2008

A supervised learning approach for imbalanced data sets

Giang H. Nguyen
University of Wollongong, giang_nguyen@uow.edu.au

Abdesselam Bouzerdoum
University of Wollongong, bouzer@uow.edu.au

Son Lam Phung
University of Wollongong, phung@uow.edu.au

A supervised learning approach for imbalanced data sets

Abstract

This paper presents a new learning approach for pattern classification applications involving imbalanced data sets. In this approach, a clustering technique is employed to resample the original training set into a smaller set of representative training exemplars, represented by weighted cluster centers and their target outputs. Based on the proposed learning approach, four training algorithms are derived for feed-forward neural networks. These algorithms are implemented and tested on three benchmark data sets. Experimental results show that with the proposed learning approach, it is possible to design networks to tackle the class imbalance problem, without compromising the overall classification performance.

Disciplines

Physical Sciences and Mathematics

Publication Details

G. Nguyen, A. Bouzerdoum & S. Lam. Phung, "A supervised learning approach for imbalanced data sets," in International Conference on Pattern Recognition, 2008, pp. 1-4.

A Supervised Learning Approach for Imbalanced Data Sets

Giang H. Nguyen, Abdesselam Bouzerdoum, and Son L. Phung
School of Electrical, Computer and Telecommunication Engineering,
University of Wollongong, Australia
E-mail: {giang, a.bouzerdom, phung}@uow.edu.au

Abstract

This paper presents a new learning approach for pattern classification applications involving imbalanced data sets. In this approach, a clustering technique is employed to resample the original training set into a smaller set of representative training exemplars, represented by weighted cluster centers and their target outputs. Based on the proposed learning approach, four training algorithms are derived for feed-forward neural networks. These algorithms are implemented and tested on three benchmark data sets. Experimental results show that with the proposed learning approach, it is possible to design networks to tackle the class imbalance problem, without compromising the overall classification performance.

1. Introduction

Although significant progress has been made in pattern classification, several issues still remain. A problem that we focus on this paper is how to learn a classification task from imbalanced data sets. The class imbalance problem, which is one of the fundamental problems in machine learning, has received much attention recently [1, 4, 9, 14, 15]. In many real-world diagnostic applications, e.g., computer security, biomedical, and engineering, uneven distribution of data patterns is very common, where number of training instances of a minority class is much smaller compared to other majority classes; as a result, the classifier tends to favor the majority class [7]. A study by Murphey et al. showed that the traditional feed-forward neural network has difficulty learning from imbalanced data sets [8]. Because of the overwhelming training instances of the majority class, the network tends to ignore the minority class and treat it as noise. In general, learning algorithms for class imbalance problems can be divided into two categories: resampling and cost-sensitive based. Resampling methods such as over-sampling and under-sampling [7, 14] modify the prior probability of the majority and minority class in the training set to obtain a more balanced number of instances in each class. The under-sampling method extracts a smaller set of majority instances while preserving all the minority instances. This method is suitable for large-scale applications where the number of majority samples is tremendous and lessening the training instances reduces the training time and makes the learning problem more tractable [15]. However, one problem associated with under-sampling techniques is that we may lose informative instances from the discarded instances.

In contrast to under-sampling, over-sampling method increases the number of minority instances by oversampling them. The advantage is that no information is lost from the training samples because all instances are employed [1]. However, the minority instances are over-represented in the training set and moreover, adding training instance means increasing training time. Weighting based methods are another type of over-sampling techniques, where more weights are assigned to the minority training instances [1, 4]. Cost-sensitive based methods are an alternative treatment of class imbalance problem. Berardi and Zhang introduced cost-sensitive neural networks by assigning different costs to errors in different classes [2]. Their method improves the classification accuracy for minority classes by assigning larger cost to them. However, adding a cost function in the learning process will modify the probability distribution [14].

In this paper, we introduce a new approach for supervised learning with imbalanced data sets. The concept of our method is similar to under-sampling. However, we employ unsupervised clustering to reduce the majority training instances, by selecting cluster centers as representative the samples. Furthermore, weighting of the minority and majority training exemplars is introduced in the cost function; the weights for each class serve as an approximation to its probability mass.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

The rest of the paper is organized as follows. Section 2 describes the proposed learning approach and derives four modified training algorithms based on gradient descent, gradient descent with momentum, resilient back-propagation and Levenberg Marquardt. Section 3 presents experimental results where the proposed approach is applied to different classification tasks. Finally, Section 4 presents concluding remarks.

2. Modified training algorithms for feedforward neural networks

Suppose that a multi-layer feed-forward neural network is to be trained using a set of M tuples $\{\mathbf{x}_m, \mathbf{d}_m\}$, where $m = 1, 2, ..., M$; $\mathbf{x}_m$ is the input vector and $\mathbf{d}_m$ is the corresponding output vector or desired vector. Let $\mathbf{w}$ be a vector consisting of all free network parameters, including weights and biases. The objective of supervised learning is to find a vector $\mathbf{w}^o$ that minimizes a cost function. A common objective function is the *mean square error* (MSE), defined as

$$E(\mathbf{w}) = \frac{1}{M \times N} \sum_{m=1}^{M} \sum_{i=1}^{N} \left(y_{im} - d_{im}\right)^2, \quad (1)$$

where N is the number of neurons in the output layer, and y_{im} is the network output. When number of training instances of different classes are uneven, the error contribution of each class to the objective function is unequal. In a two-class problem, the majority class has significant effect in the optimization process. Hence, we propose a more efficient algorithm for training feed-forward neural networks. In this approach, a pre-processing step is introduced to obtain a more balanced number of samples in each class. To this end, unsupervised *clustering* is applied to training samples of the majority classes to extract cluster centers that yield a compact representation of the majority classes.

Here, clustering is applied independently to all the training samples representing a particular class. Therefore, each cluster represents samples from a single class, and each class is represented by several clusters. In this approach, we deal with imbalanced data sets by simply assigning the same number of clusters to each class. After clustering, the data set is reduced to K exemplars, each is represented by a cluster *centroid* $\mathbf{c}_k$ and *size*. Here, the cluster size z_k is simply the number of training samples in the cluster. In the following, we present four training algorithms that integrate the cluster sizes and centroids into the learning rule.

2.1 Modified cost function

In the supervised learning stage, training samples are replaced by a set of cluster centroids which is then presented to the network along with the target outputs. To compensate information lost during the clustering process, weights for each class are introduced in the cost function:

$$E_p(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} p_k \left(y_{ik} - d_{ik}\right)^2, \quad (2)$$

where d_{ik} is the i-th element of the target or desired output vector $\mathbf{d}_k$, and p_k is the cluster weight which represents an approximation to the probability mass,

$$p_k = \frac{z_k}{\sum_{i=1}^{N_{cl}} \omega_i \, \gamma_{ki}}, \quad (3)$$

where N_{cl} is number of classes in the training set, ω_i is the size of class i, and γ_{ki} is the degree of membership of cluster k in class i:

$$\gamma_{ki} = \begin{cases} 1, & \text{if } \mathbf{c}_k \in \text{class } i \\ 0, & \text{otherwise.} \end{cases}$$

2.2 Modified training algorithms

Numerous optimization algorithms for minimizing E can be derived to train feed-forward neural networks. In this paper, we implemented four algorithms, namely gradient descent (GD), gradient descent with momentum and variable learning rate (GDMV), resilient back-propagation (RPROP), and Levenberg-Marquardt (LM) based on our propose approach. Each training algorithm updates network weights and biases according to $\mathbf{w}(t+1) = \mathbf{w}(t) + \Delta\mathbf{w}(t)$. Because details of the standard algorithms can be found in [13, 6, 12, 5], we only summarize their main characteristics herein.

- Gradient descent: weights are updated along the negative gradient $\Delta\mathbf{w}(t) = -\alpha \nabla E_p(t)$, where α is scalar learning rate, $\alpha > 0$.
- GD with momentum and variable learning rate: weight update is a linear combination of gradient and previous weight update

  $$\Delta\mathbf{w}(t) = \lambda \, \Delta\mathbf{w}(t-1) - (1-\lambda) \, \alpha(t) \, \nabla E_p(t),$$

  where λ is momentum parameter, $0 < \lambda < 1$, and $\alpha(t)$ is the adaptive scalar learning rate.
- Resilient back-probagation: weight update depends only on the sign of gradient

  $$\Delta w_i(t) = -\text{sign}\{\frac{\partial E_p}{\partial w_i}(t)\} \times \Delta_i(t),$$

  where $\Delta_i(t)$ is adaptive step specific to weight w_i.

Table 1. Comparison of standard and modified algorithms on benchmark data sets

(a) Liver disorder data set

	Classification rate		g-mean		K-coefficient	
Method	Standard	Modified	Standard	Modified	Standard	Modified
GD	75.65	75.94	71.37	74.33	0.472	0.503
GDMV	78.84	79.13	74.11	77.71	0.545	0.569
RPROP	78.26	79.71	75.97	77.87	0.543	0.576
LM	80.00	80.87	76.75	78.50	0.577	0.595

(b) Hepatitis data set

	Classification rate		g-mean		K-coefficient	
Method	Standard	Modified	Standard	Modified	Standard	Modified
GD	94.84	95.48	91.39	93.31	0.848	0.874
GDMV	95.48	96.13	91.82	96.18	0.866	0.872
RPROP	95.48	96.77	94.90	97.03	0.878	0.919
LM	94.19	94.19	89.87	91.61	0.829	0.852

(c) Pima Indian Diabetes data set

	Classification rate		g-mean		K-coefficient	
Method	Standard	Modified	Standard	Modified	Standard	Modified
GD	81.38	81.51	77.86	78.83	0.562	0.569
GDMV	81.38	81.38	77.74	78.57	0.572	0.588
RPROP	80.99	81.51	78.64	78.12	0.575	0.585
LM	81.64	82.29	79.23	79.18	0.585	0.604

- Levenberg-Marquardt: the weight update rule is given by

$$\Delta \mathbf{w}(t) = [J^T \mathcal{P}\, J + \mu I]^{-1} \nabla E_p,$$

where $\mathcal{P}$ is the expanded cluster weight matrix. Refer to [11] for details of computing $\mathcal{P}$.

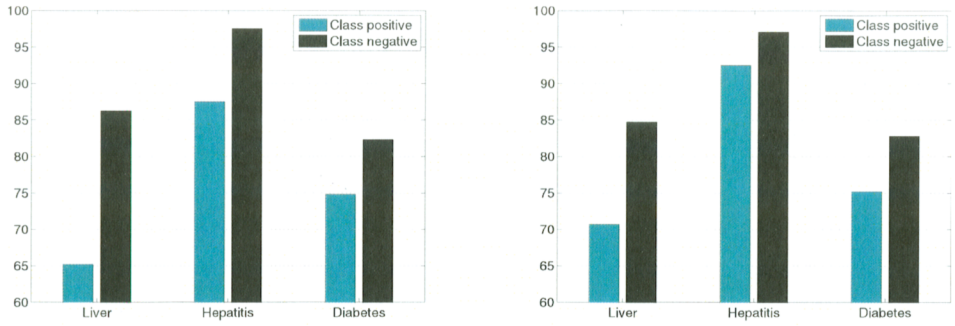
3. Experiments and Analysis

In this section, we apply the proposed learning approach to three benchmark problems, taken from UCI database repository [10]. The benchmarks used were the liver disorder, hepatitis and Pima Indian diabetes data sets. Details of these data sets are summarized in Table 2. Our aim is to study the generalization capability of the proposed approach, compared to the standard approach for neural networks training. The comparison is based on a five-fold cross validation in the classification tasks. For each fold, data set is partitioned into 60% as training set, 20% as validation set and 20% as test set. Several networks are trained and the best performing network on the validation set is selected for testing; its performance is evaluated on the test set. The average classification rate on the test set, over the five folds, is used as an estimate of generalization performances. Since overall classification rate is not most suitable tool for imbalanced data, other means of measuring the generalization performances are also performed including the geometric mean and Kappa coefficient [3].

Table 2. A brief summary of data sets used in the experiments

Data sets	Size	Attribute	Class distribution
Liver	345	6	145/200
Hepatitis	155	19	32/123
Diabetes	768	8	268/500

The comparison results of different training algorithms over all data sets are shown in Table 1. The modified training algorithms and the standard training algorithms achieve almost similar classification rates. For examples, in the hepatitis data set, CRs of RPROP and Mod-RPROP are 95.48% and 96.77%, respectively. However, the modified algorithms have higher values of g-mean and the Kappa coefficient than their counterparts. In the hepatitis data set, g-mean values of RPROP and Mod-RPROP are 94.90% and 97.03% and the Kappa coefficient of RPROP and Mod-RPROP are 0.878 and 0.919, respectively. This finding suggests that the modified training algorithms exhibit good classification rates in all classes. Figure 1 shows the average



(a) Standard training algorithms

(b) Modified training algorithms

Figure 1. The average classification rates of each class over all training algorithms

CRs by class over four training algorithms. The classification rates of positive class is increased, for example, 5.25% improvement in liver data set, 5% in the hepatitis data set, and 0.38% improvement in the diabetes data set.

4. Conclusions

In this paper, a new training approach for feedforward neural networks on imbalanced data sets that combines unsupervised clustering and supervised learning has been presented. The proposed approach can be applied to existing training algorithms. Experimental results show that the proposed approach can effectively improve the classification accuracy of minority classes while maintaining the overall classification performance.

References

[1] R. Alejo, V. Garcia, J. M. Sotoca, R. A. Mollineda, and J. S. Senchez. *Improving the performance of the RBF neural networks trained with imbalanced samples*, volume 4507 of *Lecture Notes in Computer Science*, pages 162–169. Springer-Verlag Berlin Heidelberg, 2007.

[2] V. L. Berardi and G. P. Zhang. The effect of misclassification costs on neural network classifiers. *Decision Sciences*, 30(3):659–683, 1999.

[3] R. Congalton, R. Oderwald, and R. Mead. Assessing landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering and Remote Sensing*, 49:1671–1678, 1983.

[4] X. Fu, L. Wang, K. S. Chua, and F. Chu. Training rbf neural networks on unbalanced data. In *Neural Information Processing*, volume 2, pages 1016– 1020, Inst. of High Performance Comput., Singapore, 2002.

[5] M. Hagan and M. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5:989–993, 1994.

[6] M. T. Hagan, H. B. Demuth, and M. H. Beale. *Neural network design*. PWS Publishing, Boston, MA, 1996.

[7] Y. Lu, H. Guo, and L. Feldkamp. Robust neural learning from unbalanced data samples. In *IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence*, volume 3, pages 1816 – 1821, May 1998.

[8] Y. L. Murphey, H. Guo, and L. Feldkamp. Neural learning from imbalanced data. In 117-128, editor, *Applied Intelligence, special issue on Neural Networks and Applications*, volume 21, 2004.

[9] Y. L. Murphey, H. Wang, G. Ou, and L. Feldkamp. Oaho: an effective algorithm for multi-class learning from imbalanced data. In *IEEE International Joint Conference on Neural Networks*, pages 406–411, 2007.

[10] D. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[11] G. H. Nguyen, A. Bouzerdoum, and S. Phung. Efficient supervised learning with reduced training examplars. In *International Joint Conference on Neural Networks*, 2008.

[12] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE International Conference on Neural Networks*, volume 1, pages 586 – 591, 1993.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*, volume I, pages 318 – 362. Bradford Books, Cambridge, MA, 1986.

[14] K. Yoon and S. Kwek. A data reduction approach for resolving the imbalanced data issue in functional genomics. *Neural Comput and Applic*, 16:295–306, 2007.

[15] Z. Zhou and X. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):63–77, 2006.