Journal of Intelligent Systems and Internet of Things
Vol. 11, No. 02, PP. 22-29, 2024
DOI: https://doi.org/10.54216/JISIoT.110202

Improving Support vector machine for Imbalanced big data classification

Alaa Abdulazeez Qanbar[1], Zakariya Yahya Algamal[2]

[1] Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

[2]Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

Emails: alaa.22csp59@student.uomosul.edu.iq; zakariya.algamal@uomosul.edu.iq

*Corresponding Author: alaa.22csp59@student.uomosul.edu.iq

Abstract

A significant proportion of one type of pattern and a relatively small quantity of another type of pattern can be found in many unbalanced real data sets. In addition, finding significant observations and excluding influential observations is effectively accomplished through diagnostic analysis. Support vector machines (SVM), a common classification technique, perform poorly on imbalanced datasets and when influential observations exist. In this research, the pigeon optimization algorithm as a metaheuristic algorithm is employed to address the influence observation issues in SVM. Experiments are done on three real sets of data. Our approach provides higher classification accuracy compared to other widely used algorithms. This approach could be used for further biological, chemical, and medical datasets.

Received: August 17, 2023 Revised: November 11, 2023 Accepted: January 11, 2024

Keywords: Pigeon optimization algorithm; meta-heuristic algorithm; imbalanced data; support vector machine.

1. Introduction

In machine learning, classification is the process of dividing data into discrete classes or categories according to particular traits or qualities. Assigning fresh observations to established classes or categories based on the patterns and information discovered from the training data is the aim of this supervised learning technique [1-3].

In classification, labeled examples that are each connected to a recognized class or category make up the training data needed to train the model. With the help of these labeled examples, the model gains knowledge and attempts to identify patterns or relationships in the input data that will allow it to distinguish between various classes with accuracy. Because real-world datasets are often imbalanced, with a minority class that has relatively few instances compared to the other classes in the dataset, the imbalanced learning problem in data mining has garnered a great deal of attention from the research community and practitioners. In supervised learning, standard classification algorithms have trouble accurately classifying the minority class. The majority of these algorithms make the assumption that there is an equal distribution of classes and misclassification costs for every class. Furthermore, these algorithms are made to output the most straightforward hypothesis that best fits the data by generalizing from sample data [4-5].

A kernel-based machine learning model for classification is called support vector machines (SVM) [6]. SVM provides a hyperplane that shows the greatest margin (or separation) between two classes. Nonlinear classification using SVM can divide a feature into two classes. SVM was initially designed to handle linear classification problems; it was then expanded to handle nonlinear problems as well. Finding the vector space hyperplane between two classes is the aim of SVM. Pattern recognition, also known as training, and verification, often known as testing, are the two key steps in SVM. A line known as a hyperplane is created during the pattern recognition process and is used as a separator between two classes that are located exactly in the middle [7].

The imbalanced data have a significant impact on SVM performance [8]. Imbalancedness may produce inaccurate classification rates using SVM for the minority class that is often the most critical one [9]. On somewhat skewed data, SVM performs better than other common classifiers. The rationale is that only SVs are utilized in the classification process; therefore, removing several majority samples that are distant from the decision boundary won't have an impact on the classification. On the other hand, a severe class imbalance can cause an SVM classifier to become sensitive, which would lower its classification performance in the positive class. It frequently produces a classifier with a significant estimation bias in favor of the majority class, which leads to a high number of false negatives [10].

Finding significant observations is effectively accomplished through diagnostic analysis. Perhaps the most often used technique for evaluating the individual effects of instances on the learning process was case elimination. We refer to this method as global influence analysis. But since case deletion removes all information from an observation, it is difficult to determine whether or not the observation has any bearing on a particular component of the model. By using the local influence strategy, one can get around this issue by once more examining the model's sensitivity to slight perturbations [11-12].

A higher-level mathematical framework or strategy known as a metaheuristic algorithm is employed to address optimization issues that are challenging, intricate, or impractical to resolve with conventional optimization methods. Usually, natural processes like simulated annealing, swarm intelligence, or evolutionary processes serve as the inspiration for these algorithms [13]. Although they seek to locate a passably decent answer in a reasonable period of time, metaheuristic algorithms do not promise an ideal solution. They are frequently used in situations when more conventional optimization techniques are unworkable or wasteful, particularly in situations involving huge search spaces or non-linear connections.

In this paper, pigeon optimization algorithm as a metaheuristic algorithm is employed to address the influence observations issues in SVM. The proposed approach will efficiently help to find the most insignificant observations with high classification performance. The experimental results show the favorable performance of the proposed approach when the number of both the sample size and features is high.

1. Support vector machine

The SVM is a classification technique that divides data into two groups. It is based on supervised machine learning techniques that are developed from statistical learning theory. In the SVM, the space between the class border and the training samples is linearly separable. In the non-linear separable, the patterns are turned into a high-dimensional space by the kernel functions [14].

The hyperplane of SVM is defined by $w^T x + b = 0$, where $w \in R^N$ is an orthogonal to the hyperplane and $b \in R^N$ is the constant. Giving some training data set d, a set of point of the form.

$$d = \{(x_i, y_i) : x_i \in R^N, y_i \in \{-1, +1\}\}_{i=1}^{n}, \qquad (1)$$

where x_i is a N-dimensional real vector, y_i either equal +1 or -1 and represents the class of the input vector. The SVM is maximizing the margin by two parallel hyperplanes $w^T x + b = +1$ and $w^T x + b = -1$, and these two equations can be simplified by one equation [15]:

$$y_i (w^T x + b) \geq 1. \qquad (2)$$

SVM finds for optimal separation value in the space $f(x) = w^T x + b$. The classifier is defined as:

$$f(x) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i x_i^T x + b), \qquad (3)$$

where sgn(.) represent the sign of function, α_i represent Lagrange multiplier, x is the test sample x_i is a training sample. Let the distance from the points of data to the hyperplane be $1/\|w\|$. The non-separable case problem in SVM training is solved using quadratic programming problems (QPP) in the manner described below [16]:

$$\min \phi(w, \xi) = \frac{1}{2}\|w\| + C\sum_{i=1}^{N} \xi_i \quad ; \quad \xi_i \geq 0, \qquad (4)$$

Such that

$$y_i (w^T k(x_i) + b) \geq 1 - \xi_i \qquad for \quad 1 \leq i \leq N .$$

The constraint in SVM can be satisfied if ξ_i is sufficiently large and C is a regularization parameter. Through the use of unique functions known as the kernel, the data in the non-linear case of SVM will be transferred from a lower dimensions space to a higher dimensional space in order to be more accurately classified. SVM uses a variety of kernel functions, including radial basis function, polynomial kernel, and linear kernel. Furthermore, the classifier can be defined in the following way using the kernel function:

$$f(x) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i K(x_i^T, x) + b). \qquad (5)$$

In this paper, the RBF of SVM is used as the kernel function as

$$K(x, y) = \exp\left(-\sigma\|x - y\|^2\right), \qquad (6)$$

where σ is the kernel parameter which controls the width of the Gaussian kernel. From equations. (4), (5) and (6), it is noted that the performance of the SVM classifier is dependent on the kernel function parameter σ and tradeoff regularization parameter C .

2. Pigeon optimization algorithm (POA)

POA is a class of evolutionary algorithms that uses principles of natural evolution and identity of the genetic evolution of organisms, where it introduced by John Holland first in 1970 [21]. POA is a heuristic search that modifies the individual functions of coded individuals as real or binary string by using operators of POA [41] [46] [47] [48] [49]. It finds the optimal solutions from a randomly created population, where repeatedly modifies the individual at each stage to be parents and uses the parents to find the offspring for next generation. The individuals are evaluated using a fitness function which is determined to problem [22].

The POA uses primary operations on the population: selection, crossover (recombination) and mutation to find optimal solution and the algorithm is stopped when either maximum number of generations has been generated or the optimal solutions has been reached by fitness function [23]. Several procedures are important for pigeon optimization algorithm. They are initialization, fitness evaluation, selection, crossover, mutation, and termination.

3. Pigeon optimization algorithm

The pigeon optimization algorithm (POA) mainly consists of two operators: the map and compass operator and the landmark operator. In the map and compass operator, pigeons sense the geomagnetic field to shape the map for homing. Suppose that the search space is N-dimensional, and then the i-th pigeon of the swarm can be represented by a N-dimensional vector $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,\mathrm{N}})$. The velocity of this pigeon, which represents the position change of this pigeon, can be represented by another N-dimensional vector $V_i = (V_{i,1}, V_{i,2}, ..., V_{i,\mathrm{N}})$. The best previously visited position of the i-th pigeon is denoted as $P_i = (P_{i,1}, P_{i,2}, ..., P_{i,\mathrm{N}})$ .The global best position of the swarm is $g = (g_1, g_2, \ldots, g_N)$. Each pigeon is flying according to the following two equations:

$$V_i(t+1) = V_i(t) \times e^{-Rt} + rand \times \left(X_g - X_i(t)\right) \qquad (7)$$

$$X_i(t+1) = X_i(t) + V_i(t+1), \qquad (8)$$

where R is a map and compass factor, while $rand$ is a uniform random number in the range $[0, 1]$, X_g is the global best solution, $X_i(t)$ denotes the current position of a pigeon at instance t , and $V_i(t)$ denotes the current velocity of a pigeon at iteration t .

In landmark operator, all the pigeons are ranked according to their fitness value. In each generation, the number of pigeons is updated by Eq. (8), where only half number of pigeons is considered to calculate the desired position of the centered pigeon, while all other pigeons adjust their destination by following the desirable destination position.

$$N_p(t+1) = \frac{N_p(t)}{2}, \qquad (9)$$

where N_p is the number of pigeons in the current iteration t .

The position of the desired destination is calculated by Eq. (10), while all other pigeons update their position toward this position by Eq. (11) [17].

$$X_c(t+1) = \frac{\sum X_i(t+1) \times \mathrm{Fitness}(X_i(t+1))}{N_p \sum \mathrm{Fitness}(X_i(t+1))} \qquad (10)$$

$$X_i(t+1) = X_i(t) + rand \times \left(X_c(t+1) \times X_i(t)\right), \qquad (11)$$

where X_c is the position of the centered pigeon (desired destination).

.

4. The proposed approach

In data analysis and machine learning, identifying and diagnosing data abnormalities is crucial. Data points or patterns that substantially depart from expected or typical behavior are called anomalies. In other words, in machine learning, diagnosing influential instances is the process of locating data points that significantly affect the performance or predictions of the model. These occurrences may have a significant impact on the model's output, alter its decision bounds, or introduce biases, among other things.

Outlier detection methods and strategies are frequently employed to identify significant cases in machine learning. Data points that substantially differ from the bulk of the data are known as outliers. Finding influential examples can be aided by identifying outliers. Outliers can be found using strategies such as clustering-based approaches, modified z-score, and z-score. Outliers have the potential to significantly affect the model's performance. Where Eq. (1) represents the output between input and hidden layer, while Eq. (2) represents the final output between hidden layer and the output layer.

It is crucial to remember that identifying impactful examples is a process that is iterative and necessitates a thorough comprehension of the behavior of the model. Depending on the particular machine learning method, dataset, and issue area, other approaches can be more appropriate.

Our contribution of this paper is that SVM is created from training data, and the final model may change if one of the training instances is removed. When a training instance is removed from the training data and considerably improve the classification, then we refer to it as "influential.". Searching for several influential training instances is time consuming. However, employing pigeon optimization algorithm as a meta-heuristic

algorithm can overcome time consuming and then improve classification accuracy. In POA, each member is coded as 0 (the training instance is considered as influential) or 1 (the training instance is not considered as influential). A representation of the purpose of POA is shown in Figure 1.

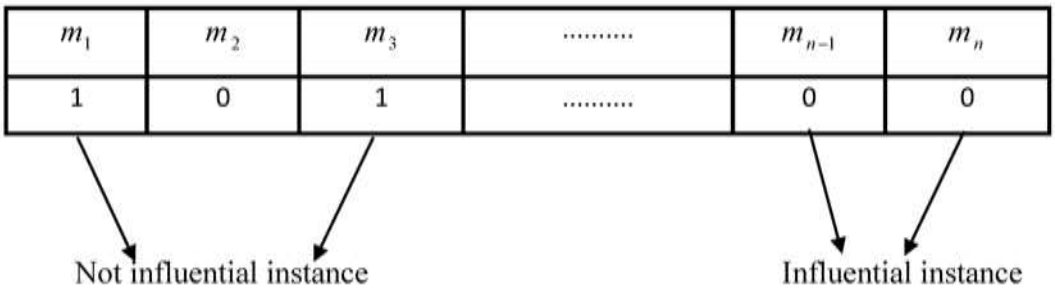

Figure 1: A representation of the purpose POA.

The proposed approach will efficiently help to find and eliminate the influential instances with high classification performance. The parameter configurations for our proposed approach are presented as follows.

- The number of pigeons is set to 25 and the number of iterations is $t_{max}=500$.
- The positions of each pigeon are randomly determined by uniform distribution with the range [0, 1].
- The fitness function is defined as the miss-classification error.
- The positions are updated using Eq. (11).
- Steps 3 and 4 are repeated until a t_{max} is reached.

5. Experimental results

To examine the performance of the proposed approach of the POA algorithm, we compare this method with the SVM and leave-one-out deletion method (LOOD). Their performances are evaluated using the classification accuracy (CA), the geometric mean (G-mean) of specificity and sensitivity, and the area under the curve (AUC).

There are three publicly accessible datasets—all binary classification problems—that are taken from the University of California at Irvine (UCI 2013) machine learning library. These datasets are described in Table 1. Using 10-folds cross validation (CV) as the training and testing methodology, each dataset is randomly split into two sets of varying sizes: 30% of the observations were used to test the algorithm, while 70% of the datasets were used for training. There are 50 iterations of this method.

Table 1: Data used description

Dataset	n	# positive	# negative
Blood	748	180	568
Diabetes	768	268	500
Breast	228	47	151

It can be seen in Table 2 that, from among the three methods, the POA approach performs the best with average results of, overall, the three datasets, 96.63% and 95.66%, in terms of classification accuracy, for both training and testing datasets, respectively. Further, as it can be observed from Table 3, LOOD overtake the standard SVM. Beside the high classification performance, the robustness is an important factor in evaluating a classifier. The standard deviation of all criteria for POA in all datasets is small. This shows that POA is a robust approach.

In terms of G-mean criterion, for both the training and testing datasets, the proposed POA outperformed LOOD and SVM in all datasets. The most remarkable result for POA concerns the Diabetes dataset. We obtain 98.93% and 97.74% accuracy for training and testing datasets, respectivelyshown in the following equation [28] :

Comparatively speaking, in all of the datasets, SVM without all instances obtains worse classification accuracy compared with other influential diagnosis approaches.

Table 2: Average CA performance of the POA, LOOD, and SVM.

Data set	Method	Training dataset	Testing dataset
Blood	SVM	91.58 ± 0.021	90.76 ± 0.022
	LOOD	94.32 ± 0.018	93.11 ± 0.015
	POA	97.71 ± 0.010	96.25 ± 0.011
Diabetes	SVM	92.31 ± 0.024	91.62 ± 0.022
	LOOD	95.18 ± 0.021	94.84 ± 0.020
	POA	98.88 ± 0.019	97.26 ± 0.018
Breast	SVM	90.79 ± 0.025	89.18 ± 0.026
	LOOD	93.24 ± 0.021	92.74 ± 0.026
	POA	96.63 ± 0.019	95.66 ± 0.024

Table 3: Average G-mean performance of the POA, LOOD, and SVM.

Data set	Method	Training dataset	Testing dataset
Blood	SVM	92.71 ± 0.020	91.54 ± 0.021
	L	94.88 ±	93.45 ±

	OOD	0.018	0.015
	POA	97.95 ± 0.011	96.81 ± 0.010
Diabetes	SVM	92.94 ± 0.024	91.87 ± 0.022
	LOOD	95.81 ± 0.020	95.36 ± 0.020
	POA	98.93 ± 0.019	97.94 ± 0.019
Breast	SVM	91.74 ± 0.025	90.28 ± 0.024
	LOOD	93.92 ± 0.022	92.88 ± 0.022
	POA	96.87 ± 0.020	95.91 ± 0.021

To further verify the effectiveness of the proposed POA, the statistical paired t-test is performed to verify whether there is a significant difference between POA and the other two approaches: SVM and LOOD in terms of the AUC. Table 4 displays the difference values between the POA and the other two approaches and the p-values (in the parentheses). The bold values indicate a statistically significant taking significance level of $\alpha = 0.05$. It can be seen that there is a statistical difference between POA and each of LOOD and SVM for the three datasets.

Table 4: Paired t-test results of POA, LOOD, and SVM in terms of AUC.

Dataset	LOOD	SVM
Blood	3.54 (0.0000)	5.84 (0.0002)
Diabetes	4.68 (0.0001)	7.36 (0.0000)
Breast	6.08 (0.00002)	8.27 (0.0001)

6. Conclusion

This article examines the effects of the influence observations on imbalanced data classification. A metaheuristic algorithm, POA, is suggested to improve the classification accuracy overall. The usefulness of

the suggested strategy was demonstrated by the experimental findings on three medical datasets. The POA is capable of maintaining high accuracy in classification and G-mean. The paired t-test was employed to assess the efficacy with respect to AUC. The findings demonstrate that, when compared to other approaches in use, the POA has greatly improved its classification ability.

References

[1] Ismael OM, Qasim OS, Algamal ZY. Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning. Journal of Chemometrics. 2020;34(11). doi: 10.1002/cem.3311.

[2] Qasim OS, Algamal ZY. A gray wolf algorithm for feature and parameter selection of support vector classification. International Journal of Computing Science and Mathematics. 2021;13(1):93-102.

[3] Guido R, Groccia MC, Conforti D. A hyper-parameter tuning approach for cost-sensitive support vector machine classifiers. Soft Computing. 2023;27(18):12863-12881.

[4] Wang Y, Xu Y. A non-convex robust small sphere and large margin support vector machine for imbalanced data classification. Neural Computing Applications. 2023;35(4):3245-3261.

[5] Wang Z, Liu Q. Imbalanced Data Classification Method Based on LSSASMOTE. IEEE Access. 2023;11:32252-32260.

[6] Vapnik V. The nature of statistical learning theory. Springer science & business media; 1999.

[7] Widodo CE, Adi K, Gernowo R. A support vector machine approach for identification of pleural effusion. Heliyon. 2023.

[8] Cervantes J, Li X, Yu W. Imbalanced data classification via support vector machines and genetic algorithms. Connection Science. 2014;26(4):335-348.

[9] Benítez-Peña S, Blanquero R, Carrizosa E, et al. Cost-sensitive probabilistic predictions for support vector machines. European Journal of Operational Research. 2023.

[10] Tang Y, Zhang Y-Q, Chawla NV, et al. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, Cybernetics, Part B. 2008;39(1):281-288.

[11] Rocha AV, Simas AB. Influence diagnostics in a general class of beta regression models. Test. 2011;20:95-119.

[12] Algamal ZY. Diagnostic in poisson regression models. Electronic Journal of Applied Statistical Analysis. 2012;5(2):178-186.

[13] Al-Thanoon NA, Algamal ZY, Qasim OS. Feature selection based on a crow search algorithm for big data classification. Chemometrics and Intelligent Laboratory Systems. 2021;212. doi: 10.1016/j.chemolab.2021.104288.

[14] Lin K-C, Chen S-Y, Hung JC. Feature Selection and Parameter Optimization of Support Vector Machines Based on Modified Artificial Fish Swarm Algorithms. Mathematical Problems in Engineering. 2015;2015:1-9. doi: 10.1155/2015/604108.

[15] Saad Y, Shaker K. Support Vector Machine and Back Propagation Neural Network Approach for Text Classification. Journal of University of Human Development. 2017;3(2):869-876. doi: 10.21928/juhd.20170610.40.

[16] Tharwat A, Hassanien AE. Chaotic antlion algorithm for parameter optimization of support vector machine. Applied Intelligence. 2017;48(3):670-686. doi: 10.1007/s10489-017-0994-0.

[17] Duan H, Qiao P. Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning. International Journal of Intelligent Computing and Cybernetics. 2014;7(1):24-37. doi: 10.1108/ijicc-02-2014-0005.