ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse9191.42020.pdf
https://doi.org/10.30534/ijatcse/2020/9191.42020

Improving Support Vector Machine Rainfall Classification Accuracy based on Kernel Parameters Optimization for Statistical Downscaling Approach

Nurul AininaFilza Sulaiman[1], Shazlyn Milleana Shaharudin[2*], Nurul Hila Zainuddin[3], Sumayyah Aimi Mohd Najib[4]
[1,2,3]Department of Mathematics, Facultyof Science and Mathematics,Universiti Pendidikan Sultan Idris, Malaysia
[4]Department Geography and Environment, Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Malaysia
[1]aininafilza@gmail.com
[2]shazlyn@fsmt.upsi.edu.my
[3]nurulhila@fsmt.upsi.edu.my
[4] sumayyah@fsk.upsi.edu.my

ABSTRACT

This study proposed a statistical downscaling model to find the best accuracy model of daily rainfall data in east-coast Peninsular Malaysia. Statistical downscaling is an approach to build relationship between the gap of climate change data from GCM and local climate by applying various mathematical model. The current studies don't contain a detailed investigation on selection the best accuracy model of statistical downscaling in study area. The proposed statistical downscaling having a main step which is classification-based Support Vector Machine (SVM). Predictor for this classification model was selected from large scale weather variables in NCEP reanalysis data. Methodologies involve in this proposed study are Principle Component Analysis (PCA) as preprocessing steps of data and SVM as classification model. The result was found that the best accuracy achieved by SVM are 65.28% by using radial basis function kernel and pair of parameter which is gamma (γ =0.5) and penalty term (C =10000).

Key words : Statistical downscaling, classification, PCA, SVM, misclassification tolerance error, gamma

1.INTRODUCTION

Downscaling is a method to produce result in high-resolution climate or climate change data from relatively coarse-resolution General Circulation Model (GCM) [1]. The gap between the simulation obtained and climate change data from GCM which is needed in temporal and spatial scale was minimized by applying downscaling technique.Downscaling also possible to model these correlations and create relationship between atmospheric condition and local climate[2]. Other than that, downscaling also can explained by using the output from GCM which is in coarse-resolution data as input and produced high-resolution data by applying various mathematical model[3].

Dynamical downscaling and statistical downscaling approaches was developed in downscaling as the solution to fit the relationship between the coarse scale GCM outputs and hydroclimatic variables. The dynamical downscaling was nested the GCM with Regional Climate Model (RCM) to find the local weather variables. Otherwise, statistical downscaling build the empirical relationship between large-scale GCM outputs and local weather variables. The reason of statistical downscaling more widely used in hydrological studies due its simplicity and less computational demanding [4]

The further studies about statistical downscaling stated that it was developed by machine learning techniques [5]. Machine learning having two types of techniques that are supervised and unsupervised learning [6]. The supervised learning is to predict the future data by trains a model on input and output data while, unsupervised learning has been described to finds the hidden pattern or structure based on input data only. The supervised learning has two steps of techniques which are classification and regression. Then, under the unsupervised learning only has clustering techniques. If the study involving the dataset of predictor and outcome variables, then the selected of using supervised technique is more relevant [7]. The study will be stress on the classification techniques or method to find the best accurancy of doing downscaling for data hydrology in Malaysia.

The impact of climate change phenomena in water resources was gained the attention of hydrology researches community. The change of climate was influenced by the spatial and temporal variability of water resources. As the study by [8] focused on using the future projection climate change to predict the water resources in Malaysia. Furthermore, rainfall pattern in Malaysia also was analyses by meteorologists with focus on events of heavy rainfalls by using machine learning tools. This is a very significant issue and the results of such studies could be used as a guide to climatologist or hydrologists to recommend actions in mitigating the flood damages and taking necessary precautions when they happen.

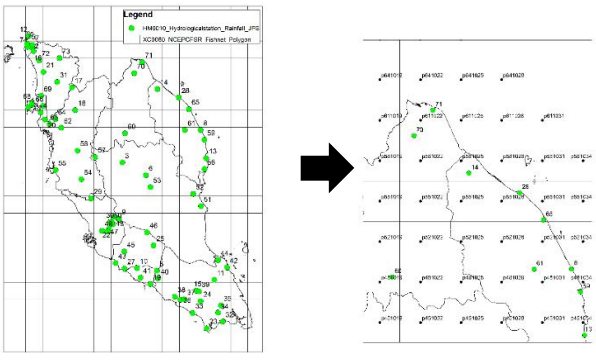
Classification is an important technique in supervised machine learning that help to extract the information from

large dataset [9].Most of the study was applied the classification method in data mining studies. For example, the study published by [10],used the classification of Random Forest and regression ofSupport Vector Machine (SVM) in order to predict the possiblelities of raining trend and predict the amount of rainfall occuring in that particular day respectively**.**Moreover, the study by [11] said that the classification method was used in their studies to identify either the day is wet or dry. The study using both step of supervised learning in machine learning techniques.

Hence, this study usethe SVM as machine learning tool due to its advantage to improve the accuracy of classification procedureespecially in data mining. The aim of this study is identifying the best accuracy of SVM in classify local precipitation data by based on relationship with climate change data from GCM. By identifying the accuracy of model, we can assist hydrologist to predict rainfall for future.

2.STUDY AREA AND DATA

The datasets used in this study were daily mean precipitation data (daily rainfall) and large-scale weather factors. Rainfall data was collected at nine rain-gauge stations in east-cost Peninsular Malaysia shown in Figure 1b. The nine rain-gauge stations were Kota Bharu, SekolahKebangsaanAmpungJabi, Kampung Merang, GuaMusang, Stor JPS Kuala Terengganu, Kampung Menerong, SekolahMenengah Sultan Omar, SekolahKebangsaanKemasek and JPS Kemaman. Daily rainfall data was recorded from 1998 until 2007 which is about 10 years and obtained from the Department of Irrigation and Drainage Malaysia. The large-scale weather factors as predictors were obtained from the National Centre for Environmental Prediction (NCEP) reanalysis dataset. The variables selected as predictors are maximum temperature ($°C$), minimum temperature ($°C$), precipitation (mm), wind (knot), relative humidity (%) and solar.



(a) (b)

Figure 1(a): All station in Peninsular Malaysia **(b):** Nine stations in East Peninsular Malaysia

3.METHODOLOGY

The proposed model for daily precipitation downscaling method consists of classification model and dimension reduction approach. The classification model will classify either the days are wet or dry since the precipitation only happen in wet days and finding the highest accuracy of classification model by variety the kernel used in analysis. The amount of precipitation on dry-days in Malaysia is less than 1mm (<1mm) [12]. The dimensional reduction approach will reduce the dimension of large data matrix to a lower dimension by retaining most of the original variability in the data [13]. The proposed of classification downscaling method was shown in Figure 2.

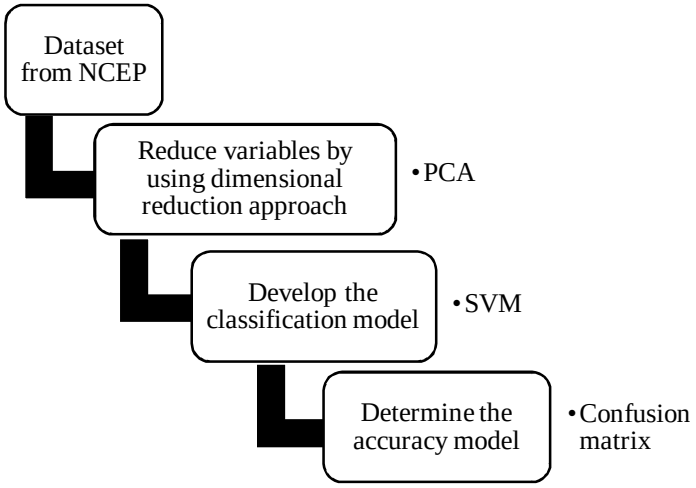


Figure 2: Flowchart of proposed classification model

A. Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is the commonly used to reduce high dimensional dataset into smaller set of components while remain the most significance of variation data [14]. PCA have two major important functions where, 1) minimize or reduce the number of variables, and 2) principle components (PC) will identify new indices that are linear combination of chosen variables[15]. The principle components (PCs) have specific properties that respect to variance [16]. Those are the steps doing PCA:

1) Step 1: standardized dataset by mean variation with has column-wise value one and occupied the same dimension as original dataset. This standardized aim to transform various variables in dataset so it can accept similar values [17].
2) Step 2: generate a correlated database. The correlated database with form of matrix correlation $p \times p$ where p represents the number of variables [15].
3) Step 3: find eigenvalue λ_j where $j = 1$,….,m. The eigenvalue can be determined by calculate the correlation matrix. The resulting of eigenvalues will be ordered from large to small.
4) Step 4: find eigenvector V_j by the help of calculated and ordered the eigenvalues λ_j of correlation matrix.

5) Step 5: determine principle components (PC) by transform the result of multiplication between standardized dataset and eigenvector matrix [17].

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful machine learning tools whichwas proposed by [18] and become more attracted of machine learning researchers and community. The algorithm of SVM has been proven effectively to be used in regression and classification methods. Based on previous studies from [19], the study reported that the SVM is generally able to result the best accuracy of classification compare than other methods. Also, SVM can performs linear and nonlinear classification with high efficiently. However, the challenging of using SVM in classification or regression is to find the best of penalty term parameter and kernel parameters. It is because of SVM is very sensitive to the parameter used.

Consider that dataset from PCs were divided into two sets which are training data and testing data. The training data with two classes $[(x_1, y_1), (x_2, y_2), \dots\dots, (x_i, y_i)]$ and the input vector is x_i, the output is y_i. The output was labelled by $y_i \epsilon \{+1, -1\}$. The classifier for the problem of binary classification is

$$f(x) = sign[w^T \cdot \phi(x) + b] \quad (1)$$

where the input vector (x) was mapped with a feature space by non-linear function $\phi(x)$. Then, the w and b are the classifier parameter. By solving the optimization problems are equivalent with determine the SVM classifier from theory,

$$\min \quad \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to } y_i[w^T \cdot \phi(x_i) + b] \geq 1 - \xi_i \quad (2)$$

$$\xi_i \geq 0, \quad i = 1, \dots, l$$

where ξ_i is a non-negative slack variables that influence objective function when data misclassified. Then, C is a penalty parameter with positive value. The optimization problems will be solved by using Lagrange multiple, α_i where $0 \leq \alpha_i \leq C$. Hence, the classifier will be eqn. (3) by a series of mathematical derivation.

$$f(x) = sign\left(\sum_{ij=1}^{l} \alpha_i y_i \phi(x_i)^T \cdot \phi(x_j) + b \right. \quad (3)$$

The kernel function, $K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j)$ was introduced to calculated the inner products. There are a quite number of kernel might be used in SVM classification but the standard kernels used are linear, polynomial, radial basis function and sigmoid. A most popular and capable kernel is the radial basis function with parameter γ.

$$K(x, x') = \begin{cases} x^T . x' & linear \\ (x^T . x' + 1)^d & polynomial \\ \exp(-\gamma ||x - x'||^2) & RBF \\ \tanh(\gamma x . x' + C) & sigmoid \end{cases} \quad (4)$$

In the final classifier, only nonzero Lagrange multiple will be take part as indicated in eq. (3). The data which having nonzero corresponding Lagrange multiple will be named as support vector. Then, the classifier will be written as,

$$f(x) = sign(\sum_{k=1}^{m} \alpha_k y_k K(x_k, x) + b) \quad (5)$$

where x_k is the support vector and m is the number of support vector. In the support vector classifier, there have two parameters to calibrate which are C(misclassification tolerance parameter) and γ (gamma). By determine those parameter, the Lagrange multiple and parameter b in eq. (5) can be find by the SVM algorithm.

4.RESULT AND DISCUSSION

In the stage of reducing dimension, PCA was used as dimension reduction approach to reduce the number of variables by using Pearson correlation [20]. This will aim the extraction of PCs that produced more significant information. The result of analysis including eigenvalue, variance, and cumulative percentage were presented in Table 1.

Table 1: Result of Principle Components Analysis

Dimension	Eigenvalue	Percent Variance	Cumulative Percentage
comp 1	2.6196	43.6604	43.6604
comp 2	1.5976	26.6265	70.2869
comp 3	0.6380	10.6332	80.9201
comp 4	0.5531	9.2179	90.1380
comp 5	0.4521	7.5349	97.6729
comp 6	0.1396	2.3271	100.0000

The result of the analysis in Table 1 shows that there having 6 components were extracted with their own value of eigenvalue and total variance. The number of components were extracted in PCA usually equivalent with the number of variables selected. The percentage of variance and eigenvalue were shown in decreasing order. The first components calculate as greater eigenvalue and most percent of variance. Followed by second component which accounts with most variance without accounts the first components. The analysis will continue like this until overall variance in data were accounted. However, based on the Kaiser criterion [21], the value of eigenvalue is more than 1.00 will be selected to interpret the component. Then, the result shown component 1 and 2 having eigenvalue with more than 1.00 which are 2.62 and 1.60 respectively. They were share total variance of 70.29%. As study by [12] they said that generally

recommended taking at least 70% of cumulative percentage of total variance as a benchmark to cut off eigenvalues in a large data set for extracting the number of components. Hence, there the reasons on selected of component 1 and 2 as components extracted for next analysis.

Table 2: Results of each loading variables of components selected

	Comp 1	Comp 2
max. temp	0.8463	-0.1868
min. temp	0.0904	0.8058
precipitation	-0.7290	0.1962
wind	-0.0262	0.8919
relative	-0.8557	-0.2814
solar	0.7943	0.0137

Based on the result in Table 2 shows that the selected components were contributed for the certain variables. The loading with above 0.5 are pick as contribute value for the components represents. As for Comp 1 shows that maximum temperature, precipitation, relative humidity and solar are variables were contributed with value of loading 0.84,0.73,0.86 and 0.79 respectively. Then, the loading variables minimum temperature and wind shows that they were contributed for comp 2 with value of 0.81 and 0.89 respectively.

The Table 3 described result of development of SVC as classification model. As shown in table below, the types of kernel were used in this study are radial basis function, sigmoid, polynomial and linear. Each kernel function has a particular parameter that must be optimized to obtain the best result performance [21]. In way to get the best result by using SVC, the selected of value of parameters also are the main point to highlight in this study. The best accuracy model can be determined based on the suitable parameter values were substituted in simulation. The selected values of parameter C are 0.01, 0.1,1,10,100,1000,10000 and 100000 for each types of kernel respectively. Then, the selected values of γ are 0.25, 0.5 and 1.0 for each values of C respectively. Those value pair of parameters involved in this study was selected based on the study from [22]that was examined using fivefold cross-validation and the best of cross-validation was picked.

Table 3: Result of Support Vector Classification

Type of kernel	Parameter		Result		
	C	γ	no. of support vector	Misclass-fication error	Accuracy model (%)
Radial	0.01	0.25	13512	0.355	64.48
		0.50	13626	0.354	64.62
		1.00	14075	0.353	64.67
	0.1	0.25	12543	0.353	64.75
		0.50	12457	0.352	64.83
		1.00	12503	0.351	64.88
	1	0.25	12294	0.352	64.78
		0.50	12165	0.350	64.95
		1.00	12144	0.349	65.12
	10	0.25	12175	0.351	64.91
		0.50	12052	0.349	65.07
	100	0.25	13076	0.358	64.21
		0.50	13076	0.358	64.21
		1.00	12941	0.357	64.35
	1000	0.25	12941	0.357	64.35
		0.50	12941	0.357	64.35
		1.00	12941	0.357	64.35
	10000	0.25	12038	0.349	65.09
		0.50*	**12050***	**0.347***	**65.29***
		1.00	12363	0.351	64.90
	100000	0.25	13549	0.354	64.64
		0.50	14440	0.350	65.03
		1.00	14937	0.352	64.83
Polynomial	0.01	0.25	14762	0.436	56.39
		0.50	14594	0.431	56.95
		1.00	14564	0.430	57.03
	0.1	0.25	14587	0.431	56.95
		0.50	14563	0.430	57.03
		1.00	14559	0.429	57.04
	1	0.25	14559	0.430	57.04
		0.50	14558	0.430	57.04
		1.00	14560	0.430	57.04
	10	0.25	14562	0.430	57.04
		0.50	14558	0.430	57.04
		1.00	14549	0.426	57.41
	100	0.25	14560	0.429	57.04
		0.50	14532	0.425	57.53
		1.00	14128	0.422	57.76
	1000	0.25	1947	0.466	53.39
		0.50	14286	0.443	55.75
		1.00	1947	0.466	53.39
	10000	0.25	13791	0.456	54.45
		0.50	1653	0.473	52.68
		1.00	446	0.467	53.28
Types of kernel	Parameter		Result		
	C	γ	no. of support vector	Misclass-fication error	Accuracy model (%)
Polynomial	100000	0.25	1392	0.467	53.25
		0.50	438	0.473	52.68
		1.00	443	0.471	52.85
Sigmoid	0.01	0.25	13191	0.372	62.85
		0.50	9660	0.441	55.85
		1.00	8762	0.460	53.98
	0.1	0.25	7890	0.437	56.26

		0.50	7648	0.452	54.77
		1.00	7806	0.467	53.33
	1	0.25	7327	0.440	56.05
		0.50	7447	0.454	54.63
		1.00	7715	0.468	53.22
	10	0.25	7259	0.440	56.00
		0.50	7429	0.454	54.63
		1.00	7705	0.468	53.21
	100	0.25	14560	0.430	57.04
		0.50	7425	0.454	54.63
		1.00	7703	0.468	53.21
	1000	0.25	14600	0.433	56.75
		0.50	14286	0.443	55.75
		1.00	1947	0.466	53.39
	10000	0.25	13791	0.456	54.45
		0.50	1653	0.473	52.68
		1.00	446	0.467	53.28
	100000	0.25	1392	0.468	53.25
		0.50	438	0.473	52.68
		1.00	443	0.471	52.85
Linear	0.01	0.25	13174	0.359	64.12
		0.50	13174	0.359	64.12
		1.00	13174	0.359	64.12
	0.1	0.25	13174	0.359	64.12
		0.50	13087	0.358	64.21
		1.00	13087	0.358	64.21
	1	0.25	13078	0.358	64.21
		0.50	13078	0.358	64.21
		1.00	13078	0.358	64.21
	10	0.25	13077	0.358	64.21
		0.50	13077	0.358	64.21
		1.00	13077	0.358	64.21
	100	0.25	13076	0.358	64.21
		0.50	13076	0.358	64.21
		1.00	13076	0.358	64.21
	1000	0.25	12941	0.357	64.35
		0.50	12941	0.357	64.35
		1.00	12941	0.357	64.35
Types of kernel	Parameter		Result		
	C	γ	no. of support vector	Misclass-fication error	Accuracy model (%)
Sigmoid	10000	0.25	12036	0.359	64.12
		0.50	12036	0.359	64.12
		1.00	12036	0.359	64.12
	100000	0.25	6885	0.400	60.04
		0.50	6885	0.400	60.04
		1.00	6885	0.400	60.04

As shown in Table 3, SVC was resulting the number support vector and misclassification error. The number of support vector are representing the data which approaching or far away from hyperplane during classification. The suitable value of number of support vector is medium value where it is represents the classification is overfitting or under fitting. Based on Table 3, the highest number of support vector is 14937 and the lowest number of support vector is 443. Then, the accuracy of model to be classification model can be determining by subtracts misclassification error with 1. The values of misclassification model were obtained by using matrix confusion between prediction and current class. The best accuracy of model achieved when the determined pair (10000,0.5) and radial kernel was set as parameter and the worst accuracy get from pair (10000,0.5), (100000,0.5) and polynomial and sigmoid as parameter respectively.

5.CONCLUSION

In this research, statistical downscaling was applied by stressing on the classification method for the data hydrology in East Peninsular Malaysia. The classification method was helped to classify the precipitation to either two or more classes. SVM is used as the classification model due its capability to classify large data set. The unique of SVM method to classify group also be a reason for widely used in variety field.

In this study, 32869 data daily with two dimension or variables is used and reduced by PCA. Then, by using SVM, the classification of data is made by classified the data into four groups and it is found that the most accuracy result of SVM as classification model is 65.28% with the lowest value of misclassification is 0.347146. The most accuracy of SVC was achieved by applying radial kernel with pair of $C(10000)$ and $\gamma(0.50)$. The best number of support vector for this study is 12050 which not be overfitting or under fitting.

ACKNOWLEDGEMENTS

This research has been carried out under Fundamental Research Grants Scheme (FRGS/1/2019/STG06/UPSI/02/4) provided by Ministry of Education of Malaysia.

REFERENCES

1. M. Devak. **Downscaling of Precipitation in Mahanadi Basin, India**, *International Journal of Civil Engineering Research*, vol .5, no. 2, pp. 111-120,2014.
2. T.Tahir, A. Hashim and K.Yusof. **Statistical downscaling of rainfall under transitional climate in Limbang River Basin by using SDSM**, in *IOP Conference Series: Earth and Environmental Science*, 2018, pp. 140.
3. K. Salvi, Kannan S. and S. Ghosh.**High-resolution multisite daily rainfall projections in India**, *Journal of Geophysical Research: Atmospheres*,vol. 118, no. 9,pp.
4. 3557-3558, 2012.

R. Xu, N. Chen, Y. Chen, and Z. Chen.**Downscaling and Projection of Multi-CMIP5 Precipitation Using Machine Learning Methods in the Upper Han River Basin,***Advances in Meteorology*, pp. 1-17, 2020.

5. D. Sachindra, K. Ahmed, M. Rashid, S. Shahid and B.J.C. Perera. **Statistical downscaling of precipitation using machine learning techniques**, *Atmospheric Research*,vol. 212,pp. 240-258, 2018.
6. P. Koturwar, S. Girase, D. Mukhopadhyay.**A Survey of Classification Techniques in the Area of Big Data**,*Computer Science*, pp. 1-7, 2015.
7. D. Ganatra and D. Nilkant.**Ensemble methods to improve accuracy of a classifier**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3434-3439, 2020. https://doi.org/10.30534/ijatcse/2020/145932020
8. A.J. Shaaban.**Regional Modeling of Climate Change Impact on Peninsular Malaysia Water Resources**, *Journal of Hydrology Engineering*, 2011.
9. T. Ketha and S.S. Imambi.**Analysis of road accidents to indentify major causes and influencing factors of accidents – a machine learning approach**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3492-3497, 2019.
10. S. Pour, S. Shahid and E. Chung.**A hybrid model for statistical downscaling of daily rainfall**,*Procedia Engineering*, vol. 154, pp. 1424-1430, 2016.
11. S. Chen, P. Yu and Y. Tang.**Statistical Downscaling of Daily Precipitation using Support Vector Machines and Multivariate Analysis**,*Journal of Hydrology*, vol. 385, no. 1-4, pp. 13-22, 2010.
12. S. M. Shaharudin, N.H. Ahmad, Zainuddin and N.S. Mohamed. **Identification of rainfall patterns on hydrological simulation using robust principal component analysis**,*Indonesian Journal of Electrical Engineering and Computer Science*,vol.11, no.3, pp. 1162-1167, 2018.
13. S. M. Shaharudin, S. Ismail, S. Mariana and A. Norhaiza. **An efficient method to improve the clustering performance using hybrid robust principal component analysis-spectral biclustering in rainfall patterns identification**,*IAES International Journal of Artificial Intelligence*, vol. 8, no. 3, pp. 237-247, 2019.
14. L. Bethere, J. Sennikovs and U. Bethers.**Climate indices for the Baltic states from principal component analysis**,*Earth System Dynamics*, vol 8, no. 4, pp. 951-962, 2017.
15. S. Omprakash and Prof S. Sumit.**A Review on Dimension Reduction Techniques in Data Mining**,*Computer Engineering and Intelligent Systems*, vol. 9, no. 1,pp. 7-14, 2018.
16. M. Denguir and S. Sattler.**A dimensionality-reduction method for test data**,in *Proceedings of the 2017 IEEE 22nd International Mixed Signals Testing Workshop (IMSTW)*, 2017,pp. 1-6.
17. V.N. Vapnik. ***The Nature of Statistical Learning Theory***, New York: Springer-Verlag, 1995.
18. O.Yamini and Prof S. Ramakrishna.**A Study on Advantages of Data Mining Classification Techniques**,*International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 9,pp. 969-972, 2015.
19. S.M. Shaharudin, N. Ahmad, and F. Yusof.**Improved Cluster Partition in Principal Component Analysis Guided Clustering**, *International Journal of Computer Applications*, vol. 75, no. 11, pp. 22-25,2013. https://doi.org/10.5120/13156-0839
20. T. Kabanda and S. Nerwiini, **Impacts of climate variation on the length of the rainfall season:an analysis of spatial patterns in North-East South Africa**, *Theoretical and Applied Climatology*, vol. 125, no. 1-2,pp. 93-100, 2016.
21. Y. Zhang and L.Wu. **Classification of fruits using computer vision and a multiclass support vector machine**, *Sensors (Switzerland)*, vol. 12, no. 9, pp. 12489-12505, 2012.
22. Y.Zhao and S. Miner, ***Data Mining Applications with R***," Amsterdam: Elsevier, Amsterdam, 2014.