2_1}, \mathbf{x}_{2_2}, \ldots, \mathbf{x}_{2_{n_2}}\}$, where $\mathbf{x}_{1_j}$ and $\mathbf{x}_{2_j}$ are draw independently from their respective classes respectively. The estimators can be written as

$$\hat{P}(\omega_j) = \frac{n_j}{n_1+n_2}, \quad j = 1, 2, \tag{9}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{j_i}, \quad j = 1, 2, \tag{10}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(\mathbf{x}_{j_i}-\hat{\boldsymbol{\mu}}_j)(\mathbf{x}_{j_i}-\hat{\boldsymbol{\mu}}_j)^t, \quad j = 1, 2. \tag{11}$$

For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the unbiased estimator is

$$\hat{\boldsymbol{\Sigma}} = \frac{(n_1-1)\hat{\boldsymbol{\Sigma}}_1 + (n_2-1)\hat{\boldsymbol{\Sigma}}_2}{n_1+n_2-2} \approx P(\omega_1)\hat{\boldsymbol{\Sigma}}_1 + P(\omega_2)\hat{\boldsymbol{\Sigma}}_2 \quad \text{for } n_1 \text{ and } n_2 \text{ large enough.} \tag{12}$$

In pattern recognition circles, LDA usually refers to Fisher's linear discriminant (FLD). The objective of the FLD is to find the optimal projection so that the Fisher criterion of between-class scatter over within-class scatter is maximized. The within-class scatter and the projection matrix are given by

$$\mathbf{S}_j = \sum_{i=1}^{n_j}(\mathbf{x}_{j_i}-\hat{\boldsymbol{\mu}}_j)(\mathbf{x}_{j_i}-\hat{\boldsymbol{\mu}}_j)^t, \quad j = 1, 2, \tag{13}$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2, \tag{14}$$

$$\mathbf{W}^F = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{15}$$

From (7)–(15), it can be seen that the only difference between FLD and Gaussian-based LDA is an unimportant divisor, n_1+n_2-2. That is, the projection direction of FLD is in fact identical to that of Gaussian-based LDA. In this study, the Gaussian-based LDA is selected as the representation of LDA.

It is clear that the imbalanced data sets will not have effects on the projection matrix if the two sample covariance matrices are identical. However, the assumption of equal sample covariance matrices is restricted to particular cases in real-life scenarios. Therefore, we should consider the effect of the imbalanced data sets on the performance of LDA in practice. It is also clear that, if the two sample covariance matrices are different, the huge imbalance in class distribution is very problematic for LDA because the prior probability of majority class overshadows the differences in the sample covariance matrix terms. That is, the imbalanced data sets may hinder the performance of LDA.

3. Experimental methodology

The main objective of our experiment is to compare the performance of LDA on imbalanced data sets with that of LDA on balanced data sets. To make this comparison objective, a common method is to calculate the area under the ROC curve (AUC) [6]. The reason why AUC is used is that it has an important statistical meaning: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test

Table 1
Description of data sets

Data set	Examples	Attributes	Class (min., maj.)	Class (min., maj.) %
Letter-a	20000	16	(a, remainder)	(3.95%, 96.05%)
Satimage-3	6435	36	(3, remainder)	(21.1%, 78.9%)
Waveform	5000	21	(1, remainder)	(33.33%, 66.67%)
Image*	2310	18	(BRICKFACE, remainder)	(14.29%, 85.71%)
Vehicle	846	18	(van, remainder)	(23.52%, 76.48%)
Pima	768	8	(1, 0)	(34.77%, 65.23%)
New-thyroid	215	5	(hypo, remainder)	(16.28%, 83.72%)
Glass	214	9	(Ve-win-float-proc, remainder)	(7.94%, 92.06%)
Wine	178	13	(3, remainder)	(26.97%, 73.03%)
Iris	150	4	(3, remainder)	(33.33%, 66.67%)

of ranks [7], which is twice the area between the diagonal and the ROC curve. Therefore, AUC has been a common method to compare classifiers [2–4].

In order to allow us to generalize from our results, we have selected ten data sets from UCI [8], which have different degrees of imbalance. Table 1 summarizes the data employed in this study. For each data set, it shows the number of examples, number of attributes and class attribute distribution. For data sets having more than two classes, we chose the class with fewer examples as the minority class, and collapsed the remainder as the majority class. It should be noted that for data set Image, which are identified with an asterisk (*), the third column of the original data set is deleted because of they are identical to a constant.

To get the balanced data sets, four sampling methods are used to rebalance the original data sets. They are briefly described as following:

- Random over-sampling is a non-heuristic method that aims to balance the class distribution through the random replication of minority-class examples. In this paper, the simplest form of over-sampling, duplication of minority class, is used to rebalance the original data sets.
- Random under-sampling is also a non-heuristic method that aims to balance class distribution through the random elimination of majority-class examples.
- Tomek links [9] can be defined as follows: given two examples E_i and E_j belonging to different classes, and $d(E_i, E_j)$ be the distance between E_i and E_j, then a (E_i, E_j) pair is called a Tomek link if there is not an example E_l, such that $d(E_i, E_l) < d(E_i, E_j)$ or $d(E_j, E_l) < d(E_i, E_j)$. If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline cases. As an under-sampling method, examples belonging to the majority class are removed.
- Synthetic minority over-sampling technique (Smote) [10] is a heuristic over-sampling method. Its main idea is to form new minority-class examples by interpolating between several minority-class examples that lie together.

4. Experimental results and analysis

In this work, we calculate the AUC for each data set using 4-fold cross-validation. The training and test sets are formed as follows. First, the test set is one of the four subsets of each cross validation, which is formed by 25% of the minority-class examples and 25% of the majority-class examples from the original data set, without replacement. The resulting test set will therefore adhere to the original class distribution. The remaining data set is as the original training set. It was worth noting that Smote, random over-sampling and random under-sampling are used to set up the approximate 1:1 class distribution in training set.

Table 2 lists the AUC for LDA on the original and the balanced data sets. It clearly shows that, except for Tomek links, the other three sampling methods lead an improvement in AUC for all data sets but Pima. This is consistent with the above theoretical analysis. That is, for the assumption of equal sample covariance matrices being restricted to particular cases in real-life scenarios, the most imbalanced data sets have a negative effect on the performance of LDA. For data set Pima, we attribute the experimental result to

Table 2
AUC for the imbalanced and balanced data sets

Data set	Original	Random over	Random under	Tomek	Smote
Letter-a	0.9754	0.9855	0.9853	0.9754	0.9800
Satimage-3	0.9860	0.9876	0.9871	0.9861	0.9877
Waveform	0.9414	0.9434	0.9425	0.9413	0.9437
Image*	0.9925	0.9939	0.9928	0.9925	0.9941
Vehicle	0.9828	0.9906	0.9902	0.9834	0.9908
Pima	0.8317	0.8303	0.8293	0.8296	0.8309
New-thyroid	0.9788	0.9977	0.9957	0.9761	0.9977
Glass	0.7497	0.8392	0.7884	0.7593	0.9147
Wine	0.9987	1.0000	0.9990	0.9987	1.0000
Iris	0.9673	0.9839	0.9825	0.9673	0.9877

Table 3
Cosine values of the angles between different projection directions

Data sets	Sampling methods	Original	Random over	Smote	Tomek	Random under
Letter-a	Original	1	0.9652	0.9855	1	0.9592
	Random over		1	0.994	0.9653	0.9931
	Smote			1	0.9856	0.9868
	Tomek				1	0.9593
	Random under					1
Satimage-3	Original	1	0.9242	0.9017	0.9882	0.8328
	Random over		1	0.9809	0.9136	0.9023
	Smote			1	0.8945	0.8752
	Tomek				1	0.8341
	Random under					1
Waveform	Original	1	0.997	0.9928	0.9961	0.9801
	Random over		1	0.9985	0.9955	0.9835
	Smote			1	0.9923	0.9857
	Tomek				1	0.979
	Random under					1
Image*	Original	1	0.9818	0.9174	1	0.4251
	Random over		1	0.9307	0.9818	0.4483
	Smote			1	0.9174	0.534
	Tomek				1	0.4251
	Random under					1
Vehicle	Original	1	0.9932	0.9781	0.9987	0.947
	Random over		1	0.9899	0.9942	0.9549
	Smote			1	0.9814	0.9721
	Tomek				1	0.9513
	Random under					1
Pima	Original	1	0.9957	0.99	0.997	0.9852
	Random over		1	0.9975	0.9916	0.9913
	Smote			1	0.9869	0.9896
	Tomek				1	0.9787
	Random under					1
New-thyroid	Original	1	0.7413	0.7348	0.9994	0.7645
	Random over		1	0.9985	0.7428	0.9783
	Smote			1	0.7473	0.9772
	Tomek				1	0.776
	Random under					1
Glass	Original	1	0.8627	0.9767	0.9851	0.4502
	Random over		1	0.8734	0.8156	0.7417
	Smote			1	0.9696	0.4399
	Tomek				1	0.4365
	Random under					1
Wine	Original	1	0.9676	0.9535	1	0.9338
	Random over		1	0.9967	0.9676	0.9712
	Smote			1	0.9535	0.9707
	Tomek				1	0.9338
	Random under					1
Iris	original	1	0.9286	0.8055	0.9967	0.946
	Random over		1	0.9629	0.93	0.9226
	Smote			1	0.8083	0.8265
	Tomek				1	0.9636
	Random under					1

the fact that the two sample covariance matrices may be equal. However, the Tomek links method rarely leads to an improvement in AUC for most data sets. This can be explained by the fact that there are so few majority-class examples are removed that the training set is yet imbalanced. Table 2 also shows that Smote and random over-sampling are more effective than random under-sampling in improving the performance of LDA. We attribute this to the fact that random under-sampling discards potentially useful majority-class examples and thus can degrade classifier performance.

In order to make the above analysis more comprehensible, Table 3 lists the average cosine values of angles between different projection directions of LDA on cross validation imbalanced and balanced data sets. The values near 1 mean that the two projection directions are near identical. Table 3 shows that, for most data sets, the projection directions of

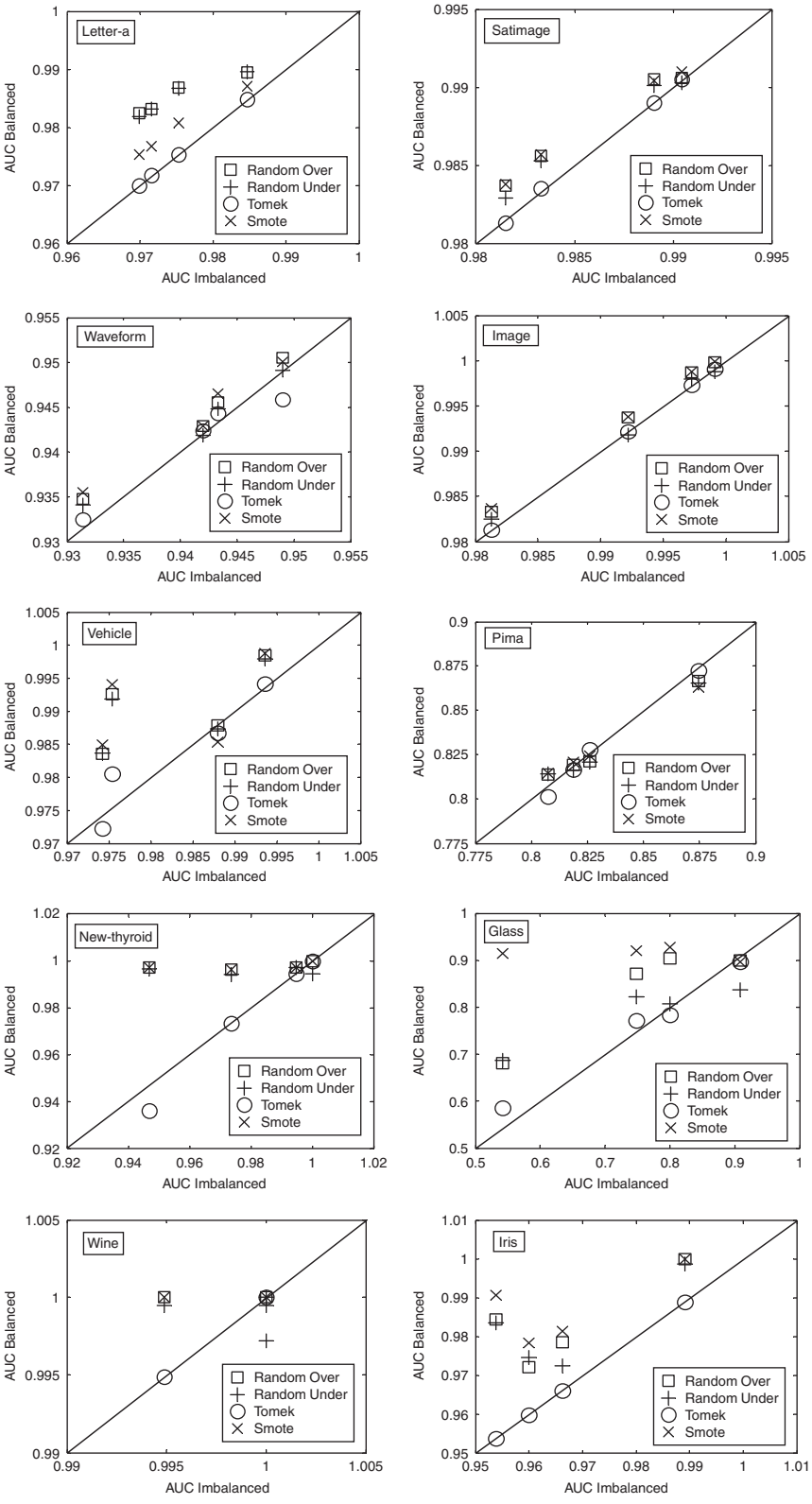

Fig. 1. AUC of LDA on the original and balanced data sets of each cross validation.

LDA on balanced data sets are different from that of LDA on the original data sets. The experimental results of Pima data set in Table 2 can be easily explained by that fact that the cosine values of angles between all the projection directions are near 1. That is, the two sample covariance matrices are nearly equal.

Fig. 1 clearly shows a comparison of the performance of LDA of each cross validation. Line $x = y$ represents when both LDA on imbalanced and balanced data sets obtain the same AUC. Plots above this line represent that LDA on balanced data sets obtain better results, and plots under this line the opposite. It can be seen that most results of each cross validation are consistent with the above analysis.

5. Conclusion

In this paper, we demonstrate that the imbalanced data sets have a negative effect on LDA theoretically. This theoretical analysis is confirmed by the experimental results: using four sampling methods to rebalance the imbalanced data sets, it is found that the performances of LDA on balanced data sets are better than those of LDA on imbalanced data sets. The experimental results also show that the two over-sampling methods are more effective than the two under-sampling methods in improving the performance of LDA.

References

[1] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mech. Intell. 22 (1) (2000) 4–37.

[2] N.V. Chawla, N. Japkowicz, A. Kolcz, Special issue on learning from imbalanced data sets, ACM SIGKDD Explorations 6 (1) (2004).

[3] N.V. Chawla, N. Japkowicz, A. Kolcz (Eds.), Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets, 2003.

[4] N. Japkowica (Ed.), Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets, AAAI Technical Report WS-00-05, AAAI, 2000.

[5] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.

[6] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.

[7] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[8] C.L. Blake, C.J. Merz, UCI repository of machine learning database, ⟨http://www.ics.uci.edu/∼mlearn/MLRepository.html/⟩, 1998.

[9] I. Tomek, Two modifications of CNN, IEEE Trans. Syst. Man Commun. SMC-6 (1976) 769–772.

[10] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.