Duo Bundling Algorithms for Data Preprocessing: Case Study of Breast Cancer Data Prediction

Janjira Jojan and Anongnart Srivihok

Abstract—Classification of imbalanced dataset is the most popular and challenged problems for researchers to solve in nowadays. This paper proposed a two-steps approach to improve the quality of class prediction imbalanced breast cancer dataset. The two-steps approach consists of two main techniques: 1) using feature selection techniques to filter out unimportant features from the dataset; and 2) using the over-sampling technique to adjust the size of the minority class to be similar to the size of the majority class. The three different classification algorithms: artificial neural network (MLP), decision tree (C4.5) and Naïve Bayes, were applied. The classification result indicated that C4.5 was the most suitable to classify this dataset which can give the highest accuracy of 83.80%.

Index Terms—Feature selection, over-sampling, classification, imbalanced dataset, breast cancer data.

I. INTRODUCTION

Data mining is a popular research tool used to classify high dimensional data. Data classification is one of an important technique in data mining that can construct a model for classifying data into the assigned classes. The model of the classification can also consider the class of unseen data in the future.

Nowadays, most of data collected are imbalanced datasets that cause the imbalanced problem of the class distributions. The imbalanced dataset [1] means the number of samples in one class is exceedingly more than the other classes. The efficiency of classifying imbalanced dataset is quite terribly because of the majority class is given precedence to classify while the minority class is not. Hence, data mining technique is necessary to preprocess this kind of data before classifying to yield better classification results.

Breast cancer is the most frequent cause of cancer incidence among women all over the world. The number of cases reported is increasing every day. Consequently, the breast cancer prediction model is needed to lighten the doctors' duties in diagnosis the grade of each patient.

In this study we purposed a two-steps approach to improve the quality of breast cancer dataset. The two-steps approach consists of two main techniques: 1) using feature selection techniques to filter out unimportant features from the dataset; and 2) using the over-sampling technique to adjust the size of

the minority class to be similar to the size of the majority class. Finally, the prediction model was built by using classification algorithms.

The rest of the paper is organized as follows. Section II provides related works which are the brief of techniques used to address imbalanced data classification. Section III and Section IV describe the research approaches and the classification algorithms used in this study, respectively. Section V shows the experimental design and evaluation methods. Section VI shows the experimental results. And finally, the summary is presented in Section VII.

II. RELATED WORKS

The two majority approaches for handling class imbalanced problems are data level approach and algorithm level approach [2]. Data level approaches aim to solve imbalanced problems by dealing the class distribution or reducing the high dimensionality of the dataset. While the algorithm level approaches will adjust the existing learning algorithms to emphasize to the minority class [3].

In this paper, we only focused on the data level approaches to handle imbalanced breast cancer dataset. There are two distinguished techniques adoring used to deal high dimensional and imbalanced dataset; they are feature selection techniques and over-sampling techniques, respectively.

Salama *et al.* [4] compared the performances of five classification algorithms: decision tree (J48); Multi Layer Perceptron (MLP); Naïve Bayes (NB); Sequential Minimal Optimization (SMO); and Instance Based for K-Nearest neighbor (IBK), on the Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) datasets. Before classifying by the above five algorithms, they applied feature selection methods to the datasets which including Chi-square test and Principle Component Analysis (PCA). The experimental results indicated that on the WBC dataset, the fusion between MLP and J48 classifiers with features selection (PCA) is superior to the other classifiers. For the WDBC dataset, the fusion of SMO and MLP or SMO and IBK is performed better than other classifiers. And for the WPBC dataset, the fusion of MLP, J48, SMO and IBK is superior to the other classifiers.

Lavanya *et al.* [5] analyzed the performance of decision tree classifier-CART (Classification and Regression Tree) with and without feature selections on various breast cancer datasets (Breast Cancer, Breast Cancer Wisconsin (Original), and Breast Cancer Wisconsin (Diagnostic)). They used 13 existing feature selection methods in WEKA software to

Manuscript received February 7, 2014; revised April 10, 2014. This work was supported in part by the Department of Computer Science, Faculty of Science, Kasetsart University, Thailand.

Janjira Jojan and Anongnart Srivihok are with the Department of Computer Science, Faculty of Science, Kasetsart University, Thailand (e-mail: g5314401215@ku.ac.th, ajjojan@gmail.com, fcsiang@ku.ac.th).

apply with all dataset. The experimental results indicated that feature selections greatly enhanced the accuracy of classification and a particular feature selection using CART enhanced the classification accuracy of a particular datasets.

Wang *et al.* [6] proposed an over-sampling technique, LLE (the locally linear embedding algorithm) - based SMOTE (Synthetic Minority Oversampling TEchnique) to classify imbalanced dataset (chest x-ray image dataset). In the beginning, LLE algorithm is applied to handle the high-dimensional data into a low-dimensional space until the input variables are more separable, they oversampled data by SMOTE. After that they learned data by applying three classification algorithms: Naïve Bayes, K-Nearest Neighbor (K-NN) and support vector machine. The Experimental results indicated that the LLE-based SMOTE algorithm achieves a performance superior to that of the traditional SMOTE.

Geo *et al.* [7] aimed to address the problems of binary-imbalanced data classification, the combination of SMOTE and PSO (Particle Swarm Optimization) by using RBF (Radial Basis Function) as classifier, was presented in this paper. Three different datasets: Pima Indian Diabetes (768 records), Haberman Survival (306 records), and Austempered Ductile Iron (2,923 records) were selected to apply with the proposed method. The experimental result demonstrated that SMOTE combined with PSO when RBF is a classifier, attained high performance in handling the classification of binary-imbalanced data.

III. RESEARCH APPROACHES

A. Feature Selection Techniques

Dash and Liu [8] concluded the meaning of feature selection that is the selection of subset of the smallest attribute which is according to the conditions: the precision of the classification will not decrease significantly, and the class distribution after feature selection is similar to the initial class distribution of full features.

Feature selection methods are classified as filter, wrapper and hybrid approaches. In this paper, the filter approach was applied to the data before the classification.

The objectives of feature selection are to improve the performance of classifiers in predicting data and to prepare the features for rapidly process predicting.

In general, there are four steps of feature selection as shown in Fig. 1. The first step is subsets generation after that evaluate them by using any of the evaluation functions which are distance measures; information measures; dependence measures; consistency measures; and classifier error rate measures. Repeat the loop until the best subset is chosen according to the stopping criterion. Finally, validate that subset again by using classifier algorithm.

B. Over-Sampling by Synthetic Minority Over-Sampling Technique (SMOTE)

Over-sampling is a technique used to reduce the degree of imbalanced class distribution. The method is increasing the size of the minority class either by duplicating or interpolating the minority samples [3].

Fig. 1. General feature selection procedure [9]

One of the most famous oversampling method is SMOTE which developed by Chawla [2]. SMOTE is a preprocessing method used to generate synthetic minority class samples by interpolating between minority instances that lie together. Fig. 2 shows the algorithm of SMOTE, it begins with searching the k-nearest neighbors for each minority instance and each neighbor. It selects the point from the line connecting the neighbor of the same class and the instance itself randomly [3].

S is the original dataset; M is the set of minority class instances.
 For each instance x in M
 Find the K-nearest neighbors minority class instances to x in M
 Obtain y by randomizing one of k instances
 Difference = x - y
 Gap = random number between 0 and 1
 n = x + difference * gap
 Add n to S
 End for

Fig. 2. SMOTE algorithm [3].

However, SMOTE algorithm can generate synthetic samples rather than duplicate minority class instances. Consequently, it can avoid the over-fitting problem of the generated prediction model [3].

IV. CLASSIFICATION ALGORITHMS

We used three different classification algorithms for our experiment as classifiers.

Artificial neural networks (ANNs): is a supervised learning inspired by the human brain. Multi layer perceptron (MLP) is one kind of ANN which focuses on searching linear equations that can separate the instances of each class [10].

The algorithm of MLP is searching the appropriate weight for generating the linear equation that can correctly separate every instance. The general method of MLP is presented in Fig. 3:

Fig. 3. Multi layer perceptron [10].

where $x_1, x_2, \dots, x_n$ are input examples and $w_1, w_2, \dots, w_n$ are weights of each example which fn is an activation function.

C4.5 Decision tree: is the improvement from the ID3 algorithm. C4.5 builds decision trees from a set of training data using the concept of information entropy. For further details see [10].

$$Entropy(S) = -\sum_{i=1}^c p_i \log_2 p_i \quad (1)$$

where P_i is the proportion of data in class i compare with all data and c is the number of class of data.

Naïve Bayes: The Naïve Bayes classifier can be computed the conditional probability which means the probability of the incidence of situations under the condition of another situation happened [11].

$$P(A|B) = P(A)P(B|A)/P(B) \quad (2)$$

V. EXPERIMENTAL DESIGN AND EVALUATION METHODS

A. Dataset

In this paper we used breast cancer dataset from the Surveillance Epidemiology and End Result (SEER) Program [12]. It is the dataset year 2009 which consists of 191 attributes and 240,616 instances. After using the data preparation and data cleaning, the final dataset was constructed, which consisted of 17 attributes and 215,950 instances. The final dataset contain 17 attributes which are presented in Table I.

TABLE I: INPUT ATTRIBUTES

No.	Name of Attributes	Types of Attribute
1	Sex	Categorical
2	Race	Categorical
3	Primary Site	Continuous
4	Radiation	Categorical
5	CS tumor size	Continuous
6	CS extension	Continuous
7	Laterality	Categorical
8	Number of primary	Continuous
9	Age at diagnosis	Continuous
10	Marital status	Categorical
11	Reason no cancer-directed surgery	Categorical
12	Regional nodes positive	Continuous
13	Histologic type ICDO3	Continuous
14	CS lymph node	Continuous
15	Diagnosis confirmation	Categorical
16	First malignant primary indicator	Categorical
17	Grade (class labels)	Categorical

The dependent attribute (grade) is a nominal categorical attribute with four categories: grade 1, grade 2, grade 3, and grade 4, where grade 1 denoted *well differentiated*, grade 2 denoted *moderately differentiated*, grade 3 denoted *poorly differentiated*, and grade 4 denoted *undifferentiated*. Table II presents the distribution of the dependent attribute.

TABLE II: THE DISTRIBUTION OF DEPENDENT ATTRIBUTE (ATTRIBUTE "GRADE") IN PREPROCESSED DATA

Category	Number of Instances	Percentage
grade 1 (well differentiated)	80,136	37.11
grade 2 (moderately differentiated)	29,893	13.84
grade 3 (poorly differentiated)	91,990	42.60
grade 4 (undifferentiated)	13,931	6.45
Total	215,950	100.00

From Table II we can consider that the number of instances distributed in each class was imbalanced especially for grade 4 and grade 2, were less than the other classes considerably. Therefore, before classifying these data, we have to adjust the size of the minority class to be similar to the size of the majority class to yield better classification results.

B. Methodology

In this paper we purposed the two-steps approach to improve the quality of breast cancer dataset before classifying. The two-steps approach consists of two main techniques: 1) using feature selection techniques to filter out unimportant features from the dataset. The filter methods with three evaluated functions; ChiSquareAttributeEvaluation, ConsistencySubsetEvaluation, and InfoGainAttributeEvaluation, were applied in this step. Finally, the best final attribute set after selected by those methods was constructed; and 2) using the over-sampling technique to adjust the size of the minority class to be similar to the size of the majority class. The dataset used in this step was the best final attribute set from feature selection in the first step with the over-sampling technique, SMOTE.

Finally, after the two-steps approach above was done, learn the data by using three different classification algorithms: multi layer perceptron (MLP), decision tree (C4.5), and Naïve Bayes.

Our proposed method and classification phase were demonstrated by WEKA software version 3.6.0.

An experimental framework of this paper is shown in Fig. 4.

Fig. 4. Experimental framework.

C. Evaluation Methods

We tended to use k-fold cross validation (10-fold in this paper) in order to minimize the bias of the training and the testing data and used three performance measures [13]: accuracy, sensitivity, and specificity, which were shown in equation (3), (4), and (5), respectively.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (3)$$

VI. EXPERIMENTAL RESULTS

As the above methodology, there are two main steps of our experiment: preprocess data and classification. In preprocessing data, there are two sub-techniques using in this step, the first technique is feature selections and the other one is over-sampling data. After data was preprocessed, the classification results were applied. Therefore in this section, we showed the several results of each experimental step.

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (4)$$

$$Sensitivity = \frac{TN}{(TN + FP)} \quad (5)$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative [13].

A. Feature Selection Results

TABLE III: FEATURE SELECTION RESULTS

Attribute Evaluators	Number of selected attributes	Selected Attributes*	Preliminary classifying results by C4.5
ChiSquareAttribute Evaluation	15	(1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (12), (13), (14), (16), (17)	57.90 %
ConsistencySubset Evaluation	9	(1), (4), (5), (6), (8), (9), (10), (14), (17)	62.40 %
InfoGainAttribute Evaluation	12	(1), (3), (4), (5), (6), (7), (8), (9), (10), (12), (14), (17)	59.09 %

Note that * means the number of attribute in the blankets related to the number of attribute in Table I (input attributes).

From Table III, we can see that set of attributes selected by ConsistencySubsetEvaluation gave the best preliminary classifying result that was 62.40%. The set of those selected attributes contain 9 attributes which consist of Sex, Radiation, CS tumor size, CS extension, Number of primary, Age at diagnosis, Marital status, CS lymph node and Grade. So, the 9 attributes above will be used in the next step of over-sampling.

B. Over-Sampling Results

The over-sampled rate for over-sampling class 2 and class 4 were 100% and 250%, respectively, gave the best

preliminary classifying result that was 83.80%. Therefore, we chose this dataset to classify by the different classification algorithms in the next step.

TABLE IV: OVER-SAMPLING RESULTS

Class	Over-sampled Rate	Preliminary classifying results by C4.5
2	100 %	76.43 %
4	200 %	
2	100 %	83.80 %
4	250 %	
2	100 %	72.67 %
4	300 %	
2	150 %	74.97 %
4	250 %	

C. Classification Results

After we got the new dataset from feature selections and over-sampling, we learned it by using three classifiers: multi layer perceptron (MLP), decision tree (C4.5), and Naïve Bayes. The results of classification are shown in Table V.

TABLE V: CLASSIFICATION RESULTS

Classifiers	Accuracy	Sensitivity	Specificity
Multi Layer Perceptron (MLP)	0.7854	0.7983	0.7795
Decision tree (C4.5)	0.8380	0.8517	0.8236
Naïve Bayes	0.7562	0.7744	0.7484

There are three measurement methods used to measure the performance of our dataset: accuracy, sensitivity, and specificity. From the above Table V, we can see that MLP, C4.5 and Naïve Bayes gave the accuracy of 78.54%, 83.80% and 75.62%, respectively.

C4.5 did not give the highest accuracy only but when we noticed the sensitivity and specificity, it gave the highest values of those two measurements also (85.17% and 82.36%, respectively). Therefore, we can conclude that for this breast cancer dataset, C4.5 classifier achieved better in classifying data than the other two remain classifiers, MLP and Naïve Bayes.

VII. SUMMARIES

In this paper we purposed a two-steps approach to improve the quality of breast cancer dataset before classifying. The two-steps approach consists of two main techniques: feature selection and over-sampling techniques. For feature selection, ConsistencySubsetEvaluation selected the best set of attributes which consisted of 9 attributes including Sex, Radiation, CS tumor size, CS extension, Number of primary, Age at diagnosis, Marital status, CS lymph node and Grade. Then, those 9 attributes were fed into the over-sampling phase to adjust the sizes of the minority class, class 2 and class 4. They were over-sampled by SMOTE with the over-sampled rates for over-sampling class 2 and class 4 were 100% and 250%, respectively.

After we got the new dataset from that two-step approach above, we learned the data with three different classification algorithms: artificial neural network (MLP), decision tree (C4.5) and Naïve Bayes which gave the accuracy of 78.54%, 83.80% and 75.62%, respectively. The results indicated that decision tree (C4.5) was the most suitable to classify this dataset.

ACKNOWLEDGMENT

The authors would like to thanks to National Cancer Institute, USA, for permitting us to use the cancer data in SEER Database for the experiment.

REFERENCES

- [1] N. Japkowics and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429-449, October 2002.
- [2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligent Research*, no. 16, pp. 321-357, January 2002.
- [3] W. Prachuabsupakij, "Multiclass imbalanced data classification using K-MEANS algorithm," Ph.D. Dissertation. Kasetsart University, Thailand, 2013.
- [4] G. I. Samala, M. B. Abdelhalim, and M. A.-E. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," *International Journal of Computer and Information Technology*, vol. 1, no. 1, pp. 2277 – 0764, September 2012.
- [5] D. Lavanya, "Analysis of feature selection with classification: Breast cancer datasets," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 5, pp. 756-763, October-November 2011.
- [6] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding," in *Proc. ICSP2006*, 2006, vol. 3.
- [7] M. Gao, X. Hong, S. Chen, and C. J. Harris, "On combination of SMOTE and particle swarm optimization based radial basis function classifier for imbalanced problems," in *Proc. International Joint on Neural Networks*, 2011, pp. 1146-1153.
- [8] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131-156, 1997.
- [9] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms,"

Yukoslav Journal of Operations Research, vol. 21, no.1, pp. 119-135, 2011.

- [10] N. Soonthornphisaj, *Artificial Intelligence*, Bangkok, Thailand: Kasetsart University, 2009.
- [11] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, CA: Morgan Kaufmann Publishers, 2001.
- [12] Sueviellance Epidemiology and End Results (SEER) Program. (October 19, 2012). SEER*Stat Database. Research data (1973-2009). National Cancer Institute, DCCPS. [Online]. Available: <http://www.seer.cancer.gov>
- [13] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113-127, June 2005.

Janjira Jojan was born in Thailand on April 19, 1988. She received her bachelor's in computer science from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 2010. She is currently a graduate student in computer science at Kasetsart University, Thailand. Her research interests include data mining, machine learning and artificial intelligence.

Anongnart Srivihok is an associate professor in the Department of Computer Science of Kasetsart University, Thailand. She completed her M.S. in computer science from University of Mississippi, USA and PhD. in information systems from Central Queensland University, Australia. Her research interests are: information systems, data mining, ontology and artificial intelligence.