A Classification Model for Imbalanced Medical Data based on PCA and Farther Distance based Synthetic Minority Oversampling Technique

NADIR MUSTAFA
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

Engr. Raheel A. Memon
Assistant Professor Computer Science
Sukkur Institute of Business Administration
Airport Road, Sukkur 65200, Sindh, Pakistan

JIAN-PING LI
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

Mohammed Z. Omer
School of Computer Science and Engineering
University of Electronic Science and Technology of China,
Chengdu, 611731, China

***Abstract*—Medical data are extensively used in the diagnosis of human health. So it has played a vital role for physicians as well as in medical engineering. Accordingly, many types of research are going on related to this to have a better prediction of the diseases or to improve the diagnosis quality. However, most of the researchers work on either dimensionality space or imbalanced data. Due to this, sometimes one may not have the accurate predictions or classifications of the malignant diseases as both the factors are equally important. So it still needs an improvement or more work required to address these biomedical challenges by combing both the factors. As such this paper proposes a new and efficient combined algorithm based on FD_SMOTE (Farther Distance Based on Synthetic Minority Oversampling Techniques) and Principle Component Analysis (PCA), which successfully reduces the high dimensionality and balances the minority class. Finally, the present algorithm has been investigated on biomedical data and it gives the desired results in terms of dimensionality and data balancing. Here, In this paper, the quality of dimensionality reduction and balanced data has been evaluated using assessment metrics like co-variance, Accuracy (ACC) and Area Under the Curve (AUC). It has been observed from the numerical results that the performance of the algorithm achieved the best accuracy with metrics of ACC and AUC.**

Keywords*—*Principle Component Analysis; Information Gain; farther Distance based Synthetic Minority Oversampling; Correlation based Feature

I. Introduction

Classification is an important task of machine learning and data mining. Classification modeling is to learn a function from training data, which makes as few errors as possible when being applied to data previously unseen. A large number of classification algorithms have been developed and used with medical applications, due to its importance for physicians in the diagnosis. Many researchers have been done to discuss the great challenges of the medical data. Imbalance class is the main challenge that influences to the classification of the medical data. In many cases, the nature of medical data follows the skewed distribution. Its instances in the majority and minority classes are not equality represented [1, 2]. Hence, the medical data becomes imbalanced when its majority class has a larger number of instances. With the traditional classification algorithms obtain a higher accuracy over majority while Versa with minority class. For this reason, new techniques and methods for dealing with class imbalance have been proposed [9]. These techniques can be classified into three methods: those that amend the data distribution by resampling techniques (data level methods) [11], and those at the level of the learning algorithm which adapt a base classifier to deal with class imbalance (algorithm level methods), and those at the features selection level which find an optimal features among the whole the features. In this paper, we proposed a combined solution to classify imbalanced data, which successfully reduces dimensionality, and balances the minority class using a combination of Principle Component Analysis (PCA) and Synthetic Minority Oversampling Techniques. The innovation of this proposal is the joint utilization of both (PCA) and FD_SMOTE techniques, which achieved superior results in our experiment. In this paper, the quality of dimensionality reduction and balanced data has been evaluated using assessment metrics like Co-variance, Accuracy (ACC), and Area Under the Curve (AUC). It has been observed from the numerical results that the performance of the algorithm achieved the best accuracy with metrics (ACC) and (AUC). Finally, the FD_SMOTE technique has been investigated on biomedical data, and it realized the desired results in terms of dimensionality and data-balancing.

This paper is organized as follows. In Section 2 background of the present study with the literature review has been presented. After that in Section 3 existing approaches have been discussed. Next in Section 4, a new method has been proposed with experimental analysis. Lastly Section 5 includes the conclusion part.

II. Backgrounds

Imbalanced data is the most important issue in all applications of the real world, and the classification accuracy based on minority class can get a higher priority than that majority class, so it is a significant work to enhance the classification precision of minority class. In this section, we will explain the basic concept of the problem and the associated solution.

A. *Imbalanced Data Problem*

Sun et. al stated that the most understandable problem in data set is the imbalance data distribution between classes [10]. Nevertheless, the earlier studies and research stated that the imbalanced data distribution is not only the main issue that reduces the performance of the existing classifiers in specifying rare samples. The other influential issue of the classifier performance is small samples size, separability and the existence of within-class.

B. *Presented Approach of Imbalanced Data Problem*

There are different approaches have been presented to tackle the imbalance class problem [7], [8,] [9], which can be categorized as a resampling approach, algorithms approach and features selection approach.

- The preprocessing approach is a combination of over-sampling technique and under-sampling technique. The Oversampling is a powerful method used to add new samples, while under-sampling is a process of removing existing samples. These techniques mostly fix the imbalance data by generating or updating some of the classifiers algorithms. The classification algorithm should include the cost sensitivity, recognition-based approaches, and kernel-based learning techniques, which perfectly provide an acceptable solution for the imbalanced data problem. The support vector machine SVM is one of the most popular algorithms that embed the previous techniques [9]. Due to a large amount of bio-medical data and class imbalance ratio, applying the algorithm alone is not a good idea. Hence new hybrid approaches are required as a combination of sampling techniques and algorithms [10].

- The algorithms approach is the most popular technique that has been used to fix the imbalanced data problem, which is the bias towards the majority class and ignoring the minority class. The correct classification of the minority class gives a better accuracy, while in many applications, misclassification of minority class results in serious problems [11]. The inaccurate classification of the benign disease leads to additional diagnosis, while the inaccurate classification of malignant disease puts the human life at serious risk. Therefore, most of the machine learning algorithms tries to enhance the inaccurate classification of the minority class.

- The feature selection approach has been presented as a good solution for bio-medical data with a large amount. The size of this data can be reduced to a lower space dimension using linear transformation or non-linear transformation which is used based on its linearity nature. Imbalanced data on minority class and high dimensionality problem causes a misclassification. This misclassification of entities that have the same attribute value could disturb the diagnoses of diseases. For example, the boundaries between a malignant headache and a brain tumor could be vague under some circumstances, which is obviously catastrophic. Therefore, it is not easy for the medical doctors to examine the abnormalities in human in the misclassified data. The hybridized of reduction dimensionality and balance data technique is necessary in most bio-medical applications in order to enhance and recover misclassifications details that may be hidden in the data [3][4].

III. The Proposed Method

The proposed method provides an accurate classification model by using a combination of the PCA and SMOTE technique. The PCA is used to reduce the high dimensionality of data by select an optimal feature from the original data set. The PCA generate a new dimension space of the data which implemented with the FD_SMOTE to balance the data of the minority class, while the imbalanced data split into train and test data, and then the balanced data applied to the different classifiers to achieve the better classification for the medical data.

A. *Principle Component Analysis*

In the proposed model the features selection is used as the key technique to find a subset of optimal features from the original data. The extracted features allow the classifier to achieve the best accuracy. Here, PCA to reduce the high dimensional point into lower dimensional point and then using filters to order the importance of the selected attributes based on a rule [5]. In this model, the dimensionality reduction has been implemented based some metrics such as mean, co-variance, eignvalue and Eigenvectors to compute the principle component. Finally, the PCA provide a new transform of PCs which generated by using correlation matrix of the data to find the best PCs among all the features. These steps well explained in the algorithm 1.

$$C = \frac{1}{N}\sum_{j=1}^{N}\varphi_j\varphi_j^T = \rho\rho^T \tag{1}$$

$$\rho = \left(\varphi_1\varphi_2\cdots\varphi_j\right) \tag{2}$$

$$\varphi = \upsilon_j - \mu \tag{3}$$

$$\mu = \frac{1}{M}\sum_{i=1}^{M}\upsilon_i \tag{4}$$

Where v_i is a vectors from the original dataset X_i, and μ is mean of Jth vectors of the data, where φ is a variance of the vectors that subtracted from mean, and Then C is a co-variance matrix which generated by multiplication of variance with its variance transpose as $\varphi \times \varphi^T$. Finally, the eignvalue λ and Eigenvectors υ can be easily substituted according to the co-variance matrix C to achieve new principle component.

B. Farther Distance based SMOTE

The SMOTE technique provides an optimal solution for imbalanced data distribution problem based on oversampling technique. The basic assumption of the SMOTE based on how to find the similarities of the feature among the minority class instance. The assumption is achieved by calculating the centriod [c] of the minority class sample and the distance [di] between all the minority sample and its centriod, then compute the average [avg] of distance matrix and the seed sample represented as a farther distance to the class center [c] and greater than the average distance [avg]. The new synthetic sample has been generated randomly by select one of the N-centriod, then multiply the difference between the seed sample and centriod with a random number σ between [0, 1] and then added to the original seed. Finally, the mathematical steps of the algorithm illustrated as follows:

$$c = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5}$$

$$d_i = \left(y_i - c\right) \tag{6}$$

$$avg = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{7}$$

$$Ss = \left\{ y_i \middle| d_i > avg \right\} \tag{8}$$

$$nss = Ss_i + (Ss_i - c) \times \sigma \tag{9}$$

The FD_SMOTE work on creation of new examples instead of duplicating the minority class samples, as shown in Figure 1, the new “synthetic” examples are being created in the neighborhood of minority classes. Where the synthetic examples are generated operating in “feature space” rather than operating in “data space”. Along the line segment, each minority class has been taken and introducing synthetic examples to join all minority class nearest neighbors. The numbers of required synthetic example vary situation to situation so according to the requirement the numbers of k minority classes are chosen to generate the nearest neighbor synthetic example. Finally, the pseudo code the proposed method illustrated as in algorithm 2.

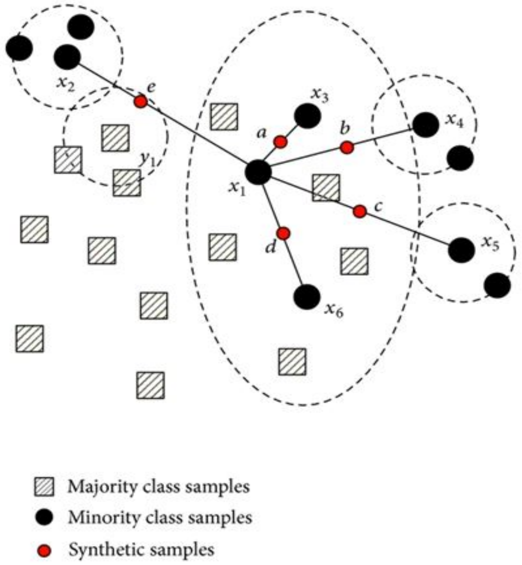

Fig. 1. FD_SMOTE Technique

Algorithm 1. Principle Component Analysis

```
    Input: Original data set {Xi | i = 1, 2, . . . , m}, which each sample has m attributes without decision attribute.
    Output: Principle Component {Yi | i = 1, 2, . . . , n},
    Victories the data into Vi ……. Vm
    for $j \to n$ do jth is all vectors
      for $i \to m$ do ith instances of Vi
        Compute the mean according to Eq.(7)
        Subtract the instances according to Eq.(6)
      end for
        Multiply the variance according to Eq.(5)
        Compute the convince according to Eq.(4)
    end for
    Compute the eignvalue $\lambda$ according to Eq.(4)
    Compute the eigenvectors $\upsilon$ according to Eq.(4)
    Output new Principle Component of features
```

Algorithm 2. FD_SMOTE resampling

```
    Input: Origin set of minority, Dmin = {Yi | i = 1, 2, . . . , n}, the balance factor $\sigma$
    Output: New et of minority, Dmaj = {Zi | i = 1, 2, . . . , m}
    Compute c , $d_i$ and avg according to Eqs. (5), (6) and (7)
    Create seed sample according to Eq. (8)
    for $i \to \sigma$ do
      fr $i \to m$ do
      Generate random number γ
         Generate new sample y according to Eq.(9)
       end for
    end for
    Output new set of minority
```

IV. Experimental Analysis

A. Collected Data

TABLE 1. Provide the characteristic of the data used in this work, which describe the name, number of features and the number of instances of the data. Its provides a different kind of the size and level of imbalance data. Also, these data are inspired from biomedical domains some of which are proprietary. Pima diabetes, Breast cancer and Thyroid disease (which contain a binary class) are all available through the UCI repository [1].

TABLE I. Data Characteristics

no	Name	Instances	Features
1.	Pima diabetes	768	9
2.	Breast cancer	699	11
3.	Thyroid disease	3163	27

B. ACC Evaluation Measures

The confusion matrix is most powerful metrics that assess the performance of machine learning algorithm as shown in TABLE 2. The confusion matrix categorized into columns and rows that describe the prediction class and actual class respectively. The confusion matrix parameters are used to show the accuracy the classification algorithm. These four parameters are classified as follows TN (True Negatives), FP (False Positives), FN (False Negatives) and TP (True Positives). The positive instance most of them correctly classified, and the rest incorrectly classified. Furthermore, the negative instance most of them correctly classified, and the rest incorrectly classified. Generally, the equation of the classification accuracy or the prediction accuracy is calculated as illustrated in the following formula 6.

$$Acc = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{10}$$

In term of the imbalanced data there two metrics are used as equal error costs and unequal error costs respectively. The error rate (Er) is calculated as most important tool that used to investigate the performance of these metrics, which calculated as illustrated in the formula 7.

$$\mathrm{E}_r = 1 - accuracy \tag{11}$$

For the existence of the imbalanced data with unequal error cost, the area under the curve (ROC) is the most suitable metric used to tackle the imbalance data problem. There are similar techniques are presented by (Ling & Li, 1998; Drummond & Holte, 2000; Provost & Fawcett, 2001; Bradley, 1997; Turney, 1996). Finally, many works are presented with the term of ROC which supports the study of decision boundaries or relative costs of TP and FP. ROC metrics is coordinated on two axis as X-axis and Y-axis to calculate the %FP = FP/ (TN+FP) of X-axis and %TP = TP/ (TP+FN) of Y-axis respectively. The ROC provide a better performance on the point (0,100), which explain the correct instance and incorrect instance of the positive and negative class.

TABLE II. Confusion Matrix

		Prediction	
		Predicted Negative	***Predicted Positive***
Actual	***Actual Negative***	TN	TN
	Actual Positive	FN	TP

C. AUC Evaluation Measures

The ROC curve can be easily shifted by manipulating the balance of training instance for each class in the training set. Area under the ROC Curve (AUC) is a helpful measure for classifier performance as it is independent of the decision criterion specified sand previous probabilities. The AUC comparison can create a strong relationship between classifiers. If the ROC curves are overlapping, the total AUC is a mean comparison among the models (Lee, 2000). But, for certain cost and class distributions, the classifier have highest AUC may reality be sub-optimal. Thus, we also calculate the ROC convex hulls, since the points lying on the ROC convex hull are possibly ideal (Provost, Fawcett, & Kohavi, 1998; Provost & Fawcett, 2001).

The Classification Performance of FD_SMOTE technique with different percentages can be observed in the Tables 1, 2 and 3. Here it can observe from the all the tables the representation of the rows or classes in the dataset, the SMOTE technique analyze the percentage (%) of the majority and minority class for all three datasets. The majority represents the patients who are not affected by a disease and their features need to model. So to balance the minority class that requires increasing the minority sample by setting the percentage of SMOTE technique in multiples of 100 as follows:

TABLE III. SMOTE (%) of PIMA Diabetic

SMOTE (%)	**Majority Class**	**Minority Class**	**Total**
SMOTE % = 0	500 66%	268 34%	768
SMOTE % = 100	500 48%	536 52%	1036
SMOTE % = 200	500 38%	723 62%	1305

TABLE IV. SMOTE (%) of Breast Cancer

SMOTE (%)	**Majority Class**	**Minority Class**	**Total**
SMOTE % = 0	458 65%	241 35%	699
SMOTE % = 100	458 49%	482 51%	940
SMOTE % = 200	458 39%	723 61%	1181

TABLE V. SMOTE (%) of Thyroid Disease

SMOTE (%)	Majority Class	Minority Class	Total
SMOTE % = 0	2559 81%	604 19%	3163
SMOTE % = 100	2559 68%	1204 32%	3767
SMOTE % = 200	2559 58%	1812 42%	4371
SMOTE % = 300	2559 58%	2416 49%	4975

The Performance evaluation of Pima diabetes data classification using FD_SMOTE technique can be observed in the tables 5 and 6. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 5 and 6 shown that the ACC, AUC metrics generated with PCA and FD_SMOTE technique are better than the ACC metrics that based feature (CFs) and information gain (InfoGs) technique in all classifiers methods. It reveals that the AUC metrics in all biomedical data is higher than other metrics.

TABLE VI. ACCURACY RESULT OF PIMA DIABETIC

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	88.1771	76.4323	76.7375
SVM	91.0156	71.0425	75.3906
N Neighbor	92.9863	76.0618	73.9583
Bagging	90.6094	74.0885	75.6510
Random Forest	91.8698	74.8698	72.7865
Naïve Bayes	89.6094	76.3672	74.8698

TABLE VII. AUC RESULT OF PIMA DIABETIC

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.998	0.723	0.815
SVM	0.971	0.719	0.827
N Neighbor	0.963	0.741	0.804
Bagging	0.989	0.805	0.820
Random Forest	0.997	0.812	0.800
Naïve Bayes	0.984	0.823	0.813

Figs. 3 and 4 illustrate the relationship of AUC and ACC of all classifiers algorithms for Pima diabetes classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

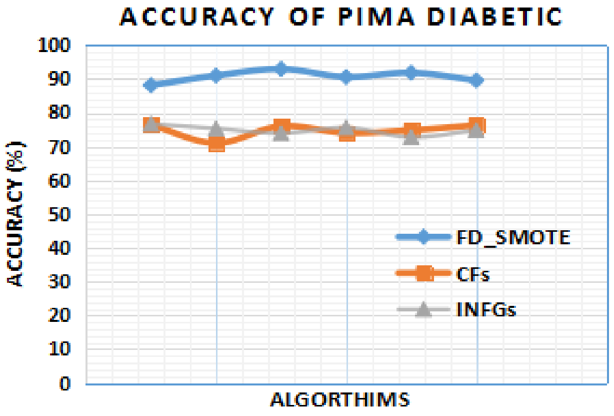

Fig. 2. ACC result of FD_SMOTE, CFs and InfoGs

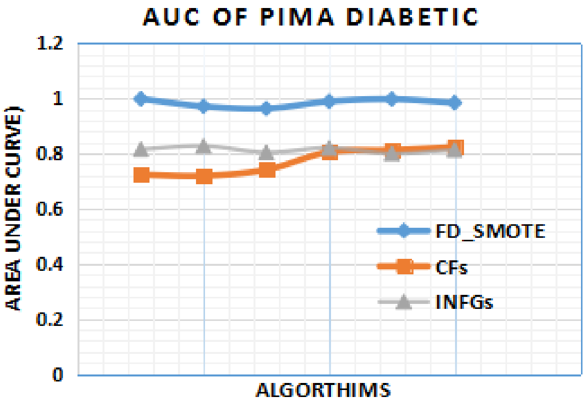

Fig. 3. AUC result of FD_SMOTE, CFs and InfoGs

The Performance evaluation of breast cancer data classification using FD_SMOTE technique can be observed in the tables 7 and 8. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 7 and 8 shown that the ACC, AUC metrics generated with SMOTE technique are better than the ACC metrics that generated based feature (CFs) and information gain (InfoGs) techniques in all classifiers methods. It reveals that the AUC metrics in all biomedical data is higher than other metrics.

TABLE VIII. ACC RESULT OF BREAST CANCER

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	93.8072	81.4235	74.4206
SVM	96.6809	82.9957	86.4235
N Neighbor	95.6809	80.1373	85.9943
Bagging	94.7340	86.2804	75.9943
Random Forest	89.8404	79.7082	75.4220
Naïve Bayes	92.1184	82.1373	90.7082

TABLE IX. AUC RESULT OF BREAST CANCER

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.847	0.555	0.555
SVM	0.795	0.577	0.551
N Neighbor	0.759	0.581	0.535
Bagging	0.893	0.561	0.563
Random Forest	0.881	0.595	0.566
Naïve Bayes	0.894	0.586	0.571

Figs. 5 and 6 illustrate the relationship of AUC and ACC of all classifiers algorithms for breast cancer classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

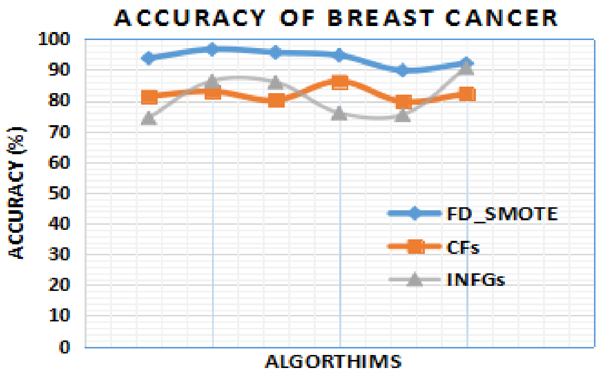

Fig. 4. AUC result of FD_SMOTE, CFs and InfoGs

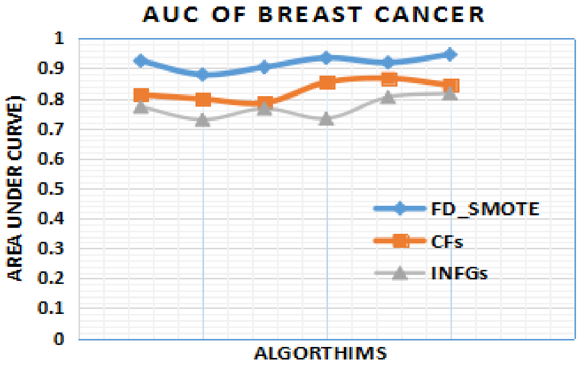

Fig. 5. AUC result of FD_SMOTE, CFs and InfoGs

The Performance evaluation of medical thyroid disease data classification using FD_SMOTE technique can be observed in the tables 9 and 10. From the relationship of the accuracy (ACC), area under the curve (AUC), here the Table 9 and 10 shown that the ACC, AUC metrics generated with SMOTE technique are better than the ACC metrics that based feature (CFs) and information gain (InfoGs) techniques in all classifiers methods. It reveals that the AUC metrics in all medical data is higher than other metrics.

TABLE X. ACC RESULT OF THYROID DISEASE

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	82.7228	56.2500	56.2500
SVM	84.1291	62.7315	65.2800
N Neighbor	77.1267	62.2685	58.7963
Bagging	84.1146	61.3426	64.3519
Random Forest	83.2176	66.2037	63.4259
Naïve Bayes	84.1291	59.9537	65.2778

TABLE XI. AUC RESULT OF THYROID DISEASE

Classifiers	FD_SMOTE	CFs	InfoGs
MultiPerceptron	0.925	0.812	0.772
SVM	0.879	0.798	0.729
N Neighbor	0.904	0.785	0.766
Bagging	0.935	0.853	0.733
Random Forest	0.919	0.867	0.804
Naïve Bayes	0.946	0.844	0.817

Figs. 7 and 8 illustrate the relationship of AUC and ACC of all classifiers algorithms for thyroid disease classification. Here it can be observed that ACC and AUC metrics of PCA combined FD_SMOTE technique has better results compared with correlation based feature (CFs) and information gain (InfoGs) techniques.

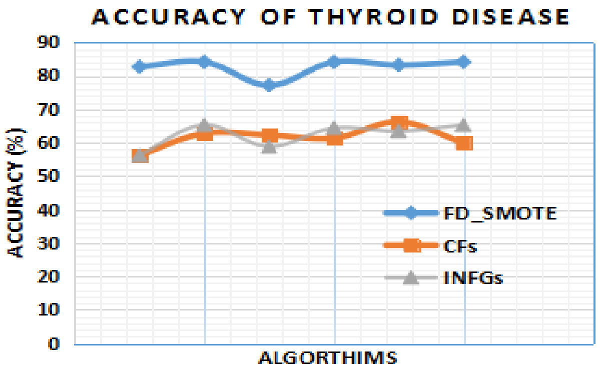

Fig. 6. AUC result of PCA and FD_SMOTE

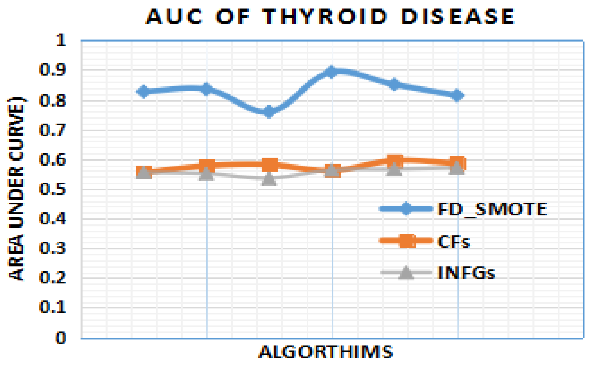

Fig. 7. AUC result of PCA and FD_SMOTE

V. CONCLUSIONS

In this paper a new algorithm has been proposed for generating an accurate classification of biomedical data. This aims to tackle the skewed data distribution and high dimensionality problem. The approach has been constructed by combing the PCA and FD_SMOTE based on farther sample. From the qualitative and quantitative analysis different classifiers based on PCA and FD_SMOTE has been used and it reveals that the new approach increases the performance of

(AUC) metrics and (ACC) metrics which used on a variety data of biomedical field. The present analysis shows that the combined technique is most effective than other existing approaches such as correlation based feature (CFs) and information gain (InfoGs). However the future plan is to investigate the present problem with rough set theory including the imbalanced data.

ACKNOWLEDGMENTS

This paper was supported by National Natural Science Foundation of China (Grant NO: 61370073), the national high technology research and development program of china (Grant No: 2007AA01z423).

REFERENCES

[1] Shuo Wang, Member, and Xin Yao, “Multiclass Imbalance Problems: Analysis and Potential Solutions”, IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

[2] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance”IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol.40, No. 1, January 2010

[3] Björn Waske, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance”IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol.40, No. 1, January 2010.

[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,“On the Class Imbalance Problem” Fourth International Conference on Natural Computation, 2008.

[5] Mike Wasikowski, Member and Xue-wen Chen, “Combating the Small Sample Class Imbalance Problem Using Feature Selection”, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.

[6] Rukshan Batuwita and Vasile Palade,“Fuzzy Support Vector Machines for Class imbalance Learning” IEEE Transactions On Fuzzy Systems, Vol. 18, No. 3, June 2010.

[7] Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh, “Class Imbalance Robust Incremental LPSVM for Data Streams Learning” WCCI 2012 IEEE World Congress on Computational Intelligence June, 10- 15,2012 - Australia.

[8] David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, ”Mine Classification With Imbalanced Data”, IEEE Geosciences And Remote Sensing Letters, Vol. 6, No. 3, July 2009.

[9] Mikel Galar,Fransico, “A review on Ensembles for the class Imbalance Problem: Bagging,Boosting and Hybrid-Based Approaches” IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42,No.4 July 2012

[10] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, , and Sven Krasser “Correspondence SVMs Modeling for Highly Imbalanced Classification” IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1, February 2009.

[11] Qun Song Jun Zhang Qian Chi “ Assistant Detection of Skewed Data Streams Classification in Cloud Security”, IEEE Transaction 2010.

[12] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz “Special Issue on Learning from Imbalanced Data Sets” Volume 6, Issue 1 - Page 1-6.

[13] S¸ eyda Ertekin1, Jian Huang, L´eon Bottou, C. Lee Giles “Active Learning in Imbalanced Data Classification”

[14] Saumil Hukerikar, Ashwin Tumma, Akshay Nikam, Vahida Attar “SkewBoost: An Algorithm for Classifying Imbalanced Datasets” International Conference on Computer Communication Technology (ICCCT)-2011.

[15] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, “Improving Learner Performance with Data Sampling and Boosting” 2008 20th IEEE International Conference on Tools with Artificial Intelligence.

[16] Benjamin X. Wang and Nathalie Japkowicz “Boosting Support Vector Machines for Imbalanced Data Sets” Proceedings of the 20th International Conference on Machine Learning-2009.

[17] http://www.ejpau.media.pl/volume17/issue3/art-03.html (Accessed on Jan 13, 2017).

[18] http://blog.sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting (Accessed on Jan 13, 2017).

[19] Beckmann, M., Ebecken, N.F.F. and de Lima, B.S.L.P. (2015) A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 7, 104-116.

[20] Hu, Y., Guo, D.F., Fan, Z.W., Dong, C., Huang, Q.H., Xie, S.K., Liu, G.F., Tan, J., Li, B.P. and Xie, Q.W.(2015) An Improved Algorithm for Imbalanced Data and Small Sample Size Classification. Journal of Data Analysis and Information Processing, 3, 27-33. http://dx.doi.org/10.4236/jdaip.2015.33004

[21] Beckmann, M., Ebecken, N.F.F. and de Lima, B.S.L.P. (2015) A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 7, 104-116.

[22] http://www.ejpau.media.pl/volume17/issue3/art-03.html (Accessed on Jan 13, 2017).

[23] http://blog.sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting (Accessed on Jan 13, 2017).