An Imbalanced Data Classification Algorithm of De-noising Auto-Encoder Neural Network Based on SMOTE

Chenggang Zhang[1], Jiazhi Song[2], Zhili Pei[3] and Jingqing Jiang [3]

[1]Inner Mongolia University for the Nationalities, College of Mathematics, 028000 Tongliao, China
[2]Northeast Normal University, College of Computer Science and Information Technology, 130000 Changchun, China
[3]Inner Mongolia University for the Nationalities, College of Computer Science and Technology, 028000 Tongliao, China

Abstract. Imbalanced data classification problem has always been one of the hot issues in the field of machine learning. Synthetic minority over-sampling technique (SMOTE) is a classical approach to balance datasets, but it may give rise to such problem as noise. Stacked De-noising Auto-Encoder neural network (SDAE), can effectively reduce data redundancy and noise through unsupervised layer-wise greedy learning. Aiming at the shortcomings of SMOTE algorithm when synthesizing new minority class samples, the paper proposed a Stacked De-noising Auto-Encoder neural network algorithm based on SMOTE, SMOTE-SDAE, which is aimed to deal with imbalanced data classification. The proposed algorithm is not only able to synthesize new minority class samples, but it also can de-noise and classify the sampled data. Experimental results show that compared with traditional algorithms, SMOTE-SDAE significantly improves the minority class classification accuracy of the imbalanced datasets.

1 Introduction

Classification problem is one of the important research contents in the field of machine learning. Some of the existing classification methods can generally perform better when classifying balanced data. However, there are large amounts of imbalanced datasets in the field of practical application, such as network intrusion, text classification, credit card cheat-detection, medical diagnosis, etc., for which minority class recognition rate is more important [1-2]. To remedy the deficiency that minority class samples have on distributed information, SMOTE algorithm put forward by Chawla etc. [2]. Not only effectively synthesizes minority class samples, but, to a large extent, it also avoids over-fitting problem. The algorithm has already achieved a favourable effect in imbalanced datasets classification, but it brings a new problem such as noise.

Auto-encoder neural network based on thought of deep learning has already obtained huge success in the field of machine learning [3]. It initializes the network weights through unsupervised layer-wise greedy learning, and learn about data features and reduces the irrelevant and redundant data through constantly adjusting the network parameters, and then it fine-tunes the network parameter using back propagation (BP) algorithm. Stacked De-noising Auto-Encoder neural network, SDAE, can train more robust expressions of input data through adding noise into original data. Thereby it can improve the generalization ability of auto-encoder neural network to input data [4-6]. Imbalanced data classification algorithm of de-noising auto-encoder neural network based on SMOTE proposed in the paper can reduce the noise problem from SMOTE, and it de-noises and classifies the sampled data, which improves minority class classification quality.

2 Related works

2.1 SMOTE Algorithm

Synthetic minority over-sampling technique (SMOTE) algorithm is a kind of typical sampling method proposed by Chawla etc. in 2002 [5]. Compared to traditional over sampling technology, it can effectively avoid over fitting phenomenon of the classifier. The main gist is to insert synthesized minority class samples at the nearest neighbours of it, thus, to increase the number of minority class samples to balance the dataset. To be specific:

Suppose oversampling ratio is N. First, randomly choose K samples from P nearest minority class neighbours of each minority class sample. Then according to (1), synthesize each minority class sample and chosen K samples respectively to generate N new minority class samples; finally, add the new samples to original sample set to form a new training sample set.

$$x_{new} = x + rand(0,1) \times (y[i] - x) \tag{1}$$

Among which, $i = 1,2,\ldots N$, $rand(0,1)$ is a random number in range $(0,1)$; x_{new} is the new synthesized sample, x is minority class sample, $y[i]$ is the $i-th$ nearest neighbour of x.

2.2 Auto-encoder Neural Network (AE)

Auto-encoder neural network is an unsupervised learning neural network which reconstructs input data as much as possible. It initializes the network weights using the greedy layer-wise training method, and fines tune network parameter using the back propagation (BP) algorithm to optimize the overall performance. Using hidden layer outputs as the new input features, SAE (stacked auto-encoder) deep structure is formed through multiplication AE.

2.3 Stacked De-noising Auto-Encoder neural networks (SDAE)

On the basis of traditional de-noising auto-encoder neural network (AE), after adding noise with a certain probability distribution to the input data, DAE proposed by Vincent etc. [6], makes auto-encoder neural network (AE) learn how to remove the noise, and reconstruct undisturbed input as much as possible. Therefore, the features generated from the learning of input corrupted with noise are more robust, which improved the data generalization ability of auto-encoder neural network model to input data.

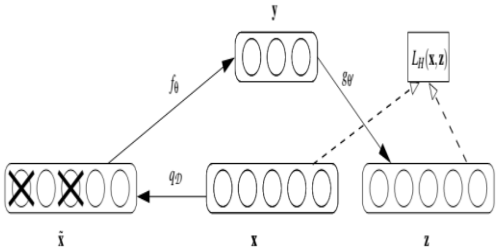

Figure 1. De-noising Auto-encoder neural network [6]

In Fig.1, original data x is added noise in a certain probability q_D to form disturbed data $\widetilde{x}$ as auto-encoder input. And f is an activation function used to compute the activation values of each neuron of the hidden layer, as defined in (2).

$$h_{w,b}(\widetilde{x}) = f(\sum w\widetilde{x} + b) \tag{2}$$

Therefore, the cost function of de-noising auto-encoder neural network is defined according to (3).

$$J_{DAE}(w,b) = \frac{1}{m}\sum(\frac{1}{2}\left\|h_{w,b}(\widetilde{x}) - z\right\|^2)) \tag{3}$$

Here, w is weights between neurons, b is bias, m is number of samples among which $sigmoid$ activation function $f(x) = 1/(1+e^{-x})$ with the range [0, 1] is used in the study. Using hidden layer outputs as the new input features, Stacked Auto-Encoder neural networks (SDAE) deep structure is formed through multiplicating DAE.

3 SMOTE-SDAE Algorithm

Combining the strengths of SMOTE and SDAE, the paper puts forward an imbalanced data classification algorithm of de-noising auto-encoder neural network based on SMOTE. First of all, use SMOTE to balance data set. Then aiming at the new data noise problem brought by SMOTE, more robust features are obtained after unsupervised layer-wise greedy training of SDAE. SDAE improves input data generalization ability of auto-encoder neural network, which improves the classification accuracy of minority class and overall samples. Detailed descriptions of the algorithm as follows:

1) *Training phase:*

a) Set parameters: $\theta = \{W, b\}$ among which W is network weight, b is bias; neuron number of the visual layer is *v*, neuron numbers of the hidden layer is *h1, h2*; q_D is Gaussian noise, *T* is total number of positive samples in data set, *N* is sample synthesis rate, *k* is number of chosen nearest neighbors, default is 5.

b) Load dataset: Let *Dadaset* is the original training set.

c) Oversample dataset: newDataset = SMOTE(*T, N, k*).

d) Add noise: Generate *D-newDataset* through adding Gaussian noise according to q_D into *newDataset.*

e) Unsupervised training: Study network parameters $\theta = \{W, b\}$ with layer-wise greedy learning.

f) Supervised training: Use *newDataset* to fine-tune network parameter with optimization algorithm L-BFGS.

2) *Test phase:* Test *N* with test set, and return *AUC,F-value* and *G-mean*.

4 Experiment and analysis

Table 1. AUC contrasts of different algorithms.

Dataset	Total Number of Samples	Number of Attributes	Negative Sample Size	Positive Sample Size	Imbalance Rate
Breast	277	9	196	81	2.42
German	1000	24	700	300	2.33
Heart	303	13	164	139	1.18
Ionosphere	351	34	225	126	1.79
Pima	768	8	500	268	1.87
Sonar	208	60	111	97	1.14
Spam base	4601	57	2788	1813	1.54
Wpbc	198	33	151	47	3.21

4.1. Dataset description

All the experimental data in the paper are eight binary classification datasets commonly used in the study of unbalanced data classification, which is obtained from UCI machine leaning databases, detailed descriptions are shown in table 1.

4.2. Evaluation index based on confusion matrix

In the study, minority class in classification learning is defined as positive, and majority class negative. Confusion matrix to evaluate two-class problem is shown in Table 2

Table 2. Confusion matrix for two-class problem.

Classification	Actual Positive Sample	Actual Negative Sample
Predict as Positive	TP	FP
Predict as Negative	FN	TN

In Table 2, TP stands for the number of minority class being classified as minority class. TN is the number of majority samples being assessed as majority class.

In learning imbalanced data, the effect of minority class on classification accuracy is far less than that of majority class. So classification learning taking classification accuracy as a criterion usually leads to low minority class recognition rate. The classifier tends to forecast a sample as majority class sample. The classic classification accuracy evaluation criteria do not apply to the classifier performance assessment of imbalanced data. Therefore, for imbalanced data classification, there have been new evaluation criteria, such as *AUC*, F-*value* and *G-mean* [7], etc. Their definitions are as follows:

AUC (Area Under roc Curve) provides a method to measure the classifier performance when it is hard to be judged because of ROC (Receiver Operating Characteristic) curve intersection. *AUC* value of a classifier is the area under the corresponding ROC curve. The larger the area, the better the performance of the classifier will be.

F-value is a classification evaluation index comprehensively incorporating *recall* and *precision*, as defined in (4).

$$F-value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \qquad (4)$$

In (4) , $precision=TP/(TP+FP)$, $recall=TP/(TP+FN)$, β ranges from $[0,+\infty]$.When $\beta=1$， in the experiment, *F-value* can be used to balance the equally important relation between *recall* and *precision*.

G-mean is the geometric mean of classification accuracy of minority class and majority class. It is defined according to (5).

$$G-mean=\sqrt{\frac{TN}{TN+FP}\times\frac{TP}{TP+FN}} \qquad (5)$$

G-mean is the maximized accuracy of two classes under the condition of maintaining the classification accuracy balance of minority class ad majority class. *G-mean* is the maximum only in the case of simultaneous high classification accuracy of minority class and majority class.

4.3. Experimental results and analysis

Experiments are conducted on the proposed SMOTE-SDAE algorithm and SVM、SMOTE-SVM、SDAE、SAE algorithms, the results of which are compared using three evaluation methods mentioned above. The following three tables show the compared experimental results of AUC, F-value and G-means of each algorithm. Libsvm toolbox is used in all the SVM algorithms in the experiment, and RBF is adopted as Kernel function with parameter. The average of 10 times 10 fold cross-validations is used as a result. The experimental environment is win7 64bit ， Matlab2012b ， CPU 3.4GHz，RAM 4G.

Table 3.AUC contrasts of different algorithms

AUC	Breast	German	Heart	Ionosphere	Pima	Sonar	Spambase	Wpbc
SVM	0.7277	—	0.8295	0.9391	0.6571	0.8555	0.7077	0.6618
SMOTE-SVM	0.7479	0.7179	0.8736	0.9679	0.7045	0.9585	**0.8878**	0.7900
SAE	0.6253	0.7133	0.7412	0.9267	0.6765	0.8073	0.8318	0.5784
SDAE	0.6536	0.7313	0.7809	0.9136	0.6890	0.8318	0.8346	0.6386
SMOTE-SDAE	**0.7845**	**0.7412**	**0.9113**	**0.9792**	**0.7371**	**0.9767**	0.8871	**0.8949**

Table 4.F-value contrasts of different algorithms

F-value	Breast	German	Heart	Ionosphere	Pima	Sonar	Spambase	Wpbc
SVM	**0.8391**	**0.8230**	0.8432	0.9538	0.7856	0.8655	0.7835	0.8750
SMOTE-SVM	0.5844	0.4681	0.7428	0.9489	0.3908	0.8414	0.4892	0.7737
SAE	0.4322	0.4946	0.7205	0.8648	0.5300	0.7612	0.8028	0.2857
SDAE	0.4191	0.8084	0.7618	0.8836	0.5236	0.7997	0.8085	0.4255
SMOTE-SDAE	0.8319	0.8144	**0.9492**	**0.9842**	**0.8367**	**0.9787**	**0.9186**	**0.9021**

Table 5.G-mean contrasts of different algorithms

G-mean	Breast	German	Heart	Ionosphere	Pima	Sonar	Spambase	Wpbc
SVM	0.5071	—	0.8128	0.9322	0.4235	0.8462	0.5726	0.4830
SMOTE-SVM	0.6582	0.5656	0.8000	0.9596	0.5082	0.8574	0.5701	0.8008
SAE	0.5725	0.6041	0.7465	0.8812	0.6287	0.7722	0.8380	0.4677
SDAE	0.5494	0.7002	0.7795	0.9098	0.6196	0.8258	**0.8426**	0.5872
SMOTE-SDAE	**0.7297**	**0.7082**	**0.8450**	**0.9689**	**0.6440**	**0.9177**	0.7747	**0.8824**

5 Conclusions

Imbalanced data classification problem has always been a research focus in the field of data mining and machine learning. Directing at the SMOTE algorithm's problems such as noise when synthesizing new minority class samples, the current paper put forward a De-noising Auto-Encoder neural network algorithm based on SMOTE, which can balance data sets through generating new minority class samples, and de-noise and classify the sampled data sets. In comparison with traditional de-noising auto-encoder neural network and algorithms like SVM, the suggested algorithm improved minority class classification accuracy. Such problems as how to better synthesize minority class samples, neural network node number selection and high-dimensional data processing remain to be studied further.

Acknowledgement

This work was supported by The National Natural Science Foundation of China (Project No. 61163034, 61373067), and The Natural Science Foundation of the Inner Mongolia autonomous region of China (Project No. 2013MS0910, 2013MS0911).

References

1. X. Zhang, MATEC Web of Conference **22**,05003 (2015)
2. Ramentol E, Caballero Y, Bello R, et al, K.A.I.S.**33**,245(2012)
3. Chan P.K., Fan W., Prodromidis A. L, et al,IEEE Intel. Syst, **14**,67(1999)
4. G. E. Hinton, and R. Salakhutdinov, Scinence,**313**,504(2006)
5. Y. Bengio, Foundatons and trends in Machine Learning,(2009)
6. P. Vincent, H. Larochelle, Y. Bengio, PA.Manzagol, J Mach Learn Res,**11**,3371(2010)
7. Buckland M, Gey F, JASIST, **45,** 12(1994).