Improvement the Imbalanced Data Classification with Restarting Genetic Algorithm for Support Vector Machine Algorithm

Keerachart Suksut, Nuntawut Kaoungku, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—The general datamining algorithm also classify the balanced dataset, when the data have imbalanced the predicted rate over minority class is still low. The random sampling techniques has been applying to solve the imbalanced data, but sometimes the random technique has selected the features is clearly different from both, when the unseen data (from minority class) has features look like the majority class, the classification model show miss classification because the model learning sample data does not complete. To improve the performance to classify the data, the genetic algorithm is applying to finding the optimal parameter, but sometimes the genetic algorithm cannot find the best set of parameters because the random initial population is not cover the best set of parameters, in this research proposed the techniques to guarantee the genetic algorithm can find the optimal parameter by using restarting technique to re-create the initial population when the new generation show powerful less than the old population. The results show that proposed technique can improve the performance to classify the minority class from imbalanced dataset more than the other techniques.**

***Index Terms*—Imbalanced data classification, optimization parameter, genetic algorithm, restarting genetic algorithm.**

I. Introduction

The datamining techniques are to find the knowledge from the stored information and database [1]. Today we have many techniques to extraction the knowledges such as the Naïve Bayes Algorithm, Decision Tree Algorithm, Neural Network Algorithm, Support Vector Machine Algorithm etc. The general techniques also classify the balanced data, when the data has imbalanced the general techniques show poor performance to classify the minority class.

The random sampling technique has been applying to solve the imbalanced data classification such as the random under sampling, random over sampling, and SMOTE technique. But the random techniques are a random data reduction from majority class or increasing the data from minority class with random technique, sometimes the random technique has selected the features is clearly different from both (majority class, minority class). The effect is when the unseen data has features look like the majority class, but the actual class is the minority class, the classification model show miss classification because the model learning sample data does not complete.

The Support Vector Machine (SVM) has recently gained popularity due its overall high performance on classifying both balanced and imbalanced data. However, predict rate over minority class is still low [2]. To improve the algorithm on classifying minority, some techniques to properly adjust learning parameters have been proposed with applying genetic algorithm to learn optional parameter values [3]. But the main problem of simple genetic algorithm is that sometimes the results is not the best parameter because the random initial population is not covering the set of the best parameter. In this research, we thus proposed techniques for optimizing parameter with restarting genetic algorithm for support vector machine learning algorithm, the aim of this research is to improve the performance of classify the minority class.

II. Background Theories

A. Imbalanced Data

The imbalanced data is the problem in machine learning where the total number of a class of data (minority class) is far less than the total number of another class of data (majority class) [4]. The effect of imbalanced data is the most machine learning algorithms and works best when the number of instances of each class are roughly equal, when the number of instances of one class far exceeds the other, the algorithm still low performance classifies the minority class.

The imbalanced degree (imbalanced ratio) is the measure for identification information is imbalanced or balanced. The imbalanced ratio will show the ratio between the number of majority class and the number of minority class [5]. The imbalanced ratio can be computed with equation (1):

$$Imbalanced\ Ratio = \frac{n_{majority}}{n_{minority}} \quad (1)$$

where $n_{majority}$ is the number of majority class

$n_{minority}$ is the number of minority class

The result of imbalanced data classification can be categorized into four cases which are started in the Table I, confusion matrix.

Manuscript received December 19, 2018; revised May 8, 2019. This work was support by grant from Rajamangala University of Technology Isan.

Keerachart Suksut is with Computer Engineering Department, Rajamangala University of Technology Isan (corresponding author, e-mail: mikaiterng@gmail.com).

Nuntawut Kaoungku is with School of Computer Engineering, Suranaree University of Technology (e-mail: nuntawut@sut.ac.th).

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

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

doi: 10.18178/ijfcc.2019.8.2.541

TABLE I: Confusion Matrix

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

where True Positive (TP) – An example that is positive and is classified correctly as positive

True Negative (TN) – An example that is negative and is classified correctly as negative

False Positive (FP) – An example that is negative but is classified wrongly as positive

False Negative (FN) – An example that is positive but is classified wrongly as negative

Accuracy is a measure for overall performance of the classification model, and the computation is as shown in equation (2):

$$Accuracy = \frac{(TP + TN)}{(TP+FN+FP+TN)} \quad (2)$$

Precision is the proportion of predicted positive class to the real positive class, computed as in equation (3):

$$Precision = \frac{(TP)}{(TP+FP)} \quad (3)$$

Recall or Sensitivity is the ration of data that are predicted as positive to the number of all positive data, computed as in equation (4):

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

F-measure is a measure that considering both precision and recall. The computation of F-measure is as shown in equation (5):

$$F - measure = \frac{(2*Precision*Recall)}{(Precision+Recall)} \quad (5)$$

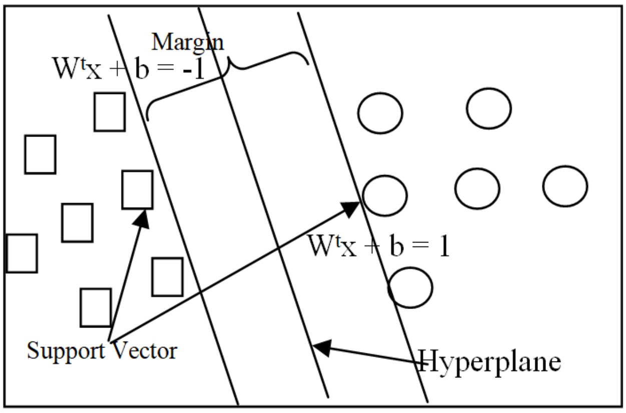

Fig. 1. Support vector machine.

B. Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is the data classification technique that currently shows the highest performance for classifies the unseen data. SVM were developed by Cortes and Vapnik [6] for binary classification, basically, looking for the optimal separating hyperplane between the two classes by maximizing the margin between the classes' closest points show in Fig. 1, the point lying on the boundaries are called support vectors, and the middle of the margin is our optimal separating hyperplane.

The hyperplane as a model to classify data of class +1 from the class -1 can be computed as the formula given in equation (6).

$$\begin{aligned} w^T x + b &\geq 1,\ when\ y_i = +1 \\ w^T x + b &\geq 1,\ when\ y_i = -1 \end{aligned} \quad (6)$$

where x is the data vector,

y is the class label,

w is the weight vector,

b is the bias value.

To apply the SVM for classifying the new data, we set values of the tree important parameters: Cost, Epsilon, and Gamma.

C. Grid Search Algorithm

The general grid search algorithm is trying all possible values of the (Cost, Epsilon, Gamma) pair. To each (Cost, Epsilon, Gamma) pair, cross-validation is used to get its Accuracy.

In cross-validation [7]. Step 1: randomly divide the training set into k disjoint subsets of roughly equal size for each fold. Step 2: use k-1 subsets as training subset, the remaining as testing subset of the classification model. The accuracy is calculated for each testing subset. The average accuracy of k testing subsets is as performance evaluation parameter.

After all positive values of (Cost, Epsilon, Gamma) pair are tested, (Cost, Epsilon, Gamma) pair with highest accuracy is the best pair. However, the grid search process needs a lot of time, which is not proper for many applications.

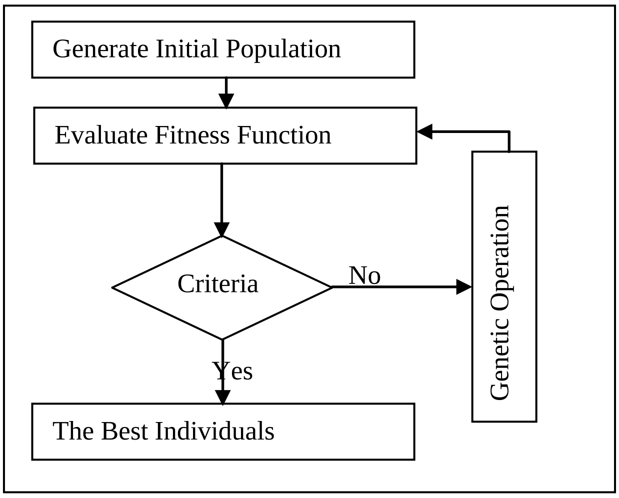

Fig. 2. Simple genetic algorithm.

D. Genetic Algorithm

Genetic algorithm [8] is one type of optimization algorithm, the algorithm can find the optimal solution based on the major characteristics of nature such as natural selection and evolutionary in which the ones who are stronger have more chance to survive than those who are weaker and can be inherit strength into the new generation. The concept of basic genetic algorithm shown in Fig. 2.

The process of genetic algorithm usually begins with a randomly selected population of chromosomes. These chromosomes are representations of the problem to be solved. Then an evaluation function is used to calculate the fitness of each chromosome. After that check for stopping criterion such as stop the genetic algorithm when it has generated the new generation over 10 times if the stop criterion is not met: generate the new generation with genetic operation, the process is the series of several operations staring from random selection for the 2-parent chromosome with selection technique such as the roulette wheel selection. Then, crossover the 2-parent chromosome for the exchange of genetic material between the parent chromosomes with crossover technique such as one-point crossover. The results of crossover process are the new 2 chromosomes. After that, perform mutation by randomly selecting chromosome for changing some gene within the chromosome. At the end of these operations, the old population is replaced by the chromosome from the new generation with replacement technique.

III. Proposed Technique

A. Research Framework

In the proposed work, we design the research framework for improvement the imbalanced data classification as shown in Fig. 3.

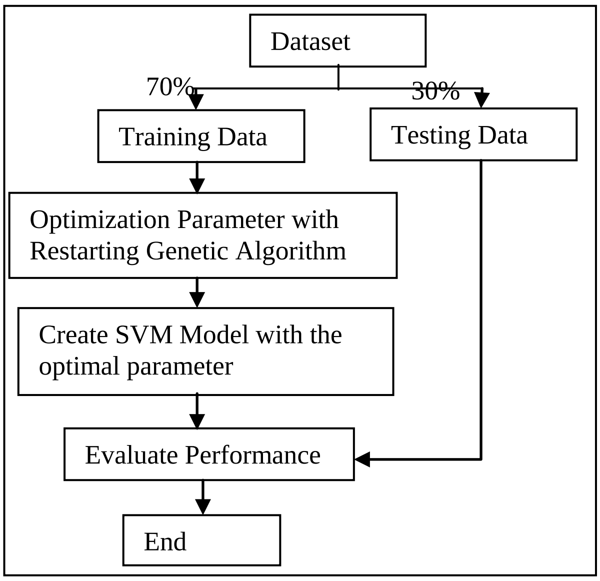
Dataset
70%
30%
Training Data
Testing Data
Optimization Parameter with Restarting Genetic Algorithm
Create SVM Model with the optimal parameter
Evaluate Performance
End

Fig. 3. Research framework.

We can describe our proposed framework as follows. Step 1: we split data into 2 sets, the training set with 70% of all data and the testing set is 30% of all data. Step 2: we find the optimal parameter for support vector machine with used the data from training set and adapt the simple genetic algorithm with restart techniques to find the optimal parameter. Step 3: we apply the optimal parameter to create SVM classification model. Step 4: we evaluate performance SVM classification model with apply the data from testing set.

B. Restarting Genetic Algorithm

In the proposed work, we adapt the simple genetic algorithm by applying the restarting technique in the initial population process to restart the initial population step when the new generation has fitness value lower than the old population and the termination criterion has not been met. The process of restarting genetic algorithm shown in Fig. 4.

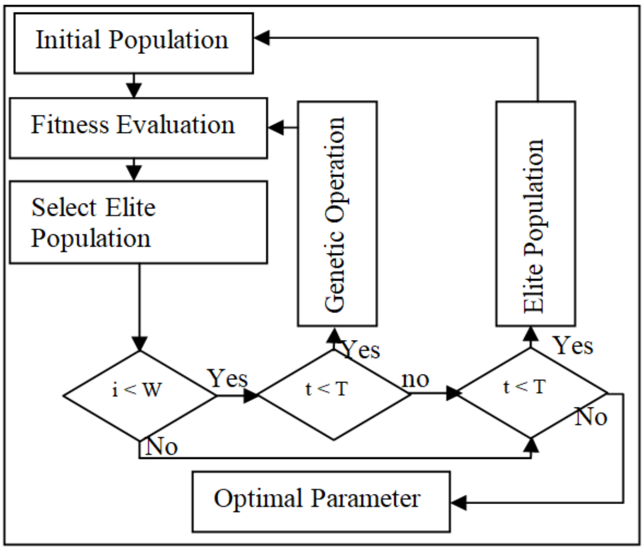

Fig. 4. Restarting genetic algorithm.

The steps of restarting genetic algorithm can be described as follows.

Step 1: Random initial population creation until obtaining the specified population size.

Step 2: Evaluate the fitness of all populations (in this work, we used the measure for fitness evaluate is Recall).

Step 3: Select the top k chromosomes that have the highest fitness values call elite chromosomes.

Step 4: Start the process of genetic operation with random selection the two chromosomes to be the parent chromosomes, then crossover the parent chromosomes and randomly pick one chromosome for mutation, replace the old population with the new generation and evaluate fitness of all new population.

Step 5: Compared the fitness values between the new population and the elite chromosome. If the new population has higher fitness value than the elite chromosome, then repeat the genetic operation process, the replacement process, and the fitness evaluation process of the new generation until stopping criterion is met. Suppose the new population has the fitness value less than the fitness value of the old population and stopping criterion is not met, we restart the random initial population process with elite chromosome and repeat all steps.

IV. Experimental Result

A. Parameter Setup

For parameter optimization based on restarting genetic algorithm, we specified the values of parameters shown in Table II.

We specified the population size for initial population process with random initial 100 population, used the measurement for fitness evaluate is Recall and select the elite chromosome that have highest Recall with 20 chromosomes, and the iteration (stop criterion) with 100 times, the worse generation means the new population has the fitness value less than the fitness value of the old population, and Cost, Epsilon, Gamma is the important parameter for SVM and its genes in chromosome for restarting genetic algorithm.

TABLE II: PARAMETERS FOR RESTARTING GENETIC ALGORITHM

PARAMETER	VALUE	PARAMETER	VALUE
PROB. CROSSOVER	0.8	COST	10^{-4}-10^{-2}
PROB. MUTATION	0.01	EPSILON	10^{-2}-10
POPULATION SIZE	100	GAMMA	10^{-3}-10
ITERATION	100	WORSE GENERATION	5
ELITE CHROMOSOME	20		

B. Dataset

In this research, we used the default of credit card clients Data Set [9]. The dataset contains 30,000 data records with 24 attributes. There are 6,636 for class yes (minority class) and 23,367 for class no (majority class). The imbalanced ratio for this dataset is 3.52

C. Result

To evaluate performance of the classification model, we used the accuracy and recall because in this research aim to improve performance for classify the minority class. We compare the classification performance of our proposed method against the SVM with default parameter, SVM with optimization parameter by using grid search, SVM with optimization parameter by using simple genetic algorithm. In case of simple genetic algorithm, we run the program repeatedly 10 times. The comparative results shown in Table III.

TABLE III: COMPARATIVE PERFORMANCE

ALGORITHM	ACCURACY	RECALL
SVM + DEFAULT PARAMETER	81.40	33.55
SVM + GRID SEARCH	77.88	34.81
SVM + GENETIC ALGORITHM (WORSE)	76.60	26.22
SVM + GENETIC ALGORITHM (BEST)	79.76	35.41
PROPOSED TECHNIQUE	79.76	35.41

It can be seen from the results in Table III that when consider overall accuracy the SVM with default parameter has highest accuracy at 81.40%, whereas proposed technique and SVM with optimization parameter by using genetic algorithm (best of 10 times for experimental) is the second best accurate model at 79.76%, SVM with optimization parameter by using grid search is the third at 77.88%, and SVM with optimization parameter by using genetic algorithm (worse of 10 times for experimental) is the worst with 76.60%.

When considering recall measurement on minority class recognition, we found that the proposed technique and SVM with optimization parameter by using genetic algorithm (best of 10 times for experimental) performs the best at recall rate 35.41%, the second best is SVM with optimization parameter by using grid search (34.81%), whereas SVM with default parameter is the third one (33.55%), and SVM with optimization parameter by using genetic algorithm (worse of 10 times for experimental) is the worst (26.22%).

It's can be seen that when considering only accuracy the SVM with default parameter show powerful than the other algorithm, However, SVM with default parameter predict rate over minority class is still low (Recall is low), our proposed technique performs better than others algorithm to classify the minority class by using the original dataset.

The proposed techniques show still low performance for predict the minority class because the dataset have higher imbalanced ratio, the number of majority class have more than the number of majority class, when create the model with original data the model to learn the features of the minority class not enough.

V. CONCLUSION

The main problem of the imbalanced data classification is that the performance for predict the minority class is still low. The random sampling techniques has been applying to solve the imbalanced data classification, but the effect of random techniques is, sometimes the random technique has select the features is clearly different from both, when the unseen data has features look like the majority class, but the actual class is the minority class, the classification model show miss classification because the model learning sample data does not complete.

The general datamining techniques also classify the balanced data, when the data has imbalanced the general techniques show poor performance to classify the minority class. To improve performance to classify the minority class by using the original data is applying the optimization techniques to find the optimal parameter for classification algorithm. The popular technique is genetic algorithm, but the main problem of genetic algorithm is that sometimes the algorithm cannot find the best set of parameters because the random set of parameters at the initial population is not cover the best set of parameters. We thus proposed the technique that apply the re-create the initial population process, when the new generation show powerful less than the old population. The proposed techniques guarantee that we can find the best set of parameters with 1 time.

The experimental result shown that when we used the imbalanced dataset with the original data (not re-sampling to pre-processing step of classification to balance amount of data in each class) the proposed techniques improved the performance for predict the minority class more than the other optimization techniques.

REFERENCES

[1] J. Han, and M. Kamber, *Data mining: Concepts and Techniques*, San Francisco: Margan Kaufmann, 2006.

[2] 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.

[3] 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.

[4] 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.

[5] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Computing*, vol. 13, no. 3, pp. 213-225, 2009.

[6] C. Cortes and V. Vapnik, "Support vector network," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[7] F. Guo-he, "Support vector machine parameter selection method," *Computer Engineering and Applications*, vol. 47, no. 3, p. 123, 2011.

[8] H. Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor: The University of Michigan Press, Michigan, 1975.

[9] UCI Machine Learning Repository Dataset. (2016). *Default of Credit Card Clients Dataset*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Keerachart Suksut is currently a lecturer at Computer Engineering Department, Rajamangala University of Technology Isan, Thailand. He received his doctoral degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2016, master degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2014, and bachelor degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2012. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

Nuntawut Kaoungku is currently a lecturer at School of Computer Engineering, Suranaree University of Technology, Thailand. He received his doctoral degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2014, bachelor degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2012, and master degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2013. His current research includes data mining and semantic web.

Kittisak 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, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer 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.

Niataya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received he r bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and 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.