EasyChair Preprint

№ 7293

Estimation of Data Parameters Using Cluster Optimization

Dileep Kumar Kadali, M Chandra Naik and
R.N.V. Jagan Mohan

January 5, 2022

Estimation of Data Parameters Using Cluster Optimization

Dileep Kumar Kadali[1], M. Chandr Naik[2], R.N.V. Jagan Mohan[3]

[1]Research Scholar, Dept. of CSE, GIET University-Gunupur, Odisha- 765022.
dileep.kadali@giet.edu and dileepkumarkadali@gmail.com

[2]Professor, Dept. of CSE, GIET University-Gunupur, Odisha- 765022.
srichandra2007@gmail.com

[3]Associate Professor, Sagi Rama Krishnam Raju Engineering College, Bhimavaram-534 204.
mohanrnvj@gmail.com

Abstract: Machine learning is the kind of process to turns out to be an undividable fraction of any methodical work is mostly because of the simplicity of use and the simplicity of generation of data. The data generation task is pricey or tedious, the data is usually generated in parallel amid various research groups and they are communal by scientists. For this activity is the initial task, is frequently to cluster the data into several categories to set up which data from what source are related to each other. In this paper, optimization is usually is implemented in use for such a clustering activity is proposed. When the data is accomplished then they are entire jumbled jointly. One way to prepare an optimization difficulty is to primary decide on a number of clusters that the data may be separated to. After that, for each cluster in not many parameters are used variables for the optimization task. The parameters have to explain a similarity role for a cluster. The activity can be achieved through an optimization solve the prediction activity can be achieved using an optimization process. This process is used a Semi-Supervised learning approach and Data Mining like the K-Nearest Neighboring process and form the path route cluster. The prediction parameter of SRGM is approximated based upon these data clusters using Least Square Estimation.

Keywords: Cluster Optimization, Data Mining, K-Nearest Neighbour, Least Square Method, Semi-Supervised Learning etc.

1. Introduction

A cluster is a clonal group of data elements or objects. It's a focal process of investigative data, and a common practice for data analysis with statistics, used in several areas, with pattern recognition, machine learning, data compression, information retrieval, bioinformatics, image analysis, and computer applications. In this, we consider a group of clustering methods, which produce a partition of those into stated different clusters, by either maximizing or minimizing some numerical standard. Such optimization methods differ from the existed methods, not essentially forming tiered classifications of the data. Differences between the approaches in all the groups rise both since of the variation of clustering norms that might be optimized and the various optimization algorithms that might be used. In the early discussion of these methods, it was assumed that the investigator has fixed the total number of groups. Techniques are possibly useful for suggesting the 'correct' number of clusters are labeled. The elementary idea behindhand the methods to be described to that associated by means of a respective partition of the n those into the essential size of groups, g, is an index c(p, q), the value of which trials approximately aspect of the 'quality of this certain cluster. For approximate indices, high values are associated with a required cluster solution, whereas for others a low value is required. Associating an index with each cluster allows them to be

compared. Varieties of such clustering criteria have been suggested. Approximately operate based on the inter-individual variations; others cluster the original data matrix.

2. Cluster Parameters Using Optimization Procedure

We consider the two-dimensional data shown in the figure. Once the data is accomplished then they are entire jumbled jointly. One way to get ready an optimization complexity is to main make a decision on a number of clusters that the data may be separated too. After that, for each cluster in not many parameters are used variables for the optimization task. The parameters have to explain a similarity role for a cluster. The activity can be achieved through an optimization solve the prediction movement can be achieved using an optimization procedure.

$$C_i(p,q) = \frac{(p-p_i)^2}{l_i^2} + \frac{(q-q_i)^2}{m_i^2} \quad (1)$$

This is an equation of an ellipse with centre at $(p_i - q_i)$ and major and minor axis values as l_i and m_i respectively. After that, each data point (p, q) is experienced for each cluster affinity (similarity) function C_i. The point is linked with that cluster for which the affinity function value is minimum. This process will connect every data point with a particular cluster. Point A (in the figure) may get linked with Cluster 2. Then, an intra-cluster distance $\left(D_i^{intra}\right)$ between all linked points of each cluster is computed as follows

$$D_i^{intra} = \frac{1}{|C_i|}\left(\sum_{(p,q)\in C_i}\sqrt{(p-p_i)^2 + (q-q_i)^2}\right) \quad (2)$$

Further, an inter-cluster distance

$$\left(D_i^{intra}\right) = \frac{1}{\binom{K}{2}}\left(\sum_{i=1}^{K}\sum_{j\neq 1}^{K}\sqrt{(p-p_i)^2 + (q-q_i)^2}\right) \quad (3)$$

Then, the following optimization problem is solved.

$$\text{Minimize}\frac{D_i^{intra}}{D_i^{intra}} \quad (4)$$

Subject to

$$(p_i, q_i) \in R,$$

where R is the particular range in which the centers of ellipses are hypothetical to lie. The minimization will ensure that intra-cluster points are as close to the center as possible and all cluster centers are as far away from each other as possible, thereby achieving the aim of clustering. Prediction is another data mining activity that is used every day. Given that the modeling activity can be achieved during optimization problem solving as conferred before is the prediction activity can be achieved using an optimization procedure.

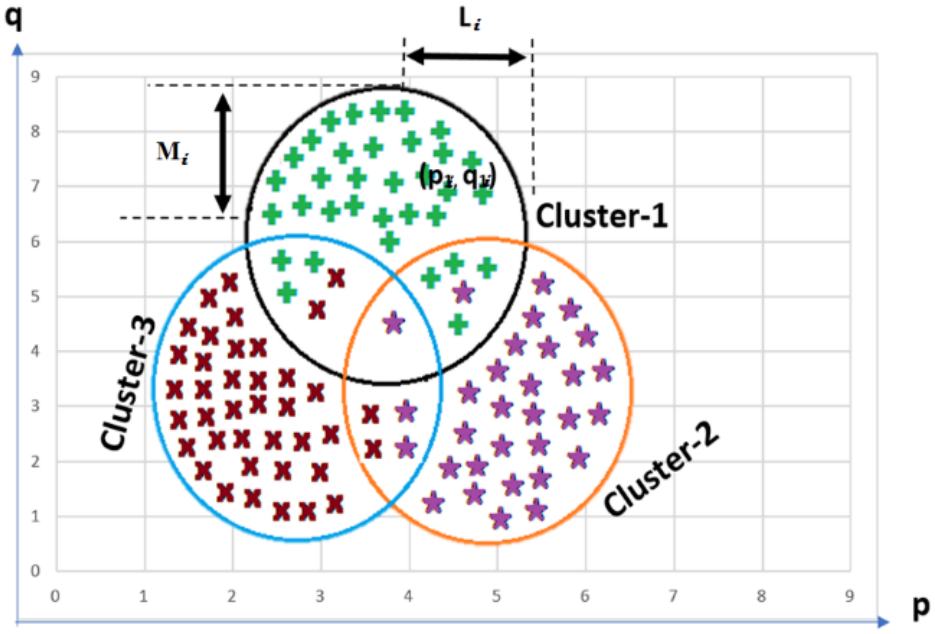


Figure: 2.1. Clustering Procedure is illustrated

3. Classifications of Parameters Estimation Using K-Nearest Neighboring

The proposed procedure consists of the implementation and utilization of machine learning approaches like the Semi-Supervised learning approach and Data Mining like the K-Nearest Neighboring process and the form of the path route cluster. Initially, we gathered the data from Data warehousing. Then, we will effort to classify the variables based on the Euclidean distance formulation.

$$D(p, q)=\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \quad (5)$$

By using this formula distance, we will categorize the new data point in the space that is occupied. To prepare the similar we need to deliberate the K value for the point. While the nearest data points are recognized, the new data cluster can be encompassed into that cluster. In addition to considering all the above data points from the cluster node and submitting them as the inputs to the route path in this process.

The following will be the algorithm of the K-NN procedure in this process of implementation. Here, we have steps to identify the operations in the K-NN procedure.

The basic process steps are as follows:

1. To store/load all the input data into the training data set.
2. For each trial in the test set data.
3. Find out for the K-nearest model to input model using a Euclidean Distance Measure.
4. For each cluster classification and compute the poise for each and every cluster as C_i/K, where C_i is count of the number of samples between the K-Nearest samples fitting to cluster i.
5. The classification for given input model is the cluster with higher confidence.

4. Semi-Supervised Learning Using K-NN Classification

Cluster evaluation techniques are seeking to partition information set into similar subgroups. It is very useful in a huge form of sets. Predictable clustering techniques are unsupervised, meaning that there is no conclusion variable nor is whatever recognized approximately the connection between the observations inside the records set. In many

situations, however, facts around the clusters are similar to the values of the capabilities. For instance, the cluster labels of a few observations may be recognized or sure observations that may be recognized to belong to the indistinguishable cluster. In different cases, one might also desire to become aware of clusters, which are associated with a selected conclusion variable. This describes namely semi-supervised clustering methods that may be implemented in these conditions.

Input: Cluster with few labeled data.

Output: All the data labeled in the cluster data.

Initial training size;
While<all the objects in the cluster data set are labeled>
Do
Test size <-Euclidean Distance between training-size;
Create Data matrix;
Group->Group [cluster, training-size, test-size];
Classifier->model [Group"K-NN"];
Result->Classify [Group, model];
Append the Result label GroupObjects;
Training size<- training size+test size;
Done
End

5. Parameter Approximation and Comparison Measures

The Software Reliability Growth Model (SRGM) be contingent upon the quality of disappointment data heavily gathered. The prediction parameter of SRGM is approximated based upon these data. The least-square technique or all-out probability approximation is commonly used for the approximation of parameters of an SRGM. The non-linear model methods are applied to numerical (medical) data to find the solution using the least square technique and the essential measures to solve it.

5.1. Mean Square Error (MSE):

Below this comparison model to be used for simulate the fault data is to changes between the predictable values are $\hat{m}(e_i)$ and the observed data O_i measured by MSE as follows,

$$\text{MSE}=\sum_{i=1}^{k} \frac{(\hat{m}(e_i) - O_i)}{k} \qquad (6)$$

where k is the total number of observations

The lower value of MSE is point to a smaller amount fitting error, therefore better goodness of fit.

5.2. Coefficient of multiple determinations (R^2)

The coefficient is ratio of the amount of squares resultant from the trend model to that from constant model subtracted from1,

$$R^2= 1 - \frac{residual\ Sum\ of\ Square}{corrected\ Sum\ of\ Square} \qquad (7)$$

R^2 is the measure of the percentage for the entire variation and mean reported for the fitted curve. The measure values range between 0 to 1. The minimum values can indicate the model because it does not fit the actual data. The higher R^2 is improved the model explains the variation in the data..

5.3. Prediction Error

At any instant time of t, the number of failures in between observation and predictions is the predictions Error. The small value of estimate error improved value is the goodness of fit the average of prediction errors is recognized as bias. The small value of Bias is the improved value of goodness of fit.

Variation: The standard deviation of Prediction Error is recognized as a variation.

$$\text{Variation}=\sqrt{\left(\frac{1}{N-1}\right)}\sum(Prediction\ Error_t - Bias)^2 \qquad (8)$$

Where small value of variation improved is the goodness of fit.

5.4. RMSPE (Root Mean Square Prediction Error) :

The RMSPE is a measure of familiarity by means of which a model predicts the observation,

$$\text{RMSPE}=\sqrt{(Bias^2 + Variation^2)} \qquad (9)$$

Where small value of RMSPE (Root Mean Square Prediction Error) is improved the goodness of fit.

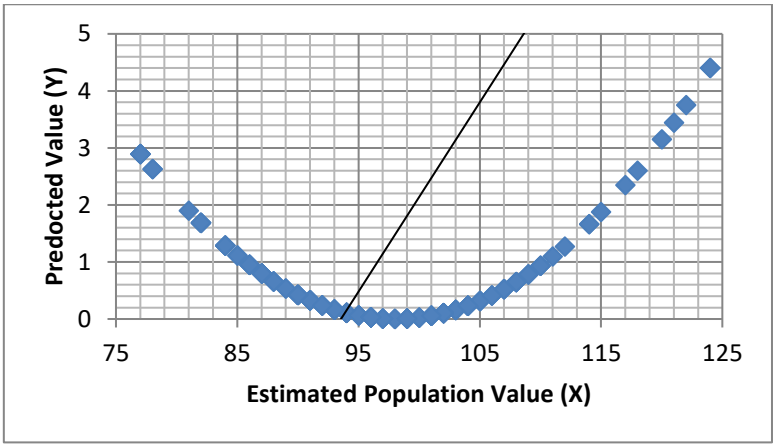
The model is discussed and proposed are validated and compared on actual life, data set refers to this paper.

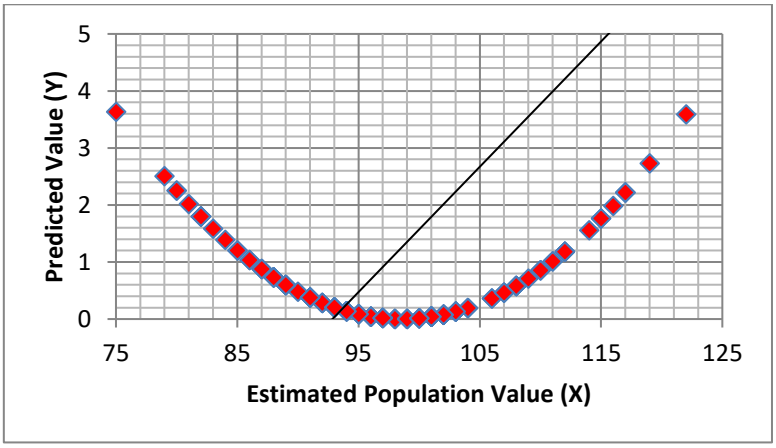
6. Experimental Result

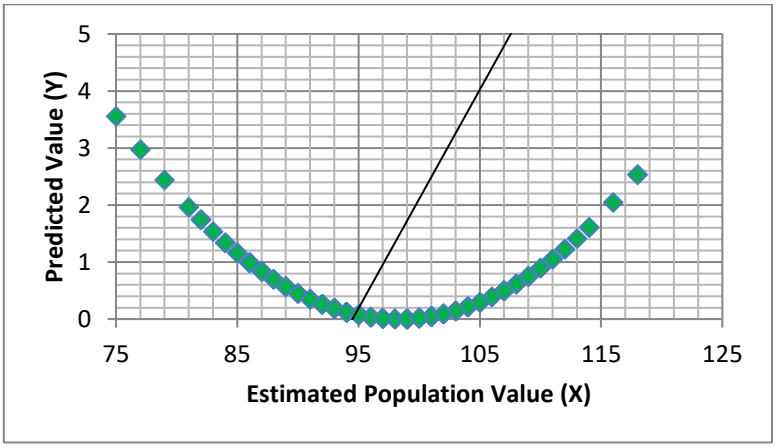
The Software Reliability Growth Model (SRGM) depends upon the quality of disappointment data heavily gathered. The prediction parameter of SRGM is approximated based upon these data. The least-square technique or all-out probability approximation is broadly used for the approximation of parameters of an SRGM.

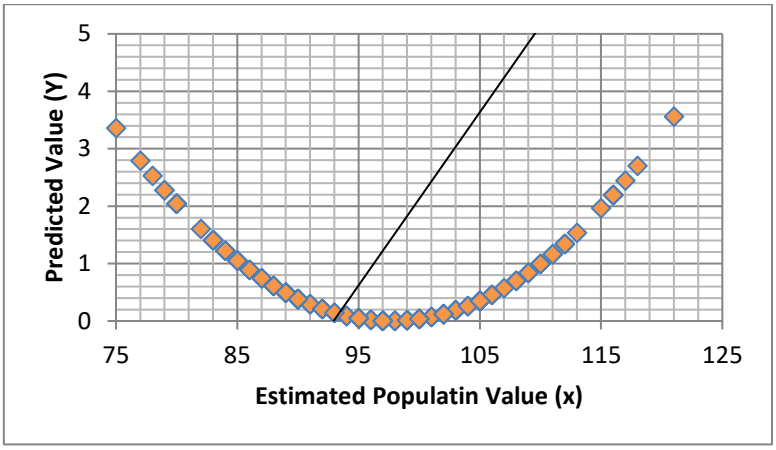
Crime Details→ / City Names↓	Murder		Rape		Robbery		Auto Theft	
	Estimated Population Value(Units)	Predicted Value (Units)	Estimated Population Value(Units)	Predicted Value (Units)	Estimated Population Value(Units)	Predicted Value (Units)	Estimated Population Value(Units)	Predicted Value (Units)
A Konduru	90	0.423	100	0.013	90	0.452	93	0.142
Achanta	92	0.239	97	0.016	103	0.144	92	0.21
Addateegala	96	0.027	103	0.128	91	0.349	116	2.196
Agiripalle	106	0.414	95	0.083	112	1.225	84	1.221
Ainavilli	105	0.316	95	0.083	98	0.001	87	0.744
Akividu	77	2.895	74	3.946	86	0.991	95	0.047
Alamuru	92	0.239	82	1.795	109	0.747	90	0.384
Allavaram	93	0.166	107	0.464	111	1.052	80	2.04
Amalapuram	109	0.784	96	0.043	70	5.239	83	1.406
Ambajipeta	89	0.535	96	0.043	86	0.991	108	0.698
Anaparthi	101	0.057	104	0.193	97	0.011	105	0.351
Atreyapuram	103	0.16	75	3.631	99	0.003	97	0.003
Attili	98	0	101	0.039	100	0.019	109	0.839
Avanigadda	115	1.879	85	1.204	104	0.211	105	0.351
Bantumilli	103	0.16	82	1.795	101	0.047	102	0.123
Bapulapadu	114	1.664	98	0.002	114	1.609	104	0.262
Bhimadole	107	0.524	81	2.018	104	0.211	115	1.963
Bhimavaram	90	0.423	88	0.73	107	0.493	88	0.611
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Undrajavaram	96	0.027	87	0.875	95	0.072	90	0.384
Unguturu	192	57.696	187	51.109	202	70.27	177	41.134
Uppalaguptam	101	0.057	101	0.039	92	0.26	86	0.89
Vatsavai	88	0.66	93	0.203	102	0.089	97	0.003
Veeravasaram	87	0.797	91	0.375	89	0.567	105	0.351
Veerullapadu	92	0.239	111	1.01	100	0.019	105	0.351
Vijayawada	104	0.232	97	0.016	96	0.035	80	2.04
Vissannapet	99	0.006	83	1.585	94	0.122	97	0.003
Vuyyuru	87	0.797	109	0.711	95	0.072	88	0.611
Y Ramavaram	74	3.779	119	2.728	108	0.613	100	0.036
Yelamanchili	78	2.626	100	0.013	97	0.011	89	0.491
Yeleswaram	96	0.027	100	0.013	97	0.011	110	0.994

Table: 6.1. Estimated Crime Data for Various Cities


Predocted Value (Y)
0
1
2
3
4
5
75
85
95
105
115
125
Estimated Population Value (X)


Predicted Value (Y)
0
1
2
3
4
5
75
85
95
105
115
125
Estimated Population Value (X)


Predicted Value (Y)
0
1
2
3
4
5
75
85
95
105
115
125
Estimated Population Value (X)



	Murder	Rape	Robbery	Auto Theft
Mean	98.04545455	98.57142857	98.31168831	97.66883117
Variance	229.5992276	193.697479	224.5688821	219.2425516
MSE	9446.807096	9600.742143	9491.008248	9391.389512
RMSE	97.19468656	97.98337687	97.42180581	96.90918177

Table: 6.2. Predicted Crime Data for Various Cities

The Predicted Crime Data for Various Cities is variation between iterations like Murder, Rape, Robbery and Auto Theft has been renowned. By using the cluster, the approach four groups have been identified Murder, Rape, Robbery, and Auto Theft. The primary selection of centroids can affect the output clusters so that the algorithm is often run several times with different initial conditions in order to get a fair interpretation of what the clusters should be.

7. Conclusion

The optimization was implemented for such a use of clustering activity is proposed. That data was accomplished and they are entire jumbled jointly. One way to prepare an optimization difficulty is to primary decide on a number of clusters that the data may be separated to. After that, for each cluster in not many parameters are used variables for the optimization task. The parameters have to explain an affinity (similarity) role for a cluster. The activity can be achieved through an optimization solve the prediction activity can be achieved using an optimization process.

8. References

1. B.Li, K.Tang, J.Li, and X.Yao "Stochastic ranking algorithm for many-objective optimization based on multiple indicators" in IEEE Transactions on Evolutionary Computation, vol. 20, No:6, pp. 924–938, 2016.
2. C. Zhou, G. Dai, and M. Wang "Enhanced dominance and density selection based evolutionary algorithm for many-objective optimization problems" in Applied Intelligence, vol. 1, pp. 1–21, 2017.
3. C.Zhu, L.Xu, and E.D.Goodman, "Generalization of pareto-optimality for many-objective evolutionary optimization" in IEEE Transactions on Evolutionary Computation, vol. 20, no. 2, pp. 299–315, 2016.

4. Dileep Kumar Kadali, R.N.V.Jagan Mohan and M. Chandra Naik, “Unsupervised based Crimes Cluster Data Using Decision Tree Classification” in Journal Solid State Technology, Volume. 63, issue.5 page: 5387-5394.
5. Dileep Kumar Kadali and R.N.V.Jagan Mohan “Optimizing the Duplication of Cluster Data for Similarity Process” in ANU Journal of Physical Science| VOL-2, JUN-DEC-2014 | ISSN: 0976- 0954, (Dileep kumar kadali et al. 2014).
6. Dileep Kumar Kadali and R.N.V.Jagan Mohan presented a paper titled “Shortest Route Analysis for High level Slotting Using Peer-to-Peer” in International conference ICRTIB-2019 – Springer [Scopus indexed] on 19th – 20th Oct 2019 at GIET University, Gunupur, Odisha.
7. Dileep Kumar Kadali, R.N.V.Jagan Mohan and M. Srinivasa Rao “Cluster Optimization for Similarity Process Using De-Duplication” in IJSRD -International Journal for Scientific Research & Development| VOL. 4, Issue 06, AUG-2016 | ISSN (online): 2321-0613, (Dileep kumar kadali et al.. 2016).
8. Dileep Kumar Kadali, R.N.V.Jagan Mohan and Y. Vamsidhar “Similarity based Query Optimization on Map Reduce using Euler Angle Oriented Approach” in International Journal of Scientific & Engineering Research, Volume 3, Issue 8, August-2012, ISSN 2229-5518(Dileep kumar kadali et al.. 2012).
9. H.M. Silva, A.M. P. Canuto, I.G.Medeiros, J.C. Xavier-Jnior "Cluster ensembles optimization using coral reefs optimization algorithm" in The 25th International Conference on Artificial Neural Networks, 2016.
10. Jagan Mohan R.N.V. and Subbarao. R. and Raja Sekhara Rao K., “Efficient K-Means Cluster Reliability on Ternary Face Recognition using Angle Oriented Approach”, Published In the Proceedings of International Conference on Advances in Communication, Navigation and Signal Processing Technically Co-Sponsored by IEEE, Hyderabad Section, March 17th-18th, 2012, Dept of ECE, Andhra University College of Engineering (A).
11. K.Deb, K.Sindhya, and J.Hakanen “Multi-objective optimization,” in Decision Sciences: Theory and Practice, pp. 145–184, CRC Press, Boca Raton, FL, USA, 2016.
12. R.Cheng, Y.Jin, M.Olhofer and B.Sendhoff “A reference vector guided evolutionary algorithm for many-objective optimization” in IEEE Transactions on Evolutionary Computation, vol. 20, No:5, pp. 773–791, 2016.
13. R.Wang, Z.Zhou, H.Ishibuchi, T.Liao and T. Zhang “Localized weighted sum method for many-objective optimization” in IEEE Transactions on Evolutionary Computation, vol. 22, no. 1, pp. 3–18, 2018.
14. S. Jiang and S. Yang “A strength pareto evolutionary algorithm based on reference direction for multiobjective and many-objective optimization” in IEEE Transactions on Evolutionary Computation, vol. 21, no. 3, pp. 329–346, 2017.
15. S.Mahdavi, S.Rahnamayan, and K.Deb “Opposition based learning: a literature review,” Swarm and evolutionary computation, vol. 39, pp. 1–23, 2018.
16. S.Mirjalili, P.Jangir, and S.Saremi “Multi-objective ant lion optimizer: a multi-objective optimization algorithm for solving engineering problems” in Applied Intelligence, vol. 46, No:1, pp. 79–95, 2017.
17. W.K.Li, W.L.Wang, and L.Li “Optimization of water resources utilization by multi-objective moth-flame algorithm” in Water Resources Management, vol. 47, no:10, pp. 3303–3316, 2018.
18. W.Wang, S.Ying, L.Li, Z.Wang, and W. Li “An improved decomposition-based multiobjective evolutionary algorithm with a better balance of convergence and diversity” in Applied Soft Computing, vol. 57, pp. 627–641, 2017.
19. X. Bi and C. Wang “A niche-elimination operation-based NSGA-III algorithm for many-objective optimization” in Applied Intelligence, vol. 48, no:1, pp. 118–141, 2018.
20. Y. Xiang, Y. Zhou, M. Li, and Z. Chen “A vector angle-based evolutionary algorithm for unconstrained many-objective optimization” in IEEE Transactions on Evolutionary Computation, vol. 21, no. 1, pp. 131–152, 2017.