Support Vector Machine with Restarting Genetic Algorithm for Classifying Imbalanced Data

Keerachart Suksut, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Algorithms for data classification are normally at their high performance when the dataset has good balance in which the number of data instances in each class is approximately equal. But when the dataset is imbalanced, the classification model tends to bias toward the majority class. The goal of imbalanced data classification is how to improve the performance of a model to better recognize data from minority class, especially when minority is more interesting than the majority data. In this research, we propose technique for balancing data with hybrid resampling techniques and then perform parameter optimization with restarting genetic algorithm. The optimized parameters are for support vector machine to induce efficient model for recognizing data in minority class, whereas maintaining overall accuracy. The experimental results show that the proposed technique has high performance than others.**

***Index Terms*—Imbalanced data, restarting genetic algorithm, support vector machine.**

I. Introduction

Currently, data mining has been applying to many fields. The concept of data mining is to find the knowledge from the stored information and database. Knowledge can be a pattern or relationship that is hidden in the data. The knowledge extraction can be done with mathematical method, statistics or other computational methods [1]. There are many types of data mining such as data classification, association rule mining, clustering, forecasting, and other analysis tasks.

Techniques in the data classification include artificial neural network (ANN), decision tree, naïve Bayes, support vector machine (SVM), and many more. The concept of ANN is simulating computer to resemble the human brain, which can learn as a human learns. The idea of decision tree induction for data classification is to partition data into subsets using tree as a data structure to store data subsets. The nodes in a tree represent data attributes used for partitioning data into subsets and the leaf nodes are classes of data. The concept of naïve Bayes is to use the probability to classify the data. Main concept of SVM is creating the hyperplane for separating data with high distance between groups of data. SVM has recently gained popularity due to its overall high performance on classifying both balanced and imbalanced data [2], [3]. However, recognition rate over minority class is still low.

To improve the algorithm on classifying minority, some techniques to properly adjust learning parameters have been proposed. For instance, Yin et al. [4], Jamshidi et al. [5] and Shiff et al. [6] applied genetic algorithm to learn optimal parameter values. But the problem of genetic algorithm is that sometime the algorithm cannot find the best parameter due to improper setting of a random initial value. In addition, most classification algorithms work effectively when the data is balanced. In this research, we thus propose techniques for balancing data and then optimizing parameters with restarting genetic algorithm for the subsequent application of SVM learning algorithm.

Manuscript received February 15, 2017; revised April 8, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

K. Suksut is with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand (corresponding author: K. Suksut; Tel.: +66879619062; e-mail: mikaiterng@gmail.com).

K. Kerdprasop is with the School of Computer Engineering. He is also with Knowledge Engineering Research Unit, SUT, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is the School of Computer Engineering. She is also with Data Engineering Research Unit, SUT, Thailand (e-mail: nittaya@sut.ac.th).

II. Background Theories

A. Data Sampling

Data sampling is a pre-processing step of classification to balance amount of data in each class. The two major sampling approaches to balance data are under sampling and over sampling. Under sampling is a technique of down sampling that reduce the amount of data in the majority class to be in the same proportion as the number of data in the minority class [7]. The basic idea is shown in Fig. 1.

Over sampling, on the contrary, is the up-sampling technique in the sense that data in the minority class is increased to be in the same amount of data in other classes. Sampling data from minority class can be either the repeated selection of data from the minority class, or the generation of data points based on some criteria.

SMOTE technique [8] applies the later scheme by creating a synthetic data by measuring the distance from the sample data to the nearest data point and then randomly create new data. The new data are created within the distance computed as in equation (1):

$$N_p = O_p + (Rand[0,1] * dist(x, y, \dots, z)) \qquad (1)$$

where *Np* is the new data of minority class, *Op* is the old data point in minority class used as the reference point for computing neighbor distance, Rand[0,1] is random number between 0 to 1,dist(x,y,…,z) is the distance between default data andneighbors.

B. Genetic Algorithm

Genetic algorithm is the search for optimal answer by using imitation of natural evolution such that the one who is stronger has more chance to survive than those who are

doi: 10.18178/ijfcc.2017.6.3.496

weaker and the stronger one can inherit strength to their children. John Holland [9] introduced this concept of genetic algorithm in 1975. After that, it has been successfully applied to many applications. The draft computation steps of genetic algorithm are shown in Fig. 2.

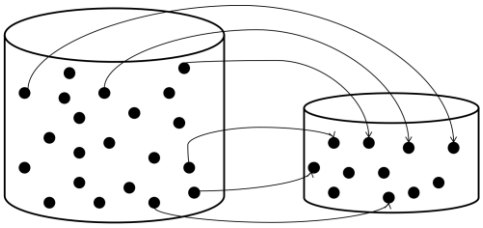

Fig. 1. Under sampling data.

Firstly, the initial population has to be randomly created. Random number of the population equals to the number of the population size. After that, the fitness value of each population is computed for selecting the best population to be used as the chromosomes to inherit as genetic material. Then, genetic operation process such as crossover and mutation will be applied to mutate chromosome for hopefully being stronger. The new generation of population that is stronger than the old one will replace the old population. The process iterates until it converges to the stopping criterion.

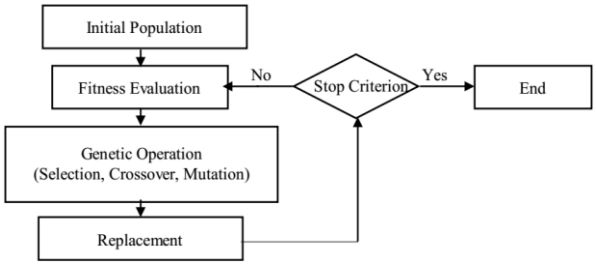

Fig. 2. Simple genetic algorithm.

C. Support Vector Machine

Support vector machine, or SVM, [10] is an algorithm for classifying data by creating a hyperplane to separate data with different classes. Optimal hyperplane for SVM is the line or plane that has maximum margin between the plane and the nearest data points on each side of the plane. This concept is shown in Fig. 3.

The hyperplane will split the data having different classes apart from each others with the maximum distance between data from each class. The weight vector is used for determining the direction and inclination of the hyperplane. Weight vector is perpendicular to the hyperplane and the data with classes 1 and -1 can be separated according to the equation (2):

$$w^T x + b \geq 1, when\ y_i = +1$$
$$w^T x + b \leq 1, when\ y_i = -1 \quad (2)$$

where *w* is weight vector, and b is bias.

Weight vector is the line perpendicular to the hyperplane and bias will determine the distance between the hyperplane and origin. Consider two dimensional data $X = (x_1, x_2)^T$, the equation of linear hyperplane is:

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + b = 0 \quad (3)$$

Given two data points on hyperplane A = (A_1, A_2) and B = (B_1, B_2), the equation for compute the weight vector is:

$$weight\ vector = -\frac{w_1}{w_2} = -\frac{(B_2 - A_2)}{(B_1 - A_1)} \quad (4)$$

The margin can be computed with equation (5) and the size of weight vector is computed as in (6):

$$margin = \frac{2}{||w||} \quad (5)$$

$$||w|| = \sqrt{w_1^2 + w_2^2} \quad (6)$$

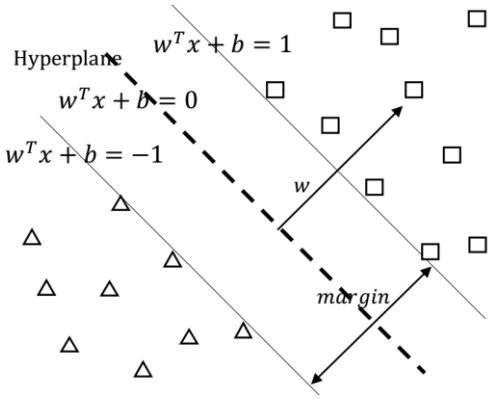

Fig. 3. Optimal hyperplane for support vector machine.

D. Adaboost (Adaptive Boosting)

Adaboost algorithm is the application of boosting technique [11] to increase classification performance. The main concept (shown in Fig. 4) is a combination of weak learners with adjusted higher weight for data that are wrongly classified. Then create new learner from miss-classified data until receiving strong learner with high predictive performance. There is an extension of Adaboost called RUSBoost in which under sampling technique has been applied before classifying data with Adaboost.

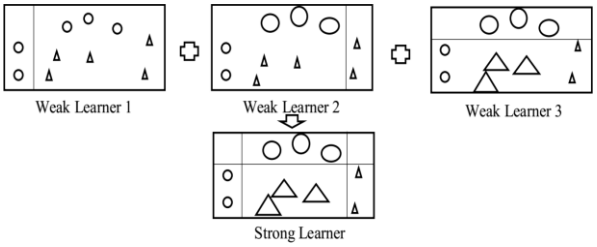

Fig. 4. Adaboost algorithm.

E. Classification Performance Evaluation

To evaluate performance of classification model on recognizing majority and minority classes of imbalanced data, we use four measurements: Accuracy, Precision, Recall and F-measure. The computation of these metrics is based on the values in confusion matrix as shown in Table I.

TABLE I: CONFUSION MATRIX FOR TWO CLASS CLASSIFICATION

		Predicted Data	
		Positive	Negative
Actual Data	Positive	TP	FN
	Negative	FP	TN

Rows in the matrix are number of actual data for each class and columns are number of predicted data for each class. The

acronyms TP, FP, FN, TN are possible outcomes of prediction made by the classification model. Suppose the data are either of class positive or negative, the outcome of prediction can be one of the following 4 cases:

Case 1: TP is the number of actual data from positive class and the model can correctly predict that data to be in a positive class.

Case 2: FN is the number of actual data from positive class but the model predict that the data incorrectly as in a negative class.

Case 3: FP is the number of actual data from negative class but the model incorrectly predict that data to be in a positive

Case 4: TN is the number of actual data from negative class and the model can correctly predict that data to be in a negative class.

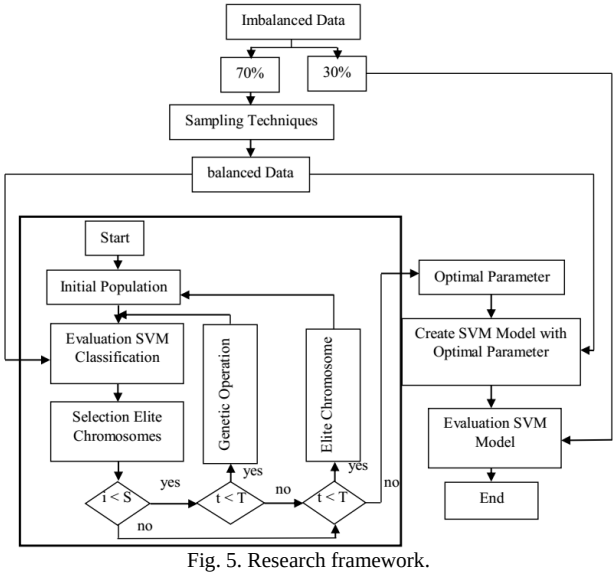

Fig. 5. Research framework.

Accuracy is a measure for overall performance of the classification model, and the computation is as shown in equation (7):

$$Accuracy = \frac{(\text{TP} + \text{TN})}{(TP+FN+FP+TN)} \quad (7)$$

Precision is the proportion of predicted positive class to the real positive class, computed as in equation (8):

$$Precision = \frac{(\text{TP})}{(\text{TP}+\text{FP})} \quad (8)$$

Recall or Sensitivity is the ration of data that are predicted as positive to the number of all positive data, computed as in equation (9):

$$Sensitivity = Recall = \frac{(\text{TP})}{(\text{TP}+\text{FN})} \quad (9)$$

F-measure is a measure that taking into account both precision and recall. The computation of F-measure is as shown in equation (10):

$$F - measure = \frac{(2*\text{Precision}*\text{Recall})}{(\text{Precision}+Recall)} \quad (10)$$

III. Materials and Methods

The design and implementation of our work to deal with imbalanced data classification are as shown in Fig. 5. Firstly, we split data into 2 subsets, 70% of them is training set and the remaining 30% is testing set. We preprocess training set with random under sampling to reduce number of data in the majority class and synthetically generate data in the minority class with SMOTE technique. We then find the optimal parameter for the subsequent classification process by introducing restarting genetic algorithm. For the Chromosome encoding, we use real-value encoding and random initial population until obtaining the specified population size. The fitness value of each chromosome is evaluated based on the accuracy from classifying data with support vector machine by using training set and parameter from each chromosome. After that, we select elite chromosomes, which are the top k chromosomes with highest fitness values, and applying the genetic operation to obtain new population.

If the new generation is less powerful than the old population, repeat the process by replacing initial population with elite chromosome and proceed until the stopping criterion has been met. After completion, create model with optimal parameters for support vector machine and evaluate model with testing set. Then, compute performance with accuracy, precision, recall and F-measure metrics.

Restarting genetic algorithm in this research is the addition of condition to re-create the initial population when the new generation has fitness value less than the old population and the stopping criterion has not been met. The steps in restarting genetic algorithm are shown in Fig. 6.

Restarting Genetic Algorithm
Input : c, epsilon, gamma,
number of generation T,
number of worst generations S
Output : optimal of c, epsilon, gamma
Method:
1. chromosome encoding
2. initial population p
3. evaluate fitness value of each chromosome
4. for t ≤ T
4.1) for new generation having fitness value less than the old population, i < S
4.1.1) genetic operation (selection, crossover, mutation)
4.1.2) replacement
4.1.3) fitness evaluation (select elite chromosome)
4.2) for new generation having fitness value less than the old population, i ≥ S
4.2.1) re-create initial population with elite chromosome
4.2.2) fitness evaluation
5. select the best chromosome

Fig. 6. Restarting genetic algorithm.

IV. Experimental Results

A. Dataset

In this research, we use 2 datasets. One is a real dataset, another is synthetic dataset. Details of data are as follows.

Synthetic dataset contains 700 data records with 16 attributes. The majority class comprises of 600 records,

whereas the minority class has 100 records.

The real dataset is Asthma data [12] containing 677 data records with 16 attributes. The majority class contains 570 records, but the minority class has only 128 records.

B. Parameter Setup

The setting of parameters c, epsilon, gamma, number of iteration, population size, probability of crossover, probability of mutation, and number of worst generations for restarting genetic algorithm are summarized in Table II.

TABLE II: PARAMETER DETAIL FOR RESTARTING GENETIC ALGORITHM

Cost	$10^{-4} – 10^{-2}$	Prob. of crossover	0.8
Gamma	$10^{-3} – 10$	Prob. of mutation	0.01
Epsilon	$10^{-2} – 10$	Iteration	100
Population size	100	Restart GA	2

C. Results

For evaluate performance of classification model, we use the accuracy, precision, recall, and F-measure metrics. We compare the classification performance of our proposed method against the powerful algorithms that have been widely used to learn model from imbalanced data. These standard algorithms are support vector machine (with default parameters), Adaboost, and RUSBoost. The comparative results on synthetic dataset are shown in Table III.

TABLE III: COMPARATIVE PERFORMANCE OF SYNTHETIC DATASET

	SVM	Adaboost	RUSBoost	Propose
Accuracy	**88.00**	87.50	78.50	85.00
Precision	**100.00**	88.89	32.56	47.92
Recall	14.29	25.00	50.00	**82.14**
F-measure	25.01	39.03	39.44	**60.53**

From Table III, when considering overall accuracy for classifying imbalanced data, we found that SVM using default parameters has highest accuracy at 88.00%, whereas Adaboost is the second best accurate model at 87.50% prediction correctness. Our proposed method is the third at 85.00% correctness, and RUSTBoost is the worst with 78.50% correctness.

When considering precision value, SVM show the best performance at 100%. Adaboost comes second at 88.89% of precision on predicting minority class. Our proposed technique is the third one (47.92%) and RUSBoost are the worst (32.56%).

For the recall measurement on minority class recognition, we found that our proposed technique performs the best at recall rate 82.14%. The second best recall model is RUSBoost (50.00%), whereas Adaboost is the third one (25.00%) and SVM is the worst (14.29%) in terms of minority class recognition. To consider both precision and recall with the F-measure metric, our proposed method is the best (60.53%). RUSBoost is the second best one (39.44%), followed by Adaboost (39.03%) and SVM (25.01%).

The results of asthma dataset are shown in Table IV.

TABLE IV: COMPARATIVE PERFORMANCE OF ASTHMA DATASET

	SVM	Adaboost	RUSBoost	Propose
Accuracy	**79.52**	78.10	66.67	70.00
Precision	**38.89**	38.71	35.78	37.76
Recall	17.95	30.77	**100.00**	94.87
F-measure	24.56	34.29	52.70	**54.02**

For the real asthma dataset, SVM is also the best model in terms of overall accuracy (79.25%) and precision (38.89%) on predicting class. The second best one is Adaboost model (accuracy = 78.10% and precision = 38.71%). Our proposed model is the third one (accuracy = 70%, precision = 37.76%). The worst model is RUSBoost (accuracy = 66.67%, precision = 35.78%).

But when considering only recall rate, RUSBoost is the best model on recognition the minority class of asthma dataset. Out proposed model performs the second best at 94.87% of recognition rate. The Adaboost and SVM models are very poor on recalling data in the minority class with the recognition rate at 30.77% and 17.95%, respectively.

To evaluate with the F-measure, our proposed model is the best one (54.02%), and the RUSBoost model is the second (52.70). Both Adaboost and SVM show poor performance at 34.29% and 24.56%, respectively.

It's can be seen that when considering only accuracy and precision, SVM shows higher performance than other techniques. But when considering about recall performance and F-measure, which is the compromising of both recall and precision metrics, our proposed technique performs better than others.

V. CONCLUSION

The major problem on building a model to classify data that distribution among classes is uneven is that the model built from traditional method tends to bias toward majority class in such a way that the model is most likely to guess the class of all new data as the majority one. This tendency of a model is not harmful when the main model measurement of interest is overall predictive accuracy. But when data in minority class is the class of concern, traditional method is not powerful enough to catch the minority cases.

To improve the algorithm on classify imbalanced data to better recognizing the minority data that are normally overshadowed by the majority class, we propose a novel method that firstly balancing data by using random under sampling data in majority class, as well as creating synthetic the data to increase the amount in the minority class with SMOTE technique. We then propose to use restarting genetic algorithm to find the optimal parameters for support vector machine. The experimental results show that support vector machine (with default parameters) performs better than other techniques in terms of accuracy and precision, but shows poor performance when evaluated with recall and F-measure. When high recall of data in minority class and F-measure are the main measurements of interest, our proposed method has been experimentally proven better than the traditional support vector machine.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[2] J. J. Liao, C. H. Shih, T. F. Chen, and M. F. Hsu, "An ensemble-based model for two class imbalanced financial problem," *Economic Modelling*, vol. 37, pp. 175-183, 2014.

[3] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32-41, 2014.

[4] F. Yin, H. Mao, and L. Hua, "A hybrid of back propagation neural network and genetic algorithm for optimization of injection molding

process parameters," *Materials & Design*, vol. 32, no. 6, pp. 3457-3464, 2011.
[5] M. Jamshidi, M. Ghaedi, K. Dashtian, S. Hajati, and A. Bazrafshan, "Ultrasound-assisted removal of Al 3+ ions and Alizarin Red S by activated carbon engrafted with Ag nanoparticles: central composite design and genetic algorithm optimization," *RSC Advances*, vol. 5, no. 73, pp. 59522-59532, 2015.
[6] S. Shiff, M. Swissa, and S. Zlochiver, "A genetic algorithm optimization method for mapping non-conducting atrial regions: a theoretical feasibility study," *Cardiovascular Engineering and Technology*, vol. 7, no. 1, pp. 87-101, 2016.
[7] K. Chomboon, "Classification technique for minority class on imbalanced dataset with data partitioning method," Suranaree University of Technology, 2016.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over- sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[9] H. Holland, "Adaptation in natural and artificial systems," Ann Arbor: The University of Michigan Press, Michigan, 1975.
[10] C. Cortes and V. Vapnik, "Support vector network," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[11] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th International Conference on Machine Learning*, vol. 96, pp. 148-156, 1996.
[12] P. Teerarassamee, "The methodology to find appropriate k for k-nearest neighbor classification with medical datasets," Suranaree University of Technology, 2015.

K. Suksut is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology, Thailand, in 2011, the master degree in computer engineering from Suranaree University of Technology, Thailand, in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

K. Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and the doctoral degree in comp. science from Nova Southeastern University, U.S.A., in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.

N. Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and the doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, logic programming, and intelligent databases.