A Preliminar Analysis of CO^{2}RBFN in Imbalanced Problems

M.D. Pérez-Godoy[1], A.J. Rivera[1], A. Fernández[2], M.J. del Jesus[1], and F. Herrera[2]

[1] Department of Computer Science, University of Jaén, Campus Las Lagunillas, 23071 Jaén, Spain
[2] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

Abstract. In many real classification problems the data are imbalanced, i.e., the number of instances for some classes are much higher than that of the other classes. Solving a classification task using such an imbalanced data-set is difficult due to the bias of the training towards the majority classes. The aim of this contribution is to analyse the performance of CO^{2}RBFN, a cooperative-competitive evolutionary model for the design of RBFNs applied to classification problems on imbalanced domains and to study the cooperation of a well known preprocessing method, the "Synthetic Minority Over-sampling Technique" (SMOTE) with our algorithm. The good performance of CO^{2}RBFN is shown through an experimental study carried out over a large collection of imbalanced data-sets.

Keywords: Neural Networks, Radial Basis Functions, Genetic Algorithms, Imbalanced Data-Sets.

1 Introduction

Radial Basis Function Networks (RBFNs) are one of the most important Artificial Neural Network (ANN) paradigms in the machine design field. An RBFN is a feed-forward ANN with a single layer of hidden units, called radial basis functions (RBFs) [5,19]. The overall efficiency of RBFNs has been proved in many areas such as pattern classification [6], function approximation [24], time series prediction [28] and multiple specific applications such as credit assessment [16], process control [14], medical diagnosis [17], and time series forecasting [27] among others. In most of these areas, the data-sets have a common and very usual characteristic: they are imbalanced data-sets [7].

The imbalance data-set problem occurs when the number of instances of one class overwhelms the others. To solve it there are two main types of solutions: solutions at the data level which is achieved balancing the class distribution and solutions at the algorithmic level, for example adjusting the cost per class.

An important paradigm for RBFN design is the Evolutionary Computation [12]. There are different proposals in this area with different scheme representations: Pittsburgh [13], where each individual is a whole RBFN, and cooperative-competitive [28], where an individual represents an RBF.

J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 57–64, 2009.

In this work we study the performance of CO^{2}RBFN [25] in the framework of imbalanced data-sets and analyse the cooperation of a well known preprocessing method, the "Synthetic Minority Over-sampling Technique" (SMOTE) [8] with our algorithm. To do so, in Section 2 a brief description of the imbalanced data-sets in classification and the solutions provided for them in the specialized bibliography is shown. The cooperative-competitive evolutionary model for the design of RBFNs applied to classification problems, CO^{2}RBNF, is described in Section 3. The analysis of the experiments and the conclusions are shown in Sections 4 and 5.

2 Imbalanced Data-Sets in Classification

The problem of imbalanced data-sets in classification [7] occurs when the number of instances of one class is much lower than the instances of the other classes. This problem is implicit in most real world applications including, but not limited to telecommunications, finance, biology and medicine.

In this work, we focus on binary imbalanced data-sets, where there is only one positive and one negative class. Furthermore, we use the imbalance ratio (IR) [22], defined as the ratio of the number of instances of the majority class and the minority class, to organize the different data-sets.

In classification, the class imbalance problem will cause a bias on the training of classifiers and will result in the lower sensitivity of detecting the minority class examples. For this reason, a large number of approaches have been previously proposed to deal with it. These approaches can be categorized into two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration [3,30] and external approaches that preprocess the data in order to diminish the effect caused by their class imbalance [4,10].

For RBFNs there are some proposals for the use of a cost function in the training process to compensate imbalance class and strategies to reduce the impact of the cost function in the data probability distribution [1]. With respect to external approaches, different studies have been done to assess the effect of the size of the training data-set on the accuracy [2]. In [20], the authors study three different ANN architectures, multilayered back propagation, RBFNs and Fuzzy ARTMAP, with the use of random oversampling and the snowball training methods. In [23], the over-sampling method SMOTE [8], has been used with RBFNs showing a good behaviour. According to this, we employ in this contribution the SMOTE algorithm in order to deal with the problem of imbalanced data-sets.

The SMOTE method (detailed in [8]) tries to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

With respect to the evaluation in imbalanced domains the most used empirical measure, accuracy, does not distinguish between the number of correct labels of

different classes, which in the framework of imbalanced problems may lead to erroneous conclusions. The metric used in this work is the geometric mean of the true rates [3], defined as

$$GM = \sqrt{\frac{TP}{TP+FN}\frac{\dot{TN}}{FP+TN}} \tag{1}$$

where TP,TN,FP and FN stand for True Positives, True Negatives, False Positives and False Negatives respectively. This metric attempts to maximize the accuracy of each one of the two classes with a good balance.

3 CO^{2}RBFN: An Evolutionary Cooperative-Competitive Hybrid Algorithm for RBFN Design

CO^{2}RBFN [25], is an evolutionary cooperative-competitive hybrid algorithm for the design of RBFNs. In this algorithm each individual of the population represents, with a real representation, an RBF and the entire population is responsible for the final solution. The individuals cooperate towards a definitive solution, but they must also compete for survival. In this environment, in which the solution depends on the behaviour of many components, the fitness of each individual is known as credit assignment. In order to measure the credit assignment of an individual, three factors have been proposed: the RBF contribution to the network output, the error in the basis function radius, and the degree of overlapping among RBFs.

The application of the operators is determined by a Fuzzy Rule-Based System. The inputs of this system are the three parameters used for credit assignment and the outputs are the operators' application probability.

The main steps of CO^{2}RBFN, explained in the following subsections, are shown in the pseudocode in Figure 1.

```
1. Initialize RBFN
2. Train RBFN
3. Evaluate RBFs
4. Apply operators to RBFs
5. Substitute the eliminated RBFs
6. Select the best RBFs
7. If the stop condition is not
     verified go to step 2
```

Fig. 1. Main steps of CO^{2}RBFN

RBFN initialization. To define the initial network a specified number m of neurons (i.e. the size of population) is randomly allocated among the different patterns of the training set. To do so, each RBF centre, $\boldsymbol{c}_i$, is randomly established to a pattern of the training set. The RBF widths, d_i, will be set to half

the average distance between the centres. Finally, the RBF weights, w_{ij}, are set to zero.

RBFN training. The Least Mean Square algorithm [29] has been used to calculate the RBF weights. This technique exploits the local information that can be obtained from the behaviour of the RBFs.

RBF evaluation. A credit assignment mechanism is required in order to evaluate the role of each RBF ϕ_i in the cooperative-competitive environment. For an RBF, three parameters, a_i ,e_i ,o_i are defined:

- The contribution, a_i, of the RBF ϕ_i, $i = 1 \ldots m$, is determined by considering the weight, w_i, and the number of patterns of the training set inside its width, pi_i. An RBF with a low weight and few patterns inside its width will have a low contribution:

$$a_i = \begin{cases} |w_i| & if \quad pi_i > q \\ |w_i| * (pi_i/q) & otherwise \end{cases} \tag{2}$$

  where q is the average of the pi_i values minus the standard deviation of the pi_i values.
- The error measure, e_i, for each RBF ϕ_i, is obtained by counting the wrongly classified patterns inside its radius:

$$e_i = \frac{pibc_i}{pi_i} \tag{3}$$

  where $pibc_i$ and pi_i are the number of wrongly classified patterns and the number of all patterns inside the RBF width respectively. It must be noted that this error measure does not consider the imbalance among classes.
- The overlapping of the RBF ϕ_i and the other RBFs is quantified by using the parameter o_i. This parameter is computed by taking into account the fitness sharing methodology [12], whose aim is to maintain the diversity in the population. This factor is expressed as:

$$o_i = \sum_{j=1}^{m} o_{ij} \tag{4}$$

  where o_{ij} measures the overlapping of the RBF ϕ_i y ϕ_j $j = 1 \ldots m$.

Applying operators to RBFs. In CO^{2}RBFN four operators have been defined in order to be applied to the RBFs:

- Operator Remove: eliminates an RBF.
- Operator Random Mutation: modifies the centre and width of an RBF in a percentage below 50% of the old width.
- Operator Biased Mutation: modifies the width and all coordinates of the centre using local information of the RBF environment. The centre is varied in order to approximate it to the average of the patterns belonging to the

Table 1. Fuzzy rule base representing expert knowledge in the design of RBFNs

Antecedents				Consequents				Antecedents				Consequents			
	va	ve	vo	p_{remove}	p_{rm}	p_{bm}	p_{null}		va	ve	vo	p_{remove}	p_{rm}	p_{bm}	p_{null}
R1	L			M-H	M-H	L	L	R6		H		M-H	M-H	L	L
R2	M			M-L	M-H	M-L	M-L	R7			L	L	M-H	M-H	M-H
R3	H			L	M-H	M-H	M-H	R8			M	M-L	M-H	M-L	M-L
R4		L		L	M-H	M-H	M-H	R9			H	M-H	M-H	L	L
R5		M		M-L	M-H	M-L	M-L								

RBF class and inside its RBF width. The objective of the width training is that most of the patterns belonging to the RBF class will be inside the RBF width.

- Operator Null: in this case all the parameters of the RBF are maintained.

The operators are applied to the whole population of RBFs. The probability for choosing an operator is determined by means of a Mandani-type fuzzy rule based system [18] which represents expert knowledge about the operator application in order to obtain a simple and accurate RBFN. The inputs of this system are parameters a_i, e_i and o_i used for defining the credit assignment of the RBF ϕ_i. These inputs are considered as linguistic variables va_i, ve_i and vo_i. The outputs, p_{remove}, p_{rm}, p_{bm} and p_{null}, represent the probability of applying Remove, Random Mutation, Biased Mutation and Null operators, respectively.

Table 1 shows the rule base used to relate the described antecedents and consequents. In the table each row represents one rule. For example, the interpretation of the first rule is: If the contribution of an RBF is Low Then the probability of applying the operator Remove is Medium-High, the probability of applying the operator Random Mutation is Medium-High, the probability of applying the operator Biased Mutation is Low and the probability of applying the operator null is Low.

Introduction of new RBFs. In this step, the eliminated RBFs are substituted by new RBFs. The new RBF is located in the centre of the area with maximum error or in a randomly chosen pattern with a probability of 0.5 respectively.

The width of the new RBF will be set to the average of the RBFs in the population plus half of the minimum distance to the nearest RBF. Its weights are set to zero.

Replacement strategy. The replacement scheme determines which new RBFs (obtained before the mutation) will be included in the new population. To do so, the role of the mutated RBF in the net is compared with the original one to determine the RBF with the best behaviour in order to include it in the population.

4 Experimentation and Results

In this study our algorithm is applied to eight data-sets with two different degrees of imbalance and compared with four different soft-computing methods [11],

Table 2. Experimentation Results

IR	data-set	CO^{2}RBFN	E-Algorithm	CO^{2}RBFN +SMOTE	C4.5 +SMOTE	Chi +SMOTE	Ishibuchi05 +SMOTE
3.19	Glass0123vs456	92.27 ± 3.27	82.09 ± 6.96	**93.78 ± 3.28**	90.13 ± 3.17	85.83 ± 3.04	88.56 ± 5.18
4.92	New-thyroid2	98.40 ± 3.72	88.57 ± 3.82	**98.46 ± 2.22**	96.51 ± 4.87	89.81 ± 10.77	94.21 ± 4.23
5.14	New-thyrid1	**98.02 ± 3.05**	88.52 ± 8.79	98.01 ±2.79	97.98 ± 3.79	87.44 ± 8.11	89.02 ± 13.52
5.46	Ecoli2	92.02 ± 3.40	70.35 ± 15.36	**93.14 ± 4.50**	91.60 ± 4.86	88.01 ± 5.45	87.00 ± 4.43
16.68	Abalone9vs18	**75.70 ± 9.52**	32.29 ± 20.61	75.34 ± 10.34	53.19 ± 8.25	63.93 ± 11.00	65.78 ± 9.23
23.10	Yeast2vs8	71.88 ± 14.13	72.83 ± 14.97	77.31 ± 12.23	**78.23 ± 13.05**	72.75 ± 14.99	72.83 ± 14.97
32.78	Yeast5	94.12 ± 4.40	88.17 ± 7.04	94.03 ± 4.58	92.04 ± 4.99	93.41 ± 5.35	**94.94 ± 0.38**
128.87	Abalone19	50.12 ± 21.81	0.00 ± 0.00	**70.18 ± 11.77**	15.58 ± 21.36	62.96 ± 8.27	66.09 ± 9.40

where SMOTE method has been used to pre-process the datasets. In order to estimate the precision we use five fold cross validation approach.

Table 2, columns 3 and 5 show the GM measure obtained with CO^{2}RBFN applied to data-sets with and without SMOTE pre-processing respectively. In column 4, the GM results for the E-algorithm [30] are described (which does not use pre-processed data due to it does not need to change the original class distribution). Columns 6, 7 and 8 describes the GM results for C4.5 [26], Chi [9] and Ishibuchi [15] with the data-sets preprocessed by SMOTE. The analysis of the results shows that when CO^{2}RBFN is applied to the data-sets with an IR lower than 20, there is no need for preprocessing since our algorithm outperforms all the other considered data mining methods, i.e., E-algorithm, C4.5 + SMOTE, Chi + SMOTE and Ishibuchi+SMOTE.

Regarding CO^{2}RBFN in data-sets with the higher degree of imbalance, we observe the necessity to preprocess the data, as the results are improved considering the application of SMOTE. Furthermore, CO^{2}RBFN is the method with the best GM results for five of the eight data-sets. For the other data-sets, CO^{2}RBFN + SMOTE is the second best choice.

In summary, CO^{2}RBFN presents a good behavior with the imbalanced data-sets, even when it is applied to data-sets without pre-processing method.

The good performance of CO^{2}RBFN in classification imbalanced tasks, is due to the definition proposed for the credit assignment of individuals. In this way and in order to evaluate an individual only one of the three parameters takes into account the accuracy of the RBF. The other ones are dedicated for distributing, exploring and optimizing the RBFs in the dataset space definition.

5 Concluding Remarks

In this contribution, we analyze CO^{2}RBFN, a hybrid evolutionary cooperative-competitive algorithm for the RBFN design applied to the classification of imbalanced data-sets with medium and high IR.

We study the effect of a preprocessing stage in the performance of CO^{2}RBFN by contrasting the results obtained using the original data-sets against the ones obtained with the SMOTE algorithm. Furthermore, we include in our experimental study some well-known data mining methods for comparison.

CO^{2}RBFN (without pre-processing) has a robust performance with imbalanced data-sets with IR less than 20, obtaining the best GM results. The performance of CO^{2}RBFN + SMOTE is suitable with a higher IR.

As future work, we will focus our studies in the framework of high imbalanced data-sets, analysing the improvement of CO^{2}RBFN considering a precision factor in the credit assignment.

Acknowledgment

This work had been supported by the Spanish Ministry of Science and Technology under Projects TIN2008-06681-C06-01 and TIC-3928.

References

1. Alejo, R., García, V., Sotoca, J.M., Mollineda, R.A., Sánchez, J.S.: Improving the performance of the RBF neural networks trained with imbalanced samples. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 162–169. Springer, Heidelberg (2007)
2. Al-Haddad, L., Morris, C.W., Boddy, L.: Training radial basis function neural networks: effects of training set size and imbalanced training sets. Journal of Microbiological Methods 43, 33–44 (2000)
3. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(3), 849–851 (2003)
4. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations 6(1), 20–29 (2004)
5. Broomhead, D., Lowe, D.: Multivariable functional interpolation and adaptive networks. Complex System 2, 321–355 (1988)
6. Buchtala, O., Klimek, M., Sick, B.: Evolutionary optimization of radial basis function classifiers for data mining applications. IEEE T. Syst. Man Cy. B 35(5), 928–947 (2005)
7. Chawla, N.V., Japkowicz, N., Kolcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1), 1–6 (2004)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligent Research 16, 321–357 (2002)
9. Chi, Z., Yan, H., Pham, T.: Fuzzy algorithms with applications to image processing and pattern recognition. World Scientific, Singapore (1996)
10. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data-sets. Computational Intelligence 20(1), 18–36 (2004)
11. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification system with genetic rule selection for imbalanced data-set. International Journal of Approximate Reasoning (2009) doi:10.1016/j.ijar.2008.11.004
12. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)
13. Harpham, C., Dawson, C.W., Brown, M.R.: A review of genetic algorithms applied to training radial basis function networks. Neural Computing and Applications 13, 193–201 (2004)

14. Huang, S.N., Tan, K.K., Lee, T.H.: Adaptive neural network algorithm for control design of rigid-link electrically driven robots. Neurocomputing 71(4-6), 885–894 (2008)
15. Ishibuchi, H., Yamamoto, T.: Rule weight specification in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems 13, 428–435 (2005)
16. Lacerda, E., Carvalho, A., Braga, A., Ludermir, T.: Evolutionary Radial Functions for Credit Assessment. Appl. Intell. 22, 167–181 (2005)
17. Maglogiannis, I., Sarimveis, H., Kiranoudis, C.T., Chatziioannou, A.A., Oikonomou, N., Aidinis, V.: Radial basis function neural netwroks classification for the recognition of idiopathic pulmonary fibrosis in microscopic images. IEEE T. Inf. Technol B 12(1), 42–54 (2008)
18. Mandani, E., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man Mach. Stud. 7(1), 1–13 (1975)
19. Moody, J., Darken, C.J.: Fast learning in networks of locally-tuned processing units. Neural Comput. 1, 281–294 (1989)
20. Murhphey, Y.L., Guo, H.: Neural learning from umbalanced data. Applied Intelligence 21, 117–128 (2004)
21. Orr, M.J.L.: Regularization on the selection of radial basis function centers. Neural Comput. 7, 606–623 (1995)
22. Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule-based systems for imbalanced data-sets. Soft Computing 13(3), 213–225 (2009)
23. Padmaja, T.M., Dhulipalla, N., Krishna, P.R., Bapi, R.S., Laha, A.: An unbalanced data classification model using hybrid sampling technique for fraud detection. In: Mueller, M.S., Chapman, B.M., de Supinski, B.R., Malony, A.D., Voss, M. (eds.) IWOMP 2005 and IWOMP 2006. LNCS, vol. 4315, pp. 341–438. Springer, Heidelberg (2008)
24. Park, J., Sandberg, I.: Universal approximation using radial-basis function networks. Neural Comput. 3, 246–257 (1991)
25. Pérez-Godoy, M., Rivera, A.J., del Jesus, M.J., Berlanga, F.J.: Utilización de un sistema basado en reglas difusas para la aplicacin de operadores en un algoritmo cooperativo-competitivo. In: ESTYLF 2008, pp. 689–694 (2008)
26. Quilan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman Publishers, San Mateo (1993)
27. Sun, Y.F., Liang, Y.C., Zhang, W.L., Lee, H.P., Lin, W.Z., Cao, L.J.: Optimal partition algorithm of the RBF neural network and its application to financial time series forecasting. Neural Comput. Appl. 143(1), 36–44 (2005)
28. Whitehead, B., Choate, T.: Cooperative-competitive genetic evolution of Radial Basis Function centers and widths for time series predictiont. IEEE Trans. on Neural Networks 7(4), 869–880 (1996)
29. Widrow, B., Lehr, M.A.: 30 Years of adaptive neural networks: perceptron, madaline and backpropagation. Proceedings of the IEEE 78(9), 1415–1444 (1990)
30. Xu, L., Chow, M.Y., Taylor, L.S.: Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm. IEEE Transactions on Power Systems 22(1), 164–171 (2007)