JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

RFE, BOXCOX, AND PCA COMPARISON FOR MULTICLASS CLASSIFICATION SUPPORT VECTOR MACHINE OPTIMIZATION

Indrawata Wardhana[1*], Vandri Ahmad Isnaini[2], Rahmi Putri Wirman[2]
[1]Automatic Control, INSA Centre val de Loire, Bourges, France
[2]Physics, UIN Sulthan Thaha Saifuddin Jambi
email: *indrawata.wardhana@insa-cvl.fr

Abstract: The technique of multiclass classification based on SVMs has been widely used. SVM optimization will be accomplished by examining the extraction features of Principal Component Analysis (PCA), Box-Cox Transformation, and Recursive Feature Elimination (RFE). The dataset contains 13,611 rows and 17 variables, generated from the UCI repository's multiclass dry bean data. Barbunya, Bombay, Cal, Dermas, Horoz, Seker, and Sira are just a few of the dry bean kinds available. The dataset was tested using SVM Linear kernel and SVM Radial Basis.According to the results, the combination of scale-center-BoxCox-SVM Radial extraction achieves the maximum accuracy of 93.16 percent and the shortest processing time of 6.10 minutes. 96.00 percent, 100 percent, 96.71 percent, 95.16 percent, 97.60 percent, 97.74 percent, and 91.95 percent, according to bean class.RFE-SVM Radial has a 91.18 percent accuracy and a processing time of 6.55 minutes. BoxCox outperforms conventional techniques in terms of prediction accuracy while requiring less training time.

Keywords: Bean, PCA, BoxCox, SVM, RFE

Abstrak: Klasifikasi Multikelas menggunakan SVM telah banyak digunakan. Pada penelitian ini akan diuji fitur ekstraksi Principal Component Analysis, Box Cox Transformation dan fitur eliminisi Recursive Feature Elimination untuk mendapatkan optimasi SVM. Dataset berasal dari data multikelas kacang kering UCI repository dengan jumlah 13.611 baris dan 17 variabel. Kelas kacang kering yakni : Barbunya, Bombay, Cal, Dermas, Horoz, Seker dan Sira. Dataset diuji menggunakan kernel SVM Linier dan SVM Radial Basis. Didapatkan hasil, bahwa kombinasi fitur ekstraksi : scale-center-BoxCox-SVM Radial memiliki akurasi terbaik yakni 93,16% dan waktu proses 6,10 menit. Klasifikasi berdasarkan kelas kacang berturut-turut 96,00%,100%, 96,71%, 95,16%, 97,60%, 97,74% dan 91,95%. RFE- SVM Radial hanya memberikan akurasi sebesar 91,18 % dengan waktu proses sebesar 6.55 menit. Penggunaan BoxCox dibandingkan dengan lainnya, memberikan hasil prediksi lebih baik dan namun tidak mempercepat waktu pelatihan.

Kata kunci: BoxCox; Kacang; PCA; RFE; SVM

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Machine learning (ML) has been widely applied to the classification of plants in agriculture [1], [2]. The classification of plant seeds is frequently employed in machine learning-based classification predictions. When determining coffee beans using ANN (Artificial Neural Networks) and KNN (K-nearest Networks), it was found that ANN outperforms KNN by more than 90% [3]. Meanwhile, the accuracy of the wheat grain test using MLP (Multilayer Perceptron) was reported to be 100 percent [4].

Dry bean classification using the BoxCox transformation and k-folds made the classification prediction more accurate. During the optimal training phase, the model is used for a random forest with 50 decision tree parameters and a depth of 10, a gradient boosting machine model with a learning rate of 1, and a light gradient boosting machine model with a learning rate of 0. Light GBM is the best for training because it is 99 percent accurate, but only 91 percent accurate when it comes to verifying that it is correct [5].

On the other hand, it was shown that KNN outperformed ANN in the classification of cherry coffee beans [6].PCA is also used to accelerate the training time of KNN [7]. The Support Vector Machine (SVM) has a 78 percent accuracy rate in detecting cancers [8] .SVM has also been used effectively for multiclass prediction [9], including a SVM classifier of dry beans [10]. Directed acyclic graphs can also be used to generate accurate binary predictions in SVM multiclass classification [2].

Principal component analysis (PCA) is a technique that is frequently employed in unsupervised dimension reduction. PCA was effective in lowering the processing time of SVM for the detection of software flaws [11]. Additionally, the PCA extraction function improves the SVM kernel's performance [12]. Furthermore, the multi-fault categorization is included [13].

The extraction feature of the BoxCox transformation makes it possible to speed up a time-consuming process by up to 70 to 80 percent in remaining usable life (RUL) forecasting[14].

Through the use of RFE on the Support Vector Machine, the accuracy of cancer classification is improved, with big results [15]. However, it has been discovered that although multiclass RFE-SVM is not very useful for product design, it can assist in reducing the dimensions [16], which also applies to nonlinear SVM [17]. Adoption of RFE-SVM can substantially increase fault diagnostic performance [18]. SVM-RFE also increases the performance of prediction [19].

METHOD

This research comprises various stages, starting from the selection of data collected from the internet, particularly the UCI website data, followed by preprocessing, feature selection, classification, and performance measurement.

Dataset

This research utilised dry bean data, which has 17 parameters with 13 thousand rows. The dataset comes from the UCI repository website [2], particularly the URL: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset. Seven types of dry beans were applied in this study, taking into account characteristics such as shape, form, variety, and architecture ac-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

cording to the market scenario.

Variables consist of: Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Eccentricity, Convex Area, Equiv Diameter, Extent, Solidity, Roundness, Compactness, Shape Factor 1, Shape Factor 2, Shape Factor 3, Shape Factor 4, and Class.

Preprocessing

Four forms of preprocessing are used in this process: center, scale, Box Cox transformation, and Principal Component Analysis (PCA). Equation 1 illustrates the Box-Cox formula.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, if\ \lambda \neq 0; \\ \log y, if\ \lambda = 0; \end{cases} \quad (1)$$

Feature Selection

The feature selection method is applied to extract the appropriate variables for this study. The information is used in conjunction with the supervised learning technique, which employs two approaches, namely the wrapper. RFE was utilised in the wrapper technique (Recursive Feature Elimination). This technique combines the 17 parameters into less, which decreases processing time by increasing the amount of the classification results.

Classification

Support Vector Machine (SVM) is well-known in classified modeling for having a simpler mathematical concept than other classification techniques. Additionally, SVM is good at solving both linear and non-linear classification tasks.

SVM finds the optimal hyperplane by maximizing the distance between classes. A hyperplane is a class-separating function.

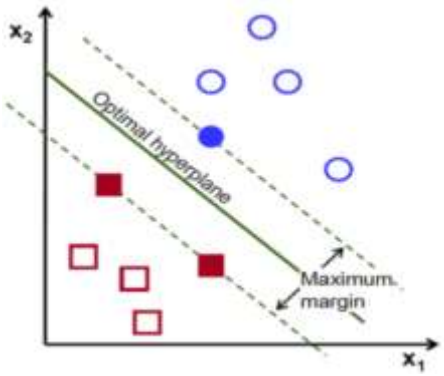

Figure 1. Hyperplane SVM.

For linear classification, use the following formula 2:

$$w^T - c = 0 \quad (2)$$

SVM-RFE

All classification problems, in general, can be reduced to a two-class classification problem. The multiple one-against-all strategy is the simplest and most generally used approach when there are more than two classes. As a result, multiple-class problems can be simplified to multiple-class problems with two classes. Consider a binary decision function with a linear kernel, which can be written as follows equation 3:

$$\begin{aligned} &\text{sgn}(f(x)) = \text{sgn}(wx + c) \\ &= \text{sgn}(\sum_{i=1}^{n} y_i \alpha_i \langle x_i . x \rangle + c \end{aligned} \quad (3)$$

ω denotes the classifier's weight vector. The optimized weight vector equals, where α is greater than zero if it is α a component of the support vector and less than zero otherwise.

SVM Radial Basis Function

The kernel used in the SVM is critical, and we use the radial basis function (RBF) kernel due to its high performance in a variety of applications.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

As defined previously, the RBF kernel is formulas 4:

$$K(x_i, x_j) = \exp(-\frac{\left\|x_i - x_j\right\|}{2\sigma^2} \quad (4)$$

The kernel width is defined by the value $\sigma > 0$.

Performance Evaluation

In this work, the classification performance has been measured using the k-fold cross-validation strategy. This stage utilizes a 10-fold cross-validation procedure that is replicated ten times. The three most critical performance measures are sensitivity, specificity, and accuracy. Three measures are defined by the formulas 3, 4, and 5:

$$Sensitivity = \frac{TP}{TP+FN}\ x\ 100\% \quad (5)$$

$$Specificity = \frac{TN}{TN+FP}\ x\ 100\% \quad (6)$$

$$Accuracy = \frac{\sum TP+\sum TN}{\sum sample}\ x\ 100\% \quad (7)$$

The abbreviations TP, TN, FP, and FN are used to show true positive, true negative, false positive, and false negative.

RESULT AND DISCUSSION

Dataset

The dataset is then partitioned into training and validation data in an 80:20 ratio. Validation by performing three iterations of tenfold cross-validation. Barbunya and Bombay beans will be processed. Cali, Dermas, Horoz, Seker, and Sira are some of the beans.

The dataset as a whole has 13,611 rows and 17 variables. After dividing, we have 10891x17 lines and 2720x17 lines for validation.

Tabel 1. Bean Class with dataset and validation

Class	Dataset	Validation
Barbuya	1058	264
Bombay	418	14
Cali	1304	326
Dermason	2837	709
Horoz	1543	385
Seker	1622	405
Sira	2109	527

SVM Kernel Linier

Scale, center, BoxCox, and PCA preprocessing tests have been performed. Thus, four data points were obtained: accuracy, kappa, p-value, and time, as shown in table 1. At 0.92, the three tests (CS, CS-PCA, and CS-BoxCox) produced nearly identical results but required less processing time. The process happens more slowly in the case of CS-BoxCox.

SVM Kernel Radial

The tests were conducted using the same approach as in the Linear SVM, namely the Scale, Center, BoxCox, and PCA preprocessing. Table 2 summarizes the results of the 4 treatments. With a computation time of 6.108614 minutes, the CS-BoxCox test has the higher precision.

Figure 2. Cost vs Accuracy (RCV).

The accuracy of SVMRadial-CS,

BoxCox (Figure 2) is stable with a cost greater than 0.5 and a maximum accuracy value higher than 0.93.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 2. SVM Linier with Scale,Center,PCA, and BoxCox

Model	CS	CS,PCA	CS,Boxcox	CS,PCA,Box
Accuracy	0.9268	0.9272	0,925	0,8912
p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16
kappa	0.9115	0.912	0,9093	0,8683
time	25.12	18.54	49,12	47,50656

Table 3. SVM Radial with Scale,Center,PCA, and BoxCox

Model	CS	CS,PCA	CS,Boxcox	CS,PCA,Box
Accuracy	0,9312	0,9279	0,9316	0,8945
p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16
kappa	0,9168	0,9128	0,9172	0,8723
time	5,046	5,568402	6,108614	7,180174

Table 4. Confusion Matrix CS-BoxCox-SVM Radial

Prediction	Barbunya	Bombay	Cali	Dermas	Horoz	Seker	Sira
Barbunya	244	0	8	0	1	1	0
Bombay	0	104	0	0	0	0	0
Cali	12	0	307	0	5	0	1
Dermas	0	0	0	663	3	4	57
Horoz	0	0	9	0	369	0	6
Seker	0	0	1	7	0	389	5
Sira	8	0	1	39	7	11	458

Table 5. CS-BoxCox - SVM Radial

Class	Barbunya	Bombay	Cali	Dermas	Horoz	Seker	Sira
Sensitivity	0.92424	1	0.9417	0.9351	0.9584	0.9605	0.8691
Specificity	0.99593	1	0.9925	0.9682	0.9936	0.9944	0.9699
Accuracy	0.96009	1	0.9671	0.9516	0.9760	0.9774	0.9195

Performance Classification

With a score of 1, the Bombay class has the best accuracy, sensitivity, and specificity. with the smallest possible classes (104 pieces). It was only when the Sira class did well that they had the best accuracy, sensitivity, and specificity scores of 0.8691, 0.9699 and 0.9195, respectively.

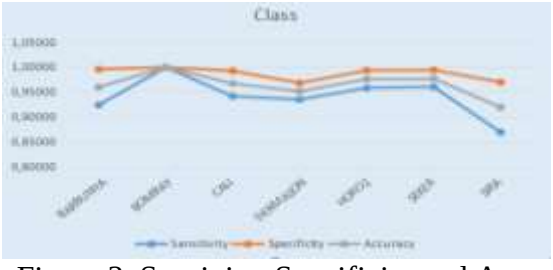

Figure 3. Sensivity, Specificity and Accuracy each class

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 2, April 2022, hlm. 231 - 238 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i2.1378
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Feature Elimination

In this study, recursive feature elimination is utilized to limit the number of factors used in prediction. Apart from class, seven variables can be utilized as predictors: convexArea, area, equivDiameter, perimeter, minorAxisLength, and class.

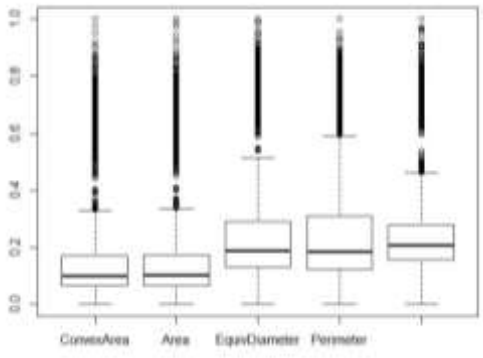

Figure 4. Boxplot of five variables obtained by RFE.

During the previous test, it was established that the SC BoxCox-SVM Radial model performed significantly better than the other models. Following that, perform the RFE-SC, BoxCox-SVM Radial test. The accuracy and kappa values obtained from this measurement are 0.9118 and 0.8931, respectively. With a cost-effective value, it indicates that the likelihood of misclassification is quite low.

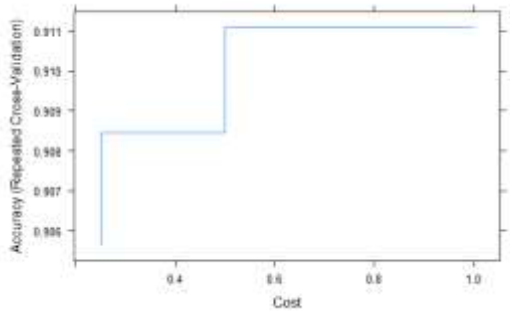

Figure 5. Cost vs Accuracy (RCV) RFE

CONCLUSION

The research discovered that using linear SVM and non-linear SVM with simple preprocessing scale and center produced accuracy values of more than 90%. With the use of PCA, the prediction process for linear SVM can be enhanced. However, combining PCA with BoxCox actually reduces the accuracy of the results. It differs from BoxCox in that non-linear can boost the accuracy of the findings while decreasing the processing speed. RFE-SVM Radial does not outperform BoxCox-Radial, but it does outperform the 91 percent accuracy achieved by using only seven variables, a ten-variable reduction.

REFERENCE

[1] H. F. Pardede, E. Suryawati, D. Krisnandi, R. S. Yuwana, and V. Zilvan, "Machine Learning Based Plant Diseases Detection: A Review," *Proceeding - 2020 Int. Conf. Radar, Antenna, Microwave, Electron. Telecommun. ICRAMET 2020*, pp. 212–217, Nov. 2020, doi: 10.1109/ICRAMET51080.2020.9298619.

[2] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Comput. Electron. Agric.*, vol. 174, Jul. 2020, doi: 10.1016/J.COMPAG.2020.105507

[3] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors 2018, Vol. 18, Page 2674*, vol. 18, no. 8, p. 2674, Aug. 2018, doi:

10.3390/S18082674.

[4] E. R. Arboleda, A. C. Fajardo, and R. P. Medina, "Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors," *2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018*, pp. 1–5, Jun. 2018, doi: 10.1109/ICIRD.2018.8376326.

[5] I. Wardhana, M. Ariawijaya, V. A. Isnaini, and R. P. Wirman, "Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 92–99, Feb. 2022, doi: 10.29207/RESTI.V6I1.3682.

[6] A. Kayabaşı, "An application of ANN trained by ABC algorithm for classification of wheat grains," 2018, Accessed: Nov. 16, 2021. [Online]. Available: http://earsiv.kmu.edu.tr/xmlui/handle/11492/1807.

[7] S. Anita and Albarda, "Classification Cherry's Coffee using k-Nearest Neighbor (KNN) and Artificial Neural Network (ANN)," *2020 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2020 - Proc.*, pp. 117–122, Oct. 2020, doi: 10.1109/ICITSI50517.2020.9264927.

[8] P. Müller *et al.*, "Scent classification by K nearest neighbors using ion-mobility spectrometry measurements," *Expert Syst. Appl.*, vol. 115, pp. 593–606, Jan. 2019, doi: 10.1016/J.ESWA.2018.08.042.

[9] S. Ramaswamy *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci.*, vol. 98, no. 26, pp. 15149–15154, Dec. 2001, doi: 10.1073/PNAS.211566398.

[10] B. Direito, C. A. Teixeira, F. Sales, M. Castelo-Branco, and A. Dourado, "A Realistic Seizure Prediction Study Based on Multiclass SVM," *http://dx.doi.org/10.1142/S012906571750006X*, vol. 27, no. 3, Feb. 2017, doi: 10.1142/S012906571750006X.

[11] A. C. Lorena and A. C. P. L. F. De Carvalho, "Comparing Techniques for Multiclass Classification Using Binary SVM Predictors," *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 2972, pp. 272–281, 2004, doi: 10.1007/978-3-540-24694-7_28.

[12] M. Mustaqeem and M. Saqib, "Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection," *Cluster Comput.*, vol. 24, no. 3, pp. 2581–2595, Sep. 2021, doi: 10.1007/S10586-021-03282-8/TABLES/7.

[13] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1–2, pp. 321–336, Sep. 2003, doi: 10.1016/S0925-2312(03)00433-8.

[14] C. Jing and J. Hou, "SVM and PCA based fault classification approaches for complicated industrial process," *Neurocomputing*, vol. 167, pp. 636–642, Nov. 2015, doi: 10.1016/J.NEUCOM.2015.03.082.

[15] Y. Zhang, R. Xiong, H. He, and

M. G. Pecht, "Lithium-Ion Battery Remaining Useful Life Prediction with Box-Cox Transformation and Monte Carlo Simulation," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1585–1597, Feb. 2019, doi: 10.1109/TIE.2018.2808918.

[16] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–233, Sep. 2005, doi: 10.1109/TNB.2005.853657.

[17] M. D. Shieh and C. C. Yang, "Multiclass SVM-RFE for product form feature selection," *Expert Syst. Appl.*, vol. 35, no. 1–2, pp. 531–541, Jul. 2008, doi: 10.1016/J.ESWA.2007.07.043.

[18] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–18, Nov. 2018, doi: 10.1186/S12859-018-2451-4/FIGURES/16.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach. Learn. 2002 461*, vol. 46, no. 1, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.