Research Article

Volume-1 | Issue-1| Jan-Jun-2024|

JOURNAL OF
Information Technology and Cryptography
Double Blind Peer Reviewed Journal
DOI: https://doi.org/10.48001/JoITC

A Proposed Clustering Algorithm for Efficient Clustering of High-Dimensional Data

S. Gopinath[1*], G. Kowsalya[1], K Sakthivel[2], S. Arularasi[1]

[1]Department of Computer Science and Engineering, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

[2]Department of Computer Science and Engineering, K. S. Rangasamy College of Technology, Namakkal, Tamil Nadu, India

[*]Corresponding Author's Email: sgopicse@gmail.com

ARTICLE HISTORY:

Received: 16th Jan, 2024

Revised: 8th Feb, 2024

Accepted: 16th Feb, 2024

Published: 25th Feb, 2024

KEYWORDS:

Algorithm,
Clustering,
High-dimensional data,
Information,
Subspace

ABSTRACT: To partition transaction data values, clustering algorithms are used. To analyse the relationships between transactions, similarity measures are utilized. Similarity models based on vectors perform well with low-dimensional data. High-dimensional data values are clustered using subspace clustering techniques. Clustering high-dimensional data is difficult due to the curse of dimensionality. Projective clustering seeks out projected clusters in subsets of a data space's dimensions. In high-dimensional data space, a probability model represents predicted clusters. A model-based fuzzy projection clustering method to find clusters with overlapping boundaries in different projection subspaces. The system employs the Model Based Projective Clustering (MPC) method. To cluster high-dimensional data, projective clustering algorithms are used. A subspace clustering technique is the model-based projective clustering algorithm. Similarity analysis use non-axis-subspaces. Anomaly transactions are segmented using projected clusters. The suggested system is intended to cluster objects in high-dimensional spaces. The similarity analysis includes non-access subspaces. The clustering procedure validates anomaly data values with similarity. The subspace selection procedure has been optimized. A subspace clustering approach is the model-based projective clustering algorithm. Similarity analysis use non-axis-subspaces. Anomaly transactions are segmented using projected clusters. The suggested system is intended to cluster objects in high-dimensional spaces. The similarity analysis includes non-access subspaces. The clustering procedure validates anomaly data values with similarity. The subspace selection procedure has been improved.

1. INTRODUCTION

Information clustering encompasses a wide range of processes and has received extensive attention from the statistics, information mining, and database fields. Within the clustering domain, multitudinous computations have been proposed. One recent collection of similar computations, model-based techniques, has sparked widespread interest because of their redundant focal points, which enable them to display the introduction structures of millions within the information.

In model-based techniques, information can be derived from a variety of colorful conceivable sources, which are naturally modelled by a Gaussian mix (Jing et al., 2007). The objective is to comprehend Gaussian generating mixes. Each Gaussian source's harshness and covariance properties. Cases include classic K-means and its variants. In any event, analogous solutions for altitudinous dimensional information.

In high-dimensional spaces, information is naturally scarce, rendering Gaussian work indecorous. According to Verleysen, with the dimension supplements, the rate of tests of a regularized multivariate Gaussian distribution collapsing around its centre quickly dwindles to 0. To put it differently, the phenomenon of the "empty space miracle" takes place when the majority of the Gaussian distribution's volume is situated in the tails rather than the center within high-dimensional space. Additionally, clusters within such high-dimensional spaces may exist in entirely distinct subspaces characterized by diverse combinations of features. In a variety of real-world operations, several focuses are associated with a specific set of measures, while others are associated with distinct metrics. In document clustering, for example, groups of libraries on distinct themes are distinguished by different subsets of catchphrases.

Keywords belonging to a particular cluster may not be present in the libraries of other clusters. To tackle this challenge, projective clustering is defined as the process of identifying clusters. A projection cluster comprises centroids, each associated with a distinct subset of features. For a set of information points in three dimensions, two unique expected clusters are defined (Moise et al., 2008).

In the domain, numerous computational approaches have been introduced to identify potential similar clusters. These approaches can be categorized into two orders. The first order, which includes algorithms like PROCLUS, ORCLUS, and FINDIT, is centered on determining specific subspaces for various clusters. On the other hand, computations in the second order address the entire information space, incorporating diverse weighting values for various cluster measures. Examples of such algorithms include EWKM, FWKM, and LAC. The majority of computations in the second order follow a k-means-like structure, which shares a similar iterative framework with the EM algorithm.

In any case, there is a shared requirement for initial models upon which these techniques might be built (Chen et al., 2008). Extended Gaussian representations, are meant for projective clustering and can help clarify common assumptions employed in well-defined projection approaches through analysis.

Currently, we derive the objective work of projective clustering based on liability proofs and provide MPC, an EM-like parameter-free computation for optimizing the objective work. MPC has been tested on both agricultural datasets and various real-world astronomical datasets, and preliminary results demonstrate its acceptability (Gan et al., 2006).

As the key consistency task for the expected cluster is repeated, the expected cluster debt metric has changed. This yields another algorithm that is not inferior to Stoner-defined parameters for obtaining dimension weights.

2. RELATED WORK

2.1 High-Dimensional Clustering Techniques

In high-dimensional information clustering, dimensionality drop procedures have been used. Point selection strategies choose the most important parcels for the clustering assignment, whereas punctuation change styles, such as PCA and SVD, attempt to epitomise the information set in a smaller number of modern measures made by straight combination of the first traits. Since these traditional techniques apply to the entire information space, problems can arise when clusters are in multiple subspaces (Domeniconi et al., 2007). Near-dimensional relaxation studies are conducted to provide different innovative measures for each cluster.

Equivalent technical issues include ensuring the dimensionality of each subspace associated with a cluster. Furthermore, the computational complexity of LDR is always changing. Bi-clustering, also known as co-clustering, has been proposed for concurrent clustering on high-dimensional information focuses and measurements. One of its common operations is within the disquisition of quality expression information, where the aim is to discover groups of rates and groups of conditions that are identical enough that the rates reveal profoundly associated exercises for each condition. In any case, many perspectives are shown virtually differently within the jotting view for circumstance. We accept the scientific classification and refine the two terms based on the research underlying them.

The goal of subspace clustering studies is to capture all thick sections of all subspaces, whereas projective clustering focuses on locating clusters projected into a particular space (Hoff, 2006). In the field of subspace clustering, crowding was the main strategy, followed by a series of calculations such as ENCLUS, MAFIA, and SUBCLU. This paper's focus is on projective clustering. We are going to concentrate on comparable procedures in the next runners.

2.2 Methods of Projective Clustering

Highlight weighting is the logical foundation of projective clustering. Each dimension in each cluster is assigned a weight that shows how essential that dimension is to the cluster. Obviously, the weight values for a specific dimension may differ amongst clusters. Based on how weights are determined, projective clustering calculations can be classified into two classes. Measurements of first order are assigned a weight of one value, resulting in a subtle inclusion weighting of the subspace (Lu et al., 2011). PROCLUS is based on the classic k-Medoids

DOI: https://doi.org/10.48001/JoITC.2023.1114-21

method and can potentially be used for weight calculations of conspiracy agents. PROCLUS tests the data, then selects a collection of medoids and repeatedly advances the clustering, with the goal of minimizing the normal outside cluster scattering. A set of measures is picked for each medoid whose normal separations from the medoid are small in comparison to factual desire.

Once the subspaces are identified, a conventional Manhattan segment spacing is employed to allocate foci to medoids. PROCLUS requires users to specify the standard number of material measures for each cluster, a task that may pose inherent ambiguity to users. FINDIT, which employs a distinct degree known as the Dimension-acquainted distinct (DOD), is structurally like PROCLUS. HARP, a colorful step-by-step clustering process, automatically determines the amount of material in each cluster without considering parameters defined by Stoner. HARP is based on the premise that if two information points are very similar in various ways, they are likely to belong to the same cluster (Bouguessa et al., 2006).

Croaker defines a subspace as a subset of features whose focal prominence is in a partition within a section. Croaker uses random calculations to calculate predictive clusters to minimize specific work quality. MINECLUS advances DOC by turning the difficulty of reaching expected clusters into the challenge of catching booby-trapped visitor sets. While PROCLUS and the other calculations mentioned above target the axis-aligned subspace of clusters, many other algorithms look for the more general axis-unaligned subspace. Here, the latest highlight is a direct combination of the original majors. ORCLUS may be a variant of PROCLUS that can discover clusters in subjectively ordered subspaces. ORCLUS chooses the eigenvectors of the set of foci by covariance network diagonalization by comparing them to the network's lowest eigenvalues.

A K-means sorted projection clustering computation, uses an SVD computation to determine subspaces whose axes are not aligned. On the other hand, EPCH uses histogram expansion to perform axes-unaligned projection clustering. Instead of finding a tricky subspace of clusters, the computation applies weights to run (0, 1) in immediate order. Because the weights may be any actual wide variety withinside the variety (0, 1), we are able to time those sensitive projective clustering computations. Naturally, the load attention for a measurement in a cluster corresponds to the scattering of values from the centre in the cluster`s measurement. In different words, an altitudinous weight famous a few scattering in a cluster measurement (Haralick & Harpaz, 2007). To all intents and purposes, all the computations on this order are primarily based totally on following not unusual place reservations. 1) the statistics extends alongside a vital measurement onto a narrower variety of values than on the opposite measures; 2) the statistics is much more likely to be constantly disseminated alongside every minor measurement.

We will explore the demonstration potential of projective clustering with respect to these two typical caveats. Several delicate projective clustering computations have recently been disclosed. An algorithm based on patch mass optimization is shown. Because a heuristic global look approach is applied, this computation may obtain near-optimal highlight weights; in any event, it will perform more gradationally than other computations. The k- means kind structure has been astronomically entered to produce an effective delicate projective clustering computation. Based on the traditional K-means clustering method, a redundant step of calculating weight values such as EWKM, FWKM, LAC, FSC, etc. is added to each cycle of these calculations. For these calculations, computation 1 appears to have a standard structure.

The absence of such a demonstration raises concerns regarding the development of more effective clustering algorithms. Consequently, we have embarked on exploring projected cluster modeling. We are motivated by the belief that this modeling process enables us to harness the full potential of cluster analysis. (Chen et al., 2010).

Point weighting is a common practice in projective clustering, where each dimension within a cluster is assigned a weighted value representing its relevance to the cluster. These weighting values can vary across clusters. Projective clustering techniques fall into two categories based on how these weights are determined: soft subspace clustering and clustering boundaries. In the initial phase, weights in the first order are discretized to either 0 or 1, establishing a binary representation of the subspace's hardpoint weighting. An illustration of this approach can be found in PROCLUS, an algorithm grounded in the classical k-Medoids technique. In PROCLUS, the procedure involves data sampling, medoid selection, and successive iterations aimed at enhancing clustering accuracy by minimizing the average dissimilarity within clusters.

After the subspaces have been linked, medoids are assigned points using an average Manhattan segmental distance. PROCLUS needs druggies to provide the average number of applicable confines each cluster, which is usually unknown to them. FINDIT is structurally comparable to PROCLUS since it employs a distance metric known as the Dimension-acquainted Distance (DOD). HARP is founded on the concept that if two data points are comparable across several boundaries, they are likely to belong to the same cluster. A subspace, according to Croaker, is a set of qualities in which the protuberance

DOI: https://doi.org/10.48001/JoITC.2023.1114-21

of points in a partition is restricted within a member. Croaker uses a randomized technique to construct projected clusters to minimize a specific quality function.

MINECLUS surpasses DOC by addressing the challenge of determining the projected clusters, resolving the analogous problem encountered in frequent item set booby-trapping. Unlike PROCLUS and the aforementioned methods, alternative approaches seek more versatile, non-axis-aligned subspaces. In these methods, new features emerge as direct combinations of the original constraints. PROCLUS version ORCLUS may search for clusters in arbitrarily familiar subspaces. PROCLUS' misdeeds were passed on to ORCLUS. KSM, a k-means type projective clustering technique, use SVD calculations to determine non-axis-aligned subspaces, whereas EPCH employs histogram construction to perform non-axis-aligned projective clustering. In contrast, the methodologies in the opposite sequence utilize weights ranging between 0 and 1, forsaking the establishment of rigid subspaces for clusters. Referred to as soft projective clustering, these approaches allow for weight values to take any real number between 0 and 1. In usual practice, the weight allotted to a dimension within a cluster correlates with the extent of dispersion of values in that dimension relative to the cluster center. In simpler terms, a dimension with a higher weight within a cluster signifies a lower level of dissipation.

Numerous soft projective clustering methods have recently been documented. One algorithm, rooted in flyspeck mass optimization, stands out for achieving nearly optimal point weights through a heuristic global search approach. However, it may operate more slowly compared to other algorithms. A prevalent approach to creating robust soft projective clustering techniques involves adopting a k-means-like structure. These algorithms, namely EWKM, FWKM, LAC, and FSC(5), extend the conventional k-means clustering procedure by integrating an extra step for computing weighting values. Algorithm 1 delineates the framework of these algorithms, illustrating their structure and incorporation of the weighting determination process.

Input: A dataset along with the desired number of clusters K;

Output: The resulting partition C and the corresponding weights W assigned to each cluster;

Find the first cluster V and set W to have equal v values;

1. Regroup the dataset into C based on V and W;
2. Recompute V based on C;
3. Recompute W based on C; Repeat until confluence is obtained.

The prevalent projective clustering algorithm follows an Expectation-Maximization (EM)-based procedure which serves as the basis for the data in Algorithm 1. However, in the approaches mentioned below, the foundational F(C, V, W) is often overlooked. The absence of such a model poses challenges in developing more efficient clustering algorithms.

This has motivated us to explore projected cluster modeling, as we believe this approach enables us to capitalize on the diverse opportunities within cluster analysis. This encompasses understanding the fundamental factors contributing to the cluster formation and tackling issues pertaining to cluster validity.

In a standard model-based clustering assessment, the objective is to identify a multivariate dissimilarity combination that accurately captures the nuances of the data. However, when dealing with high-dimensional data and the intricacies of projective clustering as discussed earlier, challenges may arise due to the curse of dimensionality. Hoff presented an illustration of "clustering shifts in cruelty and friction" by employing a nonparametric mixture of arrangements of freely chosen elements in one of the experiments, showcasing the application of model-based clustering in high-dimensional information clustering.

The show is taught using a Markov chain Monte Carlo handling; in any event, the computational risk is limited. A nonparametric consistency estimation modelling system in which the data is represented as a mix of direct manifolds. A Bayesian methodology is utilized for detecting sets of points that conform to or are situated within lower-dimensional linear structures. PCA (Principal Component Analysis) computes reduced-dimensional spaces associated with individual clusters. However, difficulties with this method revolve around its reliability in determining the dimensionality of these spaces and its computationally intensive clustering approach.

3. A PROBABILISTIC FRAMEWORK FOR PROJECTED CLUSTERING

The characteristics of anon-axis-aligned subspace are regular combinations of the original information space's measures. Since they are worrisome to decipher, regularly making the clustering comes about less precious for multitudinous genuine operations, similar as record clustering, as it were anticipated clusters in axis- aligned subspaces are homogenized within the taking after preface.

3.1 Basic Notation and Definitions

It is assumed that the dataset is normalized, ensuring each xij is within the range [0, 1] for j = 1, 2,...,D. The degree of

DOI: https://doi.org/10.48001/JoITC.2023.1114-21

membership of xi in the kth cluster ck, denoted as uki, is subject to the following criteria:

$$\boldsymbol{o \le u_{ki} \le 1; \sum_{k=1}^{k} u_{ki} = 1, i = 1, 2, \ldots\ldots, N}.(1)$$

$$\begin{cases} \sum_{j=1}^{D} w_{kj} = 1, & k = 1, 2, \ldots, K \\ 0 \le w_{kj} \le 1, & k = 1, 2, \ldots, K; j = 1, 2, \ldots, D. \end{cases} \quad (2)$$

The weight wkj is now specific to quantify the contribution of the jth dimension to ck. A higher weight indicates a more significant influence. Additionally, we introduce the D × D matrix sk, which is defined as:

$$s_k = \begin{bmatrix} \sqrt{w_{k1}} & & & \\ & \sqrt{w_{k2}} & & \\ & & \cdots & \\ & & & \sqrt{w_{kD}} \end{bmatrix}.$$

3.2 Probability Model

We will examine the distribution of each measurement to capture the fundamental structure of clusters in a high-dimensional environment. The likelihood density function can be expressed as:

$$G(y_j|\mu_{kj};\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(y_j - \mu_{kj})^2\right),$$

where kj and k represent the Gaussian mean and covariance. The preceding expression is transformed into this

$$G(x_j|v_{kj}, w_{kj};\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{w_{kj}}{2\sigma_k^2}(x_j - v_{kj})^2\right). \quad (5)$$

$$\int G(x_j|v_{kj}, w_{kj};\sigma_k)\mathrm{d}x_j = \frac{1}{\sqrt{w_{kj}}},$$

It is important to note that Gaussian mixing can be an important theory for information transfer, as shown by many model-based group calculations. In this case, clusters of information are assumed to originate from different imaginable sources, and the information from each source is modeled by a Gaussian method. In any case, Gaussian powers do not fit in high-dimensional space because of the shame of dimensionality.

The likelihood function is formulated relying on two underlying assumptions. Firstly, it assumes that the distribution of points along each measurement within the subspace is independent of others. Although this assumption might not always be valid across all applications, it can often be reasonable in probabilistic models, enabling an approximation of the joint distribution of uncorrelated factors by the product of their marginals. Secondly, it presumes that variations in points are mutually independent of each other. Since

At this point we assume N inputs x1, x2, . . . , xN spread independently and indistinguishable from the population after mixing thickness:

$$F(x;\Theta) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{D} \sqrt{w_{kj}} G(x_j|v_{kj}, w_{kj};\sigma_k)$$

with

$$\sum_{k=1}^{K} \alpha_k = 1, \alpha_k \ge 0, k = 1, 2, \ldots, K, \quad (6)$$

3.3 Clustering Criterion

The purpose of using the probability model for clustering is to approximate the given amount of data. If = (k, wk, wk, k)|1 and lt; k and lt; K) is an estimate, the distance between F(x,) and F(x,) can be calculated as follows:

SThe first, F(x;) ln F(x;)dx, is a meaningless constant; so, the following objective criterion must be applied at most:

$$\uparrow Q_1(\hat{\Theta}) = \int F(\mathbf{x};\Theta)\ln \hat{F}(\mathbf{x};\hat{\Theta})d\mathbf{x}$$
$$= \sum_{k=1}^{K}\int p(k|\mathbf{x})F(\mathbf{x};\Theta)\ln \hat{F}(\mathbf{x};\hat{\Theta})d\mathbf{x}$$

With

$$p(k|\mathbf{x}) = \frac{\hat{\alpha}_k \prod_{j=1}^{D}\sqrt{\hat{w}_{kj}}\,G(x_j|\hat{v}_{kj},\hat{w}_{kj};\hat{\sigma}_k)}{\hat{F}(\mathbf{x};\hat{\Theta})}, \quad 1 \le k \le K \quad (7)$$

$$\uparrow Q_1(\hat{\Theta}) = \sum_{k=1}^{K}\int p(k|\mathbf{x})F(\mathbf{x};\Theta)$$
$$\times \ln\frac{\hat{\alpha}_k \prod_{j=1}^{D}\sqrt{\hat{w}_{kj}}\,G(x_j|\hat{v}_{kj},\hat{w}_{kj};\hat{\sigma}_k)}{p(k|\mathbf{x})}d\mathbf{x}. \quad (8)$$

According to the law of large numbers, maximization (8) corresponds to learning the maximum likelihood of the dataset DB of all inputs x1, x2, . . . , xN.

$$\downarrow Q_2(\hat{\Theta}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}p(k|\mathbf{x}_i)$$
$$\times\left(\frac{1}{2}\sum_{j=1}^{D}\left(\frac{\hat{w}_{kj}}{\hat{\sigma}_k^2}(x_{ij}-\hat{v}_{kj})^2 - \ln\frac{\hat{w}_{kj}}{2\pi\hat{\sigma}_k^2}\right) - \ln\frac{\hat{\alpha}_k}{p(k|\mathbf{x}_i)}\right).$$

(9)

For an input x_i, the following probability $p(k|x_i)$ is thought of as the fuzzy involvement u_{ki} in clustering. Given that

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\frac{1}{2}u_{ki}\left(\sum_{j=1}^{D}\ln 2\pi\right)$$

$\frac{1}{N}$ and

$$\downarrow J(U,V,W,Z)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\frac{u_{ki}}{2}\sum_{j=1}^{D}\left(\frac{w_{kj}}{\sigma_k^2}(x_{ij}-v_{kj})^2 - \ln\frac{w_{kj}}{\sigma_k^2}\right) - u_{ki}\ln\frac{\alpha_k}{u_{ki}}\right)$$

(10)

4. MODEL-DRIVEN ALGORITHM DESIGNED FOR PROJECTIVE CLUSTERING

This section discusses our projection clustering calculation, MPC, which minimises (10) while meeting the restrictions of (1), (2), and (6), which may be a forced nonlinear optimisation problem. This can be transformed into an unconstrained optimisation problem procedure using Lagrangian multiplication

$$\min J_1(U,V,W,Z) = J(U,V,W,Z)$$
$$+\sum_{k=1}^{K}\lambda_k\left(\sum_{j=1}^{D}w_{kj}-1\right)+\xi\left(\sum_{k=1}^{K}\alpha_k-1\right)$$
$$+\sum_{i=1}^{N}\zeta_i\left(\sum_{k=1}^{K}u_{ki}-1\right),$$

(11)

4.1 The Optimization Method

To realize the neighborhood with the smallest objective work, a common strategy is to use a partial optimization of each parameter of the work. Following this strategy, J1 can be minimized in (11) by optimizing U, V, W, and Z in a sequential structure very similar to EM computer science. In each cycle, we start by setting V = , W = and Z = and decide U based on J1(U, , ,). The four fractional optimization problem can be understood by consensus after the hypotheses have been established.

4.2 MPC Algorithm

As discussed in computation 2, the MPC computation performs projection clustering by minimizing the objective function. This framework can be seen as an extension of the conventional FCM algorithm, featuring an additional step for each centroid to compute the weights (W) for each cluster. Such an approach is commonly utilized in prevalent sensitive subspace clustering algorithms. Initialization 1.1 Choose K cluster centres at random. V is denoted as V(0). 1.2 Label all weights from W to W as W (0); 1.3 Assign a nonzero constant to all values and label them Z(0). 2. Print with the U, V, and W heads It should be noted that MPC does not require client-specific parameters for focus weighting, although other existing projection accumulation methods do: in cases l in PROCLUS, FWKM, EWKM, and so on. A quasi-pending coefficient, such as a weight equation within MPC, can be determined numerically. Step 2.4 of Calculation 2 is specifically developed for this purpose. All factors, however, are given and can thus be considered constant in the ratio. Then we can solve utilizing numerical strategies.

5. PROJECTED CLUSTERING INCORPORATING OUTLIER ANALYSIS

The proposed framework, developed for high-dimensional clustering as discussed, incorporates subspace access into similarity studies Wang et al. (2008). It addresses irregular data values in a manner akin to the clustering process. Optimization is applied to streamline the preparation for subspace definition. This framework is explicitly tailored for grouping data with high-dimensional values. Irregularity analysis plays a pivotal role in enabling representation-based projection clustering. Furthermore, the framework has been enhanced with an organizational handling feature. It is structured into six main modules: data cleaning preparation, subspace definition, subspace layout, coupling with MPC, MPC with exceptions, and coupling with feature and specificity studies. The data cleaning module is specifically designed to rectify variance noise. A subspace selection module is introduced to select high-quality subsets. Property layout is done under the subspace layout module. Clustering is performed using the performance-based projection clustering method. Exception studies are coordinated with the MPC model. The Assets and Irregularity Survey is linked to the MPC's updated presentation.

6. CONCLUSION

Projective clustering techniques are employed for the clustering of high-dimensional data. Among these techniques, representation-based projection clustering stands out as a subspace clustering method. It specifically utilizes non-axial subspaces in similarity investigations. Irregular exchanges fall into expected clusters. The accuracy of the cluster in the frame is progressing. The featured mode option is optimized to handle unregulated features. Exceptional studies are given in the group handle. Cluster initialization proceeds with the preparation of subspace selection.

REFERENCES

Bouguessa, M., Wang, S., & Sun, H. (2006). An objective approach to cluster validation. *Pattern Recognition Letters*, *27*(13), 1419-1430. https://doi.org/10.1016/j.patrec.2006.01.015.

Chen, L., Jiang, Q., & Wang, S. (2008, December). A probability model for projective clustering on high dimensional data. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 755-760). IEEE. https://doi.org/10.1109/ICDM.2008.15.

Chen, L., Jiang, Q., & Wang, S. (2010). Model-based method for projective clustering. *IEEE Transactions on Knowledge and Data Engineering*, *24*(7), 1291-1305. https://doi.org/10.1109/TKDE.2010.256.

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*, *14*, 63-97. https://doi.org/10.1007/s10618-006-0060-8.

Gan, G., Wu, J., & Yang, Z. (2006). A fuzzy subspace algorithm for clustering high dimensional data. In *Advanced Data Mining and Applications: Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006 Proceedings 2* (pp. 271-278). Springer Berlin Heidelberg. https://doi.org/10.1007/11811305_30.

Haralick, R., & Harpaz, R. (2007). Linear manifold clustering in high dimensional spaces by stochastic search. *Pattern Recognition*, *40*(10), 2672-2684. https://doi.org/10.1016/j.patcog.2007.01.020.

Hoff, P. D. (2006). Model-based subspace clustering. *Bayesian Analysis*, *1*(2), 321-344. https://doi.org/10.1214/06-BA111.

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*(8), 1026-1041. https://doi.org/10.1109/TKDE.2007.1048.

Lu, Y., Wang, S., Li, S., & Zhou, C. (2011). Particle swarm optimizer for variable weighting in clustering high-dimensional data. *Machine Learning*, *82*, 43-70. https://doi.org/10.1007/s10994-009-5154-2.

DOI: https://doi.org/10.48001/JoITC.2023.1114-21

Moise, G., Sander, J., & Ester, M. (2008). Robust projected clustering. *Knowledge and Information Systems*, *14*, 273-298. https://doi.org/10.1007/s10115-007-0090-6.

Wang, Q., Ye, Y., & Huang, J. Z. (2008, July). Fuzzy k-means with variable weighting in high dimensional data analysis. In *2008 The Ninth International Conference on Web-Age Information Management* (pp. 365-372). IEEE. https://doi.org/10.1109/WAIM.2008.50.

.

DOI: https://doi.org/10.48001/JoITC.2023.1114-21