Rebalancing clinical data with probabilistic random oversampling

Thanakorn Pasangthien, Boonsit Yimwadsana
Faculty of Information and Communication Technology, Mahidol University

Abstract

Data analysis has become a popular tool to obtain knowledge and create useful application for various business areas. However, often, this is not the case for healthcare industry. This is because the data collected by hospitals and health centers are often bias. Either most data consist of similar patients or most data sets do not contain sufficiently interesting disease data. As a result, prediction and classification on healthcare data usually suffers from the problem of data bias. Since most machine learning algorithms assume sufficiently balanced data which essentially based on general statistics, the prediction performance of the algorithms is largely affected by this bias. There are several well-known methods which offer solution to this data imbalanced problem such as SMOTE and random oversampling. However, most of them do not make use of knowledge hidden in the data. We propose probability model-based random oversampling technique which makes use of the knowledge (probability distribution) of the data so that we can perform random oversampling better than existing methods. The data generated will be based on the probability models of the data portion we are interested in. This will reduce the chance of generating a rigid or strict sample of data which can strongly]alter the statistical information of the data in the experiment. We tested our method using widely known data set such as the UCI diabetes and beast cancer data set. We found that our technique outperforms SMOTE and random oversampling technique in terms of sensitivity and specificity performance.

Keywords: *Data imbalance; oversampling; prediction;*

Revised: 10 May 2022, Revise: 25 July 2022, Accepted: 30 September 2022

Correspondence: Boonsit Yimwadsana, Faculty of Information and Communication Technology, Mahidol University, 999 Phutthamonthon Sai 4 Rd, Salaya, Phutthamonthon, Nakhon Pathom 73170, E-mail: boonsit.yim@mahidol.ac.th

Introduction

Digitization of data not only allow us to automate processes, improve data quality, and improve data exchange, it also allows us to use the data for further analysis in order to gain new knowledge or answer some questions about our business. Healthcare is one of the early business areas that adopted the use of digital data and digital technology. In Thailand, most hospitals and health centers use some forms of electronic health record systems together with laboratory information systems now, and the general public do expect that hospitals and health centers have their illness information in digital form. Digital data bring convenience to both care providers and patients. It also allow modern medical diagnostic equipment to integrate seamlessly forming digital platforms for hospitals and health centers.

Since the beginning of healthcare data digitization, small and large hospitals and health centers now have a large amount of data. The benefits of having a lot of data is obvious as data can be analyzed to provide information and knowledge. Disease could be diagnosed or predicted for each person more accurately. More appropriate treatments can be presented to a specific type of patients. However, comparing to other business areas, the acceptance of data analysis on clinical data by physicians are still low. There are several reasons such as the lacking of staff who are capable of

performing data analysis, restricting data access policy, the poor quality of data, and the complexity of data analysis.

Thanks to recent advancement in data analysis, fields such as data science and machine learning have provided great knowledge and tools that help data analysis process become more convenient and accurate However, afte surveying around 20 physicians and 10 patients, none of them agrees that they would like to allow automated data analysis system to make final decision on patient treatment. Even though there could be several reasons such as whom is to blame when something goes wrong or the protection of physicians' job, it comes down to the trust in the automated data analysis system.

Physicians and patients still do not trust the correctness of decisions or knowledge that come from automated data analysis system even though in other fields, such as finance and marketing, automated data analysis systems show a lot of success in recent years. One of the main reasons is that the data that are collected in clinical facilities (e.g., hospitals and health centers) are poor. One of the main reasons for poor quality data is data imbalance which greatly affect the accuracy of data analysis system.

Data imbalance in clinical data often occurs because the information of different diseases that physicians record is either rare or too common. The distribution of disease occurrences is very sparse. As a result, machine learning techniques designed for data with general normal distribution do not perform well. Since most data are overwhelmed with a few common diseases and have small amount of rare diseases, the accuracy of machine learning methods such as prediction and classification is generally high because the methods can predict or classify common diseases very well. However, the prediction and classification accuracy of rare diseases are quite poor.

There are several techniques often used by data scientist that deal with the issue of data imbalance such as random undersampling, random oversampling, and Synthetic Minority Oversampling Technique (SMOTE). After applying these techniques, which either generate more data from minority class of data (rare diseases) or remove data from majority class of data (common diseases), the sensitivity of the machine learning classification methods increases. However, since the data used for the testing of machine learning classification methods are synthetically created without natural sampling behavior, these machine learning methods do not perform well in real-world settings.

We propose a novel data balancing technique that uses simple probability model fitting technique on the minority dataset to learn the natural behavior of minority data class, and then use probability model-based random number generator to synthetically generate random samples of the minority class until the data set is balanced. The rest of the paper is structured as follows: background section discusses background knowledge of commonly used data balancing techniques, methods section explains our methods to deal with data imbalance, result and discussion section discuss our experimental results and the conclusion summarizes our work and provides recommendation for future works.

Background

In Thailand, there are several popular health record information systems (HIS) such as HOSxP, HoMC, SSB, Hospital OS, MRecord, MedTrak, iMed, JHOS, etc. used in hospital and health centers across the country. In addition, they use laboratory information system (LIS) to collect data from modern medical examination and diagnostic test (e.g. X-ray, MRI and specimen analysis) This large amount of data set allows hospitals to operate with high efficiency and smoothness bringing patient satisfaction. HIS and LIS systems become the heart of almost all hospital and health centers. Over the years, HIS and LIS collect so much information that healthcare workers start to ponder if the data can be used more than just data to look up in the database. In other fields, particularly business and marketing, large amount of data are valuable assets since they provide untapped information and knowledge about customers and products. In healthcare, the untapped information can be about patients, treatments, drugs and diseases. It is only natural that methods and techniques applied to business and marketing fields (which are mostly data mining, data science and machine learning) should be able to work well with healthcare data. Unfortunately, this is not so because healthcare data usually contain the data from people who get sick, not the data from healthy people. In addition, given that the healthcare data contain mostly data from patients, the data also contain mostly data from people who have common diseases which are mostly treatable and not severe. However, untreatable disease data are rare and these rare diseases are often untreatable and severe. We need to understand these diseases in order to be able to reduce the number of casualties from the rare diseases.

In order to classify or predict whether a new patient will have a rare disease or not, classification techniques commonly introduced in data mining, data science and machine learning field are often not effective because they assume that the data set contains approximately the same amount of data from each class. If one class contains very few data (e.g. 1:100), the ability to detect or classify that minority class will not be good. This imbalance characteristic creates bias in the training dataset which can influence many machine learning algorithms since the algorithms will keep on learning data from majority class again and again. This leads to the ignoring of the data in the minority class entirely in most situation. Since rare diseases are usually untreatable or not usually have available treatments, it is important to be able to identify, predict or classify patient quickly. However, given the little amount of data, this task is very challenging.

Several techniques are introduced to balance the data from imbalanced data set. They are random sampling method and Synthetic Minority Oversampling Technique (SMOTE) method [1-2].

Random sampling methods consists of random oversampling method and random undersampling methods. The random oversampling method randomly duplicates or creates new synthetic sample data in the minority class, whereas the random undersampling method deletes or merges random data in the majority class. Both types of sampling methods can be effective when used in isolation, although can be more effective when both types of methods are used together. Both approaches can be repeated until the desired class distribution is balanced in the training data set. The random sampling methods are referred to as "naive resampling" methods because they do not use any knowledge about the data and no heuristic is used. This makes resampling method very simple to implement and fast to execute, which is desirable for very large and complex datasets. However, since the random oversampling method duplicates the data samples, overfitting naturally occurs. In the other hand, random undersampling method removes data samples. Some information is lost. These affect the performance of classification technique greatly.

SMOTE [3] works by selecting examples that are close in the feature space, drawing linear lines between the data samples and creating new data samples at points along those lines. The method can be viewed as a special case of random oversampling method and it has been extremely popular among researchers. Specifically, a random sample is first chosen from the minority class in SMOTE, and then k of the nearest neighbors for that example are found (typically we set k equals to 5). Then, a randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two chosen samples in feature space. This procedure can be used to create as many synthetic samples for the minority class until the data set is balanced. In practice, the majority class data in an imbalanced data set is first trimmed using random undersampling, then SMOTE is later applied to oversample the minority class to balance the class distribution. The main advantage of SMOTE is that it can mitigate from overfitting usually occurs in random oversampling. However, this method is not effective for high dimensional data set. In addition, while generating synthetic records, SMOTE does not consider the neighbor samples from other classes. The generated samples could therefore be considered as noise for the minority class instead of being useful for data balancing.

Methods

Our principle is to balance our imbalanced dataset without creating too much noise in the data and use the knowledge about the minority class. Similar to SMOTE, we try to create samples near existing real samples in the minority class. Instead of just simply generating random data points along the n-dimensional line between two neighbor samples (which can be far apart), we use the method of parametric model estimation [4] to determine the probability distribution of samples in the minority class in all data dimensions. Using the parameters of the related probability distributions, we can generate random variables based on the founded probability distributions with the specific parameters hence creating samples from the knowledge about the minority data set that we have until our data set is balanced.

The parametric model estimation could be performed using the following methods:

1. The method of moment adjusts the parameters of a distribution so that the moments of the distribution coincide with the sample moments of the data (e.g. mean, variance, or mean square). The parameter adjustment method is based on expert knowledge and experience. Frequency-based statistical analysis could help dealing with this parameter adjustment.

For instance, the mean and variance of the data portion that contains rare disease could be used to generate normal random variables with the desired mean and standard deviation.

2. Maximum likelihood method is the most popular method. The likelihood function could be defined as the joint probability mass function or pdf of the data, interpreted as a function of the unknown parameters (1). By adjusting the parameters of the probability distribution where all the random variables take the value of the data samples in the minority class, the parameters that shows the highest value of Likelihood function will be the parameters for the specific probability distribution (2).

$$\mathcal{L}_{x_1,x_2,\ldots,x_n}(\theta) = p_\theta(x_1, x_2, \ldots, x_n) \quad (1)$$

$$\theta_{model} = \arg\max_\theta \log \mathcal{L}_{x_1,x_2,\ldots,x_n}(\theta) \quad (2)$$

Once the parameters are determined for each attribute, we can generate random numbers using standard random number generation methods [5,6] to populate the data set so that the data set is balanced.

Assume that we deal with 2 data classes (A and B) where A is the majority class and B is the minority class (test positive for disease). We perform our experiments on the two popular public health data sets: breast cancer data set [7] and diabetes data set [8]. The breast cancer data set contains 569 records with 32 attributes. The data set is usually used for predicting whether a tumor sample is malignant or benign. According to the data set, the proportion of malignant data to the entire data set is 37%. The diabetes data set contains 2000 records with 9 attributes. The data set is usually used for predicting whether a patient has diabetes or not. According to the data set, the proportion of patient who has diabetes is 34.2%. Even though the data is imbalanced, but the imbalance is not severe. This is the reason why they are widely used for teaching and training purpose. In our research, we deliberately reduce the amount of minority class to around 10% of the data set). We compare the rare-disease detection performance after applying 3 data-oversampling methods which are

1. Random oversampling
2. SMOTE
3. Probabilistic-model-based random oversampling

We are interested in the detection performance (e.g. accuracy) of rare diseases or minority class data samples. Sensitivity and specificity are widely accepted metrics as the gold standard metrics for describing the detection performance of a disease. Sensitivity or true positive rate is the proportion of those who received a positive result on this test out of those who actually have the condition. Specificity or true negative rate is the proportion of those who received a negative result on this test out of those who do not actually have the condition. We are interested in finding out whether the data imbalance level (portion of the minority class to the majority class) affects prediction performance and whether our proposed method can be used to improve the prediction performance of our machine learning algorithms. In our experiment, we use logistic regression, random forest and artificial neural network (ANN) to perform prediction since they are highly popular and have good performance. The sensitivity and specificity values of the method that has the high accuracy are shown in the results section.

Results and Discussion

Our experiment results are shown in Table 1 which shows the sensitivity and specificity of prediction algorithms on the breast cancer data set [7] and diabetes data set [8]. Only the best results out of the three algorithms (logistic regression, random forest, and ANN) are shown in Table 1.

Table 1. sensitivity and specificity of data balancing

	No technique	Random Oversampling	SMOTE	Probabilistic-Model-based Random Oversampling
Sensitivity	0.75	0.88	0.92	0.94
Specificity	1.00	0.95	1.00	1.00

techniques: Random Oversampling, SMOTE, and Probabilistic-Model-based Random Oversampling

The result shows that the sensitivity performance score of the classification of minority data class increases when oversampling techniques are applied. However, SMOTE and our proposed technique perform similarly (tested with hypothesis testing of two means) and they are much better than the plain random oversampling technique.

Conclusion

This probability-based model random oversampling on minority class can perform slightly better than SMOTE, and at the same time generate a little bit less noise than the SMOTE. This probability-based model random oversampling technique consists of simple and well-known methods based on parametric model estimation of probability density which makes use of the knowledge of the data set. The method also allows error to occur due to probabilistic nature of data. This allows the data to resemble real-world data in comparison to SMOTE and random oversampling. However, parametric model estimation of probability distribution should depend on the selected probability model. This model selection process can be automatic but it requires an advanced probability modeling technique. It might also be interesting to find out how different probability distributions affect the performance of our proposed method.

It is important to note that our proposed technique should be used when the minority data class is large enough to confidently obtain the probability model of the data. If the minority data set is too small (e.g. the disease is really rare), our proposed method will not work because we don't have sufficient data to construct a reliable probability model. Acknowledgment

This research project is partially supported by the Faculty of Information and Communication Technology, Mahidol University.

References

[1] Ma, Y., and He, H., "Imbalanced Learning: Foundations, Algorithms, and Applications", Wiley-IEEE Press; 1st edition, 2013

[2] Fernนndez, "Learning from Imbalanced Data Sets", Springer, 1st ed. 2018.

[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, 2002.

[4] Madan, D., "Estimating Parametric Models of Probability Distributions", SSRN, 2015.

[5] Law, A., Kelton, D., "Simulation Modeling and Analysis", 3rd edition, McGraw-Hill Higher Education, 2000.

[6] Law, A., "Simulation Modeling and Analysis", 5th edition, McGraw-Hill Series in Industrial Engineering and Management, 2014.

[7] Breast Cancer Wisconsin (Diagnostic) Data Set, UCI Machine Learning Database.

[8] Pima Indians Diabetes Database, UCI Machine Learning Database