

COVER BASED REFINED ROUGH K-MEANS ALGORITHM

Kaustubh Devdatta Deshpande* and Venkatesan M.**

Abstract: Today large majority of researchers have focused on the research on accurate pattern mining methods. Clustering techniques are most preferable methods for this pattern mining even in case of huge and growing data. Most of the real-time data applicable to pattern mining is of non-crisp nature. This means data can't be clearly separable into disjoint classes. Thus rough sets are most preferable for representation of data where data objects can be part of multiple clusters. With rough clusters, every cluster is represented by upper and lower approximation. The data objects belong to the upper approximation may or may not belong to clusters and data objects which belong to lower approximation definitely belong to clusters. Refined rough K-Means algorithm was proposed by Pawlak in-order to improve the rough K-Means clustering algorithm proposed by Lingras. We have proposed an extension of refined rough K-Means algorithm which is based on covers. The aim of algorithm is to provide better clustering.

Keywords: Clustering, Rough sets, Covers, Rough K-Means Clustering Algorithm

1 INTRODUCTION

Today data that computer systems process can be in any format such as text, numbers, images, audio, video, facts etc. The huge amount of data is getting accumulated from different organizations, which keeps on growing as time passes. This data includes can be operational or transactional data, non-operational data, metadata. Thus information we retrieve from this data is not but relationship or association or patterns among all these data. Information is transformed into knowledge which is used for benefit of organization. For this purpose data mining techniques are used. Data mining or knowledge discovery, is a computer based procedure of burrowing through and examining huge data and afterward obtaining information from data. Data mining tools are used to drive this process in order to discover different useful as well as meaningful patterns and rules. Data mining involves several steps [1] such as access as well as preparation of data, data mining algorithm to drive this process and data analysis. Data mining consists of 6 important tasks [1] such as Clustering process, Anomaly detection, Association rule learning, Classification process, Regression process, Summarization process. Here in this paper we are going to focus only on Clustering process. Clustering process [1] is referred as method of grouping set of data objects into different groups. It is used when there is no class information available for any data object in advance which is the most common in real-time situations. There are many Clustering methods available such as Partition Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method, and Constraint-based Method. We are going to discuss the algorithm which is extension of Partitioned based method. The Classical K-Means Clustering algorithm is the Partition Based algorithm which forms a base layer for the algorithm which is proposed in this paper. The Classical K-Means Clustering algorithm [2] is the most popular algorithm and has huge no of real-time applications. Some of the major advantages of K-Means are simplicity and computational as well as memory efficiency. K-Means clustering algorithm is widely applicable as well as most efficient on large data sets [23]. K-Means is a basically an approach of partitioned based clustering technique [24]. In partitioned based clustering partitions are subset of original

* **Mail Address:** 686/5A, "Atharva" Bunglow, Deepali Nagar, Nasik, Maharashtra. **Email:** Id:kaustubh.deshpande5@gmail.com

** Professor Vit University Vellore Tamilnadu India.

data set and represents clusters. So data set is divided into partitions which never overlap. Thus there exists no element common in any two or more partitions. Many researchers have extended this Classical K-Means in different directions and they have come with different results.

In real-time scenarios, we have to deal with imperfect knowledge. The imperfect knowledge is a big problem as well as now a days it is a vital issue for researchers in the field of Artificial Intelligence. For purpose of understanding as well as manipulating imperfect knowledge, there exists many ways. The fuzzy set is one of these ways. In 1965 this concept was introduced by Lotfi A. Zadeh and Dieter Klaua as an extended classical set theory. In classical set theory the membership of every element belonging to any set is measured in binary values where as in fuzzy set theory it is measured in real unit interval $[0, 1]$. It is considered as most successful theory. Later Pawlak proposed a new way towards imperfect knowledge, known as Rough Sets [3] based on equivalence relation. Now a days to handle indiscernibility of objects, the Rough Set Concept is considered as an excellent way. Among different extensions of Classical Rough Sets, the most popular extension is Covering Based Rough Sets. In 2006, W Zhu and F.Y.Wang ([4],[5],[6]) had introduced four different kinds of covering based rough sets with different properties analysed.

In 2004, Lingras[4] proposed rough K-Means algorithm based on classic K-Means as well as Rough set notion. In contrast to classical K-Means where clusters are non-overlapping, Rough K-Means algorithm works with overlapping clusters. In 2006 George Peters proposed refinements in Lingras Rough K-Means Algorithm. This refined algorithm is known as Refined Rough K-Means algorithm [5], aimed to tackle with limitations of Rough K-Means algorithm. We have proposed an extension of Refined Rough K-Means algorithm in this paper.

2. BACKGROUND

2.1 Rough Sets

The Classical K-Means algorithm gives non-overlapping or crisp clusters, but in many scenarios we require overlapping clusters. The rough sets are used to get overlapping clusters. The rough set theory is an extension of classical set theory. In fuzzy set theory, membership degree decides whether objects are similar or distinct, but in rough set theory objects are either definite or possible members of any cluster [9]. In rough sets, the membership [10] of any object is decided by other objects, but it never happens in case of fuzzy sets. In fuzzy sets, main problem is to interpret the membership degree of objects [10]. The Rough Clustering deals with some properties of rough set. But we require upper as well as lower approximations which follow some important rough set properties [9] as given below:

1. An object w can be contained by at most one lower approximation which ensures that any two lower approximations never contain common objects.
2. Every object w which is part of lower approximation of any set is also part of upper approximation of that set.
3. If any object w is not contained in any lower approximation then it must be part of at least two upper approximations.

2.2 Covering Based Rough Sets

Covering Based rough sets are based on covers and given as (U, C) , where U is universe and C is covering set. The rough sets differs from covering based rough sets, where rough sets use partitions or equivalence classes instead of covers. The sets in the partition used by the rough sets are disjoint in nature where as it is not necessary to be always true for covering. In case of covering the data items can be part of multiple

subsets. The covering based subsets generates upper and lower approximations by applying covering which exposes more rough, thus covering based rough sets are considered more powerful than rough sets [22].

Let V be universal data set, C be the covering for V , Y be a feature vector and $Md(y)$ is the minimal description of data object y , then we can define different covering types as below:

2.2.1 First Type [12]:

$$CLA(Y) = \cup \{L \in C \mid L \text{ is subset of } Y\} \quad (1)$$

$$FH(Y) = CLA(Y) \cup \{Md(y) \mid y \in Y - CLA(Y)\} \quad (2)$$

2.2.2 Second Type [12]:

$$CLA(Y) = \cup \{L \in C \mid L \text{ is subset of } Y\} \quad (3)$$

$$SH(Y) = \{L \mid L \in C, L \cap Y \neq \emptyset\} \quad (4)$$

2.2.3 Refined Second Type [13]:

$$CLA(Y) = \{y \in V \mid (\cap Md(Y)) \in Y\} \quad (5)$$

$$CUA(Y) = \{y \in V \mid (\cap Md(Y)) \cap Y \neq \emptyset\} \quad (6)$$

2.2.4 Third Type [13]:

$$CLA(Y) = \cup \{L \in C \mid L \text{ is subset of } Y\} \quad (7)$$

$$TH(Y) = \cup \{Md(y) \mid y \in Y\} \quad (8)$$

3. LITERATURE SURVEY

In 1967, James Mac Queen introduced the concept of “K-means” [14] where as in 1957 Stuart Lyod proposed the basic K-Means algorithm. There are many benefits [13] of K-Means algorithm including flexibility, simplicity, easy to understand and implement, whereas the selection of initial centroid affects the performance of algorithm.

In 2004, Sanjiv Bhatia presented Adaptive K-Means Clustering Algorithm [15], which is based on cluster rearrangement in order to get better partition when each time new element comes into picture. In 2010, Madhu and Shrinivas proposed Enhancing K-Means clustering algorithm with improved initial center algorithm [16] which is a new method based on exploring better initial centroids as well as time efficient way of allocating data points to clusters. Azhar Rauf and Sheeba, proposed in 2012 an Enhanced K-Mean Clustering Algorithm [19] which is based on calculation of initial centroid in order to reduce number of iterations and time complexity. In 2013, Chunfei Zhang and Zhiyi Fang introduced an Improved K-Means clustering algorithm [18] based on improved initial focal point and K value. In order to scale with big data sets The Esteves, Hacker and Rong introduced Competitive K-Means clustering algorithm [24] which has improved cluster analysis accuracy as well as has decreased the variance. In 2014, Tengfei Zhang¹, Long Chen and Fumin Ma come with an algorithm based on rough k-means clustering with variable weighted distance measure [21] between data objects and cluster centres in order to improve performance. Bangoria Bhoomi M. proposed an Enhanced K-Means clustering algorithm [20] based on the changes in existing K-Means algorithm in order to decrease time complexity for numeric data. In 2016, Tengfei Zhang & Fumin Ma has introduced an improved rough k-means clustering algorithm which has used weighted distance measure with Gaussian function. Please note that, if your email address is given in your paper, it will also be included in the meta data of the online version.

In this paper we have demonstrated that, how covers can improve the performance of the refined rough K-Means algorithm, so that we can get better results than the original algorithm.

4. ARCHITECTURE

This paper demonstrates the proposed strategy with proposed building design as appeared beneath as a model which comprises of six steps:

Step 1: Input dataset is taken for the processing. This input dataset must be numerical i.e. non-categorical data. If categorical attribute is present in input dataset then the corresponding attribute values can be transformed into numerical representations.

Step 2: In this stage, data pre-processing is performed on input data. This pre-processing includes data cleaning, binning strategies, normalization and transformation. This pre-processing makes our input data into intended format needed for processing.

Step 3: This pre-processed data is given as an input to feature extraction stage. At the point when the information is too huge to ever be processed and it is suspected to be repetitive, then it can be converted into a reduced feature set. The feature extractions techniques includes PCA, kernel PCA, Non-linear dimensionality reduction etc.

Step 4: Finally we apply Covering Based Refined Rough K-means algorithm over this input data or extracted features. Depends on the input parameter algorithm divides data into overlapping sets called “clusters”.

Step 5: The final output of algorithm is clusters which represents knowledge.

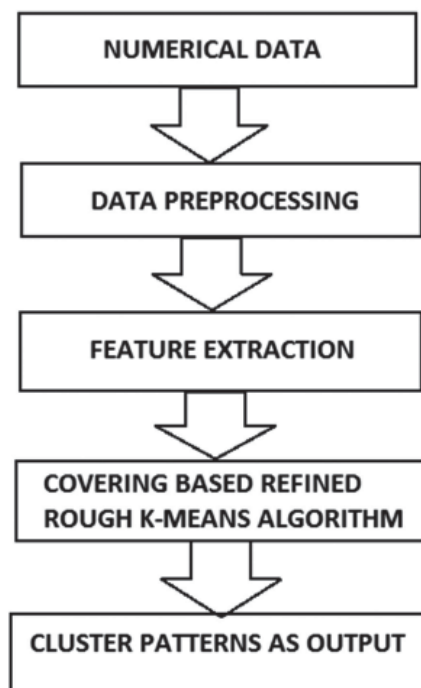


Figure 1. Five staged Architecture of Covering Based Refined K-means Algorithm.

5. THE REFERENCES SECTION

The proposed algorithm is based on refinements [8] suggested by George Peters in Lingra’s Rough K-Means clustering algorithms as well as rough covering.

In classical set theory, elements of the set uniquely determines set itself. Hence it follows crisp notion. Ex: set of odd numbers. In real-time scenarios, we generally come across vague notions, which means we can't partition objects into two classes. Ex: Classifying paintings on basis of beauty. Thus Rough Set Theory was proposed. If we want to classify elements as set X w.r.t. any indiscernibility relation R then we will get:

1. Lower Approximation of X w.r.t. R
2. Upper Approximation of X w.r.t. R
3. Boundary Region of set X .

Pawlak [8] has defined several properties for rough sets. Some of them applies to clustering.

1. Any data element can belong to at most one lower approximation.
2. If any data element belongs to a certain lower approximation then it also belongs to corresponding upper approximation also.
3. If data element is not part of any lower approximation then it is part of two or more upper approximations.

Following is the proposed algorithm for covering based refined rough K-Means:

1. Let W be the universal data set. Initialize W with data objects. You should therefore check your references thoroughly when you receive the final pdf of your paper. Let V be the covering set i.e. nonempty subset of W . Initialize the covering set V . Here $\cup v_i = W$ where $v_i \in V$.
2. Let M be number of clusters, c_i be i^{th} cluster and Y_i be subset of W where Y_i corresponds to cluster c_i . Initialize each Y_i where $1 \leq i \leq M$.
3. For $y \in W$, calculate minimum description for y as below:

$$Md(y) = \left\{ L \in V \mid y \in L \wedge \left(\begin{array}{c} \forall S \in V \wedge y \in S \wedge S \subseteq L \\ \Rightarrow \\ S = L \end{array} \right) \right\} \quad (9)$$

Let CLA_i be covering lower approximation and CUA_i be covering upper approximation for cluster C_i . Calculate CLA_i and CUA_i for each cluster C_i . There exists four different types of rough clustering which are first type, second type, third type and enhance second type clustering. Our observation suggests to employ refined second type or third type covering along with refined rough K-Means clustering algorithm. So here we can calculate CLA_i and CLU_i using refined second covering type as given below:

$$CLA(Y) = \{y \in V \mid (\cap Md(y)) \in Y\} \quad (10)$$

$$CUA(Y) = \{y \in V \mid (\cap Md(y)) \in Y\} \cap Y \neq \emptyset \quad (11)$$

Let m_i be mean for cluster C_i . Calculate new means using following equation:

$$\overline{m}_k = QL * \sum_{y_n \in CLA_k} \frac{\overline{y}_n}{|CLA_k|} + Qu * \sum_{y_n \in CUA_k} \frac{\overline{y}_n}{|CUA_k|} \quad (12)$$

Where,

$$QL + Qu = 1 \quad (13)$$

4. For every cluster C_i assign a data element to it which best represents that cluster.

(a) Assign this data element to lower and upper approximation of this cluster.

(i) Evaluate the least distance between every cluster C_i to every data element $\overline{y_L}$. Assign data element $\overline{y_L}$ to lower as well as upper approximation of the cluster C_H .

$$\text{dist}(\overline{y_L}, \overline{m_H}) = \min_{n,k} \text{dist}(\overline{y_L}, \overline{m_H}) \quad (14)$$

Where $\overline{y_L} \in \overline{CLA_k}$ and $\overline{y_L} \in \overline{CUA_k}$

(ii) Ignore $\overline{y_L}$ and $\overline{m_H}$. Repeat step[I] for those clusters left for processing in step[I]. If there exists no clusters to be processed in step [I] then go to step (b).

(b) Identify the nearest mean $\overline{m_H}$ for each of data point $\overline{y'_m}$ which is not selected in step (a).

$$d_{m,h}^{\min} = \text{dist}(\overline{y'_m}, \overline{m_H}) = \min_{k=1 \dots K} \text{dist}(\overline{y'_m}, \overline{m_H}) \quad (15)$$

Allocate $\overline{y'_m}$ to $\overline{CUA_H}$.

(c) Find out the means $\overline{m_i}$ which are too nearer to $\overline{y'_m}$.

$$F' = \left\{ t : \frac{\text{dist}(\overline{y'_m}, \overline{m_k})}{\text{dist}(\overline{y'_m}, \overline{m_H})} \leq \delta \text{ and } h \neq k \right\} \quad (16)$$

, where δ is a given relative threshold.

If $F' \neq \emptyset$ then allocate $\overline{y'_m}$ to \overline{CUA} , else allocate $\overline{y'_m}$ to $\overline{CLA_i}$.

5. Check convergence of algorithm. If algorithm converges then stop else go to step [6].

6. RESULTS

To endorse the effectiveness of the covering based refined rough K-Means clustering algorithm, this paper uses a Total 15-64 labor to population (%) data set, Employment Time Series data set and Household Income Disparity data set for experiment. The total 65+ labor to population data describes the age group above 65, percentage of all labor to total population. The Household Income Disparity represents different measures of disparity for 1990 and 2000. The variables for 1990 are based on data values of 1990, but aligned to the boundaries of U.S. Counties in 2000. Employment Time Series dataset represents key labor force measured for male and female employees with respect to several different measures. All attributes of each of the datasets are numerical in nature. The dataset name, number of attributes and number of data objects are given below in table 1, whereas the selection of initial centroid affects the performance of algorithm.

Table 1. Dataset Information

Heading level	Example	Font size and style
Total 65+ labor to population (%)	29	189
Employee Time Series	24	90
Household Income Disparity	16	3110

We have measured the performance of our clustering algorithm by a well-known performance metric called Davies Bound Index. The Davies Bound index is the one of the most popular metrics to measure performance of clustering algorithms. Davies Bound Index represents the mean of resemblance between every cluster to respective most identical cluster. The Optimal clustering techniques tries to minimize Davies Bound index value. Suppose C_j represents cluster of data objects, Z_j represents size of the cluster 'j' i.e. no of elements that forms cluster 'j', x_i denotes the data object 'i' with N dimensions and E_j represents center of cluster 'j'. Now we can calculate the value S_j denoting scatteredness of the data objects inside the cluster j as below:

$$S_j = \frac{1}{Z_j} \sum_{i=1}^{z_j} \|x_i - E_j\|_q \quad (17)$$

We have considered the value for 'q' as 2 in order to evaluate Euclidian distance. Let us consider value of $H_{a,b}$ indicates how cluster C_a and cluster C_b are separated which should be as large as possible and it can be expressed as below:

$$H_{a,b} = \|x_a - E_b\|_q \quad (18)$$

$$H_{a,b} = \sum_{i=1}^n \left(|e_{i,a} - e_{i,b}|^q \right)^{\frac{1}{q}} \quad (19)$$

Here data element $e_{i,a}$ indicates i^{th} element of cluster C_a . Suppose $R_{a,b}$ be a value indicates wellness of the clustering technique used. The value of $R_{a,b}$ can be calculated as below:

$$R_{a,b} = \frac{S_a + S_b}{H_{a,b}} \quad (20)$$

The value of $R_{a,b}$ defines a new value D_a as below:

$$C_a = \max_{a \neq b} R_{a,b} \quad (21)$$

Consider we have F no of clusters with our dataset. With help of this, we can define the Davis Bound Index value as below:

$$DB = \frac{1}{F} \sum_{k=1}^F C_k \quad (22)$$

The value of DB must be as small as possible. We have tested refined rough K-Means clustering with all four covering types and came with several observations. The experiments we performed has shown that the application of refined second type covering as well as application of third type covering when applied with refined rough K-Means have provided lower Davies Bound index values as compared to the Davies Bound index values obtained from original refined rough K-Means algorithm. We have tested each of above data sets with different cover sets as well as different values of Q_L , δ , and Q_u with each of covering techniques. We found that most of the times refined rough K-Means with refined second type covering or third type covering gives better clustering performance. We observed that the application of refined second type covering as well as third type covering provides better allocation of data elements to clusters which results into performance improvements. The performance results of refined second type clustering and third type clustering with refined rough k-means algorithm are almost similar in each test. We have considered Q_L value to be 0.45

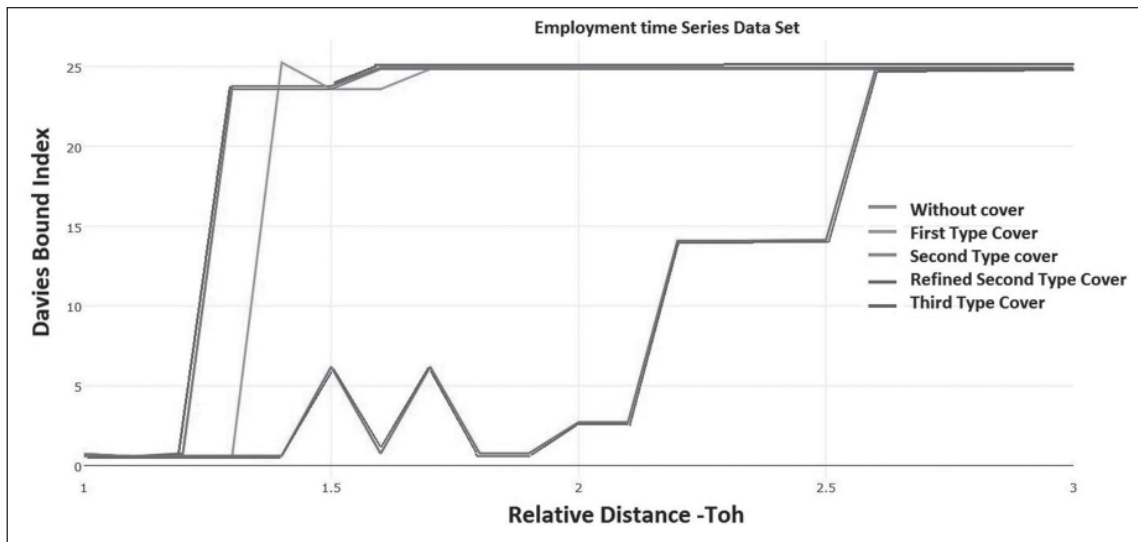


Figure 2. Davies Bouldin Index verses relative distance matric for Employee time series dataset.

The graph in figure-2 shows comparison among normal refined rough K-Means and covering based refined rough K-Means algorithm with respect to employee time series data set. The x axis represents relative distance δ and Y axis represents Davies Bound index. It shows clearly that refined second type as well as third type has lower values of Davies Bound Index. Thus performance has improved.

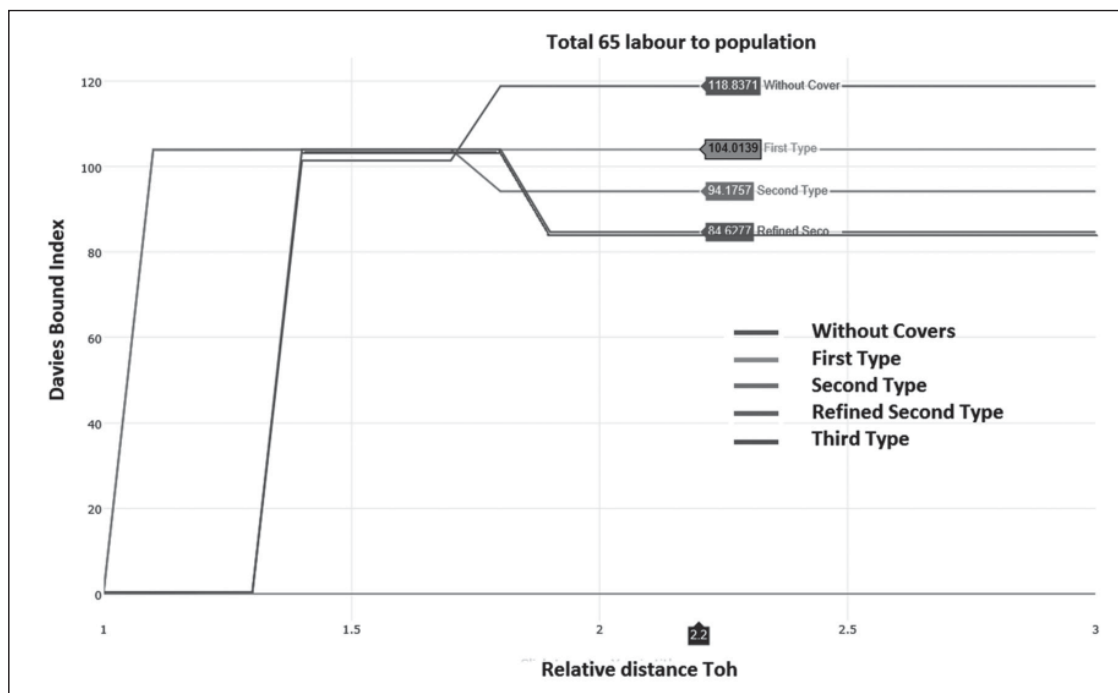


Figure 3. Davies Bouldin Index verses relative distance matric for Total 65 labor to population dataset.

The graph in figure-3 shows comparison among normal refined rough K-Means and covering based refined rough K-Means algorithm with respect to total 65 labor to population dataset. It shows how refined rough K-Means gives better performance with refined second type covering.

The figure-4 indicates performance for experiments performed over Household Income Disparity dataset. It shows performance of refined second type covering or third type covering with refined rough K-Means algorithm is most of the times better than normal refined rough K-Means as well as Refined rough K-Means with firsts type or second type covering when relative distance (δ) less than 25.

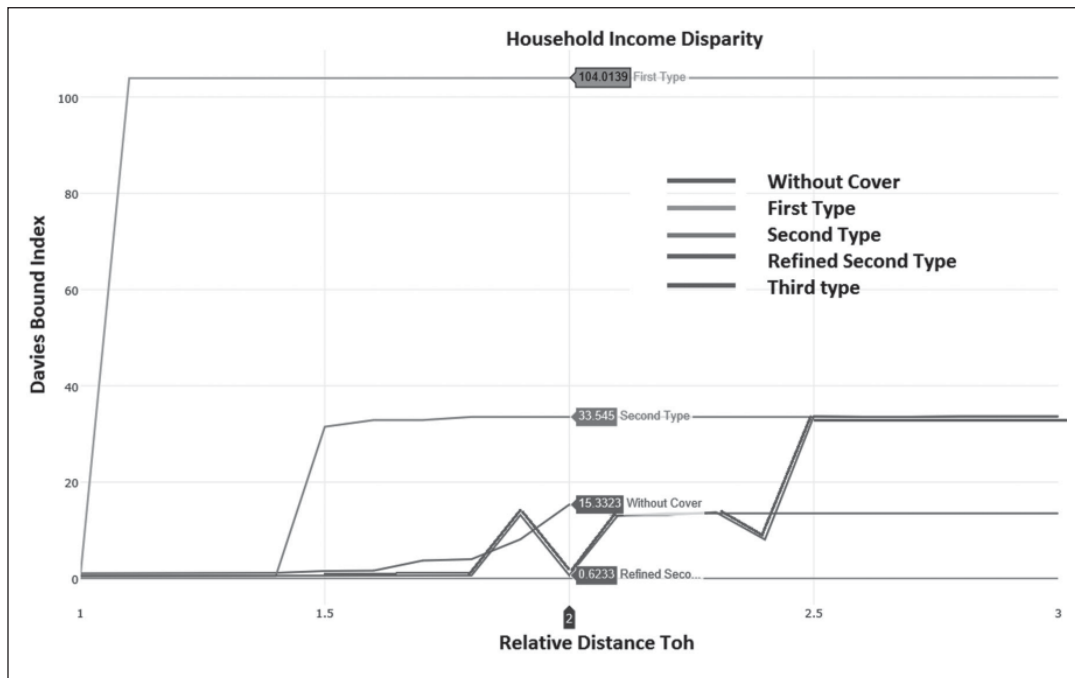


Figure 4. Davies Bouldin Index versus relative distance matrix for Household Income Disparity dataset.

Every time we have randomly generated the cover, feature vectors. The number of clusters as well as relative distance δ need to be manually defined. We have considered Q_L value to be 0.45. This random arrangement helps to eliminate the worries regarding the consideration of the values of cover and feature set.

7 CONCLUSION

The covering technique can be used as a choice for selecting initial allocation of data points to clusters. This initial allocation of data points to cluster in many cases improves the performance of clustering as we have shown in our results. The performance of covering based rough clustering heavily depends on the selection of covers and values of X_i . So generating and exploring optimal cover as well as value of X_i is a crucial task. The value of Q_L , Q_U and δ also affects the performance. So we need to carefully select these values.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195--197 (1981).
2. K. Kameshwaran, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of computer Science and Information Technology, Vol. 5(2), pp. 2272-2276, 2014.
3. Laurence Morissette, Sylvain Chartier, "Tutorials in Quantitative Methods for Psychology", 2013, Vol. 9(1), p. p.15-24.
4. Dr Debadutta Mohanty," ROUGH SET ON GENERALIZED COVERING APPROXIMATION SPACES", International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.
5. W. Zhu and F. Y. Wang, "Properties of the first of covering based rough sets", sixth IEEE International Conference on Data-Mining Workshop.
6. W. Zhu, "Properties of the second type of covering based rough sets", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web, Intelligence and Intelligence Agent Technology (WI-IATW' 06), July2006.
7. W. Zhu, "Properties of the fourth type of covering based rough sets" proceedings of the sixth International Conference on Hybrid Intelligent Systems (HIS'06) April-2006.

7. P. Lingras, C. West, "Interval set clustering of web users with rough k-means", *J. Intell. Inform. Syst.* 23 (2004) 5–16.
8. Georg Peters, "Some refinements of rough k-means clustering", *Pattern Recognition* 39 (2006) 1481 – 1491.
9. Pawan Lingras and Georg Peters, "Applying Rough Set Concepts to Clustering", Springer, 2012.
10. Y.Y. Yao, "A Comparative Study of Fuzzy Sets and Rough Sets", *Information Sciences*, Vol. 109, No. 1-4, pp. 227-242, 1998.
11. Zhanhong Shi, Zengtai Gong, "The further investigation of covering-based rough sets: Uncertainty characterization, similarity measure and generalized models", *Information Sciences* 180, 3745–3763, 2010.
12. William Zhu and Fei-Yue Wa, "On Three Types of Covering-Based Rough Sets", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 19, NO. 8, AUGUST 2007.
13. Pritesh Vora, Bhavesh Oza, "A Survey on K-mean Clustering and Particle Swarm Optimization", *International Journal of Science and Modern Engineering (IJSME)*, ISSN: 2319-6386, Volume-1, Issue-3, February 2013.
14. Shalove Agarwal, Shashank Yadav and Kanchan Singh, "K-mean versus k-mean++ clustering Techniques", in *IEEE* 2012.
15. Sanjiv K. Bhatia, "Adaptive K-Means Clustering", *American Association for Artificial Intelligence*, 2004.
16. Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, Vol. 1 (2), 2010, 121-125.
17. Tapas Kanungo, Nathan S. Netanyahu, Angela, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 24, NO. 7, JULY 2002.
18. Chunfei Zhang, Zhiyi Fang, "An Improved K-means Clustering Algorithm", *Journal of Information & Computational Science* 10: 1 (2013) 193–199.
19. Azhar Rauf, Sheeba, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", *Middle-East Journal of Scientific Research* 12 (7): 959-963, 2012.
20. Bangoria Bhoomi M., "Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Values", *International Journal of Computer Science and Information Technologies*, Vol. 5 (1), 2014, 876-879.
21. Tengfei Zhang, Long Chen and Fumin Ma, "An Improved Algorithm of Rough K-Means Clustering Based on Variable Weighted Distance Measure", *International Journal of Database Theory and Application* Vol.7, No.6 (2014), pp.163-174.
22. Y.SaiSandeep, P. Vidya Sagar Reddy, C.Manoj, K. Ajanthaa Lakkshmanan, "Identifying the Vague Regions by Using Covering Based Rough Sets", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 7, July 2013.
23. Jia Qiao, Yong Zhang, "Study on K-means Method Based on Data-Mining", *IEEE*, 2015..
24. Rui Máximo Esteves, Thomas Hacker, Chunming Rong, "Competitive K-means", *IEEE*, 2013.
25. Tengfei Zhanga & Fumin Mab, "Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function", *International Journal of Computer Mathematics*, 2016.