Improving Cluster Efficiency by Density based Approach using Hybrid Optics

J. S. Kanchana[1] G. Aayeesha Siddiqa Hussainibi[2] T. R. Amirtha Nandhini[3] P. A. Archanamai[4]
[1]Associate Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]K.L.N. College of Engineering

Abstract— Clustering is the one of the important methods used in data mining, which has wide range of applications in Image processing, Pattern recognition and Compression. The clustering algorithms are categorized into five methods, such as Partitioning, Hierarchy, Grid-Based, Density Based and Model Based. A new approach based on the Hybrid OPTICS is proposed, which is the Density based method used to cluster the data with high dimension and variable densities. OPTICS is an Ordering Points To Identify the Clustering Structure used to find the density based clusters. It addresses the one of the DBSCAN problem of detecting meaningful clusters in the data of varying densities. OPTICS requires two parameters, ε is used to describes the radius and MinPts is used to describe the number of points require to form the cluster. A point p is a core point if at least MinPts points are found with Epsilon neighbourhood N ε (p). Using this parameters Hybrid OPTICS computes core distance and reach ability distance to find local and global minimum. Hybrid OPTICS automatically determines the border set based on global and local minimum and then performs cluster analysis for each local cluster based on local minimum and integrate the all local minimum to obtain the global minimum. The efficiency of the Hybrid OPTICS is validated using real data set when compared with existing SDE algorithm.

Key words: Hybrid Optics, OGFS, SDE framework

I. INTRODUCTION

Data Mining is the process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis. Clustering is the process of making a group of abstract objects into classes of similar objects.

The primary goal of clustering algorithm is to group the objects of a database into a set of meaningful subclasses that can used either as a stand-alone tool to get insight into the distribution of a dataset.

There are three interconnected reasons why the effectivity of clustering algorithms is a problem. First, almost all clustering algorithm require values for input parameters which are hard to determine, especially for real-world datasets containing high dimensional objects. Second, the algorithm are very sensible to these parameter values, often producing very different partitioning of the dataset even for slightly different parameter setting. Third, high-dimensional real-datasets often have a very skewed distribution that cannot be revealed by a clustering algorithm using only one global parameter setting.

The purpose of introducing hybrid algorithm is to overcome the cluster analysis which does not produce a clustering of a dataset explicitly; but instead creates an augmented ordering of the database representing its density-based clustering structure.

II. RELATED WORK

This section reviews sampling techniques and feature selection methods.

A. Statistical Optimal Sample Size

Sampling in data mining is a precondition method to preselect samples of high quality from the original data. Increasing sample size generally leads to a higher accuracy; however, there is a trade-off between accuracy and efficiency of algorithms when the accuracy improvement becomes saturated at large sample sizes, as shown in Fig. 1. The optimal sample size s is determined based on the concept that the sample quality will saturate when the sample size is increased beyond a certain threshold. The authors in proposed a method, called Statistical Optimal Sample Size (SOSS), which can determine the sample size on large data sets by measuring the sample quality based on the data distribution. In researchers evaluated their measure on four large UCI KDD datasets. They found that the resulting tree sizes with SOSS are significantly smaller than those with the full size, and the accuracy rate using SOSS is higher.

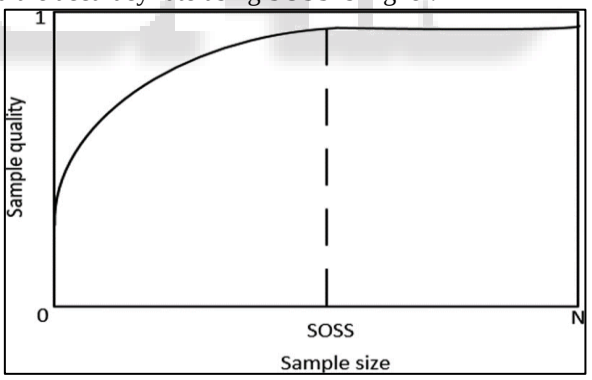

Fig. 1: Learning curve between sample size and sample quality.

B. Feature Selection

As one important pre-processing step in data clustering, feature selection is a process of choosing a representative and effective subset from original features in the high-dimensional data space according to the specified evaluation criterion, such that the preserved feature subset is most useful in capturing the intrinsic properties. Feature selection methods can be divided into three groups: filter approaches, wrapper approaches, and embedded approaches. The filter approaches generate relevance scores on features based on the intrinsic properties of the dataset, and generally high scoring features are chosen as the input to the clustering algorithm. They have low computational cost, but ignore feature dependencies. The wrapper approaches select features

with a performance measure from a predetermined learning model and take into account the feature dependencies, while their computational cost is high. The embedded approaches incorporate feature search and learning model, which is why they are faster than the wrapper approaches but slower than the filter methods. For the filter approaches, there are three major unsupervised methods, variance score laplacian score and sparsity score. The variance score is a simple unsupervised method that selects features with high variance. The variance for the r^{th} feature is obtained by the formula

$$V_r = \frac{1}{N}\sum_{i=1}^{N}(x_{ir} - \mu_r)^2, \tag{1.1}$$

where μr is the mean value of the r^{th} feature vector. The laplacian score is proposed to select features by comparing their local preserving abilities without considering the redundant features. The Laplacian score for the r^{th} feature is given as follows:

$$L_r = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_{ir} - x_{jr})^2 S_{ij}}{\sum_{i=1}^{N}(x_{ir} - \mu_r)^2 \sum_{j=1}^{N} S_{ij}}, \tag{1.2}$$

where μr is the mean value of the r^{th} feature vector and S_{ij} is the similarity measure between the neighbour nodes x_i and x_j. The sparsity score is another unsupervised feature selection algorithm without utilizing the labelled data. In this paper, we choose the filter approaches, because it can search through the feature space efficiently and perform well in the experiments. Meanwhile, we take into account the feature dependencies by introducing information entropy. Their combination with information entropy is a better choice. The sparsity score is proposed as an l_1 graph-preserving feature selection method where features are ranked according to their respective sparsity preserving abilities. This method is based on sparse representation, which aims to acquire and represent primary messages of signals with the least non-zero coefficients in signal processing. This algorithm calculates the corresponding sparsity score for overall d dimensions of features and ranks the features to select the smallest(m <d) ones. A sparse representation reconstructive coefficient vector s_i for each x_i in R^dcan be obtained using the optimization method of l_1-norm minimization

$$\min_{s_i} \|s_i\|_1 \quad \text{s.t. } x_i = X^* s_i, 1 = 1^T s_i, \tag{1.3}$$

is equal to X after removing x_i, and $s_i = [s_{i1}, s_{i2}, \ldots, s_{in}]^T$ is an n-dimensional vector. To overcome noise and small sample size problems, two modified object functions are proposed as follows:

$$\min_{s_i} \|s_i\|_1 \quad \text{s.t. } \|x_i - X^* s_i\| < \theta, 1 = 1^T s_i,$$

$$\min_{[s_i^T t_i^T]^T} \|s_i^T t_i^T\|_1 \quad \text{s.t.} \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} X & 1 \\ 1^T & 0^T \end{bmatrix} \begin{bmatrix} s_i \\ t_i \end{bmatrix}, \tag{1.4}$$

where u is the error tolerance, and ti is a d-dimensional vector as compensation and reconstruction.

The Fisher method is an effective supervised filter method by calculating a score for each feature based on the inter-cluster separation and the intra-cluster variance. Feature selection and kernel learning for local learning-based clustering (LLCFS) considers the relevance of each feature based on the combination of weighted features and local learning-based clustering. In addition, the problem of feature selection can also be improved from the other viewpoint, such as structure, graph, spectral clustering, etc. Feature selection via eigenvector centrality (ECFS) proposed a graph-based method, where each node is a feature. It assesses the importance of node based on the eigenvector centrality. Infinite latent Feature selection (ILFS) is a ranking approach based on a probabilistic latent graph, which takes into account all the possible subsets of features and finds the relevance of each feature. The authors in proposed a method with graph regularized data reconstruction preserving the original data structure and approximately reconstructing each data point. Clustering-guided sparse structural learning (CGSSL) is proposed with a joint framework including cluster analysis and sparse structural analysis. Spectral clustering methods can be used to perform reduction in dimension with the spectrum of the similarity matrix of dataset. In, the authors joined the correlation among feature stream into the online feature selection process, called online group feature selection method (OGFS). The effectiveness is demonstrated in the application of image classification and face verification.

III. PROPOSED SYSTEM

In Hybrid OPTICS Algorithm, the density based method used for clustering the data with variable densities. OPTICS is used to find the density based clusters. It uses two parameters ε and MinPts. Using this parameters, the algorithm computes the core distance to find the local and global minimum. It performs the cluster analysis and integrate all local and global minimum. Four modules are analysed here

- Density Entropy Computation.
- Distance Metrics Calculation.
- Hybrid OPTICS Algorithm.
- Performance Analysis.

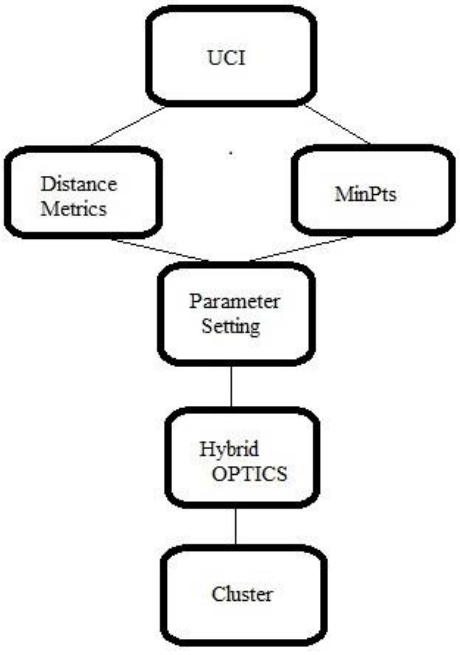

Fig. 2: Clustering using Hybrid OPTICS

A. Density Entropy Computation

Entropy a decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). Entropy is defined as a turning toward or transformation In data mining, the term entropy is used in decision tree which builds top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. It is used to create some work or energy to be pumped in order for the process to occur.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (2.1)$$

B. Distance Metric Calculation

Similarity between two objects is calculated using a distance measure. Since, clustering forms groups; it can be used as a pre-processing step for methods like classifications. Distance metric if it satisfies the properties: Non- negativity, coincidence, symmetry and sub additivity. The Euclidean distance (or) Euclidean metric is the ordinary straight-line distance between two points in Euclidean space.

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i-1}^{n} (q_i - p_i)^2}. \quad (2.2)$$

C. Hybrid OPTICS Algorithm

It takes two parameters ε and MinPts. The OPTICS algorithm takes maximum distance to find the core sample of high density and expands the cluster. Using this distance it forms cluster. It keeps cluster hierarchy for a variable neighbourhood radius.

```
OPTICS (SetOfObjects, ent,MinPts, OrderedFile)
OrderedFile.open();
FOR i FROM 1 TO SetOfObjects.size DO
Object := SetOfObjects.get(i);
IF NOT Object.Processed THEN
ExpandClusterOrder(SetOfObjects, Object, ent,
MinPts, OrderedFile)
OrderedFile.close();
END; // OPTICS
ExpandClusterOrder(SetOfObjects,Object,ent,MinPt,
OrderedFile);
neighbors := SetOfObjects.neighbors(Object, ent);
Object.Processed := TRUE;
Object.reachability_distance := UNDEFINED;
Object.setCoreDistance(neighbors, ent, MinPts);
OrderedFile.write(Object);
IF Object.core_distance<> UNDEFINED THEN
OrderSeeds.update(neighbors, Object);
WHILE NOT OrderSeeds.empty() DO
currentObject := OrderSeeds.next();
neighbors:=SetOfObjects.neighbors(currentObject, ent);
currentObject.Processed := TRUE;
currentObject.setCoreDistance(neighbors,ent, MinPts);
OrderedFile.write(currentObject);
IFcurrentObject.core_distance<>UNDEFINED THEN
OrderSeeds.update(neighbors, currentObject);
END; // ExpandClusterOrder
```

D. Performance Analysis

The performance of the Hybrid OPTICS algorithm is compared with the existing SDE framework. In which the efficiency is enhanced in terms of clustering speed is analysed below.

No. of. Clusters	SDEAlgorithm (Time in Sec)	Hybrid OPTICS Algorithm (Time in Sec)
Cluster 1	0.1	0.06
Cluster 2	0.12	0.07
Cluster 3	0.13	0.06
Cluster 4	0.11	0.07
Cluster 5	0.12	0.06

Table 1: Performance Analysis

Table I shows the performance comparison of existing SDE framework with the Hybrid OPTICS Algorithm.

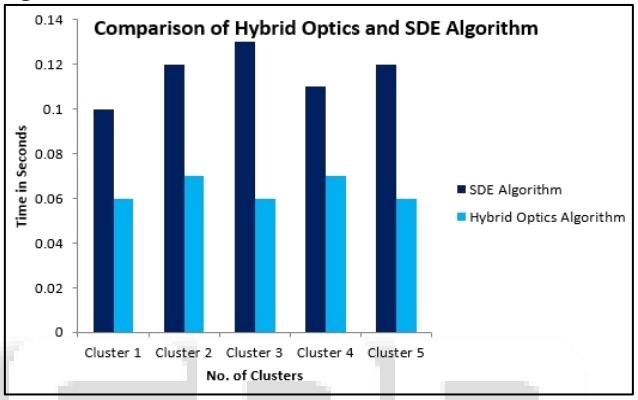

Fig. 3: Performance Analysis

IV. EXPERIMENTAL RESULTS

To illustrate the capability of the Hybrid OPTICS algorithm to handle data with complex distribution, a plenty of experiments can be carried out on the real-time datasets. The performance of the Hybrid OPTICS algorithm is compared with existing SDE framework which takes account of the local minimum alone. But proposed Hybrid OPTICS algorithm takes account of both local and global minimum and separates sparse data from dataset and clusters the dense data. The UCI repository's dataset is used to for the implementation of the Hybrid OPTICS algorithm.

For the parameter settings, the Epsilon value and MinPts are calculated for distance metrics computation and define the size of the cluster. These values are passed into theHybrid OPTICS algorithm which clusters the dense data and separates the sparse data.

In addition, the performance of Hybrid OPTICS with different dataset sizes in UCI repository is compared with existing SDE.

Clusters	Entropy	Distance Metrics
Cluster 1	14.97745422	20.18441183
Cluster 2	12.95315684	17.17905205
Cluster 3	22.04819836	25.1875356
Cluster 4	20.13788295	25.05931886
Cluster 5	30.64671306	35.3282881

Table 2: Clustering Range

Table II describe the clustering range used in the Hybrid OPTICS Algorithm.

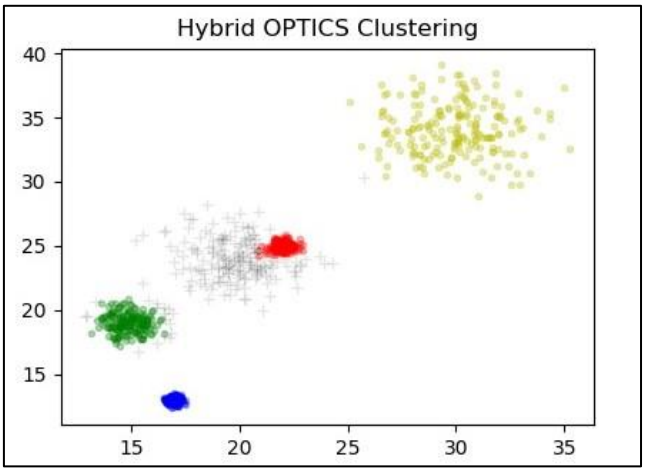

Fig. 4: Clustering using Hybrid OPTICS

Fig. 4. Illustrates the experimental results of implementing the Hybrid OPTICS Algorithm used for the UCI repository dataset.

V. CONCLUSION

The new framework based on Hybrid OPTICS is proposed for density based clustering. The OPTICS algorithm is likely DBSCAN but it cannot handle the global noises. But Hybrid OPTICS detects both local and global noises, the noises have great influence on the data distribution. In Hybrid OPTICS, two parameters are used to compute the core distance to find the local and global minimum. The effectiveness of the Hybrid OPTICS algorithm has been tested on some real data sets such as UCI. Hybrid OPTICS algorithm performs better in terms of speed and obtaining global minimum, when compared with existing SDE.

VI. FUTURE ENHANCEMENT

Picking an accurate threshold to select the features on different dataset is a major challenge for existing algorithm. In future, new Hybrid framework based on evolutionary algorithm will be proposed to pick up the accurate threshold for feature selection.

REFERENCES

[1] Sheng Li, Lusi Li, Jun Yan and Haibo He "SDE: A Novel Clustering Framework Based on Sparsity-Density Entropy", IEEE Trans. Knowl. Data Eng, vol. 30, NO. 8, Aug. 2018

[2] J. Wang, et al., "Online feature selection with group structure analysis", IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 3029–304, Nov. 2015.

[3] M. Alkasassbeh, G. A. Altarawneh, and A. Hassanat, "On enhancing the performance of nearest neighbor classifiers using hassanat distance metric", Canadian J. Pure Appl. Sci., vol. 9, no. 1.pp. 3291–3298, 2015.

[4] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang, "Graph regularized feature selection with data reconstruction", IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 689–700, Mar.2016.

[5] B. Tang and H. He, "A local density-based approach for outlier detection", Neurocomput., vol. 241, pp. 171–180, 2017.

[6] Ester M., Kriegel H.-P., Sander J., Xu X.: "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pp. 226-231.

[7] N. Passalis and A. Tefas, "Entropy optimized feature-based bagof-words representation for information retrieval", IEEE Trans. Knowl. Data Eng., vol. 28, no. 7, pp. 1664–1677, Jul. 2016.

[8] M. Liu and D. Zhang, "Sparsity score: A novel graph-preserving feature selection metho", Int. J. Pattern Recognit. Artif. Intell., vol. 28, no. 4, p. 1450009, 2014.

[9] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high dimensional data", IEEE Trans Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[10] Kaufman L., Rousseeuw P. J.: "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons,1990.

