Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding

Juanjuan Wang[1], Mantao Xu[2], Hui Wang[2], Jiwu Zhang[2]
([1]Department of Biomedical Engineering, Shanghai Jiaotong University, Shanghai, 200030)
([2]Kodak Health Group Global R&D Center, Shanghai, 201206)
E-mail: wangjuanjuan@sjtu.edu.cn

Abstract

The classification of imbalanced data is a common practice in the context of medical imaging intelligence. The synthetic minority oversampling technique (SMOTE) is a powerful approach to tackling the operational problem. This paper presents a novel approach to improving the conventional SMOTE algorithm by incorporating the locally linear embedding algorithm (LLE). The LLE algorithm is first applied to map the high-dimensional data into a low-dimensional space, where the input data is more separable, and thus can be oversampled by SMOTE. Then the synthetic data points generated by SMOTE are mapped back to the original input space as well through the LLE. Experimental results demonstrate that the underlying approach attains a performance superior to that of the traditional SMOTE.

1. Introduction

Imbalanced data classification often arises in many practical applications in the context of medical pattern recognition and data mining. Most of the existing state-of-the-art classification approaches are well developed by assuming the underlying training set is evenly distributed. However, they are faced with a severe bias problem when the training set is a highly imbalanced distribution (i.e., the data comprises two classes, the minority class C_+ and the majority class C_-). The resulting decision boundary is severely biased to the minority class, and thus leads to a poor performance according to the receiver operator characteristic (ROC) curve analysis. For this purpose, many classification algorithms have been investigated intensively, such as the undersampling technique over the majority class, the oversampling technique over the minority class, the cost-sensitive learning algorithm, and feature selection.

The synthetic minority oversampling technique (SMOTE) [1,2] is an important approach by oversampling the positive class or the minority class. However, it is limited to a strict assumption that the local space between any two positive instances is positive or belongs to the minority class, which may not always be true in the case when the training data is not linearly separable. However, mapping the training data into a more linearly separable space, where the SMOTE algorithm can be conducted, can circumvent this limitation. Once the positive class is oversampled synthetically in the linearly separable space, the newly generated data should be transformed back into the original input space. The transformation mapping from input data space into the linearly separable space should be feasibly invertible in practice. For this purpose, the locally linear embedding (LLE) algorithm [3] lends us a tool in design of an invertible mapping from the original input space to the linearly separable space. In this work, we present a new oversampling technique based on SMOTE and LLE. The training data is first mapped into a lower-dimensional space by LLE, where data is more separable. Then the SMOTE can be applied to generate a desirable number of synthetic data points for the positive class. Finally, these new data points are mapped back to the original input space.

The structure of this presentation is organized as follows: In section 2, we review the LLE algorithm. The LLE-based SMOTE algorithm is presented in section 3. Section 4 reports the performance comparison result of the LLE-based SMOTE algorithm and the conventional SMOTE algorithm. Finally, the conclusion is drawn in Section 5.

2. Locally Linear Embedding

The features extracted from medical images often have high dimensionality, and thus result in an intractable geometric complexity in data classification. Moreover, they are non-linearly separable in *Euclidean* space. The pioneer solution is a class of manifold learning algorithms [3,4], LLE, which reduces the high dimensionality by mapping the input data into a low-dimensional manifold, where data become more separable. For a given dataset $X = \{\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_N}\}$ in a d-dimensional space R^d, the LLE algorithm seeks an l-dimensional dataset Y in R^l, which has the same local geometry structure in its k-Nearest-Neighbor graph (*kNN*) as X does. In other words, any point $\mathbf{x} \in X$ is mapped to a point $\mathbf{y} = F(\mathbf{x}) \in Y$, such that,

0-7803-9737-1/06/$20.00 ©2006 IEEE

if **x** is linearly spanned by its *k* nearest neighbors $X_{kNN} = \{\mathbf{x}_j \mid 1 \le j \le k\}$

$$\mathbf{x} = \sum_{j=1}^{k} w_j \mathbf{x}_j \qquad (1)$$

then

$$\mathbf{y} = \sum_{j=1}^{k} w_j \mathbf{y}_j \qquad (2)$$

where w = (w1,…, wk) represents the coefficients of linear combination and $y_j = F(x_j)$. In practice, the LLE algorithm can be implemented in three steps: construct k-Nearest-Neighbor graph for X, estimate a weight matrix W for X, and extract the low-dimensional data Y, which are described as follows:

(1) Construct a k-NN graph $G_{kNN}(X)$ for X: for each $x_i \in X$, its k nearest neighbor is represented as X_{kNN} $(\boldsymbol{x}_i)$ $= \{\mathbf{x}_{\Gamma_{ij}} \mid 1 \le j \le k\}$.

(2) Estimate the weight matrix W such that x_i is best linearly spanned by X_{kNN} (x_i) as

$$\mathbf{W} = \underset{W}{\mathbf{argmin}} \sum_{i=1}^{N} \| \mathbf{x} - \sum_{j=1}^{k} W_{i\Gamma_{ij}} \mathbf{x}_{\Gamma_{ij}} \|^2 \qquad (3)$$

where, for any i, j, and $j \neq \Gamma_{ij}$, $W_{ij} = 0$ and

$$\sum_{j=1}^{k} W_{i\Gamma_{ij}} = 1 \qquad (4)$$

(3) Extract the embedding data Y by minimizing

$$\begin{aligned} \varepsilon(Y) &= \sum_{j=1}^{k} \| \mathbf{y}_i - W_{ij} \mathbf{y} \|^2 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} \mathbf{y}_i^T \mathbf{y}_i \end{aligned} \qquad (5)$$

where $\mathbf{M} = (\mathbf{I}\text{-}\mathbf{W})^T(\mathbf{I}\text{-}\mathbf{W})$ and **W** can be represented through sparse matrices. It can be proved [3,4] that the eigenvectors of **M** corresponding to the smallest nonzero eigenvalues are the resulting embedding data *Y*.

3. LLE-Based SMOTE

A common practice in the classification of an imbalanced data source is to oversample the minority class. Thanks to Chawla's [1] novel oversampling approach, the so-called SMOTE, in which the minority class is oversampled by using the *k-NN* graph instead of randomized sampling with replacement. Motivated by its successful application in handwritten character recognition [1], SMOTE has received a considerable interest in the pattern recognition community [2,5]. We denote the desirable number of synthetic data points created by SMOTE as *m*. The previous SMOTE algorithm [1] oversamples the minority class C_+. by using its *kNN* graph. First, for each of vector **x** in C_+, $m/|C_+|$ number of end points are randomly chosen from its *k*-nearest positive neighbors (i.e., the *kNN* in C_+). Then the synthetic data points are created through a randomized interpolation between **x** and the $m/|C_+|$ number of end points selected in $X_{kNN}(\mathbf{x})$, respectively, which is demonstrated in Fig.1.

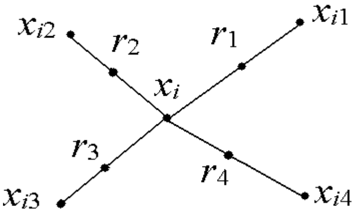

Figure 1. An illustration on how to create the synthetic data points in the SMOTE algorithm.

However, the randomized interpolation may incur as an additive noise for the original input data or violate the inherent geometrical structure of minority class and majority class, whereby the evaluation of the resulting classifiers becomes quite difficult. Instead of using the randomized interpolation scheme above for each **x**, we generate new synthetic data points by seeking the vector **r** on each line segment from **x** to each $\mathbf{x_j}$ in $X_{kNN}(\mathbf{x})$ such that it has the maximum average distance from the majority class C_- as in (6).

$$\mathbf{r} = \underset{\mathbf{r} \in \overline{\mathbf{x}\mathbf{x}}_j}{\mathbf{argmax}} \frac{1}{k} \sum_{\mathbf{x}_- \in C_-} \| \mathbf{r} - \mathbf{x}_- \| \qquad (6)$$

This intuitively allows for a good separation of synthetic data **r** from the majority class.

Even if the synthetic data can be interpolated deterministically according to (6), any oversampling of the minority class in the original input space is restricted by an assumption that the local space between any pair of positive data points is positive. However, this strict assumption is not always true when the original data is not linearly separable. In order to relax this assumption, the LLE technique can be applied to mapping the original data into a new linearly separable feature space. Then the SMOTE algorithm oversamples the minority class in the new feature space instead. A significant advantage of LLE over the other state-of-the-art learning algorithms is that a new synthetic vector z generated in the new feature space can be mapped back to the original input space according to

$$\mathbf{w} = \underset{\mathbf{w}}{\mathbf{argmin}} \sum_{i=1}^{N} \| \mathbf{z} - \sum_{j=1}^{k} w_j \mathbf{y}_j(\mathbf{z}) \|^2 \qquad (7)$$

```
Function LLE-BasedSMOTE($X$, $C_-$, $C_+$, $k$, $l$, $k_+$) Return $S$

Input:
    $X$: the original training set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $X \in R^d$
    $C_-$: majority class or negative class
    $C_+$: minority class or positive class
    $k$: the number of nearest neighbors for linear spanning in LLE
    $l$: the reduced dimension by LLE
    $k_+$: a threshold for choosing a negative neighbor in LLE

Output: $S$: the oversampled training set

Main Program:
$Y \leftarrow$ LLE ($X$, $C_-$, $C_+$, $k$, $k_+$, $l$)
Z $\leftarrow$ generates new positive vectors by using SMOTE algorithm over embedding set $Y$
$S \leftarrow X$
For each vector $\mathbf{z}$ of Z
    $\mathrm{Y}_{kNN}(\mathbf{z}) \leftarrow$ Find $\mathbf{z}$'s $k$ nearest neighbors in set $Y$
    $\mathbf{w}_\mathbf{z} \leftarrow$ Compute linear combination weights of $\mathrm{Y}_{kNN}(\mathbf{z})$ according to (7)
    $\mathbf{z}' \leftarrow$ Map $\mathbf{z}$ from $l$-dimensional embedding space back to the original input space according to (8)
    $S \leftarrow S \cup \{\mathbf{z}'\}$
End

Procedure LLE ($X$, $k$, $k_+$, $l$) Return $Y$
Initialize weight matrix $\mathbf{W}$
For each vector $\mathbf{x}$ in $X$
    $j \leftarrow 0$
    $X_{kNN}(\mathbf{v}) \leftarrow \phi$
    $Y \leftarrow \phi$
    $X^0_{kNN}(\mathbf{x}) \leftarrow$ Find $\mathbf{x}$'s $k$ number of nearest neighbors according to Euclidean distance
    For each $\mathbf{v} \in X^0_{kNN}(\mathbf{x}) \cap C_-$
        If $|X^0_{kNN}(\mathbf{v}) \cap C_+| \geq k_+$
            $X_{kNN}(\mathbf{v}) \leftarrow X_{kNN}(\mathbf{v}) \cup \{\mathbf{v}\}$
        End
    End
    $X_{kNN}(\mathbf{v}) \leftarrow$ Add $\mathbf{x}$'s $k$–$|X_{kNN}(\mathbf{v})|$ number of nearest positive neighbors to $X_{kNN}(\mathbf{v})$
    $\mathbf{w}_\mathbf{x} \leftarrow$ compute linear spanning weights of $T$ in terms of (3)
    $\mathbf{W} \leftarrow$ assign the row of $\mathbf{W}$ corresponding to $\mathbf{x}$ with $\mathbf{w}_\mathbf{x}$
End
$Y \leftarrow$ Compute the embedding data according to $\mathbf{W}$ in (5)
```

Figure 2. Pseudocodes of the LLE-based SMOTE algorithm

and

$$\mathbf{z}' = \sum_{j=1}^{k} w_j \mathbf{x}_j(\mathbf{z}) \tag{8}$$

where $\mathbf{y}_j(\mathbf{z})$ is $\mathbf{z}$'s k nearest neighbor in embedding set Y, and $\mathbf{x}_j(\mathbf{z})$ is the corresponding vector of $\mathbf{y}_j(\mathbf{z})$ in the original input space. The application of LLE completely fulfills the strict assumption required by the oversampling techniques, whereby any classifier can be designed in the original input space. The underlying LLE-based SMOTE algorithm is demonstrated in Fig.2.

In contrast to the LLE algorithm in the previous section, we present an alternative scheme for selecting k nearest neighbor vectors, which participate in the computation in (3) and (5). For each $\mathbf{x}$ in X, its nearest neighbors $X_{kNN}(\mathbf{x})$, is constructed by incorporating the information of two classes for X, i.e., the minority class C_+ and the majority class C_-, where $X = C_+ \cup C_-$. We first seek the k number of nearest neighbors for $\mathbf{x}$, $X^0_{kNN}(\mathbf{x})$ according to *Euclidean* distance and set $X_{kNN}(\mathbf{x})$ empty. Once $X^0_{kNN}(\mathbf{x})$ is constructed for each $\mathbf{x}$, for any negative vector $\mathbf{v}$ in $X^0_{kNN}(\mathbf{x})$, if the number of its positive neighbors in $X^0_{kNN}(\mathbf{v})$ is greater than k_+, we add $\mathbf{v}$ to $X_{kNN}(\mathbf{x})$. Finally, because the size of $X_{kNN}(\mathbf{x})$ is obviously less than k, the k-$|X_{kNN}(\mathbf{x})|$ number of nearest positive neighbors of $\mathbf{x}$ are added to $X_{kNN}(\mathbf{x})$. The implementation of this alternative LLE scheme is demonstrated in Fig.2.

4. Experimental Results

We evaluate the proposed LLE-based SMOTE algorithm by conducting leave-one-out validation tests [6] on three datasets and applying three classifiers: naïve Bayesian, K-NN classifier and support vector machine. As a comparison benchmark, the conventional SMOTE algorithm [1] is also evaluated in the experimental test. The three datasets are collected from several chest x-ray image databases in automatic computerized detection of pulmonary. Each data vector has 33 features extracted from a region of interest (ROI) that is located and segmented by a series of image enhancement and segmentation algorithms. The description of datasets is presented in Table 1.

Table1: Description of datasets

Datasets	Sample size	Size of minority class	Size of majority class
Set1	1180	44	1136
Set2	1530	44	1486
Set3	1164	47	1117

The ROC curve serves as a tool in evaluating classification performance obtained by using LLE-based SMOTE and SMOTE, of which the true positive rate as a function of false positive is plotted. In the principle of medical diagnosis, the larger the area below the resulting ROC curve is, the better the classification performance is attained.

The minority class is only oversampled as two times as large as its original size. The three parameters in Fig. 2 are defined as: k = 33, l = 7, and k_+=9. The classification results obtained by the three classifiers are reported respectively in Fig.3 and in Table 2.

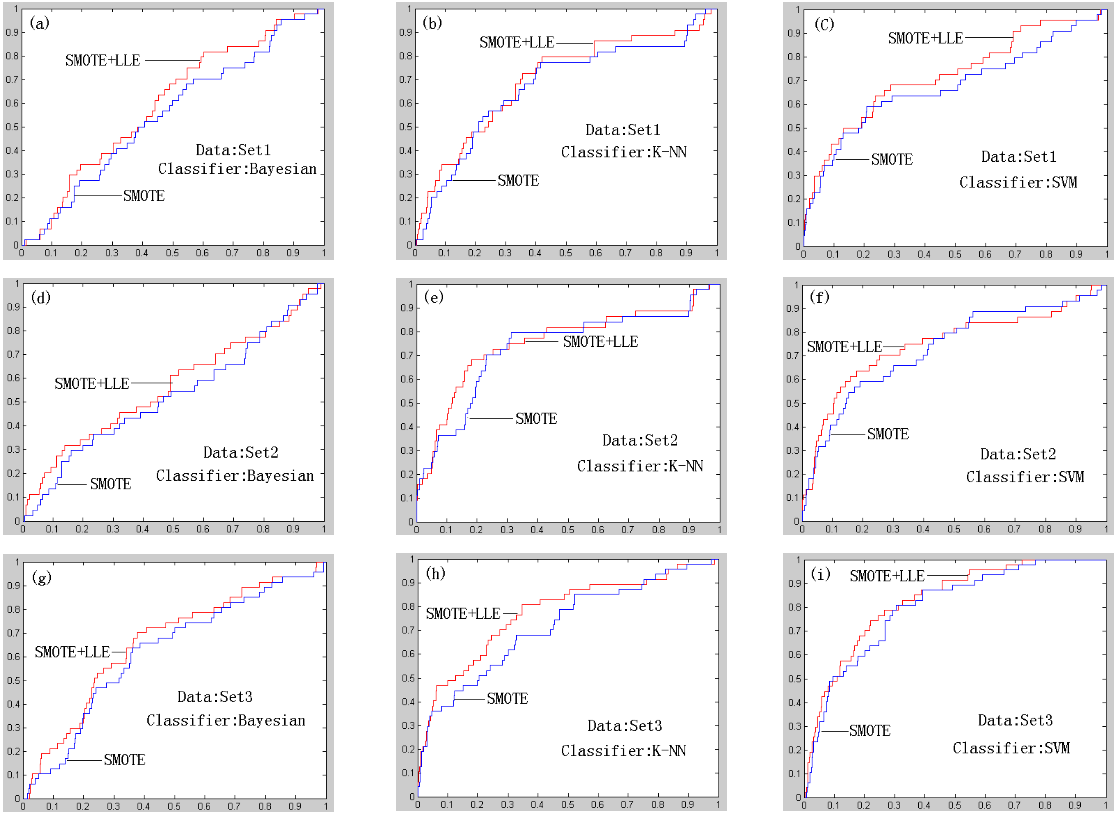

Figure 3. ROC curves obtained by the three classifiers: Naïve Bayesian classifier, k Nearest Neighbor classifier (K-NN) and Support Vector Machine.

Table 2: Areas of ROC curves obtained by the three classifiers by incorporating LLE-based SMOTE (SMOTE-II) and SMOTE.

Method		Data set		
		Set1	Set2	Set3
Bayesian	SMOTE	0.5530	0.5267	0.6202
	SMOTE-II	0.5987	0.5652	0.6597
K-NN	SMOTE	0.6703	0.7324	0.7189
	SMOTE-II	0.6964	0.7539	0.7731
SVM	SMOTE	0.6756	0.7299	0.8006
	SMOTE-II	0.7214	0.7477	0.8272

5. Conclusions

In this paper, we present an oversampling technique, LLE-based SMOTE, in classification of imbalanced data. The underlying oversampling algorithm is implemented by incorporating the LLE technique into the SMOTE algorithm. Experimental results demonstrate that the LLE-based SMOTE algorithm attains a performance superior to that of the conventional SMOTE.

Acknowledgements

This work is partially supported by the National Nature Science Foundation of China, under Award 30370420.

References

[1] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 2002, 16: 341-378.

[2] Akbani, R., Kwek, S., & Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets. ECML 2004: 39-50.

[3] Sam, T.R. & Lawrence, K.S. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 2000, 290(5500): 2323-2326.

[4] de Ridder, R., Loog, M., & Reinders, M.J.T.. Local Fisher Embedding. ICPR 2004, 2: 295-298.

[5] De-chuan, Z. & Zhi-hua, Z. Neighbor Line-based Locally linear Embedding. Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2006.

[6] Zhi-jie, X., Jie, Y. & Meng, W. A New Non-linear Dimensionality Reduction for Color Image. Journal of Shanghai Jiaotong University, 2005, 39(2): 279-283.