Hybrid Kernel Machine Ensemble for Imbalanced Data Sets

Peng Li   Kap Luk Chan   Wen Fang
Biomedical Engineering Research Center, Nanyang Technological University
Research Techno Plaza, 50 Nanyang Drive, XFrontiers Block, Singapore
Email: $lipeng@pmail.ntu.edu.sg$, $\{eklchan, fa0001en\}@ntu.edu.sg$

Abstract

A two-class imbalanced data problem (IDP) emerges when the data from majority class are compactly clustered and the data from minority class are scattered. Though a discriminative binary Support Vector Machine (SVM) can be trained by manually balancing the data, its performance is usually poor due to the inadequate representation of the minority class. A recognition-based one-class SVM can be trained using the data from the well-represented class only. However, it is not highly discriminative. Exploiting the complementary natures of the two types of SVMs in an ensemble can bring benefits from both worlds in addressing the IDP. Experimental results on both artificial and real benchmark data sets support the feasibility of our proposed method.

1. Introduction

The imbalanced data problem (IDP), also known as the class imbalance problem, has received considerable attention in recent years from the machine learning community [5]. In some imbalanced data sets, the class with large size of samples is compactly clustered and the class with small size of samples are scattered. For example, in patient monitoring, the morphologies of normal patient signals are similar to each other and the data can be easily collected. The signals corresponding to the abnormalities of the patients may exhibit various morphologies and are more difficult to collect compared to normal signals. Such a problem also exists in many other applications such object detection, network intrusion detection and information retrieval, etc. This kind of IDP can be addressed using a discriminative model, such as a Binary Support Vector Classifier (BSVC) [12] by manually balancing the data or compensating the class imbalance using different costs to the two classes. However, its performance is usually still poor due to the inadequately represented minority class. A recognition-based model such as a One-class Support Vector Classifier – νSVC [11], may do better than a discriminative model for such a problem by training the νSVC using the data from the well-represented class only. It avoids the problem caused by the inadequate representation of the minority class in BSVC. However, such a recognition-based model is not highly discriminative since the information from the minority class is left unused. Exploiting the complementary nature of such two different types of kernel machines, an ensemble constructed from them is expected to perform better than that of using either of them separately. Hence we propose to integrate these two Hybrid Kernel Machines into an Ensemble (HKME) to address this kind of IDP aforementioned. Trained using different data, these two kernel machines perform differently on this kind of imbalanced data sets. The nature of HKME is in-between the two-class classifier and one-class classifier. Hence the HKME can be regarded as a one-and-half classifier. The performance of the HKME is evaluated using an artificial data set and two real benchmark data sets.

2. Related Work

Some attempts have been reported to deal with the IDP, which can be classified into the following 3 approaches [5]. The first approach is re-sampling the training data set to make it balanced. This can be implemented either by undersampling in which the data from the majority class are down-sampled so that the size of the majority class dataset matches the size of the minority class dataset [5, 7], or by oversampling in which the data from minority class are over-sampled so that the size of minority class dataset matches the size of the majority class dataset [5]. There are also some attempts to combine these two approaches [2]. But the problem of undersampling is that some of the information may be lost if down-sampling is not conducted properly and the distribution of training data set is changed by re-sampling. So whether this is beneficial to classification remains unknown.

The second approach is to compensate for the class imbalance by altering the costs of the two classes in the training of classifiers. For example, using different penalty constants for different classes of data was used in BSVC in [9].

The third approach is to use recognition-based one-class classifiers instead of discrimination-based learning by leaving the data from one of the two classes totally unused (usually the minority class). The problem in one-class classification is different from those in conventional two-class classification where it is assumed that only information of one of the classes, **the target class**, is available and no information about the other class, **the outlier class**, is available. The task of one-class classification is to define a boundary around the target class, such that it accepts as much of the targets as possible and excluding the outliers as much as possible. For example, Japkowicz proposed to use an autoencoder to solve the IDP [5]. However, the recognition-based approach is usually outperformed by discrimination-based approach as a consequence of excluding the information from the minority class in the training of the model [9], except for seriously imbalanced data sets.

3. Proposed Method

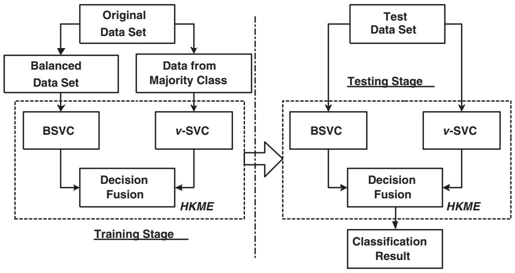

Figure 1. The flowchart of HKME.

The proposed HKME is illustrated in Figure 1, which consists of a $BSVC$ and a νSVC.

3.1. Discriminative BSVC

$BSVC$ is a discriminative classifier. Given a two-class (labelled by $y_i = \pm 1$) training set $X = \{x_i \in R^d | i = 1, 2, \cdots, N\}$ with N samples, the data are mapped to another feature space where the data can be separated by an optimal separating hyperplane expressed as

$$f(x) = \sum_{i=1}^{N} y_i \beta_i K(x_i, x) + b \tag{1}$$

where b is a bias item, β_is $(i = 1, 2, \cdots, N)$ are the solution of a quadratic programming problem that finds the maximum margin, $k(\cdot)$ is a kernel function. BSVCs have been increasingly used in many applications [12] and they have good generalization ability by finding an optimal separating hyperplane which minimizes the classification errors made on the training set while maximize the "margin" between different classes. But SVM also suffers from the IDP [1].

3.2. Recognition-based νSVC

νSVC is a kind of SVM [11] which can be used as a one-class classifier. It is an recognition-based model because only data from one-class is used in νSVC and no information about the other class is used in the training. Given a set of target data, they are mapped into a higher-dimensional space. The mapped target data are separated from the origin (corresponding to the outliers) with maximum margin using a hyperplane, which can be found by solving a quadratic programming problem [11]. The decision function corresponding to the hyperplane is similar to Equation 1. In IDP, the νSVC can be used to recognize the well-represented target data. But it is not highly discriminative since the data from the other class is totally unused.

3.3. Hybrid Kernel Machine Ensemble

In this framework, the HKME consists of two different base classifiers, a two-class BSVC and a one-class νSVC with Gaussian Radial Basis Function kernels. On one hand, the ν SVC can be trained using only the data for majority class, so it can avoid the problem of inadequate representation of the minority data but at the cost of discriminatory ability. On the other hand, a $BSVC$ can be trained using balanced data set using oversampling or undersampling. Since the νSVC and BSVC are trained using different data sets, the training sets of such two kernel machines can be considered diverse. Furthermore, the different nature of the two SVMs can further help to increase the diversity of such an ensemble. Since neither two-class BSVC nor one-class νSVC can solve the IDP well alone, exploiting the complementary nature of these two different types of models, a combination of them is expected to perform better than that of using either of them separately for the classification of this kind of imbalanced data set. Hence constructing a $HKME$ by integrating these two hybrid kernel machines in an ensemble is proposed to address this kind of IDP. This is the novelty of this proposal.

Several fusion rules are investigated for constructing the HKME for this kind of IDP, including Average (AVG), Decision Template (DET) and stacking [6, 8], etc. Let $C_i(x) = \{C_{i1}(x), C_{i2}(x), \cdots, C_{ik}(x)\}$ be a set of individual classifiers in an ensemble, each of which gets an input feature vector $\mathbf{x} = [x_1, x_2, \cdots, x_d]^T$ and assigns it to a class label y_i from $Y = \{-1, +1\}$, the goal is to find the a class label for $\mathbf{x}$ based on the posterior probability outputs of k classifiers $C_1(x), C_2(x), \cdots, C_k(x)$. As for SVM, the posterior probability can be estimated using a sigmoid function.

- Averaging: It calculates the average of the outputs of the k individual classifiers and assigns the input x the class with the largest posterior probability [6].

- Decision template: The decision template DET_j for class $y_j \in \{-1, +1\}$ is the average of the outputs of individual classifiers in the training set to class y_j [8]. The ensemble DET assigns the input x with the label given by the individual classifier whose Euclidean distance to the decision template DET_j is the smallest.

- Stacking: Taking the output of individual classifiers $C_i(x)$ as input of a upper layer classifier and the final decision is determined by the upper layer classifier. The upper layer classifiers used here include linear discriminant classifiers (LDCs) and quadratic discriminant classifiers (QDCs) assuming normally distributed classes [8].

4. Experimental Results and Discussions

The following experiments are conducted to evaluate the performance of our proposed HKME for the IDP aforementioned. A measure called Balanced Classification Rate (BCR) is used to evaluate the performance of HKME in this study. It is the algebraic mean of A_+ and A_-, $BCR = \frac{A_+ + A_-}{2}$, where A_+ and A_- denote the classification accuracy rate of positive class and negative class respectively. This measure has been used in evaluating the performance of classifiers in imbalanced data sets [4]. Only when both A_+ and A_- have large value can BCR have a large value. Therefore, the use of BCR can have a balanced assessment of the classifiers in this kind of imbalanced data sets as the BCR favors both lower false positives and false negatives.

4.1. Artificial Data Set

The first experiment was conducted using a checkerboard data set. The data are within a unit square in the two-dimensional space as shown in Figure 3. The majority class occupies the two diagonal squares of the checkerboard and the minority class uniformly occupies in a 2×2 square around the majority class. The data distribution is roughly in agreement with the assumption that our proposal is based upon. The proposed $HKME$ is compared with the other generally used methods to address the IDP, including oversampling, down-sampling, $SMOTE$ [2] and $BSVC$ using different costs to the two classes. The number of negative data was fixed as 256, the number of positive data were decreased so that the imbalance ratio is increased from $1:1$ to $32:1$. The number of test data consists of 1000 points from each class. The parameters of all the BSVCs are optimized using 3-fold cross validation. The parameters of the νSVC are optimized using artificially generated outlier data. The experiment was repeated 10 times and the average value of the BCRs by different schemes are reported in Figure 2 in which only AVG fusion rules was used.

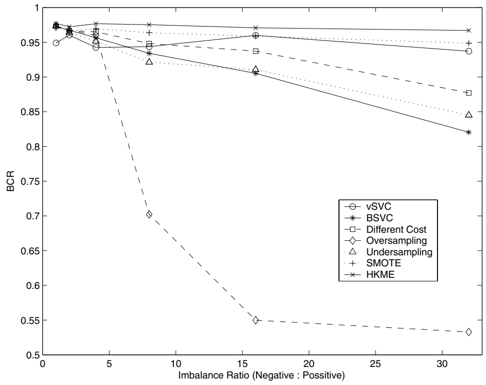

Figure 2. The result on checkerboard data set with different imbalance ratio.

It is observed from Figure 2 that $BSVC$ (trained using original data set) perform well when the imbalance ratio is not very high, but its performance deteriorates with the increase of imbalance ratio. $HKME$ using AVG rule performs the best among all the approaches. The $BSVC$ using different costs to two classes perform quite well compared original $BSVC$. Undersampling performs better than original $BSVC$, but is outperformed by using different costs. SMOTE performs reasonably well. Oversampling performs the worst among all the approaches due to overfitting.

The good performance of $HKME$ may come from the fact that it benefits from the strength of both of its individual classifiers, the discriminative $BSVC$ and recognition-based νSVC. This can be explained using their decision boundaries as illustrated in Figure 3. νSVC performs well due to its ability to model compactly clustered target class. But it has to reject some target samples to form a tighter boundary, so it tends to push the decision boundary towards the majority class. However, discriminative $BSVC$ tends to push the decision boundary toward the minority class. The ensemble of these two SVM tends to compensate these two different trends and strike a compromise. As shown in the figure, the decision boundary of $HKME$ is located in between two classifiers, which is closer to the ideal decision boundary (two squares in the checkerboard).

4.2. Real Benchmark Data Sets

In order to show the performance of the proposed $HKME$ on real data, the following experiments were conducted using 2 real data sets. One is Wisconsin Breast Can-

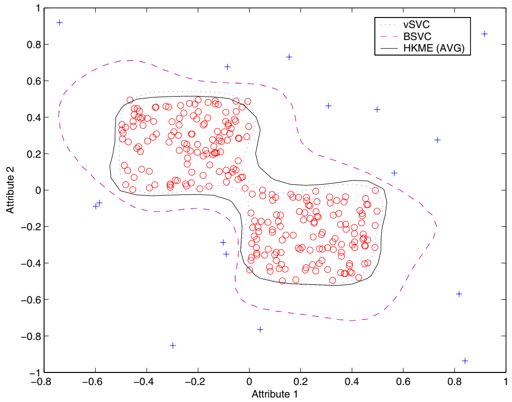

Figure 3. Comparison of the decision boundaries of νSVC, BSVC, and HKME.

Table 1. BCR (average ± standard deviation in %) achieved using (A) Breast Cancer (B) and Blood data set.

(A)

Imbalance Ratio	$1:10$	$1:30$	$1:50$
νSVC	94.3 ± 1.8	94.3 ± 1.8	94.3 ± 1.8
$BSVC$	93.1 ± 2.5	85.1 ± 3.2	85.1 ± 3.2
Different Costs	**95.6 ± 1.0**	92.2 ± 4.5	92.2 ± 4.5
Oversampling	50.2 ± 0.2	50.1 ± 0.2	50.1 ± 0.2
Undersampling	95.3 ± 1.5	92.2 ± 2.4	92.2 ± 2.4
SMOTE	88.0 ± 3.3	77.6 ± 6.1	77.6 ± 6.1
$HKME$ (AVG)	92.8 ± 1.0	90.8 ± 2.9	90.8 ± 2.9
$HKME$ (DET)	94.0 ± 1.5	93.6 ± 1.3	93.6 ± 1.3
$HKME$ (LDC)	93.2 ± 1.4	93.2 ± 1.5	93.2 ± 1.5
$HKME$ (QDC)	**95.1 ± 1.2**	**95.0 ± 1.2**	**95.0 ± 1.2**

(B)

Imbalance Ratio	$1:5$	$1:10$	$1:20$
νSVC	82.0 ± 9.8	77.5 ± 6.8	77.5 ± 6.8
$BSVC$	77.0 ± 10.6	71.5 ± 8.2	71.5 ± 8.2
Different Costs	**86.0 ± 9.7**	80.0 ± 12.2	80.0 ± 12.2
Oversampling	59.5 ± 12.3	52.0 ± 6.3	52.0 ± 6.3
Undersampling	82.0 ± 12.3	**84.5 ± 9.6**	**84.5 ± 9.6**
SMOTE	75.5 ± 10.4	72.0 ± 14.6	72.0 ± 14.6
$HKME$ (AVG)	**85.5 ± 12.3**	82.0 ± 10.1	82.0 ± 10.1
$HKME$ (DET)	**85.5 ± 8.6**	82.5 ± 8.2	82.5 ± 8.2
$HKME$ (LDC)	84.5 ± 8.3	**84.0 ± 8.4**	**84.0 ± 8.4**
$HKME$ (QDC)	83.5 ± 11.0	82.0 ± 7.9	82.0 ± 7.9

cer (Breast) from UCI database [10]. The other is Blood Disorder data set (Blood) from Biomed dataset in the Statlib data archive [3]. These data sets were splitted into training and test data sets randomly. The majority classes were used to train νSVC. The number of target data was fixed and the number of minority class was reduced to change the imbalance ratio. The experiments were repeated 10 times, the average results are reported in Table 1.

It is observed that all the $HKME$s performs well in these two data sets and show performance improvement over both νSVC and $BSVC$ and other schemes in all the cases, among which the LDC fusion rule performs the best. The reason may be that the distribution of the data in these data sets is roughly in agreement to the assumption in $HKME$. For example, in the Blood data set, the majority class is the observations made on normal healthy patients while the minority class is those that exhibiting abnormalities due to a rare genetic disease [3]. Hence the νSVC performs reasonably well. So is the HKME.

5. Conclusion

A novel hybrid kernel machine ensemble is proposed to address a kind of IDP in which the majority class is well represented while the minority class is inadequately represented by the training data. The generally used discriminative $BSVC$s suffer from the poor representation of the minority class. The recognition-based νSVCs can model the majority class well, but it is not highly discriminative due to the exclusion of the minority class in their training. The integration of such two different types of kernel machines can improve the classification over the use of either of them. Experimental results on both artificial and real benchmark data sets show the good performance of proposed method.

References

[1] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *ECML*, pages 39–50, 2004.

[2] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Artifical Intelligence Research*, (16):321–357, 2002.

[3] L. Cox, M. Johnson, and K. Kafadar. Exposition of statistical graphics technology. In *ASA Proceedings of the Statistical Computation Section*, pages 55–56, 1982.

[4] M. Gal-Or, J. H. May, and W. E. Spangler. Assessing the predictive accuracy of diversity measures with domain-dependent asymmetric misclassification costs. *Information Fusion Journal*, 6(1):3748, 2005.

[5] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–450, November 2002.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[7] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One sided selection. In *ICML*, pages 179–186, Nashville, Tennessee, 1997. Morgan Kaufmann.

[8] L. I. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001.

[9] B. Raskutti and A. Kowalczyk. Extreme re-balancing for SVMs: a case study. *SIGKDD Explor. Newsl.*, 6(1):60–69, 2004.

[10] C. B. S. Hettich and C. Merz. UCI repository of machine learning databases, 1998.

[11] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[12] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.