LNAI 6440

Longbing Cao
Yong Feng
Jiang Zhong (Eds.)

Advanced Data Mining and Applications

6th International Conference, ADMA 2010
Chongqing, China, November 2010
Proceedings, Part I

1 Part I

Springer

Springer Link

ADMA: International Conference on Advanced Data Mining and Applications

Advanced Data Mining and Applications

6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part I

- Editors
- (view affiliations)

- Longbing Cao
- Yong Feng
- Jiang Zhong

Conference proceedings **ADMA 2010**

Part of the Lecture Notes in Computer Science book series (LNCS, volume 6440)

Also part of the Lecture Notes in Artificial Intelligence book sub series (LNAI, volume 6440)

- Papers
- Volumes
- About

Table of contents

Page of 2
Next

Next

Other volumes

1. Advanced Data Mining and Applications
   6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part I
2. Advanced Data Mining and Applications
   6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part II

About these proceedings

Introduction

With the ever-growing power of generating, transmitting, and collecting huge amounts of data, information overloadis nowan imminent problemto mankind. The overwhelming demand for information processing is not just about a better understanding of data, but also a better usage of data in a timely fashion. Data mining, or knowledge discovery from databases, is proposed to gain insight into aspects ofdata and to help peoplemakeinformed,sensible,and better decisions. At present, growing attention has been paid to the study, development, and application of data mining. As a result there is an urgent need for sophisticated techniques and toolsthat can handle new ?elds of data mining, e. g. , spatialdata mining, biomedical data mining, and mining on high-speed and time-variant data streams. The knowledge of data mining should also be expanded to new applications. The 6th International Conference on Advanced Data Mining and Appli- tions(ADMA2010)aimedtobringtogethertheexpertsondataminingthrou- out the world. It provided a leading international forum for the dissemination of original research results in advanced data mining techniques, applications, al- rithms, software and systems, and di?erent applied disciplines. The conference attracted 361 online submissions from 34 di?erent countries and areas. All full papers were peer reviewed by at least three members of the Program Comm- tee composed of international experts in data mining ?elds. A total number of 118 papers were accepted for the conference. Amongst them, 63 papers were selected as regular papers and 55 papers were selected as short papers.

Keywords

Clustering HPC adaptive algorithms classification data mining data types graphs knowledge discovery machine learning online communities pattern mining sensor data sequences spatial datasets web mining

Editors and affiliations

- Longbing Cao (1)
- Yong Feng (2)
- Jiang Zhong (3)

1. Faculty of Engineering and Information Technology, University of Technology Sydney, , Sydney, Australia
2. College of Computer Science, Chongqing University, , Chongqing, China
3. College of Computer Science, Chongqing University , , Chongqing, China

Bibliographic information

- DOI https://doi.org/10.1007/978-3-642-17316-5
- Copyright Information Springer Berlin Heidelberg 2010
- Publisher Name Springer, Berlin, Heidelberg
- eBook Packages Computer Science
- Print ISBN 978-3-642-17315-8
- Online ISBN 978-3-642-17316-5
- Series Print ISSN 0302-9743
- Series Online ISSN 1611-3349
- Buy this book on publisher's site

An Improved Rough Clustering Using Discernibility Based Initial Seed Computation

Djoko Budiyanto Setyohadi, Azuraliza Abu Bakar, and Zulaiha Ali Othman

Center for Artificial Intelligence Technology University Kebangsaan Malaysia Bangi, Selangor Darul Ehsan, 43000 Malaysia
djokobdy@gmail.com, {aab,zao}@ftsm.ukm

Abstract. In this paper, we present the discernibility approach for an initial seed computation of Rough K-Means (RKM). We propose the use of the discernibility initial seed computation (ISC) for RKM. Our proposed algorithm aims to improve the performance and to avoid the problem of an empty cluster which affects the numerical stability since there are data constellations where $|\underline{Ck}| = 0$ in RKM algorithm. For verification, our proposed algorithm was tested using 8 UCI datasets and validated using the David Bouldin Index. The experimental results showed that the proposed algorithm of the discernibility initial seed computation of RKM was appropriate to avoid the empty cluster and capable of improving the performance of RKM.

Keywords: Discernibility, Initial Seed Computation, Rough K-Means.

1 Introduction

Clustering is a process of classifying objects into classes based on similarities among data. The process of assigning an object to its cluster is fully based on the data similarity; therefore the characteristics of data may influence the clustering result. The performance of K-Means, as the most widely used clustering algorithm, depends on two key points, namely the initial clustering and the instance order [1]; in which initial clustering itself fully depends on the data distribution. Since the characteristic of the data influences the performance of K-means, many improvements of K-means are being developed. Rough K-means clustering (RKM) [2] is one of the well known extended K-means algorithm.

RKM is the clustering algorithm which addresses the problem of vague data. Its capability to cluster vague data comes from the integration of Rough Set Theory in the process of clustering. While in the original K-Means the cluster is viewed as a crisp cluster only, in RKM the cluster is deployed as an interval clustering. Here, the object is divided in the lower approximation where the object is certainly a member of the cluster, and the boundary area where the object is a member of more than one cluster [2]. Looking at its characteristics, RKM can be considered as a powerful algorithm for clustering vague data. Vague data can be clustered in a boundary area which is useful for further processing.

L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part I, LNCS 6440, pp. 161–168, 2010.

Despite its advantages, RKM has a drawback especially on the numerical stability problem [3][4][5]. The problem arises because RKM equation requires that each cluster must have at least a member. This situation is also found in the original K-means and is solved by an initial seed computation [1][6][7][8][9][10]. Unlike in the original K-means, the empty cluster in the RKM will generate the numerical stability problem since there are data constellations where $|\underline{Ck}| = 0$, which refers to the computation of cluster centroid [2][11]. Therefore, several researchers have made improvements on the numerical stability problem [3][4][5].

According to the numerical stability problem, Peters [3] refined RKM by forcing at least one of the objects should be a member of the cluster. Hence one of the objects which is the closest to the centroid of the lower approximation will be assigned to the closest cluster. Miao [4] avoided the empty cluster by using the non-object outlier to the proper cluster and proposed the use of angle measurement to decide the member of clusters. Obviously, all of the previous work on RKM refinement, including that of Zhao [5] and of Lingras [11], focused on the membership function refinement. Although previous researchers had improved the RKM, they ignored the other source of the numerical stability problem i.e. the initial seed, since K-means clustering certainly relies on the chosen initial centroid [1][6][7]. Moreover, when the algorithm is applied, the boundary area should be restricted to avoid a numerical stability problem [3][4][5]. Therefore, to fill the gap of the previous work that heavily focused on refining the membership function to avoid numerical stability this work highlights the initial seed computation to avoid a numerical stability problem.

Many ISCs have been developed since the process of the K-means clustering is deterministic mapping from initial solution to local minima of final result [1][11]. The previous research showed that the use of the ISC did not only improve the performance of K-means but also was able to avoid the empty cluster problem that plagues K-Means. Hence we propose the use of ISC to avoid a numerical stability problem in RKM as an extension of K-means.

In this paper, we review the required characteristics of the previous ISC works, from which we further develop the algorithm based on the discernibility approach of Rough Set Theory which is suitable for the purpose of RKM i.e. processing the vague data. To verify the proposed algorithm, we use David Bouldin (DB) Index, which is a well-known validity measurement in clustering analysis [12].

2 Initial Seed Computation (ISC) on K-Means Clustering

Determining the initial seed points is very important in K-means since the initial centroid will determine the final centroid [1]. The main issue of the initial seed is that the initial centroid should be chosen properly. Currently, there are many studies focusing on the ISC for improving K-means algorithm in order to improve the result of clustering [1][6][7][8][9][10], which is also applicable in RKM to solve its numerical stability problem. Furthermore, the previous characteristics will be discussed below.

There are three properties commonly used as the guidance for initial seed determination, namely, centrality, sparseness and isotropy. According to the characteristics, Arthur [6] suggested a method to choose the area of initial seed, that is, after the first initial seed is selected the next initial seed should be chosen in the uncovered space of its cluster. Kang [7] improved the performance of clustering by following the properties that the initial centorid should be distant enough to each other but close to the final centroid. Redmond et al. used Density Generated Kd-tree [9], a top-down hierarchical scheme, to isolate data. Suppose we have a set of n points, ($x_1 \ldots x_n$), it will be divided roughly by splitting the data along the median value of the co-ordinates in that dimension, median (x_{1mmax} ; x_{2mmax} ; . . . ; x_{nmmax}). Then the properties of each partition will be the basis ISC. Although they use different approaches to satisfy the properties initial seed, their results are able to improve the performance of K-means significantly. The results from the previous research do not only improve the performance of K-means clustering but also can avoid an empty cluster. However, the previous methods have some limitations, such as the high processing time since they contain complex computation with *O(N2)*. Referring to the characteristics of initial seed, we propose the use of the discernibility concept of Rough Set for ISC. Discernibility is one of the important characteristics of Rough Set Theory [13], and therefore it is suitable for the main purpose of RKM.

3 Discernibility of Rough Set Theory (RST)

Discernibility, an important property in RST, is the relation of two objects which can be defined as $\text{discern}_A(B) = \{(x,y) \in U2 : a(x) \neq a(y), \forall\ a \in B\}$. Given an IS A = (U, A), and a subset of attributes B $\subseteq$ A, the discernibility matrix of A is M_B, where each entry $m_B(i,j)$ consists of the attribute set that discerns between objects x_i and x_j. The discernibility function of A over attributes B $\subseteq$ A can be defined as $f[B] = V \Lambda\, m_{[B]}(E_i, E_j)$ where $i, j \in \{1,...,n\}$ and n is the number of classes in the IS. Suppose, given a subset of attributes $A \subseteq A_t$ and a pair of objects $(x_i, x_j) \in U$ x U, i,j $\in\{1, 2, ..., |U|\}$, the quantitative discernibility relation $dis(A)(x_i, x_j)$ is defined as the complement of a quantitative indiscernibility.

$$dis(A)(x_i, x_j) \qquad = 1 - ind(A)(x_i, x_j). \tag{1}$$

satisfies to the properties $dis(A)(x_i, x_i) = 0$ and $dis(A)(x_i, x_j) = dis(A)(x_j, x_i)$ where the quantitative discernibility relation is reflexive and symmetric. The discernibility level, which is possible to represent the granularity of objects, can be used to measure the discernibility among objects. The higher level of discernibility implies that the objects are likely to be treated as discernible. This discernibility will be the basis for our ISC.

4 Proposed Discernibility Based Initial Seed Computation

The objective of ISC is to place the initial centroid close to the intrinsic centroids. Using its objective, the clustering algorithm should rapidly converge to the global optimal structure and the problem of the empty cluster can be avoided. On the other hand, the good cluster should ensure that each centroid should be in the middle of the cluster (minimizing intra-cluster variance), and the centroids should be distant enough to each other (maximizing inter-cluster variance). The proposed algorithm tries to satisfy this objective. We first discretize the entire objects in data sets. This process will be followed by the calculation of the discernibility level which is considered as the degree of a distant object. Then, we select the best combination of initial seed space. The best combination is measured by the highest discernibility level which indicates the highest degree of object distance.

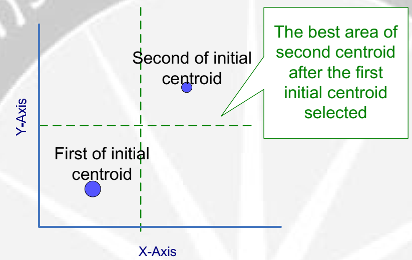


Fig. 1. The example of best search space combinations of initial centroid point for data set of two attributes and two clusters

The degree of distance is very relative and difficult to determine precisely. Therefore, we adopt the discernibility of RST to introduce the discernibility concept for initialization of seed points. We use the discernibility level terminology to measure the appropriate degree of the distance of initial centroid of RKM. Since the discernibility approach for data separation has the computation load problem, the proposed technique is focused on the discernibility relation of a binary table which is generated from any appropriate discretization method. This approach aims to partition objects roughly in the data set. Referring to the properties of initial seed [6][7], the best chosen initial seed can be achieved when a discernibility degree (α) is maximum as follows $\alpha_A(x_1,\ldots, x_n) = \mathrm{MAX}[\alpha_A(x_i, x_j)]$. Fig. 1 illustrates an example of how a discernibility concept is used. Suppose the data set has two attributes and is identified to have two clusters, we can divide the spaces of initial seed into four spaces using the binary classification. Each attribute will have two spaces as illustrated in Fig. 1.

Using one of the properties of the initial seed which is proved [6][7], the best combination space of initial seed area is achieved when the space area is the highest possible distance. This condition can be achieved when the discernibility degree is maximum. As illustrated in Fig. 1, assuming the first of initial centroids space is in quadrant one then, the best second initial centroids space is in quadrant four. Further the discernibility degree α between two objects can be calculated as in Eq.(2).

$$\alpha_A(x_i, x_j) = \frac{|\{a \in A \mid I_a(x_i) \neq I_a(x_i)\}|}{|A|} \tag{2}$$

Where |A| denotes the cardinality of a set and two objects x_i and x_j, for $\alpha_A(x_i, x_j)$ for $x_i \neq x_j$. Therefore, the initial seed points for the cluster are the points x_i, x_j that give the highest degree α for n object. It can be defined as the maximum degree of any degree α two objects as formulated in Eq(3) where $i \neq j$, $i,j = 1,2, \ldots, n$.

$$\alpha_A(x_1, \ldots, x_n) = \text{MAX}[\alpha_A(x_i, x_j)] \tag{3}$$

For example, for more than two objects, the discernibility degree α can be computed as $\alpha_A(x_1, x_2, \ldots x_n) = \text{MAX}\{\alpha(x_1,x_2), \alpha(x_1, x_3), \ldots, \alpha(x_1, x_n)\}$. In order to satisfy that all objects have the highest discernibility, the discernibility degree is obtained as the maximum of $\alpha_A(x_1, x_2, \ldots x_n)$. By using this approach, the complexity is related to the binary searching of a space area; therefore, the complexity of discernibility ISC is only about O($n$log$n$). Another advantage of using ISC for RKM is that the maximum discernibility will lead the process in order to keep the consistency and its convergence. Thus, this computation can control the influence of the threshold value. The proposed method is outlined in the following steps.

Step 1. Initial seed computation
1. Convert the information system into the binary classification.
2. Calculate the discernibility degree α for every pair of data points.
3. Find the maximum α among the objects x_i, x_j
4. Select x_i, x_j as the initial seed for RKM algorithm.

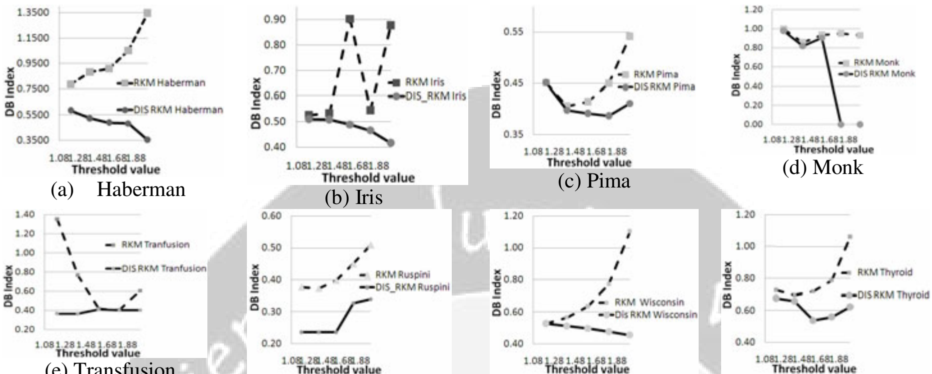
Step 2. RKM Clustering
1. Initiate the centroid using the selected object from the first step
2. Calculate the new means of RKM as in [11]
3. Calculate membership of each object of RKM as in [11]
4. Repeat from step 2.2. until convergence.

5 Experiment and Discussion

In this study we tested our proposed method called DIS_RKM upon eight UCI datasets. The datasets are Iris, Monk, Pima, Wisconsin, Ruspini, Haberman, Transfusion, and Thyroid. Two measures were used in the experiment: i) the percentage of data points in the Lower Approximation (%Lower) and ii) the DB_Index. The DB Index was calculated using the formulation in Eq(4).

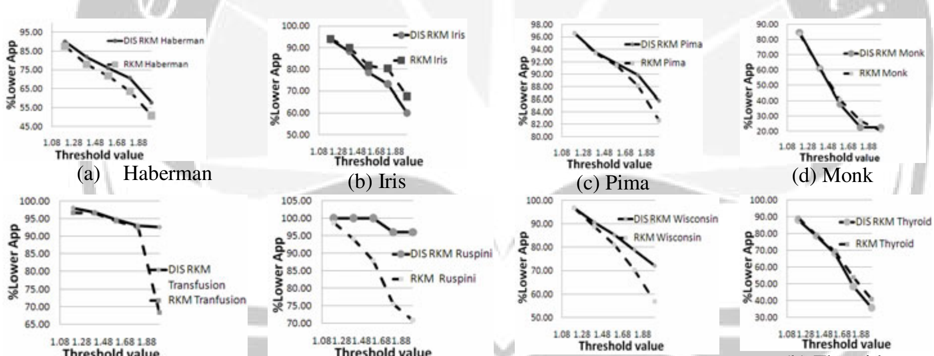
$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{j \neq i}\left\{\frac{S(U_i) + S(U_j)}{d(U_i, U_j)}\right\} \tag{4}$$

The good RKM is that the values of %Lower and DB Index decrease as the threshold values increase until the centroid move. The proposed algorithm was implemented using Java Netbean 6.8, Processor Intel T9660 Ram 3 GByte and Windows operating system. The experiment used the threshold value to simulate the consistency and stability of the method.



(a) Haberman (b) Iris (c) Pima (d) Monk (e) Transfusion (f) Ruspini (g) Wisconsin (h) Thyroid

Fig. 2. Change in DB Index with threshold in eight UCI data sets



(a) Haberman (b) Iris (c) Pima (d) Monk (e) Transfusion (f) Ruspini (g) Wisconsin (h) Thyroid

Fig. 3. Change in %Lower App. with threshold in eight UCI data sets

Fig. 2 and Fig. 3 showed the experimental results obtained using our proposed discernibility ISC for RKM called DIS_RKM and the random initial seed setting for RKM via the original RKM. The DB Index was obtained by using different threshold values. The results showed that the DIS_RKM yielded a lower DB Index than RKM in all of data sets and all of threshold values (Fig. 2), except on Pima data sets where in threshold value 1.2. Since the lower DB Index value represented a better cluster, we concluded that DIS_RKM was able to improve the performance of RKM.

Referring to the membership function in RKM, the threshold values increased, the %Lower decreased and the size of the boundary area increase is means that more vague values are moved to the boundary area. Based on DB Index simulation, the compactness of the cluster improved since the intra-cluster distance was reduced and inter-cluster distance was retained when threshold value increased. However, we can see that the original RKM was not consistent with the changes of thresholds. The DB Index values decreased to a certain level but increased when it reached the threshold 1.6. The situation also proved the robustness of DIS_RKM. By simulating the threshold we can see that DIS_RKM was consistent since it led to the similar centroid where the DB Index values decreased as the threshold values increased. Larger threshold values indicated the larger size of a boundary area. In our proposed approach, the distance between centroid was retained while the threshold changed, but the size and compactness of the lower approximation and boundary area changed. The decreasing values of DB Index were achieved by reducing the average distance among the objects within clusters while the distance among the clusters was retained when the threshold value increased. This condition can be achieved if the centroid did not move when the threshold increased.

The threshold determined directly the membership of the object which influenced the boundary area. Therefore, changing of thresholds will change the boundary area. Theoretically, it is sufficient to control the vague objects of the cluster while maintaining the DB Index. Based on Fig. 3 the %Lower measure showed that DIS_RKM was comparable with the RKM where the increasing values of threshold were followed by the decreasing %Lower approximation. Referring to the turning point of the threshold change, the point where the direction of DB Index changed from a lower to a higher point, the range of DIS_RKM was longer than that of the original RKM. It indicated that the proposed algorithm was more robust. Moreover discernibility based ISC was able to improve the random seed setting by the original RKM. In this experiment we also showed that the higher the threshold, the lower the DB_Index and the %Lower approximation where the size of the boundary area was controlled properly.

According to the issue of the numerical stability DIS_RKM outperforms in four data sets (Haberman, Tranfusion, Wisconsin, Ruspini,) and comparable in data sets (Pima, Monk) due to a slower decline of %Lower DIS_RKM. This result indicated that the ability of DIS_RKM to avoid an empty cluster was better than the original RKM. Furthermore, the final centroid of DIS_RKM converged at a certain point.

In our experiment, the DIS_RKM did not change the rule but improved it by making the RKM simpler since the lower approximation component was decreased gradually. Therefore, we were able to control the boundary area as needed, except on the Ruspini data set where the separation of the cluster was deterministic (not vague). In Ruspini dataset, we were able to show that the ISC improved the performance of RKM in a natural way, not through forcing the data of the cluster. The 100% value of %Lower indicated that no data were forced to the boundary even though the threshold values increased and the data clearly belonged to the lower approximation space.

6 Conclusion

We have introduced the discernibility ISC to improve RKM. This approach is based on the properties of good clusters that the centroids should be distant enough to each other. To implement the ISC we proposed the use of binary discernibility in order to reduce the high computation problem. We observed that the proposed ISC was able to improve the robustness, to enhance the performance and to avoid the empty cluster problem of RKM, particularly when the threshold increased. The experimental results showed improved and consistent clusters as compared to the initial random cluster centers.

Acknowledgments. I would like to thank Atma Jaya University Yogyakarta, Indonesia for the financial support for my research project, UKM for conference funding. I would also like to express my gratitude to Prof. P. Lingras for providing me with the original code of RKM and his RKM papers.

References

1. Peña, J.M., Lozano, J.A., Larrañaga, P.: An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters archive 20(10), 1027–1040 (1999)
2. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-means. Journal of Intelligent Information System 23(1), 5–16 (2004)
3. Peters, G.: Some Refinement of K-means Clustering. Pattern Recognition 39, 1481–1491 (2006)
4. Miao, D.Q., Chen, M., Wei, Z.H., Duan, Q.G.: A Reasonable Rough Approximation of Clustering Web Users. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) Web Intelligence Meets Brain Informatics. LNCS (LNAI), vol. 4845, pp. 428–442. Springer, Heidelberg (2007)
5. Zhou, T., Zhang, Y.N., Lu, H.L.: Rough k-means Cluster with Adaptive Parameters. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, August 19-22,, pp. 3063–3068 (2007)
6. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proc. ACM-SIAM Symp. Discrete Algorithms (2007)
7. Khang, P., Cho, S.: *K*-means clustering seeds initialization based on centrality, sparsity, and isotropy. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 109–117. Springer, Heidelberg (2009)
8. Khan, S.S., Ahmad, A.: Cluster Center initialization algorithm for K-means clustering. Pattern Recognition Letter 25(11), 1293–1302 (2004)
9. Redmond, S.J., Heneghan, C.: A method for initializing the k-means clustering algorithm using kd-trees. Pattern Recognition Letters 28(8), 965–973 (2007)
10. He, J., Lan, M., Tan, C.-L., Sung, S., Low, H.-B.: Initialization of cluster refinement algorithms: A review and comparative study. In: Proc. IEEE Int. Joint Conf. Neural Networks, pp. 297–302 (2004)
11. Lingras, P., Chen, M., Miao, D.: Rough Cluster Quality Index Based on Decision Theory. IEEE Transactions On Knowledge And Data Engineering 21 (2009)
12. Halkidi, M., Batistakis, Y., Vazirgianni, M.: On Clustering Validation Techniques. Journal of Intelligent Information Systems 17(2/3), 107–145 (2001)
13. Pawlak, Z.: Rough Set: Theoretical Aspect of Reasoning about Data. Kluwer Publications, Dordrecht (1991)