Jurnal Teknologi

Full Paper

AN OPTIMIZED SUPPORT VECTOR MACHINE WITH GENETIC ALGORITHM FOR IMBALANCED DATA CLASSIFICATION

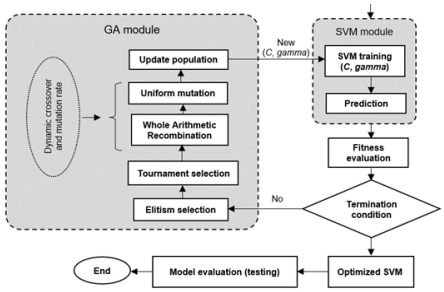
Haziqah Shamsudin[a], Umi Kalsom Yusof[a*], Yan Haijie[a], Iza Sazanita Isa[b]

[a]School of Computer Sciences, USM, 11800, Penang, Malaysia
[b]Centre for Electrical Engineering Studies, Universiti Teknologi MARA, Cawangan Pulau Pinang, UiTM, 14300, Penang, Malaysia

Article history
Received
29 December 2022
Received in revised form
28 March 2023
Accepted
10 April 2023
Published Online
25 June 2023

*Corresponding author
umiyusof@usm.my

Graphical abstract



Abstract

In supervised machine learning, class imbalance is commonly occurring when the number of examples that represent one class is much lower than other classes. Since an imbalance data may generate suboptimal classification models, it could lead to the minority examples are misclassified frequently and hardly achieving the best performance. This study proposes an improved support vector machine (SVM) method for imbalanced data namely as SVM-GA by optimizing SVM algorithm with Genetic Algorithm (GA) over a synthetic minority oversampling technique. Besides considering the best sampling method in optimized SVM, the experimental result shows that the proposed method improves by 97% compared to the baseline model and selected optimized models. The proposed model had significant performance by outperformed the baseline model and other models based SVM with Grid search and Randomized search in most of the cases, especially for the datasets which have extremely rare cases.

Keywords: Machine learning, data classification, sampling method, support vector machine, genetic algorithm

Abstrak

Dalam pembelajaran mesin yang diselia, ketidakseimbangan kelas biasanya berlaku apabila bilangan contoh yang mewakili satu kelas jauh lebih rendah daripada kelas lain. Oleh kerana data ketidakseimbangan boleh menjana model klasifikasi suboptimal, ia boleh menyebabkan kelas minoriti sering dikelaskan salah dan tidak mencapai prestasi terbaik. Kajian ini mencadangkan kaedah algoritma mesin vector sokongan (SVM) yang lebih baik untuk data yang tidak seimbang iaitu SVM-GA dengan mengoptimumkan algoritma SVM dengan Algoritma Genetik (GA) berbanding teknik oversampling minoriti sintetik. Selain mempertimbangkan kaedah pensampelan terbaik dalam SVM yang dioptimumkan, hasil eksperimen menunjukkan bahawa kaedah yang dicadangkan mencapai 97% berbanding model asas dan model yang dioptimumkan terpilih. Model yang dicadangkan mempunyai prestasi yang signifikan dengan mengatasi model asas dan model lain berdasarkan SVM dengan carian Grid dan carian rawak dalam kebanyakan kes, terutamanya untuk set data yang mempunyai kes yang sangat jarang berlaku.

Kata kunci: Pembelajaran mesin, pengkelasan data, kaedah pensampelan, mesin vector sokongan, algorithma genetic

1.0 INTRODUCTION

Recent developments in science and technology have enabled the growth and availability of raw data to arise at an explosive rate [1]. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry [2]. The imbalanced class problem refers to a classification problem where classes are not equally represented and have been described as a situation when the number of observations of one class (the majority class) far exceeds that of the other class (minority class) [3].

If the imbalance ratio (IR) > 1.5, the dataset suffers from an imbalanced class problem. If IR> 9, a dataset is highly imbalanced [4]. Imbalanced class problems in the dataset will lead to misclassification errors and bias towards the majority class when doing any classification [5].

The minority class usually represents the most important concept to be learned, and it is difficult to identify due to possibility of associated with exceptional and significant cases [6] because the data acquisition of these examples is costly [7]. Thus, there is a necessity of providing solutions for class imbalance problems in which traditionally solved using data level approaches, algorithm level approaches or hybrid approaches.

At the data-level approach, the training instances are modified in such a way to produce a more or less balanced class distribution that allows classifiers to perform similarly to standard classification. On the other hand, algorithm-level tackles the issues of the imbalanced class problem by adapting base learning methods to be more attuned to class imbalance issues. Lastly, the hybrid approach tackles the issues at both levels; data-level and algorithm-level [4, 8].

However, the preprocess in data level approaches such as oversampling technique to make the data distribution balanced to reduce the impact of the skewed class distribution in the subsequent classifier learning stage has drawbacks of generating meaningless samples, and computationally costly for some large datasets. In contrast, under-sampling has higher efficiency but may lose a few patterns and hence cause other unexpected mistakes.

On the other hand, at the algorithm-level the research interest in machine learning with unbalanced data has shown a growth trend and presently become interested discussion topics for the classification domain. In addition, researchers have used intelligence-based optimization algorithms such as Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) due to their global parallelism with acceptable running time. Although some approaches based on imbalanced data have found more success to some degree and turned out to be increasingly popular, there are still some limitations for the models sorely relied on one technique.

Researchers has commonly used SVM to tackle the classification of highly imbalanced data in real world applications due to superior performance in practical applications, i.e, medical diagnostics [9], human activity [10], financial market [11], time series [12], autonomous vehicle [13], text information filtering [14] , etc. SVM, a technique widely recognized for optimizing anticipated solutions, was first proposed by Vapnik *et al*. [16] as a kernel-based machine learning model for classification and regression tasks [15, 16]. The SVM's remarkable capacity for generalization, as well as its discriminative power and optimal solution, has recently gained significant attention from the data mining, pattern recognition, and machine learning communities [17].

With its exceptional ability to solve practical binary classification problems, SVM has been demonstrated to outperform other supervised learning methods [18, 19]. Because of its solid theoretical foundation and impressive generalization capabilities, SVM has become one of the most commonly used classification methods in recent years. The introduction of GA in intelligent optimization algorithm has significant role in optimization problems since the approach does not require too much information, and not easily fall into the local optimum like PSO [20].

Since GA is performed by parallel search instead of a point-to-point search, the effective areas exploration in the space through the population provides not easy to fall into a local minimum. As example, Yan *et al*. [21] proposed a GA optimized SVM for NOMA-based downlink satellite networks by fixed crossover rate and mutation rate in optimal parameters C and gamma of GA selection. The study proved that the proposed method was better than the traditional SVM with randomly selected parameters and had applicability for NOMA scheme.

Meanwhile, Abdullah *et al*. [22] used a GA to optimize SVM parameters for an acute leukemia diagnosis by not only chose C, $gamma$ as the body of an individual of GA, but also included the feature mask of datasets. The study concluded that the proposed model had high accuracy and suitable for implementation in medical application of acute leukemia. Similar work proposed by Yao *et al*. [23] that optimized SVM based on modified GA for fast classification of tea leaves has also proved superior comparison towards SVM optimized by PSO and CV and provided a promising way for the classification of tea products.

According to a review conducted by Sourabh *et al*. [24], the Genetic Algorithm (GA) remains relevant despite being an older algorithm, as it has great potential for improvement and state-of-the-art methods. GA's versatility and adaptability to a wide range of scenarios have enabled it to maintain its effectiveness in converging well. Hence, GA is still considered a valuable tool for various applications.

Although a significant result has been presented by

Klempka *et al*. [25] in SVM parameters optimization with grid search, the work was having limitation to converge at a slower speed particularly for large dataset and consumed higher running time. A combination of GA optimization algorithm rather than a single SVM is expected to provide better results as they can incorporate learning more stable and greatly faster than grid search [26]. Alternatively, heuristics methods such as genetic and evolutionary algorithms are faster and more efficient to approximate the solution of computationally expensive problems. It has been applied in solving numerical problems and prediction [26, 27].

As example, the data level approaches such as Synthetic Minority Oversampling Technique (SMOTE) [28] are easy to implement and flexible but may suffer from problems of overfitting. To counter the limitations, an improved algorithm for SMOTE have been proposed such as Borderline-SMOTE algorithm [29] and adaptive synthetic sampling algorithm (ADASYN) [30]. Meanwhile, other methods that combined both oversampling and under-sampling techniques are the SMOTE+Tomek [31] algorithm and SMOTE+ENN [31] to generate synthetic samples and employed Tomek algorithm to clean up the noise.

As summarized in Table 1, although the techniques performed well at specific research problem for imbalance data, most of the instances exist flaws and inadequacies including but not limited to overfitting, unimpressive improvement and high complexity [31]. Similarly, there are some disadvantages in SVM parameters optimization methods such as being time-consuming and CPU intensive. Most importantly, the SVM parameters optimization methods mainly work on balanced datasets. Therefore, the combination of sampling methods and optimized SVM with GA would be more preferred due to the simplicity and flexibility of sampling methods as well as the advantages of GA.

Table 1 GA for SVM parameters optimization approaches.

Techniques	Dataset	Findings
GA & optimized SVM [21]	Simulation data	Effective user classification in NOMA based satellite systems
GA & SVM parameter optimization and feature selection [22]	Medical images	Accuracy: 99.1870% Sensitivity: 98.1481 Precision: 99.4845 F-measure: 98.8118
Modified GA optimized SVM for rapid classification of tea leaves [23]	Chinese tea samples from the local market.	Accuracy of training set and test set is 99.73% and 98.4% respectively
Comparison of using GA and cuckoo search for multicriteria optimization with limitation [25]	Simulation data	Optimization methods (GA and CS) found acceptable solutions; however, CS was the slower algorithm
SVM parameter optimization using Grid Search and GA [26]	Datasets from UCI and BioInformatics Group Seville	SVM parameter optimization using GA is more than 15.9 times faster and more stable than using grid search

Therefore, using one technique alone to cope with imbalanced data problems is insufficient but combining techniques for imbalanced data with Support Vector Machine (SVM) parameters optimization techniques are very spare to improve the performance of proposed approach. To the best of our knowledge, no prior work on improved sampling method for imbalanced dataset and optimized SVM-GA incorporated sampling-based machine learning techniques to improve the classification accuracy of minority class as well as the overall performance, which is the novelty of this current study.

2.0 METHODOLOGY

Figure 1 shows a schematic representation of the proposed method namely as SMOTE-SVM-GA which were divided into four stages: problem modelling and formulation, selection of best sampling method, implementation of optimized SVM with Genetic Algorithm and result comparisons.

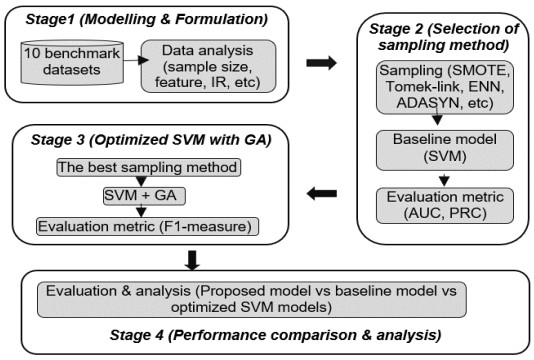


Figure 1 Proposed research framework of SMOTE-SVM-GA

The proposed approach was implemented using Python 3.7 on macOS with 16 GB RAM and Intel Core i7 and Python as programing language whilst Scikit-learn was used for preprocessing and evaluation metrics.

2.1 Dataset

The experiments used 10 different highly imbalanced datasets including abalone19, abalone9-18, ecoli4, glass2, glass5, page-blocks-1-3vs4, shuttle-c0-vs-c4, vowel0, yeast5 and yeast6 from KEEL [32] data repository. The collection of datasets in the KEEL repository is highly recommended for examining

imbalanced classification problems and is still widely used to this day [34].

As summarized in Table 2, the datasets were composed by two classes (relative lack of minority sample data and the absolute lack of minority sample data) and vary in their degree of numbers of samples, number of attributes, and imbalance ratio. Relative lack defines that the minority sample is not small in absolute number like dataset shuttle-c0-vs-c4, but its number of samples is small compared to the majority. In this case, the data is also not conducive for discrimination of the minority class since the majority class sample will blur the boundary of the minority class sample, and difficult to distinguish the minority class sample from the majority class.

In the table, the No. of Instances represents the total sample size of the data. The number of attributes represents the total number of features or columns in the data, while Majority/Minority indicates the total sample of class labels which represents majority samples and minority samples. Lastly, IR is imbalance ratio, required to understand the severity of the imbalanced class problem, whether it's highly imbalanced or low imbalanced. The formula to calculate the value of IR is defined in Equation 1:

$$IR = \frac{majority}{minority} \quad (1)$$

where, if IR>1.5, the dataset is consider as imbalanced and if IR>9 it is consider as highly imbalanced [4]. The dataset in Table 2 is sorted based on the IR value from smallest to largest.

Table 2 Description and characteristics of highly imbalanced datasets

Dataset	No. of Instances	No. of Attributes	Minority/ Majority	IR
vowel10	988	13	90/898	9.98
Glass2	214	9	197/17	11.59
Shuttle-c0-vs-c4	1829	9	1706/123	13.87
Ecoli4	336	7	314/20	15.70
Page-blocks-1-3vs4	472	10	444/28	15.86
Abalone9-18	731	8	689/42	16.40
Glass5	214	9	205/9	22.78
yeast5	1484	8	1440/44	32.73
yeast6	1484	8	1449/35	41.40
abalone19	4174	8	4124/32	129.44

2.2 Selection of Sampling Methods

Eight common sampling methods in scikit-learn were applied to the imbalance dataset for selecting the most effective sampling method. The sampling methods are SMOTE, BorderlineSMOTE, ADASYN, ENN, TomekLinks, Nearmiss-3, SMOTEENN and SMOTETomek. These sampling methods and SVM have been used to perform the prediction of imbalanced datasets along with a baseline model SVM performed at the initial dataset without any resampling. Selection of most effective sampling method made based on metric performance evaluation of ROC curve, PR curve and AUC.

2.3 Optimized SVM-GA Algorithm

Based on the basis of genetic biology and evolution theory, the GA starts the entire search process from the initial population of P (i=1,2,...,n). The initial population is randomly generated, and each individual in the population is a possible solution to the problem. These individuals continue to evolve in the iterative process and finally, the best solution comes out after specific iterations [24, 35]. The process of evolution is mainly achieved through operations such as selection, crossover, and mutation along with some parameter settings as depicted in Figure 2.

Genotype Representation is the first problem to be solved in GA since improper representation can result in poor performance. Due to the random nature of GA, searching effect is very poor for complex problems and therefore for problem that involved continuous variables rather than discrete variables, real value representation is the most natural. If the problem refers to direction and similar problems, integer representation is suitable [36].

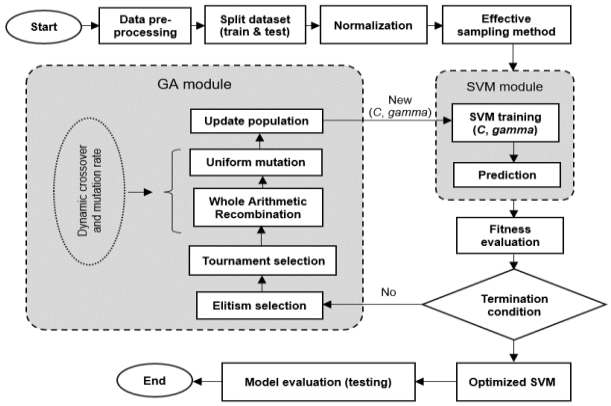


Figure 2 Design procedure of proposed SVM-GA

If the solution requires order and does not contain duplicate values, permutation representation is preferred. In this work, given that parameters of SVM are continuous variables and to speed up the converge time of population, real value representation is selected. Each variable consists of two positions; the value of parameter C is in one position while $gamma$ is in another position.

Population since population setting is also important, the population was setting as 20 in terms of the trade-off between time and performance and initial the population with random real values between 10-3 and 103 to justify if the population is too large, the

GA will be slow; if it is too small, it will not be enough to obtain a good mating [36].

Fitness Function the fitness function takes the candidate solution of the problem as input and outputs the best individual since the function act differently on different problems. It not only should be appropriate to the problem but also refers to the calculation efficiency. In this work, the GA generates different SVM algorithms through different individuals (*C*, *gamma*) to run on the new generated dataset. The F1-measure of each SVM corresponds to the fitness of the individual of all the 10 datasets. Since fitness evaluation is computational and costly, the multiprocessing and concurrent libraries were used in parallel as reasonable approach.

Parent Selection includes fitness proportional selection, ranking selection, uniform parent selection, and tournament selection. In this study, the selection pressure was selected due to its simplicity and easy to control the tournament size, K of tournament selection [37]. The function randomly selects K individual from the population and choose the best as a parent whilst the remained parent was selected in the same process [35]. In this study, the tournament size is set by the remainder after the population size is divided by 5.

Crossover and Mutation crossover rate and mutation rate are important research factors for optimizing GA to get the global optimal value to avoid the problems of premature maturity and falling into the local optimum [35, 38]. In this work, the crossover and mutation were set as Equation 2 and Equation 3 respectively.

$$crossover\ rate = 1 - \frac{current\ generation}{total\ generation} \tag{2}$$

$$mutation\ rate = \frac{current\ generation}{total\ generation} \tag{3}$$

Besides, the Whole Arithmetic Recombination is set as the crossover operators, child1 and child2 as expressed in Equation 4 and Equation 5, given that a is operator variable whilst x_1 and x_2 are input training respectively. To jump out of the local optimum and explore the entire parameter space unbiasedly, the mutation operator is implemented by Uniform Mutation which randomly selects a real value in a range from a distribution [35].

$$child_1 = \alpha x_1 + (1 - \alpha) x_2 \tag{4}$$

$$child_2 = (1 - \alpha) x_1 + \alpha x_2 \tag{5}$$

Survivor Selection. elitism selection is considered in this study as survivor selection due to its capability to selects the fittest in every iteration and discards other individuals as loss prevention of the current fittest of in the population. From a review by [38], out of eight selection methods available, three of it are elitist based method that always include the best individual in the new population. When using an elitist method, the best fitness in the population cannot decrease over time [38, 39, 40].

Termination Condition. the maximum generation of the GA was setting at 1600 that defined as stoppable cycles for the repeating algorithm [35, 41]. The overall summarization parameter settings of the GA are shown in Table 3.

Table 3 Overall parameter settings of GA

Parameter	Value
Genotype Representation	Real value representation
Population Size	20
Fitness Function	Mean of F1-measure on 10 datasets
Parent Selection	Tournament selection
Crossover	Dynamic, Whole Arithmetic Combination
Mutation	Dynamic, uniform mutation
Survivor Selection	Elitism selection
Termination Condition	1600 generations

2.4 Optimized SVM With Other Techniques

The grid search and randomized search were implemented for optimizing the parameters of SVM. In grid search, the range of the required parameter changes were initially set, and next changed these parameter values in the space according to their combination. Each of which gets the value of the objective function. Finally, the approximate optimum of the objective function is found by searching the entire parameter settings space. Meanwhile, the randomized search is similar to the grid search, but instead of trying all possible parameter combinations, it samples a fixed number of parameter settings from the specified distribution. The number of parameter settings that were sampled is given by n_iter of 300 [42].

3.0 RESULTS AND DISCUSSION

The performance based on the metrics of F1-measure, ROC curve and PR curve were used to evaluate effectiveness of the proposed model's predictions compared to baseline model and other models.

3.1 Results of Best Sampling Methods

As summarized in Table 4, SMOTE and SMOTE+Tomek have the same value of AUC and rank first in seven datasets. Although the baseline model is also the number one in seven datasets, it is inferior to these two models. Considering the overall performance as well as time expenditure of these models from the two curves and according to the popular Occam's razor principle in machine learning which says that simple theories are better than complex ones with other

things equal, SMOTE is selected as the most effective sampling method among these methods.

3.2 Results of Proposed Model and Baseline Models

Based on result showing in Figure 3, the proposed model (SVM+GA) when integrated with SMOTE is superior to other models by achieving the highest F1-measure for seven datasets. Most obviously, it greatly improves the classification performance for glass2 and glass5 datasets, which gets 0.625 and 1 from F1-score while the baseline model gets zero scores at these two datasets. The reason we can infer is that these two datasets both have rare cases with a small size which may lead to small disjuncts.

From the results, compared to the classification without oversampling, the classification combined with oversampling is significantly better when the number of clusters of the minority class increases. Out of all oversampling techniques, it is interesting to note that the performance of SMOTE outperformed all the other methods because of its good performance in the majority of the datasets.

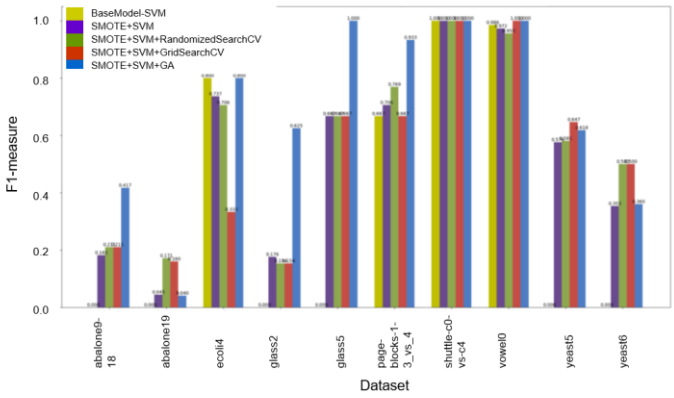


Figure 3 Performances of proposed and baseline models on 10 datasets

Table 4 Result s of ROC (AUC) and time of all SVMs-based sampling methods and a baseline SVM model without any resampling in 10 datasets

Dataset	SMOTE		Borderline SMOTE		ADASYN		ENN		TomekLinks		NearMiss		SMOTE+ENN		SMOTE+Tomek		baseModel	
	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)	AUC	*t*(s)
abalone19	**0.81**	0.07	0.77	0.07	0.80	0.07	0.72	0.02	0.70	0.02	0.67	0.01	0.80	0.06	**0.81**	0.07	0.75	0.01
Abalone9-18	**0.84**	2.29	0.61	1.29	0.82	2.20	0.82	0.13	0.79	0.12	0.64	0.01	0.85	1.64	**0.84**	2.34	0.88	0.09
Ecoli4	**0.98**	0.01	0.98	0.01	0.97	0.01	0.99	0.01	0.99	0.01	0.99	0.01	0.98	0.01	**0.98**	0.01	0.99	0.00
Glass2	0.67	0.01	0.63	0.01	0.63	0.01	0.73	0.01	0.66	0.01	0.69	0.01	0.65	0.01	0.67	0.01	0.62	0.01
Glass5	**0.99**	0.01	0.99	0.01	0.99	0.01	0.98	0.01	0.99	0.01	0.28	0.01	0.98	0.01	**0.99**	0.01	0.99	0.00
Page-blocks-1-3vs4	**0.99**	0.04	0.99	0.02	0.99	0.01	0.99	0.01	0.99	0.01	0.82	0.01	0.99	0.02	**0.99**	0.02	0.99	0.00
Shuttle-c0-vs-c4	**1.00**	0.02	1.00	0.01	1.00	0.02	1.00	0.01	1.00	0.01	1.00	0.01	1.00	0.03	**1.00**	0.03	1.00	0.01
vowel10	**0.99**	0.06	0.99	0.06	0.99	0.05	0.99	0.02	0.99	0.02	0.02	0.01	0.99	0.05	**0.99**	0.07	0.99	0.01
yeast5	**0.89**	0.08	0.87	0.07	0.89	0.11	0.85	0.02	0.88	0.02	0.83	0.01	0.88	0.08	**0.89**	0.10	0.87	0.01
yeast6	0.59	0.09	0.35	0.07	0.58	0.10	0.53	0.02	0.56	0.02	0.20	0.01	0.51	0.07	0.59	0.10	0.56	0.01

3.3 Results of Proposed Model and Baseline Models

Based on results tabulated in Table 5, the proposed model improves the prediction capability of a traditional SVM in terms of F1-measure, indicating a good classification performance on highly imbalanced data. The proposed model allows itself to recognize minority samples, even when the traditional SVM failed to recognize any minority sample (as in the case of abalone19, abalone9-18, glass2, glass5, yeast5 and yeast6 dataset).

Table 5 Comparisons of F1-Measure Results between Models

Dataset	Baseline model	SMOTE +SVM	SMOTE +SVM +RS[b]	SMOTE +SVM +GS[c]	Proposed model
abalone19	0.000	0.045	0.171	0.160	0.040
Abalone9-18	0.000	0.182	0.211	0.211	0.417
Ecoli4	0.800	0.737	0.706	0.333	0.800
Glass2	0000	0.176	0.154	0.154	0.625
Glass5	0000	0.667	0.667	0.667	**1.000**
Page-blocks-1-3vs4	0.667	0.706	0.769	0.667	0.933
Shuttle-c0-vs-c4	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
vowel10	0.986	0.972	0.955	**1.000**	**1.000**
yeast5	0.000	0.576	0.581	0.647	0.618
yeast6	0. 000	0.353	0.500	0.500	0.360

In addition, the overall average F1-measure results of different datasets are tabulated in Table 6. The proposed model improved with a range of improvement by 0.108 to 0.334 in terms of overall average F1-measure at an acceptable time

expenditure as compared to other models. This indicates that ability to produce a better percentage of F1-measure means to increase the ability in producing an accurate prediction of imbalanced data and significantly solves imbalanced data problems encountered in real scenarios.

From Table 6, the proposed model able to reach highest performance in the condition of highly imbalanced class problem. Compared to previous work, the model able to classify well between both class and reduce the bias on the performance, where it performs fairly well for both classes.

Table 6 Comparisons of overall average F1-measure results

Method	Average F1-measure	Time (s)
Proposed-model (SMOTE+SVM+GA)	0.679	5933.28
Baseline model	0.345	0.05
SMOTE+SVM	0.541	0.71
SMOTE+SVM+RandomizedSearch	0.571	413.79
SMOTE+SVM+GridSearch	0.534	14177.01

Besides, the baseline model produces an inferior performance despite the least time expenditure. Other three models provide similar performances with a medium improvement for baseline model while grid search requires an intolerable time expenditure.

4.0 CONCLUSION

In this study, the traditional SVM with multiple different sampling methods performed prediction on ten highly imbalanced datasets. They are evaluated on the basis of the comprehensive results of the F1-measure and time expenditure.

As such, the most effective sampling method was obtained. And then, this most effective sampling method was combined with an optimized SVM with Genetic Algorithm (SVM-GA) for addressing the same datasets as mentioned before.

The proposed model had outperformed the baseline model and other models based SVM with Grid search and Randomized search in most of the cases, especially for the datasets which have extremely rare cases. However, there were few datasets did not perform well and achieved a lower detection rate.

In this study, there are also some limitations in terms of the number of samples and number of attributes. The datasets used mostly have less than 5000 sample size and lesser number of features. The proposed model might not work well in more complex environment. In future, the model should be evaluated in a more complex datasets to measure the effectiveness of the proposed model in more complex environment.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The authors wish to thank the Aims for Excellent research group, School of Computer Science, USM Penang and Centre for Electrical Engineering Studies, UiTM Penang for any assistance provided in the preparation of this paper.

References

[1] Gautheron, L., Habrard, A., Morvant, E., & Sebban, M. 2019. Metric Learning from Imbalanced Data. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. 923-930. https://doi.org/10.1109/ICTAI.2019.00131.

[2] S. Maheshwari. 2017. A Review on Class Imbalance Problem: Analysis and Potential Solutions. *International Journal of Computer Science Issues.* 14(6): 43-51. https://doi:10.20943/01201706.4351.

[3] Lee, Han Kyu and Seoung Bum Kim. 2018. An Overlap-sensitive Margin Classifier for Imbalanced and Overlapping Data. *Expert Systems with Applications.* 98: 72-83. https://doi.org/10.1016/j.eswa.2018.01.008.

[4] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera. 2018. Learning from Imbalanced Data Sets. 10: 978-3. https://doi.org/10.1007/978-3-319-98074-4.

[5] J. Zheng. 2010. Cost-sensitive Boosting Neural Networks for Software Defect Prediction. *Expert Systems with Applications.* 37(6): 4537-4543. https://doi.org/10.1016/j.eswa.2009.12.056.

[6] V. López, A. Fernández, S. García, V. Palade, F. Herrera. 2013. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on using Data Intrinsic Characteristics. *Info. Sciences.* 250: 113-141. https://doi.org/10.1016/j.ins.2013.07.007.

[7] G. M. Weiss, Y. Tian. 2008. Maximizing Classifier Utility when There are Data Acquisition and Modeling Costs. *Data Mining and Knowledge Discovery*. 17(2): 253-282. https://doi.org/10.1007/s10618-007-0082-x.

[8] N. Rout, D. Mishra, M. K. Mallick. 2018. Handling Imbalanced Data: A Survey. *Advances in Intelligent Systems and Computing.* 628: 431-443. https://doi:10.1007/978-981-10-5272-9_39.

[9] A. H. Khandoker, M. Palaniswami, C. K. Karmakar. 2009. Support Vector Machines for Automated Recognition of Obstructive Sleep Apnea Syndrome from ECG Recordings. *IEEE Trans. on Info. Techn. in Biomedicine*. 13(1): 37-48. https://doi.org/10.1109/TITB.2008.2004495.

[10] Y. Kim, H. Ling. 2009. Human Activity Classification based on Micro-doppler Signatures using a Support Vector Machine. *IEEE Trans. on Geoscience and Remote Sensing.* 47(5): 1328-1337. https://doi.org/10.1109/TGRS.2009.2012849.

[11] Q. Jin, K. Guo, Y. Sun. 2017. Stock Price Forecasting using Support Vector Regression: Based on Network Behavior Data. *Proceedings, IEEE Int. Conf. on Big Data, Big Data.* 4148-4153. https://doi.org/10.1109/BigData.2017.8258436.

[12] N. Sapankevych, R. Sankar. 2009. Time Series Prediction using Support Vector Machines: A Survey. *IEEE Comp. Int. Magazine*. 4(2): 24-38. https://doi.org/10.1109/MCI.2009.932254.

[13] Y. Liu, X. Wang, L. Li, S. Cheng, Z. Chen. 2019. A Novel Lane Change Decision-Making Model of Autonomous

Vehicle Based on Support Vector Machine. *IEEE Access.* 7: 26543-26550. https://doi.org/10.1109/ACCESS.2019.2900416.
[14] Jing, O. 2020. Research on English Text Information Filtering Algorithm Based on SVM. *2020 IEEE Int. Con. on Power, Intelligent Computing and Systems.* 1001-1004. https://doi.org/10.1109/ICPICS50287.2020.9202016.
[15] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez. 2020. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing.* 408: 189-215. https://doi.org/10.1016/j.neucom.2019.10.118.
[16] V. Vapnik. 1999. The Nature of Statistical Learning Theory. Springer Science & Business Media.
[17] T. K. Bhowmik, P. Ghanty, A. Roy, S. K. Parui. 2009. SVM-based Hierarchical Architectures for Handwritten Bangla Character Recognition. International Journal on Document Analysis and Recognition (IJDAR). 12(2): 97-108. https://doi.org/10.1007/s10032-009-0084-x.
[18] J. Cervantes, F. Garcia-Lamont, A. López-Chau, L. Rodríguez-Mazahua, J. S. Ruíz. 2015. Data Selection based on Decision Tree for SVM Classification on Large Data Sets. *Applied Soft Computing.* 37: 787-798. https://doi.org/10.1016/j.asoc.2015.08.048.
[19] M. Dudjak, G. Martinović. 2021. An Empirical Study of Data Intrinsic Characteristics that Make Learning from Imbalanced Data Difficult. *Expert Systems with Applications.* 182: 115297. https://doi.org/10.1016/j.eswa.2021.115297.
[20] Y. Tian, Q. Zhang, D. Liu. 2014. v-Nonparallel Support Vector Machine for Pattern Classification. *Neural Computing and App.* 25(5): 1007-1020. https://doi.org/10.1007/s00521-014-1575-3.
[21] X. Yan, K. An, C. X. Wang, W. P. Zhu, Y. Li, Z. Feng. 2020. Genetic Algorithm Optimized Support Vector Machine in NOMA-based Satellite Networks with Imperfect CSI. *IEEE Int. Con. on Acoustics, Speech and Signal Processing.* 8817-8821. https://doi.org/10.1109/ICASSP40776.2020.9053003.
[22] N. A. Abdullah, M. A. Ibrahim, A. S. Haider. 2020. GA as a Key Parameter of SVM Parameter Optimization and Feature Selection for Acute Leukemia Diagnosis Genetic Algorithm as a Key Parameter of SVM Parameter Optimization and Feature Selection for Acute Leukemia diagnosis. *University of Aden Journal of Natural and Applied Sciences.* 24(2): 385-393. https://doi.org/10.47372/uajnas.2020.n2.a07.
[23] M. Yao, G. Fu, T. Chen, M. Liu, J. Xu, H. Zhou, X. He, L. Huang. 2021. A Modified Genetic Algorithm Optimized SVM for Rapid Classification of Tea Leaves using Laser-induced Breakdown Spectroscopy. *J. of Analy. Atomic Spectrometry.* 36(2): 361-367. https://doi.org/10.1039/d0ja00317d.
[24] S. Katoch, S. Singh Chauhan, V. Kumar. 2021. A Review on Genetic Algorithm: Past, Present, and Future. *Multimedia Tools and Applications.* 80: 8091-8126.
[25] R. Klempka, B. Filipowicz. 2017. Comparison of using the Genetic Algorithm and Cuckoo Search for Multicriteria Optimisation with Limitation. *Turkish J. of Elect. Eng. and Comp. Sci.* 25(2): 1300-1310. https://doi.org/10.3906/elk-1511-252.
[26] I. Syarif, A. Prugel-Bennett, G. Wills. 2016. SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control).* 14(4): 1502. https://doi.org/10.12928/telkomnika.v14i4.3956.
[27] K. Nath Das. 2014. Hybrid Genetic Algorithm: An Optimization Tool. *Global Trends in Intelligent Computing Research and Development.* 268-305. https://doi.org/10.4018/978-1-4666-4936-1.ch010.
[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.* 16): 321-357. https://doi.org/10.1613/jair.953.
[29] H. Han, W. Y. Wang, B. H. Mao. 2005. Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning. *Lecture Notes in Comp. Sci.* 3644: 878-887. https://doi.org/10.1007/11538059_91.
[30] H. He, Y. Bai, E. Garcia, S. Li. 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).* 1322-1328. https://doi: 10.1109/IJCNN.2008.4633969.
[31] G. E. A. P. A. Batista, R. C. Prati, M. C. Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter.* 6(1): 20-29. https://doi.org/10.1145/1007730.1007735.
[32] G. P. Wang, J. X. Yang, R. Li. 2017. Imbalanced SVM-based Anomaly Detection Algorithm for Imbalanced Training Datasets. ETRI Journal. 39(5): 621-631. https://doi.org/10.4218/etrij.17.0116.0879.
[33] J. Alcalá-Fdez, L. Sánchez, S. Garciá, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, F. Herrera. 2009. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. *Soft Computing.* 13(3): 307-318. https://doi.org/10.1007/s00500-008-0323-y.
[34] S. Szeghalmy, A. Fazekas. 2023. A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors.* 23(4): 2333. https://doi.org/10.3390/s23042333.
[35] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, V. B. Surya Prasath. 2019. Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach. *Information.* 10(12): 390. https://doi.org/10.3390/info10120390.
[36] T. Tarkowski. 2022. Genetic Algorithm Formulation and Tuning with Use of Test Functions. arXiv preprint arXiv:2210.03217. https://doi.org/10.48550/arXiv.2210.03217.
[37] N. M. Razali, J. Gerathy. 2011. Genetic Algorithm Performance with Different Selection Strategies in Solving TSP. *Proceedings of the World Congress on Engineering.* 2(1): 1-6. Hong Kong, China: International Association of Engineers.
[38] M. Lynch. 2010. Evolution of the Mutation Rate. *Trends in Genetics.* 26(8): 345-352. https://doi.org/10.1016/j.tig.2010.05.003.
[39] N. Brouwer, D. Dijkzeul, L. Koppenhol, I. Pijning, D. Van den Berg. 2022. Survivor Selection in a Crossoverless Evolutionary Algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference Companion.* 1631-1639. https://doi.org/10.1145/3520304.3533950.
[40] A. Pétrowski, S. Ben-Hamida. 2017. Evolutionary Algorithms. John Wiley & Sons. 7. https://doi.org/10.1002/9781119136378.
[41] M. Safe, J. Carballido, I. Ponzoni, N. Brignole. 2004. On Stopping Criteria for Genetic Algorithms. *Advances in Artificial Intelligence-SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence,* Sao Luis, Maranhao, Brazil, Springer. 405-413. https://doi.org/10.1007/978-3-540-28645-5_41.
[42] J. Bergstra, Y. Bengio. 2012. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.* 13: 281-305. https://dl.acm.org/doi/10.5555/2188385.2188395.