ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 7, Issue, 04, pp.15031-15033, April, 2015

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

REVIEW ARTICLE

IMBALANCED DATASETS

Deshmukh Deepika, D. Satya Bhavani , D. Ashritha and *D. Hemanth Kumar Reddy

Department of CSE, MGIT, Hyderabad, India

ARTICLE INFO

Article History:
Received 19th January, 2015
Received in revised form
02nd February, 2015
Accepted 19th March, 2015
Published online 30th April, 2015

Key words:

Imbalanced datasets,
ROC curves, Classification,
Rough Set Theory,
Oversampling,
Reduct.

ABSTRACT

The class imbalance problem emerged strong as it extended more into real domains. A dataset is said to be imbalanced if the classification categories are not approximately equally represented. Accuracy, which is considered as the major performance measure of classifier is not appropriate for imbalanced datasets as the cost of errors vary markedly. This paper proposes an efficient oversampling technique based on SMOTE (Synthetic Minority Oversampling Technique) which is used for generating synthetic samples in the process of balancing the data*et al*ong with an editing technique for efficient feature extraction based on Rough Set Theory. Performance measures which are more appropriate for imbalanced datasets such as ROC curves and cost curves are considered along with accuracy.

INTRODUCTION

The issue of data mining on highly imbalanced data sets (high ratio of majority and minority classes) became more predominant as it shows its effect on many real domains. The class imbalance problem occurs when the instances of some classes outnumber the instances of other classes. In this scenario, existing classifiers tend to be overwhelmed by the classes with more instances and ignore other classes. In real world applications, this becomes more drastic as the ratio of small to large classes will be as huge as 1 to 100, 1 to 1,000 or even 1 to 10,000. As mentioned earlier this problem is evident in many domains which include Intrusion Detection System (IDS), text classification, medical diagnosis, risk management and many others. For example, in medical diagnosis of a certain cancer, if the cancer is regarded as the positive class, and non-cancer (healthy) as negative, then misclassifying a cancer patient as negative (false-negative) is much more expensive than the false-positive error. The patient could lose his/her life because of the delay in the correct diagnosis and treatment. An oversampling technique called SMOTE (Synthetic Minority Oversampling Technique) is used to generate synthetic samples of minority class in order to balance the dataset. The major issue faced by researchers with imbalanced problem is regarding the evaluation measures. As previously mentioned, common evaluation measures such as accuracy can yield misleading conclusions as there can also be an imbalance in costs of making different errors, which could vary per different cases. So, more accurate and appropriate measures such as ROC curves and cost curves are considered. Rough Set Theory helps in the efficient feature extraction. Feature selection property of rough set theory helps in finding reducts of the oversampled dataset. In this way it not only reduces the size of dataset but also improves the classifier performance. Facts stated here are referred from (Nitesh V. Chawla). Detailed description of SMOTE and rough set based feature selection is presented in sections 2,3,4.

Corresponding author: Hemanth Kumar Reddy, D.
Department of CSE, MGIT, Hyderabad, India

SMOTE (Synthetic Minority Oversampling Technique)

Oversampling by replication of existing minority samples can result in more specific regions in feature space. This can lead to overfitting on minority class examples. SMOTE is used to overcome those difficulties and it generates synthetic examples by operating in "feature space" rather than in a "data space". SMOTE algorithm considers the minority class instances and oversamples it by generating synthetic examples joining all of the k minority class nearest neighbors. The value of k depends upon the amount oversampling to be done. The process begins by selecting some point yi and determining its nearest neighbours yi1 to yik. Random numbers from r1 to rk are generated by randomized interpolation of the selected nearest neighbors. Synthetic samples of minority can be generated as follows:

1. Consider the feature vector (minority sample) and its nearest neighbor and take the difference between them.
2. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.
3. This causes the selection of a random point along the line segment between two specific features.

The present algorithm used is taken from reference (Chawla *et al*., 2002).

Rough based feature selection

In 1982, a polish scientist zidslaw I. Pawlak introduced theory of rough sets. In 1985, he derived rough dependency of attributes in information systems. Feature selection is the important step in Rough Set Theory. Feature selection reduces the dimensionality by considering only the subset of attributes (features) that are majorly used in classification process. The final subset of features obtained is said to be a "reduct". An exhaustive solution is to use boolean reasoning laws to find out all reducts, and then to choose the one with minimal attributes. Concepts of rough sets were taken from reference (Pawlak, 1991; Han and Wang, 2004).

Implementation

pseudo-code of SMOTE algorithm in the process of generating synthetic samples is given below. Algorithm takes three inputs. T is the number of minority class samples, N is the degree of oversampling (in %) to be done and k is number of the nearest neighbors.

Algorithm Smote (T, N, k)

//Input: T, N, k
//Output: (N/100) * T synthetic minority class samples

Step 1: This step selects the number of minority class samples to be generated when the value of N i.e., amount of oversampling to be done is less than 100%. In this case minority class samples are randomized as only random percent of them will be SMOTEd.

1. if N < 100

then Randomize the T minority class samples

T = (N/100) ∗ T
N = 100
Endif

Step 2: This step is aimed to compute all the k-neighbors for each sample to be replicated.

// N = (int)(N/100) (∗ The amount of SMOTE is assumed to be in integral multiples of 100. ∗)
// k = Number of nearest neighbors
//numattrs = Number of attributes
//Sample( )( ): array for original minority class samples
//newindex: keeps a count of number of synthetic samples generated, initialized to 0
//Synthetic( )( ): array for synthetic samples (∗ Compute k nearest neighbors for each minority class sample only. ∗)
for i ← 1 to T
Compute k nearest neighbors for i, and save the indices in the nnarray
Populate (N, i, nnarray)
endfor Populate(N, i, nnarray) (∗ Function to generate the synthetic samples. ∗)
17. while N != 0
18. Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.

Step 3: Finally, this step computes the interpolation as explained above in section 2.

forattr ← 1 to numattrs
Compute: dif = Sample(nnarray(nn))(attr) − Sample(i)(attr)
Compute: gap = random number between 0 and 1
Synthetic(new index)(attr) = Sample(i)(attr) + gap ∗ dif
End for
new index++
N = N – 1
End while
return (∗ End of Populate. ∗)

Oversampled dataset obtained after applying the SMOTE algorithm is considered and the reduct algorithm is applied for the feature selection. The approach of this hybrid technique is referred from (EnislayRamentol, 2009; Japkowicz and Holte, 2001).

Algorithm Reduct (V, f)

//Input: V is the number of attributes and f is the discernibility function

//Output: A reduct set R

Step 1: Reduct set is initially set to NULL and the discernibility matrix is considered.

Discernibility matrix

Given an information table S, its discernibility matrix M = (M(x, y)) is a |U|×|U| matrix, in which the element M(x, y) for an object pair (x, y) is defined by:

M(x, y) = {a ∈ At | Ia(x) != Ia(y)}.

For example, consider the Information system presented in the below table

Table 1. A sample database

	A	b	c	d	E
X1	1	0	2	1	1
X2	1	0	2	0	1
X3	1	2	0	0	2
X4	1	2	2	1	0
X5	2	1	0	0	2
X6	2	1	1	0	2
X7	2	1	2	1	1

Discernibility matrix of the above information system is presented in the below table

Table 2. Discernibility matrix for table1

	X1	X2	X3	X4	X5	X6
X2	-					
X3	b,c,d	b,c				
X4	B	b,d	c,d			
X5	a,b,c,d	a,b,c	-	a,b,c,d		
X6	a,b,c,d	a,b,c	-	a,b,c,d	-	
X7	-	-	a,b,c,d	a,b	c.d	c,d

Step 2: For every attribute in the attribute set, the number of times it occurs in the discernibility function is counted and the count variable is incremented.

Step 3: If the count exceeds the minimum count then the attribute is added to the reductset.

Step 4 : This process continues until all the attributes are considered and finally the reduct set is obtained.

Design Flow

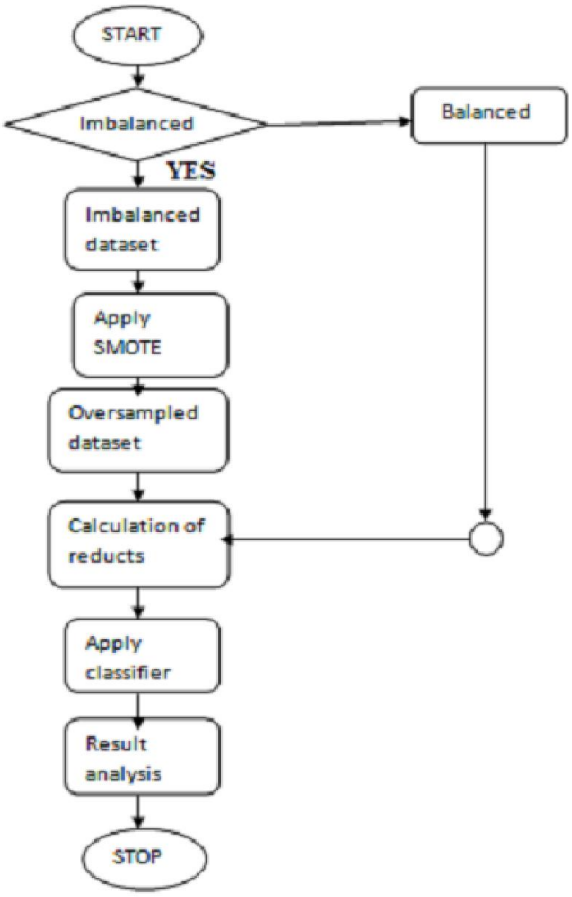

Fig. 1. Data Flow Diagram for balancing the imbalanced dataset

RESULTS

The test results of this efficient hybrid technique applied on different datasets and also the comparative analysis with the existing methods is presented in the table below

Table 3. Comparative analysis

Dataset	Attribute	Instance	Original	SMOTE	SMOTE+RST
Vowel	13	988	97.06	94.94	96.78
Sick	23	8124	92.97	94.08	95.06
Weather	5	14	57.14	93.8	95.1
Yeast	8	1484	61.3	70.04	76.09
Ecoli	7	240	81.59	79.77	78.18
Glass	9	214	89.76	88.29	92.32

From the above table it can be inferred that the AUC (Area Under Curve) in the last case is more compared to other methods for most of the datasets. So we can say that this hybrid technique yields best results.

Conclusion

This work put forth a hybrid preprocessing technique for imbalanced datasets, the SMOTE-RST process, which iteratively applies SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset by generating new synthetic instances and uses rough Set Theory techniques to carry out editing on the synthetic and majority instances. From the above experimental result obtained it is evident that this hybrid approach outperforms the existing SMOTE algorithm and also gives higher classifier accuracy than when directly applied on imbalanced datasets.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research*, 16:321-357.

Enislay Ramentol, yaile caballero, A journal on SMOTE-RSB, 23 December 2009.

Han, S.Q. and J. 2004. Wang, Reduct and attribute order, *Journal of Computer Science and Technology*, 19; 429-449.

Japkowicz, N. and R. Holte, 2001. Workshop report: Aaai-2000 workshop on learning from imbalanced data sets. AI Magazine, 22(1).

Nitesh V. Chawla, ?. Chapter on datamining for imbalanced datasets:An overview, Springer.

Nitesh V. Chawla, ?. Nathalie japkowicz, A journal on special issue on learning from imbalanced datasets, volume6, Issue1.

Pawlak, Z. 1991. Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, MA.

****