Improving Under Sampling with Neural Networks for Class Imbalance Problem

Man-Sun Kim[1]

[1] Research Center for Ubiquitous Information Appliances
Chon-nam National University,
Department of Computer Science, Chon-nam National University
300 Yongbong-Dong, Puk-Ku, Gwangju, 500-757, KOREA
mansun@kongju.ac.kr

(Paper received on June 20, 2007, accepted on September 1, 2007)

Abstract. Data in real world tasks are usually imbalanced, i.e. some classes have much more instances than others. It is one of the reasons that cause the decrease of generalization ability of machine learning algorithms. Therefore, in this paper, we handle the class imbalance problem and proposes an under-sampling method based on SOM (Self Organizing Map), one of neural networks. We apply our methods to UCI Repository data sets that have a class imbalance problem. Finally we show the improvements of new sampling method compared with other sampling methods.

1 Introduction

In a pattern recognition problem, if the number of data belonging to a class is extremely larger or smaller than that of data belonging to another class, there happens the imbalance of data. Class imbalance is often observed in response modeling, remote sensing, image classification, etc. For example, in data used in response modeling, the number of customers who have responded to marketing and purchased goods is much smaller than the total number of customers. What is more, frequently most of important information is included in purchase customers, namely, in the minor class [1,2,3,4,5,6].

In a dataset containing the class imbalance problem, data belonging to the major class are distributed excessively compared to data belonging to the minor class. In such a case, the major class infiltrates into the area of the minor class and has a negative effect on the performance of classification algorithms. Thus, it is essential to solve the class imbalance problem in order to enhance classification performance.

As solutions for the imbalance of data, two types of methods have been proposed as follows [7]. One is using a learning algorithm revised by adding a part that reflects the imbalance of data. This type of methodologies include the method of imposing different penalties on patterns misclassified into the minor class and the major class and the method of adjusting the boundary surface of separation. The other resolves data imbalance by restructuring learning data through sampling. This type of methodologies include largely three methods: over-sampling, under-sampling and ensemble[8]. Sam-

Special Issue in Neural Networks and Associative Memories
Research in Computing Science 28, 2007, pp. 151-159

pling-based methods are advantageous over revised algorithm-based methods. Sampling-based methods are easily expandable because they do not revise the algorithm itself but deal with how to compose data used in learning. The reason is that, while revised algorithm-based methods depend on one algorithm, sampling-based methods, if their high performance is proved through a specific algorithm, can be applied immediately to other pattern recognition algorithms.

Sampling is one of techniques for adjusting the size of a training dataset[9,10,11,12,13]. In general, under□sampling and over□sampling are used. Over□sampling replicates minor category data, so it does not add new information through the sampling. In addition, it is efficient in term of time complexity when handling a large volume of data. Under□sampling uses only a part of major category data, so the sample may not represent the characteristics of the whole major category. The ensemble method divides the dataset of the major class into k subsets and forms a learning dataset by combining each subset with the data of the minor class. The classifier is trained with each of the k learning dataset and the data are combined in several gathering methods of ensemble. This method shows improvement in performance experimentally, but it has disadvantages such as low accuracy of used problems, the possibility of distorted distribution of data due to the use of simple separation, and difficulty in maximizing the advantage of ensemble because the population of ensemble depends on the imbalance ratio. Under-sampling uses only a part of the major class data and thus it cannot represent the entire data of the major class. Particularly when the data difference between the minor class and the major class is large, data distribution is distorted severely. Thus, although under-sampling enhances performance, the results of individual experiments vary significantly. Many of recent under-sampling methods under study select the major class using specific strategies [8,17].

In order to solve the problems in under-sampling, the present study proposes a method of selecting major class data useful in learning using the mechanism of biological competition. The method clusters data of similar characteristic using a clustering method rather than random sampling of major class data and then builds learning data limitedly. The proposed method obtains meta-data reexpressed in competition relation by SOM (self-organizing map), which is a method of unsupervised learning, and deletes major class data around the minor class. This method is advantageous in that it can reflect the characteristics of the entire data using data distribution, and decreases data size through sampling.

This paper is composed as follows. Chapter 2 reviews recent research trends, and Chapter 3 discusses a novel under□sampling techniques based on neural network. Chapter 4 conducts an experiment and analyzes the results, and Chapter 5 draw conclusions.

2 Related works

Kubat and Matwin[14] also selectively under-sampled the majority class while keeping the original population of the minority class with satisfied results. The majority examples were divided into four categories: some noise overlapping the positive class decision region, borderline samples, redundant samples and safe samples. The result of

Kubat's experiment showed considerable improvement in performance, but not if the degree of imbalance was high.

Batista et al[15] used a more sophisticated under-sampling technique in order to minimize the amount of potentially useful data. The majority class instances are classified as "safe", "borderline" and "noise" instances. Borderline and noisy cases are detected using Tomek links, and are removed from the data set. Only safe majority class instances and all minority class instances are used for training the learning system.

Japkowicz [16] discussed the effect of imbalance in a dataset. She mainly evaluated two strategies: under-sampling and resampling. Two re sampling methods were considered. Random resampling consisted of resampling the smaller class at random until it consisted of as many samples as the majority class and "focused resampling" consisted of resampling only those minority instances that occurred on the boundary between the minority and majority classes. Random under-sampling was also considered, which involved under-sampling the majority class samples at random until their numbers matched the number of minority class samples; focused under-sampling involved under-sampling the majority class samples lying further away. She noted that both the sampling approaches were effective, and she also observed that using the sophisticated sampling techniques did not give any clear advantage in the domain considered.

Kim et al[17] proposed a focused sampling method which is more superior than previous methods. To solve the problem, The proposed method must select some useful data set from all training sets. To get useful data set, The proposed method devide the region according to scores which are computed based on the distribution of SOM over the input data. The scores are sorted in ascending order. They represent the distribution of the input data, which may in turn represent the characteristics of the whole data. A new training dataset is obtained by eliminating unuseful data which are located in the region between an upper bound and a lower bound. The proposed method gives a better or at least similar performance compare to classification accuracy of previous approaches.

3 The proposed methods

Because under-sampling uses only a part of major category data, the sample cannot represent the whole major category and distorts data distribution severely. To solve these problems, this study proposes an under-sampling method based on SOM (Self Organizing Map), one of neural networks.

For under-sampling, the present study used SOM, a method of unsupervised learning. As a result, we can reexpress high-dimensional data distribution two-dimensionally. Each data reexpressed into meta-data is allocated to a grid on the two-dimensional map. Because a grid can contain one or multiple data, it is very efficient to express data compactly. In order to minimize heterogeneity among grids containing information on different classes, grids adjacent to the minor class are selected and the selected major class grids are removed. Because the selected major class grids contain a large number of data, this process can reduce the number of data samples considerably.

3.1 SOM learning selecting the best-matching unit (BMU)

In Kohonen's learning, each neuron calculates the connection strength vector and the distance to the input vector. In addition, each neuron competes with others for the privilege of learning, and the closest neuron (best-matching unit) wins the competition. This is the application of the biological competition mechanism to learning. The best-matching unit is the only neuron that can send output signal. In Kohonen's learning, only the winner can issue output, and the winner and its adjacencies can adjust their connection strength. Because clustering is not the objective of this study, we do not need learning for adjusting connection strength. That is, only the method of selecting the best-matching unit is applied to under-sampling.

When the construction of a Kohonen network require three tasks, which are generally not necessary in other types of neural networks. One is initializing the connection strength vector of neurons on the layers adequately with random values. In the present study, it was initialized with [1,1,1,.....,1] according to the number of attributes. Another task is using normalized values between 0 and 1 for the connection strength vector and input vector. The other task is determining the size of the 2-dimensional grid. In general, the size of the grid is determined based on [ceil(5*$\sqrt{datasamples}$)] and is a multiple of the number of attributes, and with the size of the grid, the number of rows of the grid is calculated. For example, if the number of data is 200 and the number of attributes is 8, grid size becomes 72 [9*8]. The three factors are very important values to be emphasized in Kohonen network, but this study used values obtained through trials and errors. This part shall be discussed later.

The process of selecting the best-matching unit is as follows. In the process, we can obtain the result that the connection strength vector of the best-matching unit is closest to the input vector.

(1) Present a new input vector.
(2) Calculate the distance between the input vector and each neuron.
(3) Find the position of the closest grid among the calculated distances.

In order to select the best-matching unit, the distance between input and output neurons is calculated. The distance is calculated using Euclidean distance. To find the best-matching unit, the unit and distance of all maps are calculated for each data vector.

```
for i=1:dlen,
  for j=1:munits,
   for k=1:dim,
     Dist(j,i)=Dist(j,i)+mask(k)*(D(i,k)-M(j,k))^2;
   end
  end
end
(dlen:data samples, munits:row of grid, dim:column of
grid Dist(j,i): distance from i to j, mask(k): mask is
the weighting vector for distance calculation, M:codebook
matrix, D: data matrix)
```

Fig. 1. Calculation of distance between the input vector and each neuron.

3.2 Elimination of adjacencies (from the units of the minor class)

The proposed method is as follows. After the position of the closest grid among the calculated distances is obtained as a result of 3.1, each unit of the minor class is named uniti. Then, adjacent units uniti-1 and uniti+1 are stored in the eliminated unit candidate set. Here, in order to store more units in the candidate set according to the characteristic of data, the condition can be changed to uniti-2, uniti+2, ...uniti-c, uniti+c.

```
Uniti : results of 3.1
C : candidate sets of elimination units
n : number of units
for i= to n
if (i=unit of the minor class)
C=C+adjacency unit(uniti-1, uniti+1)
end if
return C
```

Fig. 2. Generation of the eliminated unit candidate set.

The generated candidate units to be eliminated are removed from the units and, as a consequence, data units of the major class, which are necessary for learning, are obtained.

The figure below shows the three steps for obtaining the units of the major class. The step of initial map generation and the step of allocation to each grid were explained in 3.1. For example, let's assume 100 input data and map size of [7 4]. Each data is allocated to each grid and forms a sub cluster. This is called a unit. As a result, the number and ratio of major class data can be reduced by eliminating units adjacent to the minor class (in Figure 3, red units). In this example, the minor class has 13 data and the major class has 87 data in the step of grid allocation. In the next step, the number of major class data can be reduced from 87 to 56 by eliminating units adjacent to the minor class. Even more adjacent units can be eliminated by the researcher's judgment. This criterion should be considered together with the characteristic of data. It is because the form of adjacent data may be different from the figure below depending on the grid distribution of the minor class and the major class. If the units of the minor class are distributed unevenly or major class data and minor class data share the same unit, units to be eliminated for high performance should be decided carefully.

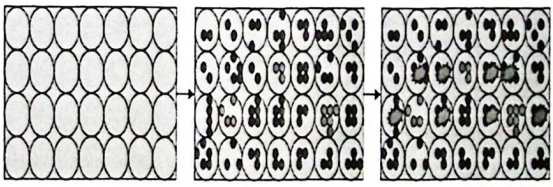

Fig. 3. Map formation step-nit allocation step-eliminate step.

In the figure above, the map formation step and the unit allocation step were explained in 3.1. In 3.2, major class units adjacent to the minor class are eliminated.

The method proposed in 3.2 is different from the method proposed by Kim et al. [17]. Their method generated data as in 3.1 but divided the result of reexpressed distribution into two sections (upper and lower). Data between the two sections were regarded as unuseful patterns and were not used in learning. This method can decrease the imbalance ratio to some degree but it may eliminate useful patterns. That is, if minor class data are not included in the two sections, the sampling method is meaningless.

4 Experiments

Because under-sampling uses only a part of major category data, the sample cannot represent the whole major category and distorts data distribution severely. To solve these problems, this study proposes an under-sampling method based on SOM (Self Organizing Map), one of neural networks. In this section, we evaluate the performance of the proposed approach in the various domain. the experiments were performed with Matlab 7.0, on 1GB RAM, Pentium IV processor. For experiment on the proposed method, we used the data of the UCI Repository [18] available on the Web.

Table 1. Data sets used in the experiments.

data set	examples	attributes	class(min.maj.)	class(min.maj.)(%)
pima	768	8	(1,0)	35.02%, 64.97%
breast	683	10	(1,0)	34.99%, 65.00%
glass	214	9	(ve_win_float_proc, remainder)	7.9%, 92.5%
wine	178	13	(3, remainder)	26.96%, 73.03%

In this experiment, we used the 5-fold hold-out method and the cross-validation method for the accurate measurement of the result of the experiment on the proposed method. The experiment divides experiment data into 5 partial datasets. One of them is used in validation, and the other 4 are used in learning. In this study, kNN was used as a basic classifier to evaluate performance. The performance scale should reflect data imbalance. In most of pattern recognition algorithms, simple accuracy measured as follows is used as a performance scale.

$$\text{Accuracy}=\frac{TP+FN}{TP+TN+FP+FN}$$

In case there happens data imbalance, however, simple accuracy cannot reflect the accuracy of the minor class sufficiently. For example, let's assume that only 10 (1%) out of 1000 customers purchased and the other 990 (99%) did not purchase. If the response model judges that none of the customer purchased, the simple accuracy of the response model is 99%, which is seemingly very high performance. However, the model failed to distinguish purchase customers, who are important, and thus it is meaningless. In this study, we use G-Mean as a performance scale for the balanced consid-

eration of the accuracy of the minor class and the major class. G-Mean (Geometric Mean) is a scale that can consider the minor class and the major class equally. It is measured as follows, and with this, the accuracy of the two classes can be understood as a geometric mean.

$$\text{G-Mean}=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{FP+TN}}$$

Table 2. The accuracy (Accu, G-Mean).

		pima	breast	glass	wine
No sampling	Accu	0.7292	0.9484	0.9679	0.9244
	G-Mean	0.6731	0.9602	-	-
kim [17]	Accu	0.4260	0.9326	0.6759	0.8202
	G-Mean	0.6457	0.9216	0.2843	0.3333
proposed	Accu	0.9055	0.9025	0.8539	0.9244
	G-Mean	0.8664	0.9860	0.6667	0.3333

As a whole, the proposed method showed high G-mean, enhancing the accuracy of both the two classes. The method proposed by Kim et al. [17] divided the result of reexpressed distribution into two sections (upper and lower), each of which contained 25% of the data. Here, discarded data may contain important information. In order to minimize the loss of important data, we eliminated data on the boundary between the minor class and the major class rather than selecting data to be eliminated by fixing the sections of the classes. Major class data on the boundary were removed by the researcher's judgement. Clarifying this part is one of tasks to be studied in the future.

In case of glass and wine, performance was very low. G-mean could not be calculated because minor class data could not be classified properly when classification was made without sampling. The reason for this phenomenon is believed to be that, in the process of imbalance data generation, one class was used as the minor class and the other classes were used as the major class because the data contained three or more classes. This shows that the result of experiment depends on the characteristic of data.

Table 3. The result of calculating imbalance ratios.

	samples	pima(768)	breast(683)	glass(214)	wine(178)
No sampling	Min	268	239	17	48
	Maj	500(65.10%)	444(65.00%)	197(92.06%)	130(73.03%)
kim [17]	Min	135	132	9	18
	Maj	250(64.93%)	209(61.29%)	99(91.66%)	71(79.77%)
proposed	Min	126	229	17	48
	maj	350(73.52%)	228(49.89%)	79(82.29%)	71(59.66%)

The table above shows the result of calculating imbalance ratios. The ratio of major class data decreased in general except PIMA. The imbalance ratio of pima data did not

have a significant effect on the accuracy of the major class and the minor class. Rather, as shown by G-mean, the overall accuracy and the accuracy of the minor class were improved.

The present used SOM to reexpress the center of class distribution. This method learns input data given without any specific instruction and expresses the characteristic of the data using specific output neurons in the space. Accordingly, we can have advantages as follows by using the results in selecting under-sampling learning data for solving the class imbalance problem.

First, the ratio of class imbalance data can be decreased. Because the distribution of the entire data is reexpressed in summarized units and sampling is made from data within a specific area, the imbalance ratio can be decreased considerably.

Second, this method can solve the problem that sampled data do not represent the whole data. In random under-sampling, the distortion of data distribution is worsened with the increase of size difference between the major class and the minor class. However, the representativeness of data can be preserved through under-sampling by a specific strategy.

Third, this method is easily expandable for learning algorithms. Because it provides the method of composing data to be used in learning, it is applicable to any learning algorithm with improved performance.

5 Conclusion

The present study proposed an efficient data selection method for solving the problem of class imbalance. In this study, first, the distances between input and output neurons was calculated to select the best-matching unit by applying SOM, which is unsupervised learning. This step finds the position of the closest grid among the calculated distances. Second, by eliminating units adjacent to minor class data, we obtained more homogeneous major class data and more heterogeneous redistribution of major and minor class data. That is, the problem of class imbalance was solved by using only useful data. The improved concentrated sampling method has advantages as follows. First, the class imbalance ratio is reduced. Second, the problem that sample data cannot represent the whole data is solved. Third, expandability for learning algorithms is high.

References

1. S. Cho, H. Shin, E. Yu, K. Ha, and D. MacLachlan, "Data Mining Problems and Solutions for Response Modeling in CRM," Entrue Journal of Information Technology, Vol.5, No.1, (2006) 55-64.
2. L. Bruzzone, D. Fernàndez Prieto, "A Combined Supervised and Unsupervised Approach to Classification of Multi Temporal Remote Sensing Images," IEEE 2000 International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, Hawaii, 24-28, Vol. 1, (2000) 162- 164.
3. R. Yan, Y. Liu, R. Jin, A. Hauptmann, "On Predicting Rare Classes With SVM Ensembles In Scene Classification," IEEE International Conference on Acoustics, Speech and Signal

Processing (ICASSP), (2003) 21-24.
4. Guobin Ou and Yi Lu Murphey, "Multi-class pattern classification using neural networks," Journal of Pattern Recognition, Vol 40, Issue 1, (2007) 4-18.
5. Vicenc Soler, Jesus Cerquides, Josep Sabria, Jordi Roig, Marta Prim, Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms, IEEE International Conference on Data Mining - Workshops (ICDMW'06), (2006) 330-336.
6. Yang Liu, Nitesh V. Chawla, Mary P. Harper, Elizabeth Shriberg and Andreas Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," Journal of Computer Speech & Language, Vol 20, Issue 4, (2006) 468-494.
7. Yanmin Suna, Mohamed S. Kamela, Andrew K.C.Wongb,YangWangc, "Cost-sensitive boosting for classi?cation of imbalanced data," Journal of Pattern Recognition.2007
8. Pilsung Kang and Sungzoon Cho, "EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems", Proceedings of International Conference on Neural Information Processing 2006
9. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. , SMOTE : Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research 16, (2002)321-357.
10. Japkowicz, N. , "The Class Imbalance Problem : Significance and strategies," International Conference on Artificial Intelligence, 2000.
11. Chawla, N.V., Hall, L. and Kegelmeyer, W. , "SMOTE : Synthetic Minority Oversampling Techniques," Journal of Artificial Intelligence Research 16, pp321-357, 2000.
12. Maloof M. A. , "Learning when Data Sets are Imbalanced and When Costs are Unequal and Unknown," ICML Workshop on Learning from Imbalanced Data Sets II, 2003.
13. Chawla, N. V. , "C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure," ICML-2003.
14. Kotsiantis S., Pierrakeas C., Pintelas P., "Preventingstudent dropout in distance learning systems using machine learning techniques," AI Techniques In Web-Based Educational-Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, (2003) 3-5.
15. Batista G., Carvalho A., Monard M. C. , "Applying One-sided Selection to Unbalanced Datasets," In O. Cairo, L. E.Sucar, and F. J. Cantu, editors, Mexican International Conference on Artificial Intelligence (MICAI), (2000) 315–325.
16. Japkowicz N., "The class imbalance problem:Significance and strategies," International Conference on Artificial Intelligence, Las Vegas, 2000.
17. Man-sun Kim, Hyung-Jeong Yang, Soo-Hyung Kim, Wooiping Cheah, "Improved Focused Sampling for Class Imbalance Problem," Journal of Korea Information Processing Society, No.14(b), Vol.4 (2007).
18. http://www.ics.uci.edu/~mlearn/databases/
19. Jigang Xie, Zhengding Qiu, "The effect of imbalanced data sets on LDA:A theretical and empirical analysis," Journal of Pattern Recognition, (2007)557-562.