Dissertation Abstract

Classification of Biomedical Data with Class Imbalance

クラスインバランスがある生物医学データの分類

Division of Electrical Engineering and Computer
Graduate School of Natural Science & Technology
Kanazawa University

Kunti Robi'atul Mahmudah
StudentID Number: 1824042010

Abstract

Pre-processing is a crucial step before feeding the data into the machine learning model. Class imbalance problem can decrease the performance or giving the model bias results. A frequently used method to resolve this problem is through oversampling. In this study, we employ some oversampling methods and features extraction methods as the pre-processing task. In the case of Chronic obstructive pulmonary disease (COPD), we proposed methods that utilize gene expression data from microarrays to predict the presence or absence of COPD. Microarray data of the small airway epithelium cells obtained from 135 samples selected from GEO dataset. Machine learning and regression algorithms performed in this study included Random Forest, Support Vector Machine, Naïve Bayes, Gradient Boosting Machines, Elastic Net Regression, and Multiclass Logistic Regression. Classification algorithms of elastic net regression and multiclass logistic regression achieved high AUC score and the other metrics which outperformed the other classifiers. In the case of binary features data, by converting binary features into numeric using feature extraction methods prior to oversampling can fully display their potential in improving the classifier's performances. It is confirmed by the result that features extraction and oversampling are synergistically contributing to the improvement of classification performance.

keywords : *binary features, class imbalance, imbalanced data, feature extraction, oversampling, dimensionality reduction, COPD, biomedical data, classification*

1. Class Imbalance

Class imbalance datasets exist in many real-world data. Class imbalance happen when thenumber of a class is far less than that in the other one as can be seen in Figure 1.1 of PCA plot of *dis* dataset retrieved from [1]. The target class is usually the minority classor the class which has samples far less than the other one and a sample in this class iscalled as positive sample while a sample from the other one is called as negative sample.This problem can lead classifier to bias toward the majority class because it will mostlikely to predict a positive sample as negative sample. Therefore, a method to deal withclass imbalance should be done first before providing the data as an input to the classifiesto improve detection of minority sample or the performance metrics.

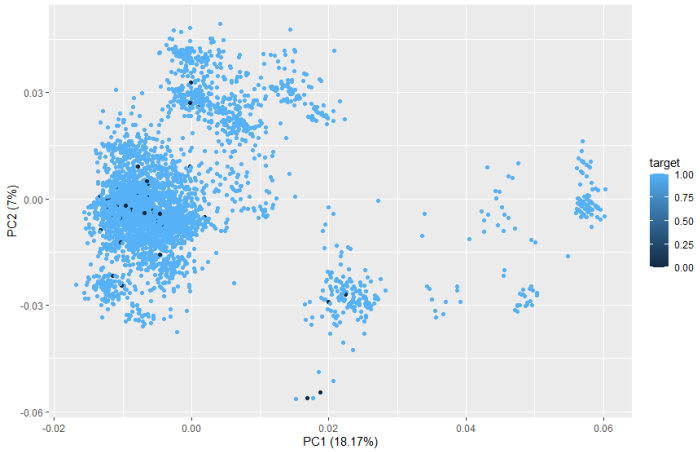

Figure 1. Class imbalanced data

Research on the class imbalance issue is critical in machine learning. These two factors outline why class imbalance problem should be handling prior to classification. The problem of class imbalance is pervasive across a wide range of important fields in the data mining community and when confronted with the problem of class imbalance, most common classification learning methods were shown to be insufficiently effective. There are many methods exist to deal with class imbalance at either in algorithm level or data level. This thesis shows methods of the data level on the classification of imbalanced data such as SMOTE, ADASYN, adaptive neighbor SMOTE, borderline-SMOTE, safe-level SMOTE, relocating-safe-level SMOTE, and DBSMOTE.

2. Binary Features Data

In the field of machine learning, it is important to understand the characteristics of input data and select the methods that are most suitable for achieving high performance in the machine learning task (regression, classification, clustering etc.). As mentioned

in the previous section of methods for dealing with class imbalance, many of the methods are specifically developed to overcome the problem in numerical data. In contrast, some types of data are represented as binary value that can take only one of two values (0/1, T/F, M/F, etc.). Such a binary feature is located on the border of numerical and categorical values. It can be treated as numerical value; however, the domain of value is quite poor and no difference from categorical value. Therefore, for a dataset consists of binary features, direct application of oversampling methods is not a good idea since such methods rely on numerically represented values for synthesizing new minority samples.

Typically, classification is conducted on a dataset that consists of numerical features and target classes. For instance, a grayscale image is usually represented as a matrix of integers varying 0-255 and it enables to apply various classification algorithms to image classification tasks. However, datasets represented as binary features are not so special and their amount is not negligible. On the other hand, oversampling algorithms such as SMOTE and its variation are often used if the dataset for classification is imbalanced. However, since SMOTE and its variant synthesize new minority samples based on the original samples, the diversity of the samples synthesized from binary features is highly limited due to the poor representation of original features. To solve this problem, a preprocessing approach is studied. By converting binary features into numerical ones using feature extraction methods, succeeding oversampling methods can fully display their potential. In this research, we study the method to handling class imbalance in binary features dataset by combining feature extraction and oversampling methods as the preprocessing task.

3. Methods to Deal with Class Imbalance

Imbalance datasets exist in many real-world data. Class imbalance happen when the number of a class is far less than that in the other one. The target class is usually the minority class or the class which has samples far less than the other one and a sample in this class is called as positive sample while a sample from the other one is called as negative sample. This problem can lead classifier to bias toward the majority class because it will most likely to predict a positive sample as negative sample. Therefore, a method to deal with class imbalance should be done first before feeding the data as an input to the classifies to improve detection of minority sample or the performance metrics. There exist many methods to deal with class imbalance at either in algorithm level or data level.

3.1. Algorithm Level

At algorithm level, a method is proposed to modify the algorithm either change or add another line of algorithm in order to solve the imbalance data issue. Ensemble learning and cost-sensitive learning are some examples of algorithm-level methods. Bagging and boosting are classic ensemble learning methods that have shown to be effective in dealing with class imbalanced problems. This method includes modifying the already available classification algorithms to make them suitable for imbalanced data sets.

In cost sensitive learning, a large misclassification cost is assigned to defective examples while a small misclassification cost is assigned to non-defective instances.

3.2. Data Level

At data level, there are some available form of re-sampling methods that were proposed. This method consists of over-sampling and under-sampling to re-balance the original dataset. The under-sampling method is done by reducing samples of the majority class whilst over-sampling method is done by adding samples of minority to the original dataset so that the new dataset become nearly balanced.

3.2.1. Down-sampling

Downsampling is a technique for reducing the number of training samples that belongs to the majority class. The majority class is subsampled to achieve basically the same number of sample as the minority class. For example, if we have a training data with 900 of it are negative samples and the remain 100 are positive sample. Down-sampling would randomly sample the first class so that we will have 100 samples of each positive and negative class.

The main drawback of downsampling if that we tend to lose a lot of important information when we remove the data.

3.2.2. Up-sampling

Up-sampling is a technique to increase the number of training samples of the minority class by randomly duplicating the minority class. Using this method, the numbers of both classes are (nearly) the same. As the same example of downsampling, after upsampling is done, in the final dataset, each positive and negative class will have a training data of 900 samples.

This balancing technique keeps the model from bias toward the majority class. The relationship (border line) between the class labels is also unaffected. Furthermore, because of the added samples, the upsampling method introduces bias into the system.

3.2.3. Under-sampling

Undersampling the majority class is one of the most popular and simplest approaches to handle imbalanced data. In order to effectively balance the class distribution, undersampling methods eliminate samples from the training dataset that belong to the majority class, such as lowering the skew from a 1:1000 to a 1:100, 1:10, 1:3, or even a 1:1 class distribution. There are some methods of undersampling such as random undersampling, near miss, condensed nearest neighbor, and Tomek Link that commonly used in the machine learning algorithm.

3.2.4. Over-sampling

On the contrary of undersampling methods, over sampling is used when the number of obtained data is insufficient. To balance the classes, oversampling approaches increase the number of objects in the minority class. There are some popular oversampling methods which most of them is developed based on SMOTE. In this study, seven methods of oversampling are explained.

Imbalance datasets exist in many real-world data. Class imbalance happens when the number of a class is far less than that in the other one. The target class is usually the minority class or the class which has samples far less than the other one, and a sample in this class is called as positive sample while a sample from the other one is called as negative sample. This problem can lead classifiers to bias toward the majority class because it will most likely to predict a positive sample as negative sample. Therefore, a method to deal with class imbalance should be done first before providing the data as an input to the classifies to improve detection of minority samples and the performance metrics. There are many methods exist to deal with class imbalance at either in algorithm level or data level. In the algorithm level, a method is proposed to modify the algorithm either change or add another line of algorithm to solve the imbalance data issue.

At data level, there are some available forms of resampling methods were proposed. This method consists of oversampling and undersampling to rebalance the original dataset. The under-sampling method is done by removing samples of the majority class, while over-sampling method is done by adding samples of minority to the original dataset so that the new dataset become nearly balanced.

Synthetic minority over-sampling technique (SMOTE) is known as the pioneer method for oversampling an imbalance dataset. The method uses k-nearest neighbors in creating new synthetic samples to balance the class distribution of the dataset [3]. Each positive sample is paired with its nearest neighbors, then along a line connecting the sample with one of the selected nearest neighbors, a synthetic sample is generated. Then, the process is repeated until the number of positive and negative samples becomes balanced. In dealing with class imbalance by creating new synthetic samples,

SMOTE is the pioneer methods where many methods in this issue were developed in order to improve the classification performances based on this technique.

ADASYN or adaptive synthetic sampling approach for imbalance learning is another technique to deal with class imbalance. The main idea of this method is to create synthetic samples based on the level of difficulty in learning the samples of the minority class [4]. A positive sample is called as difficult to learn if it has more negative samples in its k-nearest neighbors. Therefore, the harder a positive sample to learn, the more synthetic samples are generated.

Borderline SMOTE is a SMOTE-based technique concentrated on the borderline of each class [5]. It stated that a sample located on or near the borderline will give more contribution to classification. It highlighted the problem that the type of samples tends to be misclassified by classification methods than those located far from the borderline. From that reason, this technique tries to strengthen the borderline of positive samples by generating synthetic samples along this region. This idea of Borderline SMOTE in selecting certain region to generate synthetic samples inspired other SMOTE-based techniques in oversampling methods such as safe-level SMOTE and relocating safe level SMOTE.

Another improvement of SMOTE, namely safe-level SMOTE [6], highlighted the drawback of SMOTE that SMOTE naively ignore nearby majority instances in synthesizing the minority samples along a joining line of a minority samples and its selected nearest neighbors. This new generated synthetic sample by SMOTE cause classification model create larger and less specific regions resulting in overgeneralization. Therefore, safe-level SMOTE will determine whether a sample in minority class is safe to use in generating synthetic samples or not. In other words, Safe-Level SMOTE is designed to generate synthetic samples around selected positive samples that considered as safe.

Relocating-safe-level SMOTE (RSLS) is as an improvement technique of Safe-Level SMOTE [7]. It highlights the fact that safe-level SMOTE ignored the possibility that some synthetic samples are generated closer to negative samples than to positive samples. It contradicts with the procedure in generating synthetic sample of safe-level SMOTE where it tries to avoid negative samples in generating synthetic samples. Therefore, RSLS introduced an additional algorithm to relocate the generated synthetic samples if it locates around negative samples.

Another variation of SMOTE–based oversampling technique is Density-based minority over-sampling technique or DBSMOTE. It is an integration of DBSCAN and SMOTE [8]. DBSCAN or Density-Based Spatial Clustering of Application with Noise [9] aiming at discovering clusters of arbitrary shape. DBSMOTE aims at generating synthetic samples of an arbitrarily shaped cluster found by DBSCAN. Inspired by Borderline-SMOTE in maintaining the detection rate of majority class, DBSMOTE

focused the work on overlapping region. However, this technique also highlights the drawbacks of Borderline SMOTE which fails in maintaining the detection rate of positive samples while improving detection rate of negative samples. To resolve this drawback, DBSMOTE developed a different approach to precisely oversampling both in the over-lapping and the safe region by synthesizing a minority instance along the shortest path retrieved in a directly density-reachable graph from each sample to the pseudo-centroid cluster.

While methods mentioned above focused on where to generate synthetic samples, Adaptive Neighbor Smote (ANS) put its focus on the number of neighbors needed to ideally synthesize a new sample [10]. In other word, ANS concentrates on deciding the appropriate value for the parameter k in k-nearest neighbors of each positive sample that is needed in synthesizing a new sample. The parameter k is selected according to the density level of each positive sample's region. By utilizing the value of k, the area of the generated synthetic samples will be more spread out inside the dense area and not sparsely distributed as in SMOTE.

4. Features Extraction Methods

Feature extraction is a process meant to reduce the number of variables in a high-dimensional dataset by creating new variables from the existing variables without losing the information of the original dataset. This process is done for some purposes such as to reduce the problem arising from high dimension of a dataset, to increase computational efficiency, or to visualize the dataset either in 2D or 3D.

Principal Component Analysis (PCA) is one of linear feature extraction methods that commonly used to reduce the dimension of a large dataset. PCA orthogonally transforms the dimension of a large set of variables into a smaller set of variables known as principal components to identify the correlation and pattern of the original dataset [11]. In this technique, a considered non-significant principal component is excluded resulting in a lower-dimensional projection while preserving the maximal variance of the dataset. By reducing the dimension of the original dataset, PCA provides an efficient method for data description, visualization, and classification.

Another linear method of feature extraction using component analysis is Independent Component Analysis (ICA). ICA is an important method in signal-based analysis such as EEG signal to help separate normal and abnormal signals. ICA aims to extract the hidden factors of a dataset by transforming the variables to a new set of variables that is maximally independent. What distinguished ICA from PCA is that PCA assumes that signals are subject to multivariate Gaussian distribution and uses orthogonal bases to decompose signals. It can be concluded that ICA's goal is to find the linear transformation of a large dataset in which the basis vectors are non-Gaussian

and statistically independent while PCA's is to find an orthogonal transformation that maximizes the variable's variance of the dataset.

Other than linear dimensionality reduction techniques, there are some available techniques that are nonlinear. One of those is *t*-distributed stochastic neighbor embedding (*t*SNE) which reduces the dimensionality of a dataset by giving each data point a location in 2D or 3D dimensional map. This technique aims at identifying the relevant pattern of a dataset while maintain its local structure [13]. For each point of a datasets, *t*SNE models the probability distribution of other points which are closest to it. One of the most important parameters to be set when using *t*-SNE is perplexity, which is the expected number of nearest neighbors each point has. The performance of tSNE is fairly robust under different settings of this parameter. Generally, the perplexity is set depend on the size of the dataset. The default value of perplexity in some packages is set to 30 for a dataset whose variables more than 30.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) is another nonlinear dimensionality reduction which builds a mathematical theory to justify the graph-based approach [14]. It is developed based on ideas from topological data analysis and manifold learning techniques, which assumes that the data is uniformly distributed on the manifold. To make this assumption true, it defines a Riemannian metric on the manifold. Compared to tSNE, UMAP provides much faster computational running time.

5. Data Description

Chronic obstructive pulmonary disease (COPD) is a progressive inflammatory lung disease that causes breathlessness and leads to serious illness including lung cancer. It is estimated that COPD caused 5% of all deaths globally in 2015, putting COPD as the three leading causes of death worldwide. This study proposes methods that utilize gene expression data from microarrays to predict the presence or absence of COPD. The proposed method assists in determining better treatments to lower the fatality rates. In this study, microarray data of the small airway epithelium cells obtained from 135 samples of 23 smokers with COPD (9 GOLD stage I, 12 GOLD stage II, and 2 GOLD stage III), 59 healthy smokers, and 53 healthy nonsmokers were selected from GEO dataset.

The second type of data used in this study is binary features data. We were utilizing binary features dataset, which means all the input data are binary (values of target class are also represented as binary, but they are treated as categorical values). Three of the datasets, SPECT, analcatdata_fraud, and ring, are benchmark datasets retrieved from Github repository of [1]. The ring dataset's input data are originally numeric with two class label (binary) which then were normalized and binarized using the k-means

clustering algorithm of the "Binarize" R packages. It was done because of the lack of available benchmark datasets of imbalanced class especially for binary features. The ring_1500vs3000, ring_100vs500, ring_100vs2000, and ring_60vs3000 are randomly drawn from the ring dataset. The same procedure was done in order to evaluate their models on different degrees of imbalance due to the lack of available datasets.

The other three datasets are biological datasets obtained at Kanazawa University Hospital. Two datasets about MRSA share basically the same feature data, which represent the existence of mutations in each strain of MRSA corresponding to each sample [2]. The main difference in the two MRSA datasets is the target class. In MRSA pathogenicity dataset, 0 and 1 represent latent and developed, respectively. In MRSA drug resistance dataset, target class represent the resistance of each strain to Clindamycin (CLDM). Similarly, C. difficile pathogenicity dataset contains feature data representing the existence of mutations in each strain of C. difficile, which causes diarrhea in human and has difficulty in antibiotic treatment. To generate the feature data, whole genome of 77 strains were sequenced by Hiseq 2500 with 150 base reads. The reads were mapped to the reference genome NC_009089 in RefSeq (the same as AM180355.1 in GenBank) by BWA. After the mapping and file conversion by SAMtools, mutations were detected by Varscan. Among two types of mutations, Indels and SNPs, we used only Indels (insertions and deletions). After the detection of Indels, if the feature data (existence of mutation) of two or more Indels were completely the same in 77 strains, the features were integrated as one since such redundant features often have harmful influence to the performance of classification by machine learning. Finally, 2610 features were generated to 77 samples, where each feature corresponds to one or more mutations. For the multiclass case, due to the difficulty of finding a benchmark dataset that meets our criteria, we only use one dataset to show that our approach can also be applied on multiclass binary features datasets.

The number of samples and other basic information of the datasets are summarized in the table below. All the input variables of these datasets are binary. All the features of these datasets are being used in the analysis.

Table 1. Basic information of the dataset used in this study.

Dataset	**# of features**	**# of samples**	**Target class ratio**
SPECT	22	267	55:212
analcatdata_fraud	10	42	29:13
car_evaluation	22	1728	1210:384:65:69
MRSA pathogenicity	1978	96	33:63

MRSA drug resistance	1976	94	75:19
C. difficile pathogenicity	2610	77	46:31
ring_1500vs3000	21	4500	1500:3000
ring_100vs500	21	600	100:500
ring_100vs2000	21	2100	100:2000
ring_60vs3000	21	3060	60:3000

By converting binary features into numerical ones using feature extraction methods, succeeding oversampling methods can fully display their potential. Although the comprehensive experiment using benchmark datasets and real medical datasets, it was observed that a converted dataset consists of numerical features is better for oversampling methods. In addition, it is confirmed that features extraction and oversampling are synergistically contributing to the improvement of classification performance.

6. Results and Discussion

In this Chronic Obstructive Pulmonary Disease (COPD) case, microarray data of the small airway epithelium cells retrieved by the Gene Expression Omnibus (GEO) database from 135 samples of 23 smokers with COPD (9 GOLD stage I, 12 GOLD stage II, and 2 GOLD stage III), 59 healthy smokers, and 53 healthy nonsmokers were selected from GEO dataset. Machine learning and regression algorithms performed in this study included Random Forest, Support Vector Machine, Naïve Bayes, Gradient Boosting Machines, Elastic Net Regression, and Multiclass Logistic Regression. After removing imbalance data effect using SMOTE, classification algorithms were performed using 825 of the selected features. High AUC score was achieved by elastic net regression and multiclass logistic regression with AUC of 89% and 90%, respectively. In the metrics including accuracy, specificity, and sensitivity, both classifiers also outperformed the others.

In the binary features datasets, combinations of six datasets, four classifiers, five feature extraction methods including “no feature extraction”, and eight oversampling methods including “no oversampling” were tested. Due to the limitation of data and software, some combinations were omitted from the experiments. For instance, ICA, BLS, ANS were not tested for C. difficile pathogenicity dataset.

The experimental results strongly confirmed our expectation, that is "conversion of binary values of features into numerical values could improve the performance of oversampling". Most of the difference between best model and the base model (without features extraction and oversampling) were greater than zero. It means that the original performance of an oversampling method tends to be improved by a feature extraction method. For example, the accuracy of the combination (SPECT, RF, no feature extraction, RSLS) was 0.8169. Using a feature extraction (tSNE), it was greatly improved to 0.9829. At this point, at least it can be said that the use of oversampling with feature extraction will show a good performance. In addition, it should be emphasized that F1 scores showed similar improvement. Moreover, in many cases, the same combination of feature extraction and oversampling methods achieved best performances in both of accuracy and F1 score. Since accuracy and F1 score frequently show a trade-off relationship for imbalanced data, this result indicates that the approach in this study can contribute to the performance improvement of a wide variety of binary feature datasets. About the applicability, it is also noticeable that this approach was effective for various ratios of imbalance (from 557:567 to 19:75) and various ratio between features and samples (from 10:1124 to 2610:77).

One question to the results is about the relationship between two performance improvements by feature extraction and oversampling. Are they synergetic or independent? The result indicates that the former is correct. For instance, the accuracies of the combinations (SPECT, RF, no feature extraction, no oversampling) and (SPECT, RF, no feature extraction, tSNE) are not so different (0.8169 and 0.8173, respectively). It means that a simple application of tSNE to the original data without oversampling does not improve the performance. Despite that, when it is used as a preprocessing algorithm before an oversampling method RSLS, it greatly contributed to the improvement of accuracy (as described above). Furthermore, we can see many cases that the simple application of a feature extraction method decreased the performance, but the combined use of it with an oversampling method improved it. For example, the F1 score of the combination (SPECT, C4.5, no oversampling) decreased from 0.8810 to 0.8512 by the application of tSNE. In contrast, the F1 score of the combination (SPECT, C4.5, RSLS) increased from 0.9003 to 0.9528 by tSNE.

Conclusions

In this study, we used microarray dataset to predict the presence of COPD by dealing with the class imbalance at first. Prior study on this dataset have tried to predict the presence of COPD regardless of the existence of class imbalance.

The model we proposed can predict the presence of COPD with an overall accuracy and AUC score of 80% and 90% respectively, based on repeated 10-fold cv 10-times. The outcomes indicate that by dealing with class imbalance before performing machine

learning algorithms and regression analysis can be used to predict the presence of COPD more accurately. Our proposed methods also have higher sensitivity and specificity values than that without dealing with class imbalance. It shows that the selected model can be used to correctly classify subjects that belong to a certain class as well as a subject that did not belong to the class. The proposed method in this study can be used to assist in determining better treatments to lower the fatality rates caused by COPD.

Focusing on the problem of binary features that are too poor to apply oversampling algorithms like SMOTE, an approach of using feature extraction methods as a preprocessing before oversampling was presented. Through the comprehensive experiments using various datasets and methods, it was revealed that this approach works well in many cases. By converting binary features into numerical ones using feature extraction methods, it was observed that a converted dataset consists of numerical features is better for oversampling methods. In addition, it is confirmed that feature extraction and oversampling are synergistically contributing to the improvement of classification performance.

References

[1] Olson, R.S.; Cava, W.L.; Orzechowski, P.; Urbanovicz, R.J.; Moore, J.H. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining* **2017,** 10, 36. https://doi.org/10.1186/s13040-017-0154-4

[2] Abapihi,B.; Faisal,M.R.; Nguyen,N.G.; Delimayanti,M.K.; Purnama,B.; Lumbanraja,F.R.; Phan,D.; Kubo,M.; Satou,K. Cross Entropy Based Sparse Logistic Regression to Identify Phenotype-Related Mutations in Methicillin-Resistant *Staphylococcus aureus*, *Journal of Biomedical Science and Engineering* **2020**, 13(7), pp.183-196.

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

[4] He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceeding of the IEEE international joint conference on neural networks, 2008; pp. 1322-1328.

[5] Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Proceeding of International conference on intelligent computing, Springer, Berlin, Heidelberg, August 2005; pp. 878-887.

[6] Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C.; Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem.

Proceeding of the Pacific-Asia conference on knowledge discovery and data mining, Springer, Berlin, Heidelberg, April 2009; pp. 475-482.
[7] Siriseriwan, W.; Sinapiromsaran, K. The effective redistribution for imbalance dataset: Relocating safe-level SMOTE with minority outcast handling. *Chiang Mai Journal of Science* **2016**. 43(1), pp.234-246.
[8] Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence* 2012, 36(3), pp.664-684.
[9] Ester, M.; Kriegel, H.P.; Sander, J.; and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* **1996**, 96(34), pp. 226-231.
[10] Siriseriwan, W.; Sinapiromsaran, K. Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling. *Songklanakarin J. Sci. Technol* 2017, 39(5), pp.565-576.
[11] Karamizadeh, S.; Abdullah, S.M.; Manaf, A.A.; Zamani, M.; Hooman, A. An overview of principal component analysis. *Journal of Signal and Information Processing* **2013**, 4(3B), p.173.
[12] Comon, P. Independent Component Analysis, A New Concept? *Signal Processing* **1994**, 36(3), 287–314.
[13] Van der Maaten, L.; and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, 9(11).
[14] McInnes, L.; Healy, J.; and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* **2018**:1802.03426.
[15] He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* **2009**, 21(9) pp. 1263–1284, 2009.

令和　3年　8月　6日

学位論文審査報告書（甲）

1．学位論文題目（外国語の場合は和訳を付けること。）

Classification of Biomedical Data with Class Imbalance

（クラスインバランスがある生物医学データの分類）

2．論文提出者　(1) 所　　属　電子情報科学　専攻

(2) 氏（ふり）　名（がな）　Kunti Robiatul Mahmudah（くんてぃ ろびあとぅる まーむだー）

3．審査結果の要旨（600～650字）

令和3年8月6日に第1回学位論文審査委員会を開催し、同日に口頭発表、その後に第2回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

生物医学データを対象とした分類問題では、実験や観察上の制約から少数の正例と多数の負例が得られることが多い。2つのクラスの例の数が大きく異なる状況は一般にクラスインバランス（クラス不均衡）と呼ばれ、分類の精度を低下させる原因であることが知られている。本研究では2つの側面からこの問題を改善することを試みた。まず、遺伝子発現データから慢性閉塞性肺疾患（Chronic Obstructive Pulmonary Disease: COPD）を予測するという具体的な問題に対し、種々の分類器と、オーバーサンプリングアルゴリズムの一種であるSMOTEの組み合わせを試すことにより、多クラスロジスティック回帰とSMOTEを組み合わせることで正解率とAUCの両方を向上できることを示した。次に、特徴ベクトルが2値のデータで表現されている場合について、PCA等のアルゴリズムで2値データを通常の数値データに変換することにより、SMOTEを始めとするオーバーサンプリングの効果を改善できることを示した。

以上の研究成果は、クラスインバランスが生じている生物医学データの分類精度を改善するための基礎となるものであり、本論文は博士（学術）に値するものと判定した。

4．審査結果　(1) 判　　定（いずれかに○印）　（合　格）・　不合格

(2) 授与学位　　博　士（　学術　）