International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 1, February 2020, pp. 1079~1084
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp1079-1084

A new model for iris data set classification based on linear support vector machine parameter's optimization

Zahraa Faiz Hussain[1], Hind Raad Ibraheem[2], Mohammad Alsajri[3], Ahmed Hussein Ali[4], Mohd Arfian Ismail[5], Shahreen Kasim[6], Tole Sutikno[7]
[1,2,3,4]Computer Science Department, AL Salam University College, Iraq
[3,4]Department of Computer Science, College of Education, Al-Iraqia University, Iraq
[5]Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Malaysia
[6]Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[7]Department of Electrical and Computer Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:

Received Apr 9, 2019
Revised Sep 29, 2019
Accepted Oct 10, 2019

Keywords:

Classification
Data mining
Genetic algorithm
Iris dataset
Parameter optimization
SVM

ABSTRACT

Data mining is known as the process of detection concerning patterns from essential amounts of data. As a process of knowledge discovery. Classification is a data analysis that extracts a model which describes an important data classes. One of the outstanding classifications methods in data mining is support vector machine classification (SVM). It is capable of envisaging results and mostly effective than other classification methods. The SVM is a one technique of machine learning techniques that is well known technique, learning with supervised and have been applied perfectly to a vary problems of: regression, classification, and clustering in diverse domains such as gene expression, web text mining. In this study, we proposed a newly mode for classifying iris data set using SVM classifier and genetic algorithm to optimize c and gamma parameters of linear SVM, in addition principle components analysis (PCA) algorithm was use for features reduction.

Corresponding Author:

Mohammad Alsajri,
Computer Science Department,
AL Salam University College,
BAGHDAD - Sidiyah: Near Al Waleed Center, Iraq.
Email: mohammad.cs88@gmail.com

1. INTRODUCTION

Classification is a manner of data analysis which used to elicit a classifier to classify important data classes. These classifiers can expect categorical data (detached, unordered) class label [1]. Also classification is an important field in the data mining and machine learning by concluding obscure classes of samples using the learning of renowned classes of samples [2-3]. As an example, rating bank loan application can be classified as safety or risky after a construction of classification model. This analysis supplied us with a better comprehension of the data at big magnitude. Many classification approaches have been suggested in machine learning, pattern recognition, and statistics. Classification can be achieved in a process of two-steps. The construction of a classification based on preceding data is achieved in the first stage. In the second stage, specifies if the accuracy of a model is admissible, and if so, we utilize the model to classify fresh data [4]. Support vector machine (SVM) Classifier is a renowned classification method employed for predicting the results of datasets [5]. The proposed model was assessed on an IRIS dataset gained from the UCI Machine Learning Database [6].

The creation of SVM model with high prediction accuracy and consistency is based on seeking the ideal parameters on SVM, since it plays an essential role. Weakness classification performance result

from indecent parameter settings, while the perfect categorization accuracy of SVM stems from seeking optimal parameters.

a. The authors submitted a new manner which optimize SVM's parameters effectively and reduce the time of optimization and calculations cost using tow nested real valued genetic algorithm (NRGA). The NRGA compared to the conventional optimization mechanisms which operate on seeking the whole parameters together [7].
b. A notation was submitted in [8] for determining SVM parameters depending on minds from design of experiences, which initiated with an extremely rough grid comprising the complete search range and repetitive revised both the grid resolution and search boarders, safeguarding the number of forms at each phase almost constants.
c. Genetic algorithm (GA) is trend to be completely pretty at finding in general perfect universal solutions. GA has been vastly adopted for parameter setting. In [9] a manner based on GA was suggested to simultaneously optimize SVM 'S parameters and attribute subset. In [10] GA is fused with asymptotic attitudes of SVM which then guides the search to the right line of perfect generalization error in the super parameter space.
d. This study [9] develops a novel manner termed PSO+SVM. PSO based approach for parameter determining and feature selection, and then a comparison is conducted of gained results with other approaches. The SVM+PSO gained a better accuracy of classification than other tests.

2. CLASSIFIERS

Classification is imperative for data mining. The learning algorithm [11] establishes a classifier in a given set of measurement, for instance, a set of characteristic data (x1, x2,…., xn), where xi denotes feature data Xi. The purpose of classification is to initiate the actuality of groups when given a set of observation (unsupervised learning) or where various categories prevail and the target is classified into one of the previous categories (supervised learning) [12]. Supervised learning has been employed in this study as the classification method.

2.1. SVM

In this part, we focus SVM, a manner using for a classification the linear and nonlinear data. The SVM algorithm operates as follows: the nonlinear mapping is used to convert the training data into a higher distance, under the fresh distance; it investigates for the linear perfect segregating hyperplane (i.e., a "decision boundary" segregating the tuples of one class from another). With a convenient nonlinear mapping to an adequately elevated distance, the data of two classes can be always segregated by a hyperplane. The SVM finds this hyperplane using support vectors ("essential" training tuples) and edges (defined by the support vectors) [13, 14].

2.2. Genetic algorithm (GA)

Genetic algorithms (GA) operate with a collection of nominee solutions named a population. Depending on the Darwinian principle of 'existence of the fittest', the GA earns the perfect solution after sequences of reduplicate calculations. GA products consecutive populations of alternate solutions that is representative by a chromosome, i.e. a solution to the problem, till acceptable results are earned. GA a general adaptive optimization search methodology based on a direct analogy to Darwinian natural selection and genetics in biological systems is a promising alternative to conventional heuristic methods. In this study, we essentially utilize GA to refine the parameters (C and γ) of the SVM model for iris dataset [15, 16]. GA as a wrapper method combined with PCA as filter method and tested using SVM to classification leaves [16]. The results showed that GA combined with SVM given computing time effectively and improve accuracy. GA also used to select important features and instances then tested using SVM and k-nearest neighbors (KNN) [17-19]. Gain Ratio (filter) combined with sequential forward selection (SFS) wrapper proposed to deal with three datasets; iris, breast, and dermatology [20, 21]. A various feature selection methods also compared, they were information gain, gain ratio (GR), symmetrical uncertainty (SU), Chi square (CS), relief, and correlation based feature selection (CFS) [19]. The result showed that CFS was the most stable with the highest accuracy for handling data with two classes.

3. METHOD

As mentioned before SVM classifier was built to classify iris dataset into different classes. The using of GA is to optimize SVM's parameters (c, gamma), in order to obtain higher and best accuracy [22]. The iris dataset has four attributes, principle components analysis (PCA) algorithm was

applied to reduce these features (feature reduction), and then only three features were chooses. Whereas principal component analysis (PCA) is a mathematical execution that converts a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables named principal components. PCA is a dimension-decreasing instrument that can be utilized to drooping a great set of inconstant to a little set that stay involves most of the information in the big set [12, 2].

The presented technique in this study used the IRIS dataset acquired from the UCI Machine Learning Repository. The dataset is in a multivariate group as it provides the statistic on the Iris plant type based on four characteristics which include width, width and petal - length, sepal - length, and values as presented in Figure 1. The dataset is composed of three groups with 50 cases each and a total of 50 cases. The dataset were first processed by removing missing data values. The type of Iris plant is the forecasted characteristic in this dataset [5].

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100

Figure 1. IRIS dataset

Step-by-step of new model in this research for iris data set classification based on linear support vector machine parameter's optimization is:

Step-1: The Iris dataset in CSV is computed as the input.
Step-2: Divide the data into test and training datasets. In this study, the dataset was partitioned into 70% training and 30% testing.
Step-3: Distinguish the training dataset based on the class values, that is, 1, 2 and 3.
Step-4: Determine the standard deviation and mean values for the individual data case based on the class values.
Step-5: Choose the SVM (C and γ) parameters as input to genetic algorithm optimization.
Step-6: Apply the optimal value of the (C and γ) parameters as an initial value to the process of classification using SVM.
Step-7: Utilize the model and generate predictions.
Step-8: Determine the prediction accuracy through the comparison of the class data of test dataset. This accuracy is evaluated depending on the ratio between 0 to 100%.

4. RESULTS AND CORRELATIONS

The suggested model presented in Section 4 was performed on the Iris dataset with and without Step-5. In each run, the obtained results were evaluated based on the accuracy of the SVM classifier. The obtained results showed that the accuracy of the SVM increased to 98.7 using Step-5 and about 95.3% without Step-5. All the results, with the optimization, are presented in Figures 2, 3, 4, 5, 6 and 7, respectively. The results of proposed method show the powerful of using genetic algorithm to optimize the (C and gamma) parameter of SVM classifier.

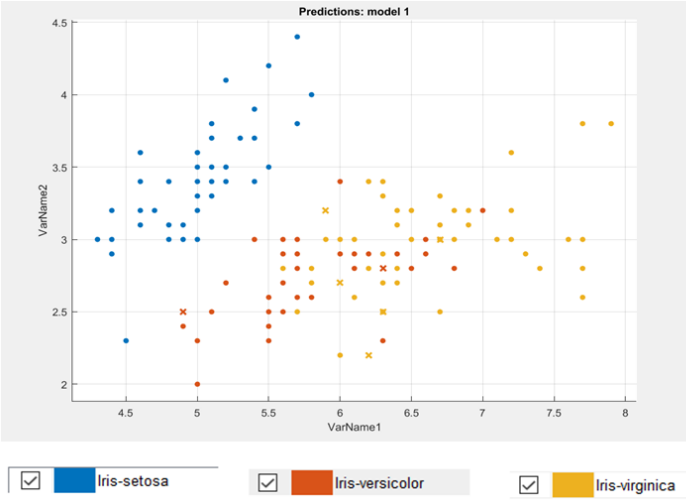

Figure 2. The scatter plot without genetic

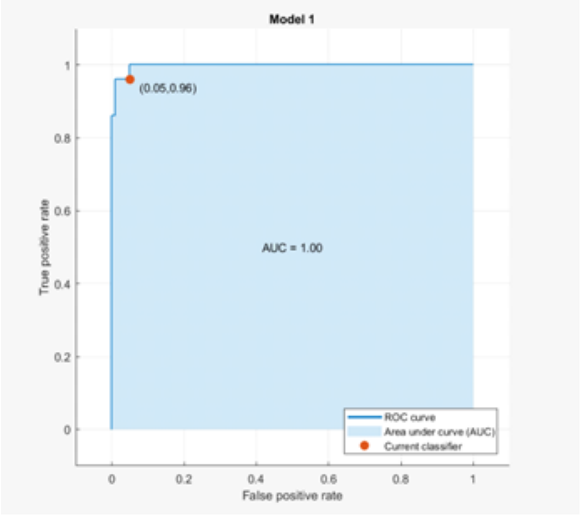

Figure 3. The ROC curve without genetic

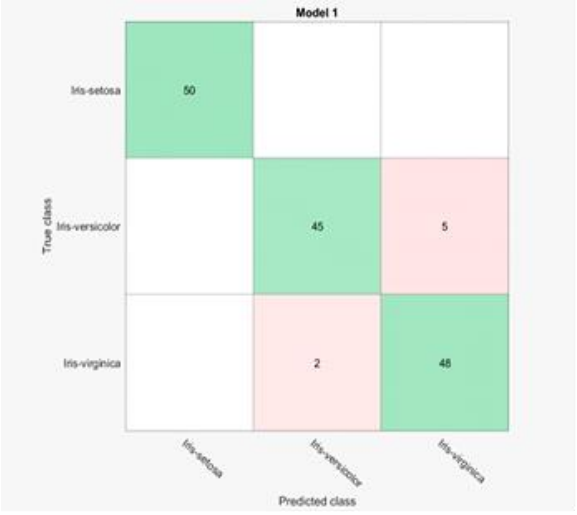

Figure 4. The confusion matrix without genetic

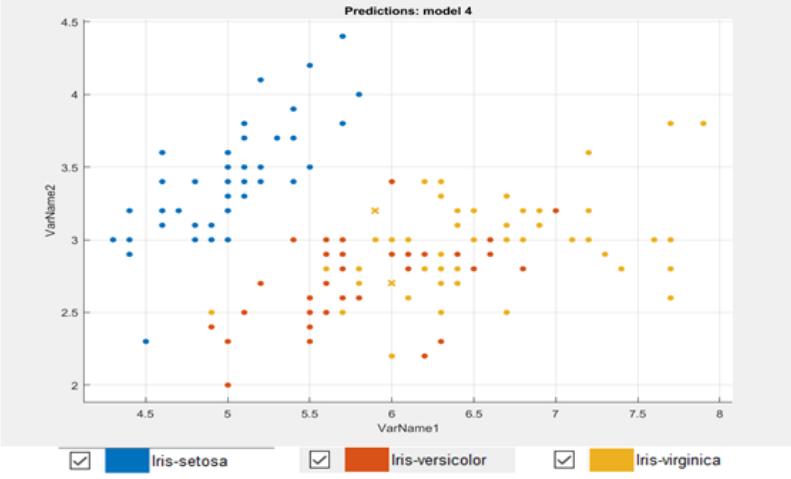

Figure 5. The scatter plot with genetic

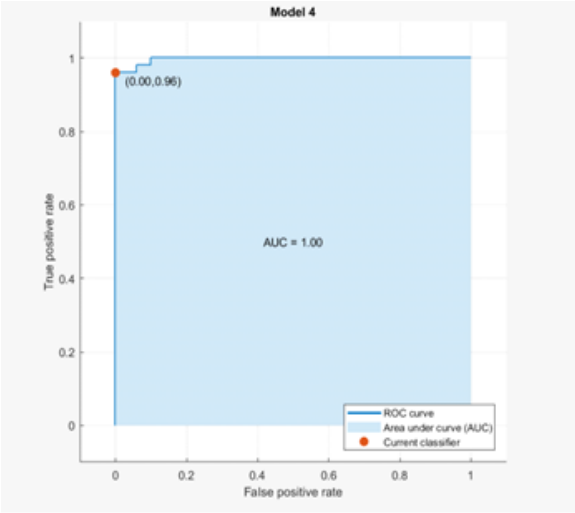

Figure 6. The ROC curve with genetic

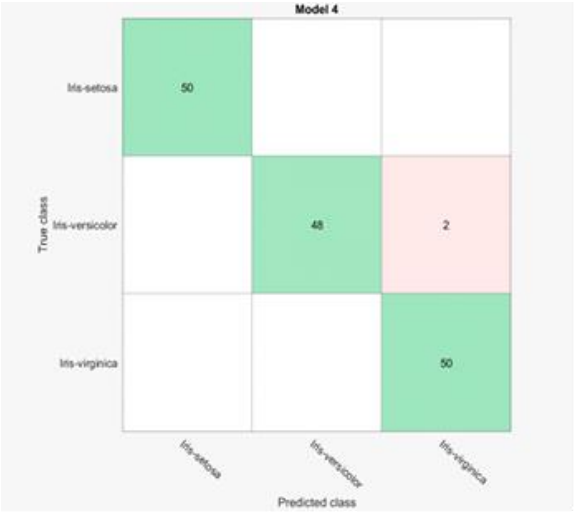

Figure 7. The confusion matrix with genetic

5. CONCLUSIONS AND RECOMMENDATION

This paper have proposed a newly mode for classifying iris data set using SVM classifier and genetic algorithm, in addition PCA algorithm was use for features reduction. This proposed mode is to optimize c and gamma parameters of linear SVM. As shown above the results obtained from applied GA on iris dataset is 98.7 and without GA is 97.78. GA was used to optimize SVM's parameters (c, gamma), in order to promotion an efficacious SVM model with high accuracy and stability, the optimal parameter seek on SVM plays a fateful role. Inadvisable parameter settings result in inferior classification performance. For the future work, this study can be extend into two part; firstly by improving the performance of GA such as hybrid GA with other method as works done by [22-24], and secondly by apply feature selection method in SVM for optimal parameter setting as proposed in [25].

REFERENCES

[1] Z. Lnlan, et al, "Using Genetic Algorithm to Optimize Parameters of SupportVector Machine and Its Application in Material Fatigue Life Prediction," *School of Mechanical Engineering, Shanghai University of Engineering Science, Shanghai, China., Advances in Natural Science*, vol. 8(1), 2015

[2] X. Z. Li and J M. Kong, "Application of GA–SVM method with parameter optimization forlandslide development prediction," *Nat. Hazards Earth Syst. Sci.*, vol. 14, pp. 525–533, 2014.

[3] Mao, K. Z., "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34(1), pp. 60-67, 2004.

[4] Abbas F. H. Alharan, Hayder K. Fatlawi, Nabeel Salih Ali, "A cluster-based feature selection method for image texture classification," Indonesian Journal of Electrical Engineering and Computer Science, Vol 14, No 3: pp1433-1442, June 2019.

[5] Maryam, N. AkhmadSetiawan, and O. Wahyunggoro., "A Hybrid Feature Selection Method Using Multiclass SVM for Diagnosis of Erythemato-Squamous Disease," *International Conference on Mathematics World Congress on*, 2017.

[6] L. Talavera., "An evaluation of filter and wrapper methods for feature selection in categorical clustering," *Adv. Intell. Data Anal. VI*, pp. 742, 2005.

[7] P. Liao, X. Zhang, and K. Li., "Parameter Optimization for Support Vector Machine Based on Nested Genetic Algorithms," *Information Engineering School, Nanchang University, Nanchang, China, Journal of Automation and Control Engineering*, 2016.

[8] C. Staelin., *Parameter selection for support vector machines*, Technical Report HPL-2002-354 (R.1), HP Laboratories Israel, 2003.

[9] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimization for support vector machine," *Expert Systems with Applications*, vol. 31(2), pp. 231-240, 2006.

[10] C. H. Wu, G. H. Tzeng, Y. J. Goo, and W. C. Fang., "A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy," *Expert Systems with Applications*, vol. 32(2), pp. 397-408, 2007.

[11] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee., "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, pp. 1817-1824, 2008.

[12] Min, J. H. and Lee, Y. C. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Syst. Appl.*, vol. 28, pp. 603-614, 2005.
[13] Gu, J. Y., Liu, J. F, and Chen, M., "A modified regression prediction algorithm of large sample data based on SVM," *Computer Engineering*, vol. 40(1), pp. 161-166, 2014.
[14] M. Z. A. Ahmed Hussein Ali, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering*, 2018.
[15] M. Z. A. Ahmed Hussein Ali, "An Efficient Model for Data Classification Based on SVM Grid Parameter Optimization and PSO Feature Weight Selection," *International Journal of Integrated Engineering*, 2018.
[16] C. Tsai, W. Eberle, and C. Chu., "Knowledge-Based Systems Genetic algorithms in feature and instance selection," *Knowledge-Based Syst.*, vol. 39, pp. 240-247, 2013.
[17] M. Karabatak and M. C. Ince, "A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases," *Expert Syst. Appl.*, vol. 36(10), pp. 12500-12505, 2009.
[18] D. Zhang, *et al*, "A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction," *BioMed Research International*, 2015.
[19] Kasim S., Hassan R., Mohd N. S., Ramlan R., Mahdin H, and Fudzee M. F. M., "A Comparative Study of Different Template Matching Techniques for Twin Iris Recognition," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7(4-2), pp. 1581-1588, 2017.
[20] Kasim S., Hassan R., Zaini N. H., Syifaa' Ahmad A., Ramli A. A, and Saedudin R. R., "A Study on Facial Expression Recognition Using Local Binary Pattern," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7(5), pp. 1621-1626, 2017.
[21] Zin N.A.M., Asmuni H., Hamed H.N.A., Othman R.M., Kasim S., Hassan R., Zakaria Z, and Roslan R., "Contact lens classification by using segmented lens boundary features," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 11(3), pp. 1129-1135, 2019.
[22] Ismail M.A., Razak K.A.B., Moorthy K., Mezhuyev V., Kasim S, and Ibrahim A.O., "Newton Competitive Genetic Algorithm Method for Optimization the Production of Biochemical Systems," *Advanced Science Letters*, vol. 24(10), pp. 7481-7485, 2018.
[23] Ismail M.A., Mezhuyev V., Deris S., Mohamad, M.S., Kasim S, and Saedudin R.R., "Multi-objective Optimization of Biochemical System Production Using an Improve Newton Competitive Differential Evolution Method," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7(4-2), pp. 1535-1542, 2017.
[24] Ismail M.A., Mezhuyev V. Moorthy K., Kasim S, and Ibrahim A.O., "Optimisation of Biochemical Systems Production using Hybrid of Newton method, Differential Evolution Algorithm and Cooperative Coevolution Algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 8(1), pp. 27-35, 2017.
[25] Ibrahim A., Hussien W., Yagoop A, and Ismail M., "Feature Selection and Radial Basis Function Network for Parkinson Disease Classification," *Kurdistan Journal of Applied Research*, vol. 2(3), pp. 167-171, 2017.