Improving the Performance of the RBF Neural Networks Trained with Imbalanced Samples

R. Alejo[1,2], V. García[1,2], J.M. Sotoca[1], R.A. Mollineda[1], and J.S. Sánchez[1]

[1] Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana (Spain)
[2] Lab. de Reconocimiento de Patrones, Instituto Tecnológico de Toluca
Av. Tecnológico S/N, 52140, Metepec, (Mexico)

Abstract. Recently, the class imbalance problem in neural networks, is receiving growing attention in works of machine learning and data mining. This problem appears when the samples of some classes are much smaller than those in the other classes. The classes with small size can be ignored in the learning process and the convergence of these classes is very slow. This paper studies empirically the class imbalance problem in the context of the RBF neural network trained with backpropagation algorithm. We propose to introduce a cost function in the training process to compensate imbalance class and one strategy to reduce the impact of the cost function in the data probability distribution.

1 Introduction

Lately, the class imbalance problem, has been considered as a fundamental problem in machine learning and data mining [1]. The class imbalance problem (for two classes) appears when the samples of a class (the minority one — *class+*), are smaller than those in the other class (the majority one — *class-*) [2]. In neural networks based on gradient descent methods for two classes, the class imbalance problem is stated as follows: The $class-$ dominates the learning process and the other one can be ignored. As a result, the convergence for minority class is very slow [3].

Several approaches have been presented to alleviate the class imbalance problem. For example, in [4], the backpropagation algorithm was altered to speed up the convergence of the multilayer perceptron (MLP) trained with imbalanced datasets.

In the MLP neural network, over-sampling techniques (which replicates samples in the *class+*) or under-sampling (eliminates samples in the *class-*) have revealed a noticeable effect in this network [5]. However, under-sampling involves a loss of information which can be detrimental. Over-sampling modifies the dataset probability distribution and increases the training time.

In recent works [1], the class imbalance problem has been dealt with *cost sensitive* techniques [6]. In the MLP, these approaches consist mainly in the application of a cost function in the training phase or in the test phase. Nonetheless, add a cost function in the training phase causes changes in data probability distribution [7].

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 162–169, 2007.

In this paper, we deal with class imbalance problem in the context of the RBF neural network (RBFNN) trained with the backpropagation algorithm to improve the classification accuracy. We introduce a cost function in the training process to compensate imbalance class and one strategy to reduce the impact of the cost function in the data probability distribution.

The rest of the paper is organized as follows. A brief explanation of learning method of the RBFNN and two strategies are shown in section 2 . Methodological aspects and empirical results are discussed in section 3. Finally, the main conclusions and possible future research are outlined in section 4.

2 RBFNN and Class Imbalance Problem

An extremely powerful type of feedforward artificial neural network is the RBFNN, which differs strongly from the MLP in activation functions and how they are used. A RBFNN in its basic form, can be defined as

$$\mathbf{F}(\mathbf{x}) = \sum_{i=1}^{K} \mathbf{w}_i h_i(\mathbf{x}) + \mathbf{b} \; = \sum_{i=1}^{K} \mathbf{w}_i \, exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{\sigma_i^2}\right) + \mathbf{b} \tag{1}$$

where $\mathbf{w}_i$ are the weights of the network, h is activation function with center $\mathbf{c}_i$ and variance σ_i^2, $\mathbf{b}$ is bias term. We can obtain a simplified version of Eq. 1 as $F_p(\mathbf{x}) = \sum^K \mathbf{w}_{ip} h_i(\mathbf{x}) + w_{0p}$. In this form the standard mean square error (MSE) can be defined as $E = \frac{1}{N}\sum^N \sum^L (y_p^n - F_p^n)^2$. y_p^n is the target output and F_p^n is the current output. Considering $h_i(\mathbf{x})$ as a differentiable activation function, the parameters $U = \{w, c, \sigma\}$ can be obtained simultaneously by an optimization procedure at backpropagation algorithm style in MLP.

Empirical studies performed with the backpropagation algorithm [3] show that imbalance problem is given by the contribution to MSE, from *class*+ in relation to *class*−, where the most contribution to MSE is produced by the *class*−. Therefore, the training process is dominated by the *class*−.

Suppose the specific case of two classes ($M = 2$), where N is the total training samples and n_m is the number of samples of each class with $N = \sum_{m=1}^{M} n_m$ and $n_1 << n_2$. Under these conditions, it is possible to express the MSE by class as $E_m(U) = \frac{1}{N}\sum_{n=1}^{n_m}\sum_{p=1}^{L}(y_p^n - F_p^n)^2$, and $E(U) = \sum^m E_m(U) = E_1(U) + E_2(U)$. If $E_1(U) << E_2(U)$, then $E(U) \approx E_2(U)$. Therefore, $-\nabla E(U)$ is not always the best direction to minimize the error of both classes [3].

Considering the contributions to $E(U)$ from both classes, the following cost function γ is introduced in $E(U)$ to compensate the imbalance of the classes.

$$E(U) = \frac{1}{N}\sum_{m=1}^{M}\gamma(m)\sum_{n=1}^{n_m}\sum_{p=1}^{L}(y_p^n - F_p^n)^2 \; ; \; \gamma(m) = n_2/n_m; \; m = 1, ..., M \; , \tag{2}$$

where $\|\nabla E_1(U)\| \approx \|\nabla E_2(U)\|$.

Empirical results illustrate that the $\gamma(m)$ function avoids that the minority class can be ignored in the training process. Nonetheless, theoretically it is possible to show that introducing the cost function in the training process appear changes in probability distribution. In some situations, it can be an appropriate strategy to deal with the imbalance problem [7]. We call this method *strategy I*. To reduce the impact of the cost function in the data probability distribution, we modify gradually the cost function in the training process diminishing its value (*strategy II*). The new cost function is defined as

$$\gamma(m)^t = \begin{cases} \gamma(m)^{(t-1)} \cdot (1-\varepsilon) & \text{if } [\gamma(m)^{t-1}] > 1 \\ 1 & \text{in the other case} \end{cases} \tag{3}$$

where t is the current iteration and $\varepsilon = 0, ..., 1$. In the first iterations of the training network, the imbalance problem is alleviated such in the last iterations the network acts as the standard MSE criterion.

3 Experiments and Discussions

The RBFNN was trained with the backpropagation algorithm, using ten hidden neurons, and learning rate of 0.1. The shutdown criterion is reached when the error is smaller than 0.0001 or a maximum of 100000 iterations have been performed.

A general criterion to measure the classifier performance is the overall accuracy. Nevertheless, in the class imbalance problems is not the most suitable measure [2]. The geometric mean (g-mean) is one of the most widely accepted criterion, and is defined as $g = \sqrt{acc^+ \cdot acc^-}$, where acc^+ is the accuracy on the *class*+ and acc^- is accuracy on the *class*−. In this work, g-mean and overall accuracy were applied to measure classifier performance.

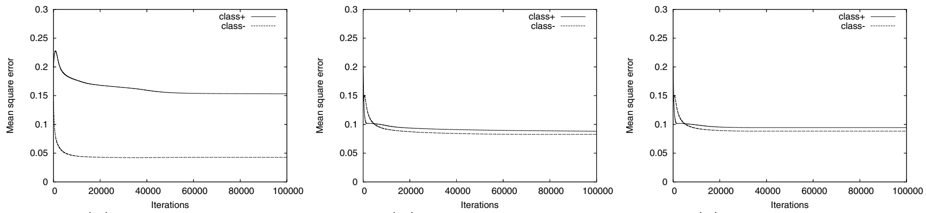
The experiments were carried out on the real datasets from UCI Database Repository (http:www.ics.uci.edu/∼mlearn) and one multispectral image (Feltwell dataset [8]). The size of the datasets and the some characteristics of the data are summarized in Table 1. A five-fold cross validation was employed in the classification tasks.

Table 1. A brief summary of the some basic characteristics of the databases

Dataset	Size	Attribute	Class	Class distribution
breast-w	698	9	2	241/457
diabetes	768	8	2	268/500
heart-s	270	13	2	120/150
liver	345	6	2	145/200
phoneme	5404	5	2	1586/3818
sonar	208	60	2	97/111
balance-scale	625	4	3	49/288/288
glass	214	9	6	70/76/17/13/9/29
Feltwell	10944	15	5	3531/2441/896/2295/1781

3.1 Behavior of the Standard and Modified Backpropagation

The MSE for *class*− and *class*+ of the training process is shown in Fig. 1 for standard backpropagation, strategies I and II.



(a) Standard (b) Strategy I (c) Strategy II

Fig. 1. MSE by classes during one training session with Phoneme dataset

In Fig. 1(a), we can observe as the MSE of the *class*− reduces rapidly in the first iterations and the MSE of the *class*+ increases considerably. Later, the MSE of the *class*+ decreases very slowly. When the strategy I is applied, the difference in the MSE of both classes is minor than in the standard backpropagation and the MSE of the *class*+ diminishes. So, the accuracy of the *class*+ goes up (see subsection 3.2), but the convergence of both classes is low. The best performance can be obtained finding a optimal value of the cost function for each class, but it is difficult task.

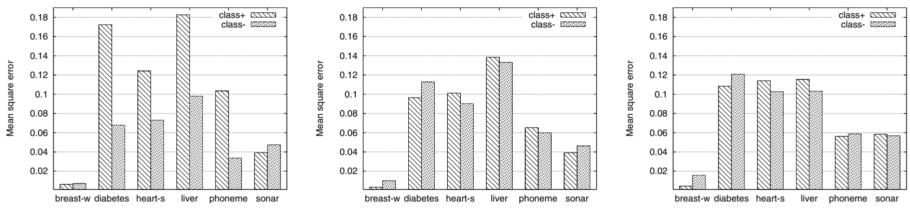
The strategy II, is proposed to avoid the change the data probability distribution (the main disadvantage of the strategy I). This strategy has a similar behavior that strategy I (see Fig. 1c). The key is the choice of an appropriate value ε and prevent abrupt changes in the cost function (see Eq. 3).

3.2 Two Class Problem

Fig. 2 illustrates the MSE by class, in the training of the standard and modified backpropagation. In diabetes, heart-s, liver and phoneme, the strategies can help that the training process is not dominated by the *class*−. Breast-w dataset is an interesting case where the imbalance does not affect significantly the network performance. Likely it happens because the class boundary is enough separated between class samples in the feature space.

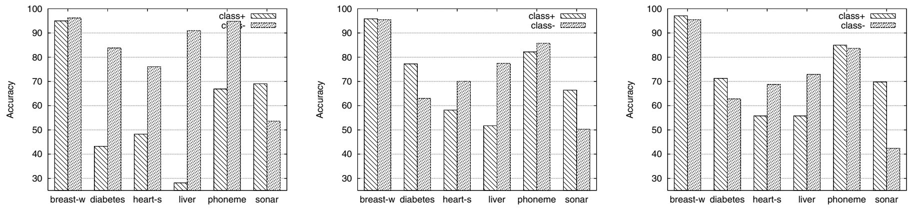
Strategy I shows a similar behavior than strategy II. It happens because the class imbalance is compensated while the network is trained. In opposition, strategy II reduces the impact of the cost function during the training process.

The variations in the MSE by class are smaller in modified backpropagation than in standard backpropagation. The evidence proves how the strategies get better the *class*+ accuracy (see Fig. 3). For example, in Liver dataset (classified by the standard backpropagation), the *class*+ accuracy is 28.13% while the *class*− accuracy is 91%. In contrast, the strategies try to equalize the accuracy



(a) Standard (b) Strategy I (c) Strategy II

Fig. 2. MSE by class for datasets with two classes



(a) Standard (b) Strategy I (c) Strategy II

Fig. 3. Classification accuracy of the minority and majority classes

by class. For liver dataset, the strategy I obtains a classification performances: 51.7%, 77.5%, and the strategy II: 55.8%, 72.9% for $class+$ and $class-$ accuracies respectively.

Diabetes, heart and phoneme datasets exhibits alike behavior than liver dataset (see Fig. 3). In sonar, the standard backpropagation get better accuracy by class than modified backpropagation. Likely this dataset is more affected by the noise or dimensionality of data than by class imbalance problem. The sonar dataset has the least level class imbalance (see Table 1).

Table 2 shows that strategies I and II have high values of g-mean in all the datasets, except in sonar. This observation suggests a good performance of the classifier in both classes when the backpropagation is modified. For example with liver dataset the strategy I improves the g-mean value in 15.54%, and the strategy II in 15.22%, with respect to the standard backpropagation (see Table 2).

In the overall accuracy context (Table 2), we can observe that there are not a substantial difference between these strategies and the standard backpropagation, i.e., the modified backpropagation do not reduce considerably the overall accuracy. In the worst case (diabetes dataset), the overall accuracy is reduced 3.9%. However, the $class+$ accuracy is increased at least in 28% (see Fig. 3).

Finally, strategy I has better overall accuracies in hearts and liver datasets and the strategy II in breast-w dataset.

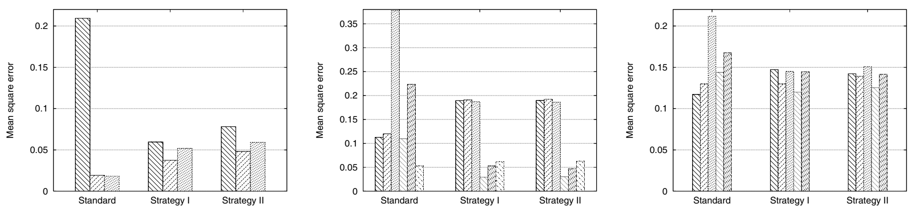
Table 2. Mean values of the g-mean and the accuracy of the classifier

	g-mean			accuracy		
Data set	Standard	Strategy I	Strategy II	Standard	Strategy I	Strategy II
breast-w	95.49	95.58	96.25	95.76	95.61	96.05
diabetes	58.43	69.70	66.10	69.66	67.97	65.76
heart-s	59.54	62.71	61.11	63.73	64.82	62.98
liver	47.42	62.96	62.64	64.59	66.65	65.70
phoneme	77.93	81.38	81.01	86.57	84.72	84.07
sonar	57.44	55.32	53.08	55.92	54.46	55.20

3.3 Multiclass Problem

To evaluate the strategies proposed in the multiclass context, three classification problems were applied (see Table 1).

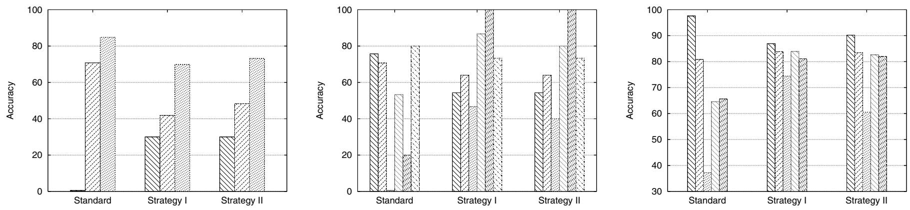
The Fig. 4(a) shows the MSE by class of the balance-scale dataset. It can observed that the MSE of class 1 by standard strategy is much greater that the other classes. This indicates that class 1 is ignored in the learning process. As a result, the accuracy class 1 is considerably affected (see Fig.5a). On the other hand, the strategies go up the accuracy of the class 1, but the accuracies of the other classes is minor. This has a negative impact in the overall accuracy. This measure reduces to 18.25% by strategy I, while strategy II the accuracy is 13.67% (see Table 3). The MSE behavior of the glass dataset is presented



(a) Balance dataset (b) Glass dataset (c) Feltwell dataset

Fig. 4. Behavior of the MSE by class in the training phase with multiclass datasets

in Fig. 4(b). For the classes with fewer samples in the TS (class 3, 4 and 5, see Table 1), the standard backpropagation presents high values of the MSE and low accuracies (specially in the class 3). When the strategies are applied to the network training, the MSE for the classes 3, 4 and 5 is reduced while the accuracy by class rise. It can seem how the MSE of the classes 1 and 2 is higher while their accuracies decrease (see Figure 5b). However, this reduction of the overall accuracy is not as important as presented in the balance-scale dataset. The strategy I reduces to 0.93% and the strategy II to 1.86% (see Table 3). In the case of the Feltwell dataset (see Fig. 4(c)), the MSE of the minority class, when the RBFNN is trained with standard backpropagation, is much higher than the MSE of other classes. This is reflected in the accuracy rate of the minority class



(a) Balance dataset (b) Glass dataset (c) Feltwell dataset

Fig. 5. Accuracy by class in the classification phase with multiclass datasets

(see Fig. 5(c)). The strategy I and II diminish the MSE of the minority class and improve the classification performance of the minority class. The strategies I and II exhibits a more stable behavior compared to the standard backpropagation. The accuracy by class is better except in class 1. In addition, the overall accuracy of the Feltwell dataset goes up to 6%. The values of g-mean shown in the Table 3 suggest a good performance of the classifier in multiclass context when the backpropagation is modified.

Table 3. Geometric mean (g-mean) and the accuracy for multiclass datasets

	g-mean			accuracy		
Data set	Standard	Strategy I	Strategy II	Standard	Strategy I	Strategy II
balance-scale	0.847	44.42	47.34	71.42	53.17	57.75
glass	5.98	68.47	65.85	65.12	64.19	63.26
Feltwell	65.95	81.95	79.05	76.80	83.65	83.33

4 Conclusions

In this paper, we deal with the class imbalance problem in the context of RBFNN. In the RBFNN trained with backpropagation, the classes with higher number of samples in the TS can dominate the training process and the classes with lower number of samples in the TS can not be learn. We include a cost function in the training to compensate the imbalance in the TS (strategy I). However, add a cost function in the training causes changes in data probability distribution. To reduce the impact of the cost function in the data probability distribution, we modify gradually the cost function until it does not have any influence (strategy II). Empirical results shows that the strategies proposed achieve a balance of the MSE and the accuracy by class. Nonetheless, when they balance the MSE by class, a tendency appear at increasing the overall MSE and decreasing the overall accuracy.

In conclusion, the strategies I and II offer evidence of their capability to increase the accuracy of the minority classes. Future investigations must be developed to study the class imbalance problem considering data probability distribution and its complexity.

Acknowledgments

This work has been supported in part by grants DPI2006-15542-C04-03 from the Spanish CICYT, P1–1B2004–08 from Fundació Caixa Castelló–Bancaixa, and SEP-2003-C02-44225 from the Mexican CONACyT.

References

1. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data. Engineering 18, 63–77 (2006)
2. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: Proc. 14th International Conference on Machine Learning, pp. 179–186. Morgan Kaufmann, Washington (1997)
3. Anand, R., Mehrotra, K., Mohan, C., Ranka, S.: An improved algorithm for neural network classification of imbalanced training sets. IEEE Transactions on Neural Networks 4, 962–969 (1993)
4. Bruzzone, L., Serpico, S.: Training of neural networks for classification of imbalanced remote-sensing data. In: IEEE Transactions on Geocience and Remote Sensing, pp. 1202–1204 (1997)
5. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intelligent Data. Analysis 6, 429–449 (2002)
6. Kukar, M., Kononenko, I.: Cost-sensitive learning with neural networks. In: 13th European Conference on Artificial Intelligence, pp. 445–449 (1998)
7. Lawrence, S., Burns, I., Back, A., Tsoi, A., Giles, C.L.: Neural network classification and unequal prior class probabilities. In: Orr, G., Müller, K.R., Caruana, R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 1524, pp. 299–314. Springer, Heidelberg (1998)
8. Serpico, S., Roli, F., Pellegretti, P., Vemazza, G.: Structured neural networks for the classification of multisensor remote-sensing images. In: Int. Geosci. Remote Sensing Symp., pp. 907–909 (1993)