www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Enhanced hybrid model for balancing dataset to improve the performance of the classifier

Dr.S.Babu

Department of Computer Science and Application, SCSVMV Deemed University ,Enathur, Kanchipuram.

ABSTRACT

Data Mining is a process by which data can be analyzed, so as to generate useful knowledge. In Data Mining, Classifiers are the widely accepted effective technique for prediction. A well balanced dataset is a vital source for the classifiers to yield the best prediction. Recent studies have shown that, imbalanced datasets exists on many real applications. Re-sampling are the techniques to handle such an issue. In this paper an Enhanced hybrid model was proposed to balance the dataset, which is the integration of both undersampling and oversampling techniques. In order to balance the dataset, the model initially uses undersampling technique to remove certain instance from majority class which has less classification information and then oversampling technique applied on minority class by using nearest neighbors. To prove the efficiency of the proposed model various experiments were conducted. To perform the same datasets with different imbalance ratio were taken from UCI and KEEL repository. The results of the experiments show that classifiers were able to outperform on the dataset which was balanced by the proposed model.

Keywords: Data Mining, Imbalanced Dataset, Oversampling and Undersampling.

INTRODUCTION

Data mining is a widely accepted interesting research area to mine the data in order to acquire the meaningful hidden knowledge from it (Haiyang, Z., 2011). Data mining not only performs the process of extracting the knowledge from data; it also performs the process like Data Cleaning, Data Integration, Data Transformation, Pattern Evaluation, Data Presentation (Tsai, C. F., et.al, 2010).

Prediction is the key concern of many real applications. In Data mining classifiers are the effective one for the same (Babu, S., et.al, 2017). By building the suitable classifier, it predicts well about which class the new instance is (Yen, S. J., et.al, 2009). Accuracy is the important aspect of the classifiers and to attain the same, different classifiers with different techniques exist. Even then the classifiers were not able to prove the best performance on imbalanced dataset.

Research focus is increasing rapidly in mining imbalanced data set because of its wide application on the real world. A dataset is said to be imbalanced, if the classification attributes are not evenly represented (Jeatrakul, P., et.al, 2010). Generally classifiers assume that, the instances are uniformly distributed on existing classes. By which the classifiers were able to perform well on the balanced dataset and not well on imbalanced dataset. In general, the datasets of the real applications are imbalanced.

The imbalance issues arise when more instances in one class (Majority Class) and very less in another class (Minority Class). Particularly it is terrible in Big Data. The reason is that the needed information is generally represented by the minority class. In such imbalanced dataset, the

Author for correspondence:
Department of Computer Science and Application, SCSVMV Deemed University ,Enathur, Kanchipuram.

classifiers are able to perform extremely well on majority class whereas not on the required minority class. (Wang, Q., 2014). Hence an effective model is needed to balance the dataset in order to improve the performance of the classifier.

REVIEW OF LITERATURE

The real time data have the problem of imbalance (Alitouche, T. A., *et.al,* 2013). Hence the recent research focus was tuned and proposed several models towards balancing the dataset. Algorithm oriented and data oriented are the approaches proposed to solve this issue. In data oriented approaches one of the following three methods are represented. **Oversampling** – Replicates certain instances in the minority class.
Undersampling – Removes certain instances from the majority class.
Hybrid – Integrates both oversampling and undersampling.

An efficient oversampling technique EMOTE to solve the imbalance issue. In this method, each misclassified instance along with their nearest neighbor which was retrieved from correctly classified instances was populated in the actual dataset. To prove the efficiency, various experiments were conducted on different dataset which has different imbalance ratio. The results prove that the classifiers were able to improve their performance on the dataset balanced by EMOTE [1].

Sampling technique,the data set is partitioned as minority and majority class sets. Then, the majority class is split into k clusters (k = 3). The majority class clusters are individually merged with minority class to create k subsets. To evaluate the proposed model, Fuzzy Unordered rule induction and decision tree algorithm are used and all the combined subsets are classified. The subset with highest accuracy is considered for further data mining process [2].

Imbalance issue, is the integration of Fuzzy Distance based Under Sampling (FDUS) and SMOTE. To evaluate the performance of the proposed method, measures like F-measure and G-mean are considered [3].

A method that combines both the SMOTE and the Complementary Neural Network (CMTNN) to solve imbalance issue. In this CMTNN is functioned as an undersampling technique and SMOTE as an oversampling technique. To prove the performance, the datasets like Pima data, SPECT heart data, German credit data and Haberman's Survival data are considered. The measures like G-mean and AUC are used for analysis [4].

SMOTE (Synthetic Minority Oversampling TEchnique) to balance the dataset. In this method, to balance the dataset oversampling used on minority class and undersampling used on majority class. Oversampling was done by creating the artificial synthetic examples. Based on the amount of oversampling needed, using kNN algorithm five nearest neighbors were taken. From that two are taken and in the direction of each, new instance was created. To analyze the performance, ROC measure was considered [5].

Input: Imbalanced Dataset S_i

Output: Balanced Dataset F_i

Step 1. Train an SVM classifier on the actual dataset S_i

Step 2. Remove certain instances from majority class based on the hyperplane and distance using (3).

Step 3. Based on the result of SVM classifier, Partition the actual dataset S_i into CC_i - Correctly Classified Instances and M_i - Misclassified instances.

Step 4. For each instance of M_i

Step 4.1 Find the' k' nearest neighbor from CC_i // k – number of neighbor

Step 4.2 Populate only the minority class samples on the actual dataset S_i

End For

Step 5. Rebuild the classifier with improved S_i and compute accuracy

Step 6. Repeat Steps 2 to 5 still improvements on the classifier's accuracy.

RESEARCH METHODOLOGY

Enhanced Minority Oversampling TEchnique (EMOTE) is a proved effective method for balancing the dataset. Even then in least case it faces the problem of overfitting which is the general issue of oversampling (Babu, S., et.al, 2017). Hence in this paper hybrid technique EMOTE + Under sampling was proposed to balance the dataset. The key idea of the proposed method is that, EMOTE is used as an Oversampling technique on minority class and undersampling used to remove certain instances from majority class which has less classification information.

The flow of the proposed method is as follows. Given training dataset

$$S = \{(x_1, y_1), (x_2, y_2), \ldots\ldots(x_i, y_i)\}$$

Where

$y_i \in \{0, 1\}$ - represents the positive and negative class label.

Suppose if an imbalanced dataset consists of 'n' instances on majority class and 'm' instances on minority class, where n > m and n + m = i, then the imbalance ratio is

$$IR = \text{n} / \text{m} \qquad (1)$$

In the undersampling process, the proposed model initially trains the SVM classifier using the dataset '*S*' and defines the classification hyperplane.

$$w . x + b = 0 \ (2)$$

After defining the hyperplane, some of the majority class instances which have less classification information are removed. This process is depends on the distance between instance 'z' and the hyperplane as follows.

$$d = \|w . z + b\| \ (3)$$

The model proportionately removes certain majority class instances which are far away from the hyperplane based on the calculated distances. This in turn leads to reduce the imbalance ratio.

In the oversampling phase EMOTE is used. The idea of EMOTE is, based on the classifier performance, the actual dataset is categorized into two different datsets - namely correctly classified and incorrectly classified. As second, the nearest neighbors of the misclassified instance from correctly classified instances are taken and populated in the actual dataset. The second step is repeated for each instance of the incorrectly classified dataset. This in turn leads to reduce the imbalance ratio [6-9].

After applying under sampling and oversampling technique, the classifier is rebuilt on the dataset which was resampled by the proposed model. The above said tasks are repeated still there is an improvement on the classifier accuracy. Based on the above description, the proposed Enhanced Hybrid Model is described in Algorithm 1.

Algorithm 1: Enhanced Hybrid Model.

Experiments and Results

To evaluate the performance of the proposed hybrid model, C# code had been written using WEKA. To use classes of WEKA in C#, WEKA.JAR is converted as WEKA.DLL using IKVM tool. The datasets like Bupa, Pima, Cmc2, Yeast, Abalone7, Glass7, Vowel, and Letter4 are taken from KEEL Repository. The characteristics of the above datasets are shown in Table 1. In general overall accuracy is the measure that reveals the performance of the classifier. But in the context of imbalanced dataset it is not apt because overall accuracy of the classifier is dominated by the majority class. Hence various measures like Sensitivity (True Positive Rate), Specificity (False Positive Rate), Gmean, F-Measure and ROC is used.

Sensitivity (True positive rate):

$$SN = TP / (TP + FN) \quad (4)$$

Specificity (True negative rate):

$$SP = TN / (TN + FP) \quad (5)$$

G-Mean (A Geometric mean of precision and recall):

$$G\text{-mean} = \sqrt{Precision\ X\ Recall} \quad (6)$$

Table 1: Characteristics of Data Sets

Data Set	Total Number of				Imbalance Ratio
	Attributes	Instances	Minority Class Instances	Majority Class Instances	
Bupa	7	345	145	200	1.38
Pima	9	768	268	500	1.87
Yeast	8	1332	84	1248	14.9
Letter4	16	20000	805	19195	23.84
Abalone7	8	4177	391	3786	9.7
Cmc2	10	1473	333	1140	3.4
Glass7	9	214	29	185	6.4
Vowel	13	990	90	900	10.0

F-Measure (A harmonic mean of precision and recall):

$$\textbf{F-Measure} = 2\ X\ \frac{Precision\ X\ Recall}{Precision+Recall} \quad (7)$$

Table 2: Performance Evaluation of C4.5

Data Set	Sensitivity (%)		Specificity (%)		Overall Accuracy (%)		
	Actual	Proposed	Actual	Proposed	Actual	Proposed	Lifts By
Bupa	53.79	98.62	79.00	98.00	68.41	98.26	29.86
Pima	55.97	98.51	85.40	98.93	75.13	98.79	23.66
Yeast	47.44	94.88	86.07	99.05	74.88	97.85	22.96
Abalone7	43.37	97.59	86.61	95.40	75.47	95.96	20.50
Letter4	63.56	96.67	93.32	100.00	86.15	99.20	13.05

To assess the proposed model, the following experiments were carried out and the results are compared with other existing methods. They are

- Performance of C4.5 on various actual imbalanced dataset and datasets resampled by proposed model.
- Performance Comparison of proposed model with Hybrid SVM Sampling.
- ROC Performance Evaluation of proposed model

Performance of C4.5 on various datasets resampled by proposed model

To evaluate the performance of the proposed model, initially classifier C4.5 was executed on various actual imbalanced dataset using WEKA and the results are recorded. As next, to balance the actual dataset, proposed hybrid model was executed on the same. After balancing the classifier C4.5 was executed on the balanced dataset and the results are recorded. Both the results are presented in Table 2. The results of the Table 2 show that the performance of the classifier was better on the dataset which was balanced by the proposed model.

Performance Comparison of proposed model with Hybrid SVM Sampling.

Hybrid SVM Sampling is a existing technique to balance the dataset. In which first undersampling is used to remove certain instances from majority class and then oversampling is used on minority class. For oversampling SMOTE is used. To prove the efficiency, various experiments are done and the results of the same are compared with the other existing methods.

To compare and analyze the proposed hybrid model with the existing Hybrid SVM Sampling model, the same datasets are considered. After balancing the dataset by the proposed hybrid model, classifier executed on it and results were recorded. From the results, Gmean and F-Measure values are calculated and presented in Table 3. The same were compared with existing methods.

In general, the high value of Gmean represents that the sensitivity and specificity of the dataset are high. This in turn reveals that the classifier performs well on both minority and majority class. The values in the Table 3 show that the Gmean value of the proposed method is high. It shows that the classifier is able to perform well on both the classes of the dataset which was balanced by the proposed hybrid model. The table values also reveal that the proposed model outperforms than other existing methods.

Table 3: Gmean values of compared methods

Dataset	SMOTE	Easy Ensemble	Hybrid SVM	Proposed
Cmc2	0.5461	0.5734	0.5823	0.8632
Glass7	0.8865	0.9237	0.9125	0.9347
Abalone7	0.4672	0.6613	0.6653	0.9457
Vowel	0.8333	0.8769	0.9013	0.9536
Yeast	0.6547	0.7532	0.7845	0.9564
Letter4	0.5945	0.6217	0.6976	0.9862

The F-Measure is the significant measure to represent the classifier performance on the positive class. The high value of F-Measure represents that the classifier performs well on required positive class. From the Table 4, it was found that the F-Measure values of the proposed model are high this in turn reveals that the proportions of correctly classified positive instances were high when compared with other existing methods.

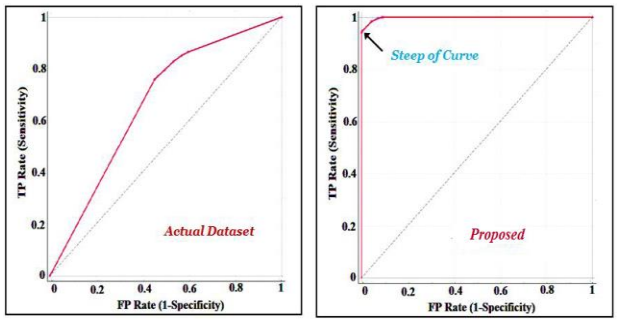
Table 4: F-Measure values of compared methods

Dataset	SMOTE	Easy Ensemble	Hybrid SVM	Proposed
Cmc2	0.4631	0.5267	0.5491	0.8436
Glass7	0.8539	0.8965	0.8867	0.9154
Abalone7	0.425	0.5913	0.5946	0.9221
Vowel	0.7769	0.8431	0.8752	0.9362
Yeast	0.6281	0.7038	0.7234	0.9344
Letter4	0.5752	0.6076	0.6614	0.9625

ROC Performance Evaluation of proposed model

The Receiver Operating Characteristic (ROC) Curves are the visualizing and summarizing technique of classifier performance. In ROC, point (0,0) states that classifier no way predicts positive instance, point (1,1) states that all instances are predicted as positive and point (0,1) is an ideal point that states that all positive instances are predicted as positive and no negative instances misclassified as positive. The Figure 1 shows the ROC curves of imbalanced and balanced Yeast dataset.

Figure 1: ROC Comparison of Yeast Dataset



The comparative analysis of ROC curves on the Figure 1 states that the performance of classifier is better on the dataset which was resampled by the proposed method. In addition, the steep of the curve also states that the number positive instances correctly classified are high and number of misclassified instances on negative class is very less when compared with actual dataset. This is the significant evidence which shows that; proposed hybrid model has greatest impact on the classifier performance in the context of imbalance presents in the relative distribution of each class.

CONCLUSION

In classification process, the issue of imbalance was not considered. Most of the existing studies tried to improve the classifier accuracy not by considering imbalance issue. Hence the performance of the classifiers built based on these studies is not fine in predicting minority class.

In this study, the Hybrid model was proposed to balance the dataset. The experiments were done with various datasets and the results were compared with the existing methods. The analysis shows that the proposed hybrid model outperforms the other existing methods. From this, it is concluded that nearest neighbors of the misclassified instances has the vital role in fine tuning to correctly classified instances. As next it is concluded that removing certain instances that has less classification information leads to reduce the proportion of oversampling required to balance the dataset. In addition it is also concluded that, the proposed hybrid model found to be more precious in the context where the relative distribution of the class variables are not in balance.

REFERENCES

[1]. Alitouche, T. A., & Bekkar, M., "Imbalanced data learning approaches review", International Journal of Data Mining & Knowledge Management Process, 3(4), 2013, 15.

[2]. Babu, S., & Ananthanarayanan, N. R. REMOTE: Enhanced Minority Oversampling TEchnique. Journal of Intelligent & Fuzzy Systems, 33(1), 2017, 67-78.

[3]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, 16, 2002, 321-357.

[4]. Haiyang, Z. "A short Introduction to data mining and its applications", 2011.

[5]. Jeatrakul, P., Wong, K. W., & Fung, C. C., "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm", In International Conference on Neural Information Processing, Springer Berlin Heidelberg, 2010, 152-159.

[6]. Rahman, M. M., & Davis, D. N."Addressing the class imbalance problem in medical datasets", International Journal of Machine Learning and Computing, 3(2), 2013, 224.

[7]. Wang, Q. A hybrid sampling SVM approach to imbalanced data classification. In Abstract and Applied Analysis,Hindawi, 2014.

[8]. Yen, S. J., & Lee, Y. S. "Cluster-based under-sampling approaches for imbalanced data distributions". Expert Systems with Applications, 36(3), 2009, 5718-5727.

[9]. Zorkeflee, M., Din, A. M., & Ku-Mahamud, K. R "Fuzzy and Smote Re-sampling Technique For Imbalanced Data Sets", Proceedings of the 5th International Conference on Computing and Informatics, ICOCI2015, Istanbul, Turkey, University Utara Malaysia., 2015.