Performance of Synthetic Minority Oversampling Technique on Imbalanced Breast Cancer Data

***K. Usha Rani**
Professor, Dept. of Computer Science
Sri Padmavati MahilaVisvavidyalayam
Tirupati– 517502
Email Id:usharani.kuruba@gmail.com

G. Naga Ramadevi
Research Scholar, Dept. of Computer Science
SPMVV
Tirupati – 517502
Email Id: ramadeviabap@yahoo.co.in

D. Lavanya
Professor, Dept. of Computer Science
SEAGI
Tirupati - 517502
Email Id: lav_dlr@yahoo.com

Abstract - Classification, one of the important data mining techniques plays an important role in classifying the data for knowledge discovery. The data sets contain some redundant and unnecessary attributes which mislead the classifiers. Feature Extraction techniques such as Principal Component Analysis (PCA) can be used to overcome high-dimensionality problem. Class imbalance is another problem which leads difficulty in learning for classifiers. Hence imbalanced datasets should be restructured using theresampling techniques. Synthetic Minority Oversampling Technique (SMOTE) can control the number of examples and distribution to achieve the purpose of balancing the dataset through synthetic new examples. In this study the performance of SMOTE with five classifiers on breast cancer data sets are analyzed.

Keywords -Breast Cancer Data, Classification, Class imbalance, PCA, SMOTE,

NOMENCLATURE

BC	- Breast cancer
DT	- Decision Tree
k-NN	- k-Nearest Neighbor
LR	- Logistic Regression
MLP	- Multi-Layer Perceptron
PCA	- Principal Component Analysis
PNN	- Probabilistic Neural Networks
RBFN	- Radial Basis Function Networks
RF	- Random Forest
SMOTE	- Synthetic Minority Oversampling Technique
SVM	- Support Vector Machine
TLGS	- Tehran Lipid and Glucose Study
WDBC	- Wisconsin Diagnostic Breast Cancer
WBC	- Wisconsin Breast Cancer
WPBC	- Wisconsin Prognostic Breast Cancer

I. INTRODUCTION

The size of data increases because of rapid development in technology and increase in number of users of technology. Medical data sets generally contain many features. Hence classification of data becomes complex because of unbounded size and imbalance nature of data. Dimensionality Reduction techniques helps to lessen the computational complexity as well as over fitting problem [1]. A data set is called imbalanced if one of the two classes having more samples than other classes. The most of algorithms are focusing on classification of major sample rather than miscellaneous and minority samples. Even though the minority samples occur rarely they are very important in real world applications especially diagnosis of cancers. Hence imbalanced datasets should be restructured using there sampling techniques [2].

Resampling is a process of drawing repeated samples from the given data or population. There are two different categories of resampling approaches: Under- sampling and Over-sampling. An oversampling technique, SMOTE is considered in this study as many researches worked on that.

For experimentation purpose only Breast Cancer Data sets are considered because it is the primary causes of death in women. In our previous study [3] a feature extraction technique, PCA is experimented and those results are considered here for comparison purpose. Here, further study is conducted with a resampling technique, SMOTE to study the performance of five popular diversified classification algorithms: k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), Logistic Regression (LR), C4.5 Decision tree algorithm and Random Forest (RF).

In section 2 related work is presented. Description of data sets is given in section 3. Proposed procedure with SMOTE is given in section 4. Section 5 and 6 contain experimental results and conclusion.

978-9-3805-4421-2/16/$31.00 ©2016 IEEE

II. LITERATURE SURVEY

NiteshV.Chawla et al [4] performed experiments on various imbalanced datasets available in the machine learning repository. The imbalanced datasets are restructured by means of SMOTE, an over sampling technique and classified the instances using the classifiers C4.5, Ripper and Naïve Bayes.

Lara Lusa et al [5] provided the capabilities of various classifiers to classify the high dimensional data. SMOTE method experimented to eliminate the biasedness towards the majority class and classified the data. The study illustrated that K-NN classifier is better than the other classifiers.

Danjuma et al [6] experimented the classifiers MLP, C4.5 and Naïve Bayes along with SMOTE for the prognosis of postoperative life expectancy in the Lung Cancer Patients.

NiteshV.Chawla et al [7] analyzed the performance of the classifier RIPPER with SMOTE on the datasets –Mammography, Sat image, Phoneme and KDDCup-99 Intrusion datasets.

Ramezankhani et al [8] evaluated the performance of SMOTE using classifiers: PNN (Probabilistic Neural Networks), Naïve Bayes and decision tree to predict the diabetes disease in a study of TLGS (Tehran Lipid and Glucose Study).

Bumju bee et al [9] predicted the Fasting Plasma Glucose status to diagnose type 2 diabetes diseases by using the classifiers Naïve Bayes and Logistic Regression on the Korean Health and Genome Epidemiology study database. To reduce the bias of the model towards the majority class, Smote was adopted and classified the data.

LanAnhT.Nguyen et al [10] considered SMOTE for class imbalance problem and experimented using the classifier SVM on the five-benchmark datasets and three protein related datasets.

MehakNaib et al [11] compared the performance of seven classification algorithms- Naïve Bayes, Logistic, MLP, Bagging, PART, J48, Random Forest along with SMOTE to predict the primary Tumors.

Dongmei ZHANG et al [12] analyzed the performance of SVM classifier and SVM classifier with and without SMOTE on the 11 datasets of UCI machine learning repository. The results showed that for some datasets SVM with Smote performs better than SVM alone.

Poolsawad et al [13] classified the clinical data set using the classifiers MLP (Multi-Layer Perceptron), RBFN (Radial Basis Function Networks), SVM (Support Vector Machines), DT (Decision Tree) and RF (Random Forest). Subsequently, data set is over sampled with SMOTE and classified the clinical data. Results exhibited that SMOTE is better than the imbalanced dataset.

Garcia et al [14] compared the performance of Imbalanced data and over sampled data using SMOTE over the three different classifiers MLP, C4.5, K-NN (k=1) on 39 datasets of KEEL dataset repository. SMOTE outperformed than Imbalanced data.

J. S. Sanchez et al [15] provided a comparison of the classifiers C4.5, SVM on imbalanced data set, oversampled data by using SMOTE with 50% and 75% percentages to oversample the instances of minority class. SVM performed better than C4.5 on imbalanced and on oversampled data.

III.DATA SETS DESCRIPTION

In this study for the analysis of classifiers performance only Breast Cancer data sets are considered as breast cancer is a leading cause of death in Women in the world as well as in our country. Four different breast cancer data sets Breast cancer (BC), Wisconsin Diagnostic Breast Cancer (WDBC), Wisconsin Breast Cancer (WBC) and Wisconsin Prognostic Breast Cancer (WPBC) are considered which are publicly available at UCI Machine learning Repository. The description of the data sets is given in Table 1.

TABLE 1: DESCRIPTION OF BREAST CANCER DATASETS

Data Sets	No. of Instances	No. of Attributes	No. of Instances under each Class	
			Major Class (-ve)	Minor Class (+ve)
BC	286	10	NR 201	R 85
WBC	699	10	B 458	M 241
WDBC	569	32	B 357	M 212
WPBC	198	34	NR 151	R 47

NR - No Recurrence **B - Benign**
R - Recurrence **M – Malignant**

IV. SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

SMOTE (Synthetic Minority Oversampling technique) [16] is a variation of oversampling technique which adds new synthetic instances to the minority class instead of replicating the existing minority instances. Synthetic examples were generated based on the feature space rather than the data space.

In the literature SMOTE is applied on the data sets with various over sampling percentages such as 50%, 75% and so on to increase the minority class so as to balance the class distribution in the data set. But here we developed a procedure such that instead of blindly applying the oversampling percentage, care is taken that minority class should not dominate majority class and at maximum it should reach the majority class because our intension is to apply SMOTE to balance the data set. Therefore based on the number of instances in a particular data set the algorithm is developed such that the minority class instances to be equal (approximately) to the majority class instances by adding synthetic instance by applying various oversampling percentages.

The amount of oversampling is maximized up to 100% or wherever the number of minority samples equals (nearly) to number of majority samples. The amount of oversampling is taken maximum up to 100% only because more than that percentage of oversampling may over fit the classifier.

Proposed Procedure:-

For each minority class sample

1. Find the k-nearest neighboring instances.
2. Randomly choose j neighbors among the k nearest neighboring instances. j is the value which depends on the amount of oversampling required.
3. Generate randomly, synthetic instances along the lines joining the minority sample and its j selected neighbors.
4. These steps to be repeated for 50%, 75%, 100% or whenever minority class instances to be equal (approximately) to the majority class instances.

Synthetic Sample generation:
For the continuous features:

- Compute the difference between the randomly chosen minority sample and its nearest neighbor.
- Multiply the difference by a random number between 0 and 1.
- Add the obtained value to the sample chosen.

Above steps can be represented simply using an equation:

$E_{new}= E_i+(E_j-E_i)\ \Delta$. E_{new} denotes new synthetic minority example. E_i is the randomly chosen minority instance in the dataset. E_j is the neighbor instance among the k nearest neighbors. Δ is a randomly chosen constant in the interval [0, 1].

For the nominal features:

Obtain the majority vote between the feature vector which is under consideration and its k-nearest neighbors for the nominal feature value. If the tie happens choose randomly.

After applying various SMOTE percentages the number of minority instances in the datasets will increase because of addition of synthetic samples. The details are given in Table 2 along with number of instances in original (imbalanced) datasets.

V. EXPERIMENTAL RESULTS

In this study for the analysis of classifiers performance on breast cancer datasets experiments are conducted with 10-fold cross-validation using open-source data mining tool WEKA (3.6.0).After preprocessing the data to handle missing values the experiments are conducted with five classifiers: k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), Logistic Regression (LR), C4.5 Decision tree algorithm and Random Forest (RF) and their performance with PCA and various percentages of SMOTE are tabulated in table 3.

For BC data set, KNN with 100% SMOTE and RF with 50% SMOTE have better accuracy and SVM, LR and C4.5 with PCA have higher accuracy rates.

For WBC data set, classifiers with SMOTE are having higher accuracy than with PCA.

For WDBC data set, except LR other classifiers are better performed with SMOTE.

For WPBC data set, KNN and C4.5 with SMOTE 100% are having better classification rate whereas SVM, LR and RF with PCA are having higher accuracies.

VI. CONCLUSION

In this study to overcome high dimensionality and class imbalance problems, PCA and SMOTE techniques are experimented with diversed five classifiers on various breast cancer data sets and the results are analyzed. The combination of these techniques will be considered for future study.

TABLE 2: DATASETS – NUMBER OF INSTANCES AFTER APPLYING VARIOUS SMOTE (VARIOUS PERCENTAGES)

Dataset	SMOTE		Class	No. of patients	Total No. of Instances
BC	Imbalanced		NR	201	286
			R	85	
	SMOTE	50%	NR	201	328
			R	127	
		75%	NR	201	349
			R	148	
		100%	NR	201	371
			R	170	
WBC	Imbalanced		B	458	699
			M	241	
	SMOTE	50%	B	458	819
			M	361	
		75%	B	458	879
			M	421	
		90%	B	458	915
			M	457	
WDBC	Imbalanced		B	357	569
			M	212	
	SMOTE	50%	B	357	675
			M	318	
		68%	B	357	703
			M	356	
WPBC	Imbalanced		NR	151	196
			R	47	
	SMOTE	50%	NR	151	221
			R	70	
		75%	NR	151	233
			R	82	
		100%	NR	151	245
			R	94	

TABLE 3: COMPARISON OF CLASSIFIERS ACCURACIES (%) WITH PCA, WITH SMOTE ON DATASETS

Classifiers	A	B	C	D
BC DATASET				
K-NN	64.68	73.17	71.06	**74.39**
SVM	**71.32**	66.46	68.76	67.92
LR	**69.93**	63.72	63.32	63.61
C 4.5	**72.37**	66.16	66.47	68.73
RF	69.58	**71.04**	69.91	70.08
WBC DATASET				
K-NN	95.85	95.60	96.47	**96.94**
SVM	96.56	97.19	97.26	**97.71**
LR	96.56	96.95	**97.26**	97.16
C 4.5	95.56	95.48	95.79	**96.17**
RF	96.56	96.82	97.04	**97.16**
WDBC DATASET				
K-NN	93.67	95.70	**96.07**	_
SVM	96.84	**97.92**	97.62	_
LR	**97.54**	94.37	96.64	_
C 4.5	93.84	**94.52**	94.12	_
RF	96.31	96.74	**97.47**	_
WPBC DATASET				
K-NN	66.66	70.14	70.81	**73.46**
SVM	**76.26**	73.75	75.54	75.51
LR	**79.29**	72.39	78.54	77.14
C 4.5	72.73	70.13	72.10	**75.51**
RF	**80.30**	78.28	77.25	75.92

A - with PCA
B - with SMOTE 1
C - with SMOTE 2
D - with SMOTE 3

SMOTE 1 – 50% for all datasets.
SMOTE 2 – 75% for BC, WBC, WPBC datasets and 68% for WDBC datasets.
SMOTE 3 – 100% for BC, WPBC datasets and 90% for WBC dataset.

References

[1] http://en.wikipedia.org/wiki/Dimensionality_reduction.

[2] VaishaliGanganwar. An overview of Classification algorithms for imbalanced datasets.*International Journal of Emerging Technology and Advanced Engineering*.Volume 2, Issue 4, April 2012.

[3] G. Naga Ramadevi, K. Usha Rani, D. Lavanya. Importance of Feature Extraction for Classification of Breast Cancer Datasets – A Study, *International Journal of Scientific and Innovative Mathematical Research*, Vol. 3, Special Issue 2,763-368, July 2015.

[4] Nitesh V. Chawla , E. Fowler Ave , W. Bowyer , Lawrence O. Hall, W. Philip Kegelmeyer, " SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research , 321–357, 2002.

[5] Lara Lusa, RokBlagus," Class prediction for high-dimensional class-imbalanced data", BMC Bioinformatics.; 11: 523, 2010.

[6] Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients", International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.

[7] Nitesh V. Chawla1, AleksandarLazarevic2, Lawrence O. Hall3, Kevin Bowyer4, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting". *Proceedings of the Principles of Knowledge* Discovery in Databases, PKDD-2003.

[8] Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D," The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes", Medical decision making, 2014.

[9] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim, "Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing Type 2 Diabetes", *IEEE Journal of Biomedical And Health Informatics*, Vol. 18, No. 2, March 2014.

[10] Nguyen, L.A.T., Dang, X.T., Le, T.K.T. Saethang, T., Tran, V.A., Ngo, D.L., Gavrilov, S. Nguyen, N.G., Kubo, M., Yamada, Y. and Satou, K,"PredictingBeta-Turns and Beta-Turn Types Using a Novel Over-Sampling Approach", *Biomedical Science and Engineering*, 7, 927-940 . 2014.

[11] MehakNaib, AmitChhabra, "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining", *International Journal of Computer Applications* (0975 – 8887) Volume 96– No.8, June 2014.

[12] Dongmei ZHANG† , Wei LIU, Xiaosheng GONG, HuiJIN ," A Novel Improved SMOTE Resampling Algorithm Based on Fractal ", *Journal of Computational Information Systems* 7: 6, 2011.

[13] N. Poolsawad, C. Kambhampati, and J.G.F. Cleland, "Balancing Class for Performance of Classification with a Clinical Dataset", *Proceedings of the World Congress on Engineering*Vol I, WCE 2014, July 2 - 4, 2014.

[14] V. García · J. S. Sánchez · R. Martín-Félez · R. A. Mollineda., "Surrounding neighborhood-based SMOTE for learning from imbalanced data sets", Progress in Artificial Intelligence, 2012.

[15] J. S. Sanchez, V. Garcia and R. A. Mollineda, "Exploring Synergetic Effects of Dimensionality Reduction and Resampling Tools on Hyperspectral Imagery Data Classification", MLDM'11 Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition,Pages511-523 , Springer-Verlag Berlin, Heidelberg, 2011.

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