Improving Penalized-Based Clustering Model in Big Fusion Data by Hybrid Black Hole Algorithm

Sarah G. M. Al- Kababchee[1], Zakariya Y. Algamal[2], Omar S. Qasim[3]

[1]Department of Mathematics, Education College, University of AL-Hamdaniya
[2]Department of Statistics and Informatics, University of Mosul, 41002 Mosul, Iraq; College of Engineering, University of Warith Al-Anbiyaa, 56001 Karbala, Iraq
[3]Department of Mathematics, University of Mosul, Mosul, Iraq
Email: sarahghanim@uohamdaniya.edu.iq; zakariya.algamal@uomosul.edu.iq; omar.saber@uomosul.edu.iq

Abstract

This paper presents an improved penalized regression-based clustering algorithm using a nature-inspired approach. Clustering is an unsupervised learning method widely used in data fusion mining, including gene analysis, to group unclassified fusion data based on their features. The proposed algorithm is an extension of the "Sum of Norms" model and aims to better estimate the data by fusing information from various sources. The performance of the proposed algorithm is evaluated on gene expression data. Results show that our approach outperforms other methods, indicating its potential impact on clustering research with data fusion.

Keywords: Black hole algorithm; Data fusion mining; Clustering fusion data, Bat algorithm; K-means.

1. Introduction

One of the problems encountered clustering at scale is complicated works in fusion data [1], pattern recognition process , image processing [2], and usual fields of sciences. The main goal of clustering data is to divide it into several clusters according to previously defined characteristics. Based on this clustering, the data are similar to each other to a large extent in the same cluster and not similar to other clusters.

Currently, known the clustering can be classified as a two parts: hierarchical clustering and partitional clustering [3]. One of the most widely known clustering algorithms is the center-based clustering algorithms. An algorithm K-means has widely used in recent years due to its simplicity and efficiency among these algorithms [4]. One of these algorithms is the Bat algorithm we used in our work, the Black hole algorithm helps us to increase the number of features for data by encoding that feature to binary, that method is useful when we have many numbers of features and also it makes our computational less.

The " sum-of-norms" (SON) clustering method or clusterpath method [5-8], consider SON is a convex centroid-based method. This method depends on over-parameterization which can control the number of clusters by using a sum-of-norm regularization and the trade-off between the model fit. Clusterpath (SON) is a convex relaxation of hierarchical clustering and k-means [9]. To find a solution for the convex problem many algorithms have been proposed [5, 7, 10, 11], for example, **ADMM** (“the alternating direction method of multipliers) and **AMA** (the alternating minimization algorithm”) [10]. Particularly difficult to separate the non-convex groups because the son with the convex base produces estimates of biased parameters. Also, some clustering methods have the same difficulty, like k-means and k-means ++[12]. To decreasing this difficulty, it is suggested to use clustering based on the penalized regression which is considered to be an extension to SON [13].

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

Using a nature-inspired approach, this research provides an improved penalized regression-based clustering technique. Clustering is a widely used unsupervised learning method in data fusion mining, particularly gene analysis, for grouping unstructured fusion data based on its properties. The suggested approach is an extension of the "Sum Of Norms" paradigm that tries to improve data estimation by fusing information from many sources. The suggested algorithm's performance is evaluated using gene expression data.

2. Related Work

Many researchers have employed swarm intelligence and evolutionary algorithms for grouping. This is because, throughout its development, iteratively identifies the optimal solution to complicated problems in order to solve them. Al-Kababchee et.al, Using an equilibrium optimization strategy is suggested as a way to improve K-means clustering. The suggested method selects the finest features while also adjusting the number of clusters to arrive at the best solution. Based on five datasets, the results demonstrate the utility of the proposed method in comparison to existing algorithms in terms of intra-cluster distances and Rand index.[14], and S Rana et.al., they introduce a novel hybrid sequential clustering method that clusters data using PSO sequentially with the K-Means technique. The suggested method fixes both algorithms' flaws, enhances clustering, and prevents getting stuck in a local optimal solution[15]. also Al-Kababchee et.al, suggest the penalized regression-based clustering for a better estimate is enhanced in this research using a nature-inspired approach. The findings of the real data application on gene expression data imply that our suggested upgrade can significantly outperform alternatives [16] , VN Wijayaningrum et.al, the issue is solved by changing the solution updating method so that a search agent has the option to follow the best search agent in addition to another randomly selected search agent. Additionally, the technique for updating each search agent's awareness likelihood in accordance with their individual capacities for problem-solving improves the balance between exploration and exploitation.[17], For SA and classification, they introduce a brand-new K-means clustering with a hybrid metaheuristic algorithm (KMC-HMA). The suggested KMC-HMA technique first preprocesses the data to exclude undesirable terms from product reviews [18], Algamal et.al., an $L_{1/2}$-norm penalized linear regression model is suggested. Furthermore, the local linear approximation strategy is applied to prevent the non-convexity of the suggested method. Using a number of benchmark data sets, the suggested method's potential application is evaluated. In comparison to other widely used penalized approaches, the suggested method not only has the best prediction power but also offers a QSAR model that is simple to understand[19], Algamal et.al., for building a reliable and effective high-dimensional QSAR model of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors, a two-stage adaptive penalized rank regression is presented. The outcomes show the efficiency of suggested approach in simultaneously generating a reliable QSAR model and choosing educational molecular descriptors.

3. Methodology

3.1 Penalized clustering

Clustering is an unsupervised application that has been used in many different fields, these include signal processing, social sciences, “marketing, economics, and business, as well as medicine and biology [20]. The task of clustering is to create a cohesive grouping of patterns, data points, or objects. Suppose **D** is the data set with n number of objects (patterns) and each object is of **m** dimensional. **D** can be represented as $D=\{X_1, X_2, ..., X_n\}$, where $D \in R^{n*m}$. Then a clustering **C** can be defined as follows: $C=\{K_1, K_2, ..., K_k\}$, such that the following conditions are satisfied $\bigcup_{i=1}^{k} K_i = D$, where K_i refers to the” i[th] cluster [21].

1- $K_i \neq \varnothing i=1,...,k.$

2- $K_i \cap K_j = \varnothing i \neq j, i, j=1,...,k$.

3- $sim(X_1, X_2) > sim(X_1, Y_1), X_1, X_2 \in K_i and Y_1 \in K_j, i \neq j.$

To evaluate the results of clustering, it is possible to use several similarity measures (“Simple Matching Coefficient (**SMC**) and cosine similarity [22]) or dissimilarity measures (Minkowski distance[23], Manhattan distance and Euclidean distance” [24]). The Euclidean distance has been used in this paper and defined as follows:

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

$$dis(X,Y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \quad (1)$$

where $dis(X,Y)$ is the Euclidean distance between two **m** dimensional objects **X** and **Y**.

When the similarity of the cluster's objects increases, the similarity "between the objects in the cluster and the objects in the other clusters diminishes. Reduce (minimize) or maximize (maximize) the similarity inside a cluster or between clusters is the aim of clustering. Calculating the distance between the two points allows one to determine how similar two data points are when the item is a data point. The greater the distance between clusters or the smaller the sum of the distances between data points and their cluster centers (intra-cluster distance) (inter-cluster distance)[25], the better the clustering effect. Therefore, the objective function can be shown as follows":

$$f=\sum_{j=1}^{k}\sum_{e\in c_j} dis(e,o_j)^2 \quad (2)$$

where $dis(e,o_j)$ the Euclidean distance between object e and center of the j[th] cluster o_j.

Let $\{x_1,x_2,...,x_n\}$ be the data set which should be clustered such that between-group dissimilarity is maximized and the within-group similarity, where $x_i \in R^m$. Suppose any data point x_i has its has centroid μ_i. The aim is to estimate μ_i's s.t μ_i's that their corresponding are equal from the same cluster.

$$\min_{\mu}\sum_{i=1}^{n}L(x_i-\mu_i)+\lambda\sum_{i=1}^{n}\sum_{i<j}h(\mu_i-\mu_j) \quad (3)$$

where L(.) represents a loss function, h(.) is a fusion penalty and $\lambda>0$ is a predefined the parameter. This model for squared error (SE) and L1-norm fusion penalty is:

$$\min_{\mu}\sum_{i=1}^{n}\left\|x_i-\mu_i\right\|_2^2+\lambda\sum_{i=1}^{n}\sum_{i<j}\left\|\mu_i-\mu_j\right\|_1 \quad (4)$$

which is a convex SON clustering model.

Specially to get rid of difficulties separating due to producing biased parameter estimates from SON with convex loss function (LF) and convex fusion penalty, **PRBC** (penalty regression-based clustering), i.e., a group- truncated lasso penalty, was used by (Pan et al. 2013). As follows:

$$\min_{\mu}\sum_{i=1}^{n}\left\|x_i-\mu_i\right\|_2^2+\lambda\sum_{i=1}^{n}\sum_{i<j}\min\{\left\|\mu_i-\mu_j\right\|_1,\tau\} \quad (5)$$

where $\tau>0$ is a predefined tuning parameter.

4. The proposed method

Several redundant data points in huge data illustrated how well clustering performed. Making a predictive clustering algorithm requires selecting a small group of related features from a large number of features, which is a crucial work. It is well known that finding the best subset of features is an NP-hard issue that takes a lot of time to compute and is expensive.

In recent years many different algorithms, been suggested to get semioptimal subsets of the solutions, that designed by simulate natural evolution, these algorithms are **GA** (genetic algorithms), **PSO** (particle swarm optimization), **ACO** (ant colony optimization), and many others.

The black hole optimization algorithm is a robust stochasticoptimization technique based on simulation of the behavior ofblack hole in outer space.The below steps explain manner of simulating BHA from blackhole phenomenon:

Step 1: Outer space is full of known and unknown stars. In realspace black hole is formed by collapsing individual stars so BHAbegins with the population of stars that located arbitrarily in theexplore space. In BHA each star has a fitness value, which is evalu-ated by a fitness function to be

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

optimized. The best star that has thebest fitness value is selected as the black hole. It is called "black"because it absorbs all the light and reflects nothing. Fig. 1 showsBHA schema. The black circle is the black hole and green circles arestars. They placed randomly in search space.

Step 2: In the real space, a black hole is an object of extreme den-sity with an intense gravitational attraction. This leads to a greatamount of gravitational force pulling stars around it. BHA has fol-lowed the same behavior. By Eq. (1) all the stars began movingtoward the black hole.

Step 3: The sphere shaped bound of a black hole in outer spaceis known as the event horizon. The event horizon radius is called asthe Schwarzschild radius. The red circle in Fig. 1 shows the eventhorizon of black hole. In the real space the Schwarzschild radius iscomputed by Eq. (2) and in BHA is computed by Eq. (3).

Step 4: Because of extreme density and strong gravitationalattraction of black hole when a star crosses the event horizon, itwill be swallowed by the black hole and disappear. In the regionof event horizon the escapee speed is tantamount to the speed ofthe light, so nothing can get away from within the event horizon.In BHA, the Euclidean distance between black hole and star is com-puted. If this distance is less than Schwarzschild radius, substituteit with a fresh star in the random location in the search space.

Step 5: In BHA if a star reaches a location with lower cost thanthe black hole, in that case theirs locations should be replaced[12,39,40].

$$X_i(t+1) = X_i(t) + rand \times (X_{BH} - X_i(t)),\ for \quad i = 1, 2, ..., N \tag{6}$$

$$R = \frac{2GM}{C^2} \tag{7}$$

$$R = f_{BH} / \sum_{i=1}^{N} f_i \tag{8}$$

where $X_i(t)$ and $X_i(t+1)$ signify the locations of the ithstar at iterations t and $t+1$, respectively. Rand indicates uniform distribution with a range from 0 to 1. N denotes the number of stars. X_{BH} points the location of the black hole in the exploration space. $M, G\ and\quad C$ signify the mass of the black hole, the gravitational constant, and the speed of light respectively. f_i denotes the fitness value of the ith star and f_B Hindicates the fitness value of the black hole.

The BHA was originally developed for continuous-valued spaces. But there exist a number of discrete combinatorial optimization problems, such as FS, in which the values are not continuous numbers but rather discrete binary integers. For this reason, we have introduced a binary version of BHA and called it BBHA. Binarization techniques can be categorized into two groups: two steps binarization and continuous-binary operator transformation. Our proposed binarization technique belongs to the first group. In the first groupwithout any modifications in the operators, only two steps is addedafter the continuous iteration. In solving FS problem the search space must be modeled as ad-dimensional Boolean lattice, where the ith star moves around the d-dimensional space.Since the problem is to select or not select of a given feature, theposition of a star only takes the values 1 or 0. Therefore, a transferfunction is needed to forces stars to move in a binary space. Transferfunctions define the probability of changing position's elementsfrom 0 to 1 and vice versa. In the proposed approach, HyperbolicTangent function is utilized to modify the position of stars as in Eqs. (9) and (10).

$$S(X_{id}(t+1)) = abs(\tanh(X_{id}(t+1))) \tag{9}$$

$$X_{id}(t+1) = \begin{cases} 1 & if \quad S(X_{id}(t+1)) > rand \\ 0 & otherwise \end{cases} \tag{10}$$

where rand is a uniform random number between 0 and 1. In Eq.(9), instead of rand threshold 0.6 can also be considered. HyperbolicTangent function belongs to the group of v-shaped transfer func-tions. It has been used here because it shows good performancecompared to the other transference functions such as sigmoid func-tion [41]. In addition, in the proposed algorithm we may face withsituation that

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

one star with small number of features has the samefitness value with black hole. In this situation we should changetheir positions.
In BBHA we only need to set number of stars. The proposed algorithm does not suffer from some of other optimization algorithms difficulties such as the slow convergence rate and adjusting severalparameters. Compared with other optimization algorithms, BBHAis easier to implement, depend on a single parameter for config-uring the model, requires much less memory, and converges morerapidly.

5. Experimental results

Our proposed algorithm, BHRBC, is compared with the PRBC and K-means method using three datasets. Table 1 presents the datasets, the features, and class for a chemical dataset. For each data set, the optimal values of the parameters $f_i = \cos(x_i)$, $x_i^j = 0.02$, and $globalf_{it} = 2$ in BBHA.

Table 1: Description of the datasets used.

Dataset	#Samples	#features	Class
Chalcone	212	3657	108 /104 (active / inactive) compounds
Hepatitis	121	2559	31 / 90 (active / inactive)
H1N1	479	2322	213 / 266 (active /weakly active)

Tables 2 and 3 compare our proposed algorithm, "BPRBC, with the BPRBC, PRBC and K-means method in terms of purity and computational time in seconds. As it can be observed from Table 3, BPRBC overtakes the standard PRBC also BPRBC. Moreover, it can be noticed that in all datasets, BPRBC obtains the highest purity with fewest selected features results compared with the others. The best results are bolded. For example, in H1N1, the improvement in purity using BHPRBC was 7.03%, in BPRBC is 6.02 and 12.16% of PRBC and K-means".
As can be seen from Table 4, in terms of computational efficiency, the BHPRBC has less time than BPRBC and PRBC. Consequently, it can be inferred that the BHPRBC outperforms the K-means algorithm. In general, BPRBC, sinusoidal map, is the fastest among all the used algorithms on all dataset

Table 2: Purity results for each used algorithm

	BHPRBC	BPRBC	PRBC	K-means
Chalcone	**0.9882**	0.9714	0.9002	0.8524
Hepatitis	**0.9725**	0.9508	0.8903	0.8324
H1N1	**0.9624**	0.9584	0.9100	0.8425

Table 3: Computational time, in seconds, results for each used algorithm

	BHPRBC	BPRBC	PRBC	K-means
Chalcone	**54**	57	70	98
Hepatitis	**43**	42	60	90
H1N1	**70**	74	88	132

To further highlight the efficiency of our proposed algorithm, Table 4 shows the average purity for 20 times replications. As can be inferred that the BHPRBC outperforms BPRBC, PRBC and K-means algorithms on all datasets. "The p-values (*) from Wilcoxon's rank-sum test (nonparametric statistical test) with a 5% significance level are adopted. The statistical test is needed to indicate that the BPRBC provides a significant improvement compared to the other algorithms. It can be seen that there is a statistical difference between BHPRBC and all the others for all datasets".

Table 4: Average purity results for each used algorithm depending on 20-time replications

	BHPRBC	BPRBC	PRBC	K-means
Chalcone	**0.9712 ± 0.005**	0.9702 ± 0.005	0.8961 ± 0.011*	0.8435 ± 0.022*
Hepatitis	**0.9604 ± 0.008**	0.9514 ± 0.008	0.8736 ± 0.012*	0.8325 ± 0.021*
H1N1	**0.9703 ± 0.008**	0.9573 ± 0.008	0.9002 ± 0.010*	0.8632 ± 0.023*

6. Conclusion

In order to determine the most useful features of clustering, we proposed this algorithm in this paper which is a black hole algorithm with penalized regression-based clustering. Feature selection plays a fundamental and important role in developing a successful clustering algorithm. Via the results of statistical analysis and experimental results on the three groups of chemical data, the proposed

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

BHPRBC compared with the BPRBC, PRBC and K-means leads to a better in terms of performance time computational and purity.

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] A. A. Esmin, R. A. Coelho, and S. Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data," *Artificial Intelligence Review*, vol. 44, pp. 23-45, 2015.

[2] Ziaul Hasan, Hassan r. Mohammad, & Maka Jishkariani. (2022). Machine Learning and Data Mining Methods for Cyber Security: A Survey . Mesopotamian Journal of CyberSecurity, 2022, 47–56. https://doi.org/10.58496/MJCS/2022/006

[3] C. K. Reddy and B. Vinzamuri, "A survey of partitional and hierarchical clustering algorithms," in *Data clustering*, ed: Chapman and Hall/CRC, 2018, pp. 87-110.

[4] X. Yan, Y. Zhu, W. Zou, and L. Wang, "A new approach for data clustering using hybrid artificial bee colony algorithm," *Neurocomputing*, vol. 97, pp. 241-250, 2012.

[5] P. A, D. D, J. FD, and B. C, "Clustering by sum of norms: stochastic incremental algorithm, convergence and cluster recovery," *International conference on machine learning*, p. 8, 2017.

[6] L. F, O. H, and L. L, " Clustering using sum-ofnorms regularization: with application to particle filter output computation," *Statistical signal processing workshop (SSP) 2011 IEEE*, pp. 201–204., 2011.

[7] H. TD, J. A, B. F, and V. J-P, "Clusterpath an algorithm for clustering using convex fusion penalties.," *international conference on machine learning*, p. 1, 2011.

[8] C. GK, C. EC, R. JMO, and L. K, "Convex clustering: an attractive alternative to hierarchical clustering," *PLoS Comput Biol* 2015.

[9] M. J, "Some methods for classification and analysis of multivariate observations," *roceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, p. 16, 1967.

[10] C. EC and L. K, "Splitting methods for convex clustering," *J Comput Graph Stat*, vol. 24, 2015.

[11] O. M. Ismael, O. S. Qasim, and Z. Y. Algamal, "A new adaptive algorithm for v-support vector regression with feature selection using Harris hawks optimization algorithm," in *Journal of Physics: Conference Series*, 2021, p. 012057.

[12] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, 2013.

[13] P. W, S. X, and L. B, "Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty," *J Mach Learn Res*, vol. 14, p. 24, 2013.

[14] S. G. M. Al-kababchee, Z. Y. Algamal, and O. S. Qasim, "Enhancement of K-means clustering in big data based on equilibrium optimizer algorithm," *Journal of Intelligent Systems*, vol. 32, p. 20220230, 2023.

[15] S. Rana, S. Jasola, and R. Kumar, "A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm," *International Journal of Engineering, Science and Technology*, vol. 2, 2010.

[16] J. Sun, W. Chen, W. Fang, X. Wun, and W. Xu, "Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization," *Engineering Applications of Artificial Intelligence*, vol. 25, pp. 376-391, 2012.

[17] V. N. Wijayaningrum and N. N. Putriwijaya, "An improved crow search algorithm for data clustering," *EMITTER International Journal of Engineering Technology*, vol. 8, pp. 86-101, 2020.

[18] M. Basha, S. Nidamanuri, A. Pureti, and V. krishna, " Clustering Based Energy Coding for Wireless Adhoc Network," *International Journal of Wireless and Ad Hoc Communication (IJWAC)*, vol. 1, pp. 34-52, 2020.

[19] Z. Y. Algamal, M. H. Lee, A. Al-Fakih, and M. Aziz, "High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm," *SAR and QSAR in Environmental Research*, vol. 27, pp. 703-719, 2016.

[20] A. Abdelaziz and A. N. Mahmoud, "A Novel Metaheuristic Optimization based Clustering with Routing Scheme for IoT Mobile Edge Computing Platform," *International Journal of Wireless and Ad Hoc Communication (IJWAC)*, vol. 4, pp. 61-71, 2022.

[21] S. Tasoulis, N. G. Pavlidis, and T. Roos, "Nonlinear dimensionality reduction for clustering," *Pattern Recognition*, vol. 107, 2020.

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023

[22] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans Cybern,* vol. 16, p. 33, 2005.
[23] P. J. Groenen and K. Jajuga, "Fuzzy clustering with squared Minkowski distances," *Fuzzy Sets and Systems,* vol. 120, pp. 227-237, 2001.
[24] J. Mao and A. K. Jain, "A self-organizing network for hyperellipsoidal clustering (HEC)," *IEEE Transactions on Neural Networks,* vol. 7, p. 13, 1996.
[25] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 1999, p. 143.

Doi: https://doi.org/10.54216/FPA.110105
Received: December 14, 2022 Accepted: March 19, 2023