IEICE TRANS. INF. & SYST., VOL.E97–D, NO.12 DECEMBER 2014

LETTER

Balanced Neighborhood Classifiers for Imbalanced Data Sets

Shunzhi ZHU[†], **Ying MA**[†a)], **Weiwei PAN**[†], **Xiatian ZHU**[††], ***Nonmembers***, ***and*** **Guangchun LUO**[†††], ***Member***

SUMMARY A Balanced Neighborhood Classifier (BNEC) is proposed for class imbalanced data. This method is not only well positioned to capture the class distribution information, but also has the good merits of high-fitting-performance and simplicity. Experiments on both synthetic and real data sets show its effectiveness.

key words: *machine learning, class imbalance, class distribution, classification*

1. Introduction

The class imbalance problem is currently an active research subject and increasingly attracting attention in the machine learning and pattern recognition area, because this problem is common in many applications [1]. The data set is said to present a class imbalance when one of the classes (the minority one) is heavily under-represented in comparison to the other classes (the majority ones). This problem is particularly important, since this imbalance causes suboptimal classification performances, especially when the cost of misclassifying a minority-class example is substantial. Existing approaches to solving the class imbalance problem mainly include data level and algorithmic level methods. Here we focus on binary classification only and study the improved neighborhood classifiers methods at the algorithmic level. In this paper, we present an improved neighborhood classifier, Balanced Neighborhood classifier (BNEC), for the classification problem with class imbalanced data. When the training set is skewed, the popular K-nearest neighbor (KNN) classifier [2] and neighborhood classifier (NEC) [3]–[5] will mislabel instances in rare categories into common ones which degrades the classification performance. Ignoring class distribution information, the performance of both classifiers is weakened by the simple majority voting method. Considering the proximity and spatial distribution of neighbors of the neighborhood, we introduce a local mean distance for each class in the BNEC to make decision.

This method is not only well positioned to capture the class distribution information, but also inherit the merits of neighborhood classifier, which has been shown to be a powerful tool for attribute reduction, feature selection, rule extraction and reasoning with uncertainty. Since the data sets in real world applications are always class imbalanced, this method is more appropriate for recognizing the minority samples. Experiments on real and synthetic data sets show that our proposed method performs well in terms of the AUC metric.

Manuscript received April 3, 2014.
Manuscript revised July 27, 2014.

[†]The authors are with Xiamen University of Technology, Xiamen, China.
[††]The author is with Queen Mary University of London, London E1 4NS, UK.
[†††]The author is with University of Electronic Science and Technology of China, Chengdu, China.
a) E-mail: maying@xmut.edu.cn

DOI: 10.1587/transinf.2014EDL8064

2. Related Work

The KNN first introduced by Fix and Hodges [6], has high classification accuracy on data with unknown distributions and has wide applications. Therefore, it has recently been recognized as one of top 10 algorithms in data mining [2]. Let $L = \{x_n \in R^m\}_{n=1}^N$ be a training data set of given m-dimensional feature space, where N is the total number of training samples, and $y_n \in \{c_1, c_2, \ldots, c_M\}$ denotes the class label for x_n. This method predicts an instance x, by its K nearest neighbors in the training set, based on the majority rule as follows.

$$c(x) = \arg\max_{c \in C} \sum_{x_i \in X_{KNN}} I(y_i = c) \tag{1}$$

where $I(x = y) = 1$ when $x = y$; and zero otherwise. Assume that the k-th nearest neighbor is x_K, then

$$X_{KNN} = \{x_j | x_j \in L, d(x, x_j) \leqslant d(x, x_K)\} \tag{2}$$

When $K = 1$, KNN has the special form 1-NN rule, which is bounded by twice the Bayes error rate [7].

Recently, another classification technique based on local information, neighborhood classifier (NEC) has been proposed by Hu et al. [3]. They introduced the neighborhoods and neighborhood relations concepts in topology to build a uniform theoretic framework for neighborhood-based classifiers. They showed the NEC outperforms KNN algorithm. The NEC is proposed for pattern classification as follows.

$$c(x) = \arg\max_{c \in C} \sum_{x_i \in X_{NEC}} I(y_i = c) \tag{3}$$

$$X_{NEC} = \{x_j | x_j \in L, d(x, x_j) \leqslant \delta)\} \tag{4}$$

Where $\delta = d_1 + w*(d_n - d_1)$, d_n is the maximum distance, and d_1 is the minimum distance. The threshold δ varies with the value w, which is dynamically assigned based on the local and global information around x.

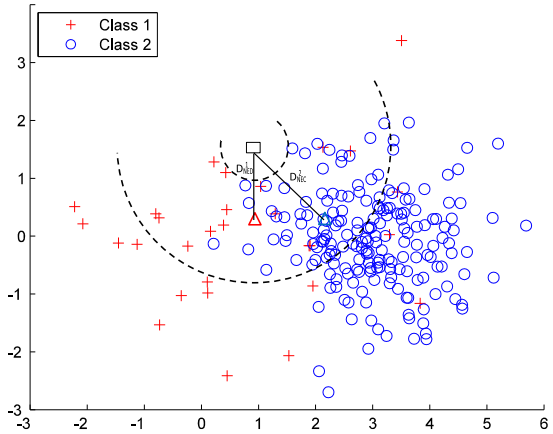


Fig. 1 An example of two-dimensional distribution. The class of the given query pattern x (square) is classified to class 1 using BNEC correctly. The red and blue triangle denote the vectors corresponding to the local mean distance of class 1 and 2, respectively. Note that, the given query pattern x is classified to class 2, using KNN and NEC.

3. Handling Imbalance Data

Since the number of the majority is much more than the minority in class imbalanced data sets, we do not use the number of the neighbors of each class to classify the query pattern. We propose a Balanced Neighborhood Classifier (BNEC), using the mean distances of the classes to handle class imbalance problem in NEC, as shown in Fig. 1. Firstly, we collect X^i_{NEC}, the neighbors of x in class c_i.

$$X^i_{NEC} = \{x_j | x_j \in L^i, d(x, x_j) \leqslant \delta^i)\} \tag{5}$$

Where $L^i = \{x_{ij} \in R_m\}_{j=1}^{N_i}$ denotes all the samples of class c_i, and N_i denotes the number of samples in subset L^i. Comparing with NEC, we should calculate threshold δ^i for each class in BNEC as follows.

$$\delta^i = d^i_1 + w * (d^i_n - d^i_1) \tag{6}$$

Where d^i_n and d^i_1 are the maximum distance and the minimum distance corresponding to class i respectively. Then, we calculate the local mean distance for each class.

$$D^i_{NEC} = \frac{1}{|X^i_{NEC}|} \sum_{x_j \in X^i_{NEC}} d(x, x_j) \tag{7}$$

Where $|X^i_{NEC}|$ is the number of samples in subset X^i_{NEC}. Finally, we assign x to the class c if the local mean distance of class c is minimum.

$$c(x) = \arg\min \ D^i_{NEC} \tag{8}$$

Algorithm 1 presents the pseudo-code of the BNEC classifier.

4. Experiments

In order to investigate the performances of our BNEC algorithm, we compare it with KNN ($K = 10$ as [3]) and NEC.

Algorithm 1 Balanced Neighborhood classifier (BNEC).

Require:
The set of labelled samples, L;
The set of unlabelled samples, U;

Ensure:
1: Calculate the distance of the test instance $x \in U$ from all training instances $x_j \in L$ using selected distance metric function, $d_t(x_j, x) = (\sum_{i=1}^m |f(x_j, a_i) - f(x, a_i)|^t)^{1/t}$;
2: Compute maximum distance d^i_n and minimum distance d^i_1 of each class;
3: Collecting instances in the neighbourhood of x using Eq. (5);
4: Compute threshold δ^i for each class using Eq. (6);
5: Calculate the local mean distances of each class using Eq. (7);
6: Assign x to the class c using Eq. (8);

$f(x, a_i)$ denotes the value of i-th attribute a_i of the sample x. Three metric functions d_1, d_2, d_∞ are used here.

$$d_1(x_j, x) = \sum_{i=1}^{m} |f(x_j, a_i) - f(x, a_i)| \tag{9}$$

$$d_2(x_j, x) = \left(\sum_{i=1}^{m} |f(x_j, a_i) - f(x, a_i)|^2 \right)^{1/2} \tag{10}$$

$$d_\infty(x_j, x) = \max_{i=1}^{m} (|f(x_j, a_i) - f(x, a_i)|) \tag{11}$$

4.1 Data Set

Since the number of the available samples of each class and the the number of dimensions can be easily controlled, the artificial data sets, Ness [8] data sets are used in our experiments. These synthetic data sets are follow normal distributions (p-dimension), which are widely used model for the pattern recognition.

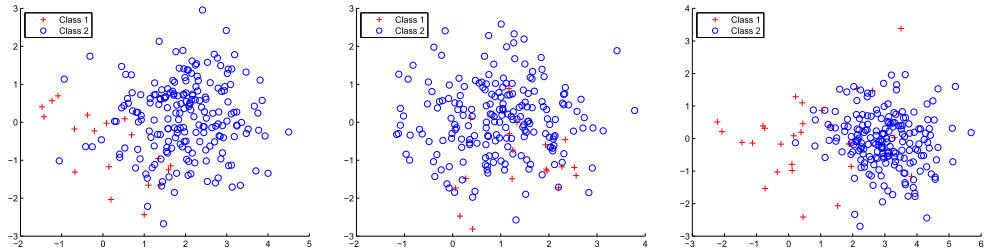
$$u_1 = 0, u_2 = [\theta/2, 0, \ldots, 0, \theta/2]' \tag{12}$$

$$\Sigma_1 = I_p, \Sigma_2 = \begin{bmatrix} I_{p/2} & 0 \\ 0 & \frac{1}{2} I_{p/2} \end{bmatrix} \tag{13}$$

where I_p denotes the $p \times p$ identity matrix, u_i is the mean vector and Σ_i is the covariance matrix from class c_i. The values of p and θ can be controlled, and θ is set to be 2, 4 and 6 in our experiments, as shown in Fig. 2. The imbalance radio is set to be 0.1. The real benchmark data sets in our experiments come from UCI repository [9], which are highly class imbalanced data. Imbalance ratio is the size of minority class divided by that of majority class. Here, each class is used as minority class, and all others are used as majority class. **Movement_libras** contains 90 attributes, 15 classes of 24 instances each. **Hayes_Roth** includes 160 instances, which are described by 9 attributes. **Hepatitis** contains 20 attributes, 155 instances. **Spectrometer** contains 102 attributes, 531 instances. The imbalance ratio of these data sets are 7.14%, 8.4%, 26%, and 9.26%, respectively.

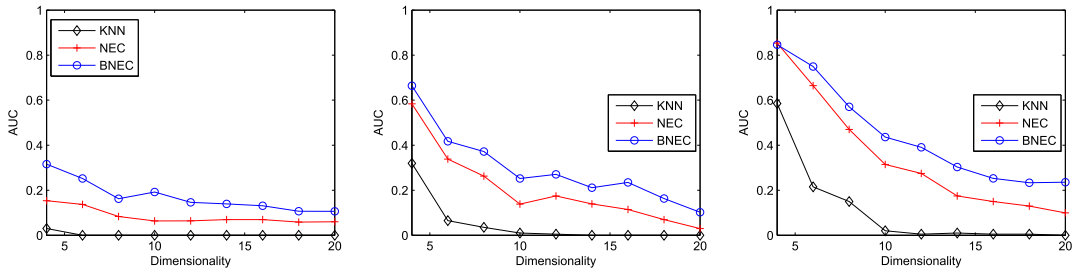
4.2 Results

In order to obtain reliable classification, the trials on each



(a) $\theta = 2$ (b) $\theta = 4$ (c) $\theta = 6$

Fig. 2 The examples of Ness data sets.



(a) $\theta = 2$ (b) $\theta = 4$ (c) $\theta = 6$

Fig. 3 The AUC values of KNN, NEC, and BNEC with increasing the dimensionality on each synthetic data set.

Table 1 Comparison of AUC performances based on KNN, NEC and BNEC.

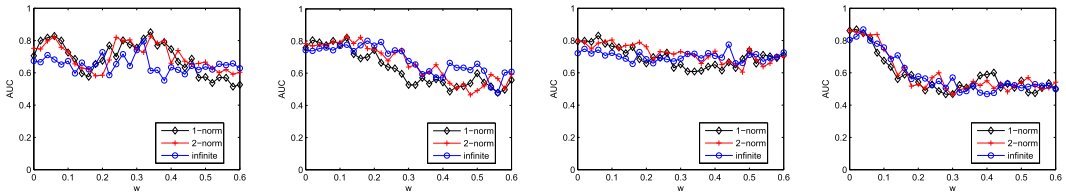
distance	data set	KNN (K=10)	NEC	BNEC
1-norm	Movement_libras	0.3013 ± 0.0415	0.6701 ± 0.0538	**0.8021** ± 0.0780
	Hayes_Roth	0.6544 ± 0.0351	0.7055 ± 0.0514	**0.7919** ± 0.0762
	Hepatitis	0.7655 ± 0.0325	0.7815 ± 0.0428	**0.8113** ± 0.0429
	Spectrometer	0.8013 ± 0.0368	0.7813 ± 0.0448	**0.8312** ± 0.0558
2-norm	Movement_libras	0.2331 ± 0.0735	0.7012 ± 0.0544	**0.7551** ± 0.0853
	Hayes_Roth	0.5810 ± 0.0451	0.7443 ± 0.0561	**0.7787** ± 0.0550
	Hepatitis	0.7623 ± 0.0550	0.7722 ± 0.0344	**0.7899** ± 0.0316
	Spectrometer	0.8223 ± 0.0421	0.7801 ± 0.0471	**0.8403** ± 0.0417
Infinite-norm	Movement_libras	0.2482 ± 0.0769	0.5473 ± 0.0521	**0.6623** ± 0.0519
	Hayes_Roth	0.6059 ± 0.0332	0.7325 ± 0.0454	**0.7643** ± 0.0492
	Hepatitis	0.7219 ± 0.0291	0.7006 ± 0.0347	0.6901 ± 0.0681
	Spectrometer	0.7692 ± 0.0318	0.8064 ± 0.0417	**0.8202** ± 0.0632

model are performed 10×5-fold cross validation[†]. We calculate the AUC values for KNN, NEC, and BNEC, with increasing the dimensionality on each synthetic data set, as shown in Fig. 3. It is interesting to note that the AUC values of all classifiers always monotonically decrease as the dimensionality increases, and the better performance of each method is usually obtain at small dimensional value.

Table 1 shows the results ($w = 0.1$ and $\delta = 0.125$ as in [3]) of the average AUC performances based on KNN, NEC, and BNEC with different norms in 1-norm, 2-norm and infinite-norm, respectively. Considering the three metric functions, BNEC has advantages of the stable performance. Especially for 1-norm, the AUC values of the BNEC with 1-norm metric functions are above 0.8, which is better than that of other two methods on all data sets. Movement_libras is high-dimensional and has relatively small sample sizes. The complexity of typical pattern discovery methods makes this problem challenging. This implies that the problem is that sample size of each class is lower than the feature size. The KNN has bad performance on these data sets. Even on this data set, BNEC has good performances.

In order to conduct more in-depth investigation on the performance of our method, we also change the size of the neighborhood in the experiment to find the optimal parameter w. In Fig. 4, all AUC values of the performances are recorded when the parameter w is gradually increased from 0 to 0.6 with step 0.02. Here we can find that there are similar trends in these four curves: Good generalization performance of the proposed parameter selection is less than 0.1. We should choose suitable parameters to train, since it shows a downward trend when the size of the neighborhood is increasing. Compared with the state-of-the-art classifiers for handling imbalanced data, we can see that BNEC is superior to Adaboosting [1], [10] and AdaCost [11] in the aspect of AUC, as shown in Fig. 5.

[†]http://en.wikipedia.org/wiki/Cross-validation_(statistics)



(a) Movement_libras (b) Hayes_Roth (c) Hepatitis (d) Spectrometer

Fig. 4 Performance curves of BNEC varying with w.

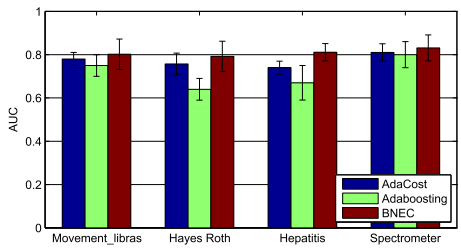


Fig. 5 The AUC results of AdaCost, Ababoosting, BNEC.

5. Conclusion

The proposed Balanced Neighborhood Classifier, which uses the samples in the neighborhood to estimate the local class probability density of the test samples, has the good merits of high-fitting-performance and simplicity. In order to well study the performance of the proposed classifier, experiments were carried out on the real and synthetic data sets, in comparisons with KNN and NEC. The comprehensive comparisons suggest that the proposed classifier has the following strengths: (a) It provides an alternative approach for pattern classification, especially for the class imbalanced data sets. (b) It is more robust than KNN and NEC. Consequently, we can draw a sound conclusion that the proposed classifier is a promising algorithm in the field of pattern classification.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61373147) and Xiamen university of technology high-level talents research funds (Grant Nos. YKJ13027R and E201411332). We thank the anonymous reviewers for their great helpful comments.

References

[1] H. He and E.A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol.21, no.9, pp.1263–1284, 2009.

[2] X. Wu, V. Kumar, J.R. Quinlan, et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol.14, no.1, pp.1–37, 2008.

[3] Q. Hu, D. Yu, and Z. Xie, "Neighborhood classifiers," Expert systems with applications, vol.34, no.2, pp.866–876, 2008.

[4] J. Zhang, T. Li, D. Ruan, and D. Liu, "Neighborhood rough sets for dynamic data mining," Int. J. Intelligent Systems, vol.27, no.4, pp.317–342, 2012.

[5] P. Zhu and Q. Hu, "Adaptive neighborhood granularity selection and combination based on margin distribution optimization," Inf. Sci., vol.249, pp.1–12, 2013.

[6] E. Fix and J.L. Hodges, "Discriminatory analysis, nonparametric discrimination," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, Feb. 1951.

[7] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol.13, no.1, pp.21–27, 1967.

[8] J. Van Ness, "On the dominance of non-parametric Bayes rule discriminant algorithms in high dimensions," Pattern Recognit., vol.12, no.6, pp.355–368, 1980.

[9] A. Frank and A. Asuncion, "UCI machine learning repository. 2010," http://archive.ics.uci.edu/ml

[10] C. Seiffert, T.M. Khoshgoftaar, and J. Van Hulse, "Improving software-quality predictions with data sampling and boosting," IEEE Trans. Syst. Man Cybern., A, Syst. Humans, vol.39, no.6, pp.1283–1294, 2009.

[11] Q. Yin, J. Zhang, C. Zhang, and S. Liu, "An empirical study on the performance of cost-sensitive boosting algorithms with different levels of class imbalance," Mathematical Problems in Engineering, vol.2013, Article ID 761814, 2013, doi:10.1155/2013/761814