K-modes Algorithm Based on Rough Set and Information Entropy

Gong Xingyu¹, Cao Ke¹, Jia Pengtao¹, Gong Shangfu¹
¹Xi'an University of Science and Technology, College of computer science and technology, Xi'an, Shanxi, China;
Corresponding author’s e-mail: gongsf@xust.edu.cn

Abstract. The traditional K-modes algorithm is susceptible to interference of redundant attributes, and only adopts the 0-1 matching method to define the distance between attribute values of each two objects, without fully considering the influence of each classify attribute on clustering result. In order to overcome these shortcomings, this paper proposes improved K-modes clustering algorithm based on rough set and information entropy. Aiming at a large number of redundant attributes in the clustering data, this paper firstly utilizes attribute reduction algorithm of rough set to eliminate redundant attributes and determine the importance of each attribute, then combines information gain to determine the weight of each attribute and finally makes performance tests of the traditional algorithm and the improved algorithm on five data sets of UCI machine learning library, such as Soybean-Small and Zoo. The experimental results show that the clustering efficiency and accuracy of improved algorithm is higher than that of traditional algorithm, and the performance of improved algorithm is better.

1. First Section Introduction

Clustering analysis is an important direction in the research of data mining. It is generally divided into partition clustering, hierarchical clustering, density clustering, grid clustering, model clustering, graph-theoretical clustering and so on. K-means algorithm is a kind of partition clustering, it can only deal with numerical data, but cannot deal with the data of classify attribute type. In view of this deficiency, Huang et al. proposed K-modes algorithm⁴, which is suitable for processing the data of classify attribute type. In 2014, Chu Lulu and Jiang Feng⁵ proposed a K-modes clustering algorithm based on weighted overlapping distances. Hong Jia⁶ et al. proposed a redundant metric method based on interdependence⁷ to calculate between different attributes degree of relevance, but its actual effect is only equivalent to giving each classification attribute a different weight. In 2016, Wang Yongsheng⁸ et al. introduced parameter weights in the mixed attribute measurement mechanism, and on this basis, constructed a rough-based Integrated classifier of the mixed attribute measurement mechanism of the set. In 2017, Zhao Liang et al.⁹ proposed a simple bayesian classifier based distance measurement method of operation results. In 2018, Wen Wu¹⁰ et al. proposed a KNN text classification algorithm based on K center points and rough sets based on the traditional KNN algorithm. Zhang Tengfei¹¹ et al. introduced a natural adaptation measure to the imbalance degree of family size, and proposed an adaptive measure based on the imbalance degree of cluster size. In 2019, Yang Youlong¹² and others established a hybrid data similarity measure and improved the K-modes algorithm process to optimize the clustering results.

None of the above methods takes into account the effects of redundant data on the clustering result. However, in the actual situation, not every classify attribute is valid for the clustering result, and the
influence degree of each classify attribute on the clustering result is also different. Therefore, based on these reasons, this paper introduces rough set theory and information entropy theory to solve the problem of redundant attributes of clustering data sets and the weight of each classify attribute. In the process of clustering, this paper utilizes the attribute reduction algorithm of rough set to remove the redundant attributes, get reduction set of training set. Then obtains attribute importance degree of each reduction set. Calculates information gain of each classify attribute in the data set, combines rough set to determine the weight of each classify attribute value, and makes better clustering analysis. Experiments prove that the proposed K-modes clustering algorithm based on rough set and information entropy is both feasible and effective.

2. K-modes Algorithm Analysis

K-modes algorithm is an evolution of K-means algorithm. K-modes algorithm adopts simple 0-1 matching method to define the distance between attribute values of each two objects, and replaces means of K-means algorithm with modes which update iteratively based on frequency method.

Definition 1: a dataset X has n objects represented as \( O = \{x_1, x_2, \ldots, x_n\} \) and r classify attributes represented as \( A = \{A_1, A_2, \ldots, A_r\} \), \( x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\} \). Supposing the classify attribute has m different attribute values, \( \text{DOM}(A_i) = \{a_{i1}, a_{i2}, \ldots, a_{im}\} \). The distance between the \( i \)th object and the \( j \)th object named \( d(x_i, x_j) \) is as formula (1).

\[
d(x_i, x_j) = \sum_{r=1}^{r} \delta(x_{ir}, x_{jr})
\]

\[
\delta(x_{ir}, x_{jr}) = \begin{cases} 
0, & x_{ir} = x_{jr} \\
1, & x_{ir} \neq x_{jr}.
\end{cases}
\]

Definition 2: dividing dataset \( X \) into \( k \) classes represented as \( Q_1, Q_2, \ldots, Q_k \). The objective function is as formula (2).

\[
F = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(x_i, Q_l)
\]

\[
y_{il} \in \{0,1\}, 1 \leq i \leq n, 1 \leq l \leq k
\]

\[
0 \leq y_{il} \leq 1
\]

The \( y_{il} \) has relationship as formula (3).

\[
\sum_{l=1}^{k} y_{il} = 1
\]

\( Q_l = \{q_{l1}, q_{l2}, \ldots, q_{ld}\} \) is the center of the \( l \)th class, \( q_{ld} \) is the mode of the \( d \)th attribute of the \( l \)th class.

With the above constraints, a traditional K-modes clustering algorithm is presented to minimize the objective function \( F \).

For the input, the number of sample objects is \( n \) and the number of classes is \( k \). The output is clustering result.

The specific steps of traditional K-modes clustering algorithm are divided into four steps.

Step (1): randomly select \( k \) objects in dataset \( X \) as initial clustering centers.

Step (2): calculate the distance \( d(x_i, Q_l) \) between all objects in dataset \( X \) and \( k \) initial centers according to formula (1), and assign each object to the class with the smallest distance to it, then get \( k \) classes as \( Q_1, Q_2, \ldots, Q_k \).

Step (3): in each attribute of each class, select the attribute value with the maximum frequency as the attribute value of the mode of this class. This value is chosen as the new clustering center.

Step (4): repeat step (2) and step (3) until the cluster center no longer changes.

The traditional K-modes clustering algorithm only takes advantage of simple matching method to measure the distance between objects and makes clustering, but it ignores influence of redundant attributes and the difference between attributes on the clustering results. In view of this shortcoming, this
paper proposes K-modes clustering algorithm based on rough set and information entropy to improve traditional K-modes algorithm.

3 Basic Concepts

3.1 Rough Set Theory
In 1982, scholar Z. Pawlak proposed rough set theory \[11-12\]. Rough set theory can deal with imprecise knowledge information and discover the hidden knowledge by analyzing and reasoning relevant data sets.

3.2 Information Entropy Theory
Definition 7: \(\forall a \in P\), the importance degree of attribute value \(a\) relative to \(Q\) is defined as \(\delta_{PQ}(a) = r_p(Q) - r_{P-(a)}(Q)\), the normalization is as formula (4).

\[
\delta_{PQ}(a) = \delta_{PQ}(a)/\sum_{a \in P} \delta_{PQ}(a)
\]  

(4)

3.2.1 Information Entropy
In 1948, Shannon proposed concept of information entropy \[13\], which solved the problem of quantitative measurement of information and widely utilized to measure uncertainty in rough set.

Definition 8: the information entropy that \(D_i\) produced is as formula (5).

\[
H(D_i) = -\sum_{i=1}^{n} (p_i) \log_2 (p_i), i \in (1,2,\cdots,n)
\]  

(5)

In formula (5), \(n\) is the number of classes divided by decision attribute \(D\) within the discourse domain \(U\). \(D_i\) is the number of corresponding elements of \(D\). \(p_i\) is the probability that samples in \(U\) belong to the \(i\)th class \(C_i\), that is \(p_i = d_i/n\).

Definition 9: for conditional attribute \(C\), \(a \in C\). If values of attribute \(a\) are set as \(a_1, a_2, \cdots, a_r\), it will divide discourse domain \(U\) into \(r\) parts as \((u_1, u_2, \cdots, u_r)\). The conditional entropy of attribute \(a\) relative to decision attribute \(D\) is as formula (6).

\[
H(a) = \sum_{i=1}^{n} \frac{|u_{ij}|}{U} \sum_{j=1}^{r} H(u_{1j}, u_{2j}, \cdots, u_{nj}) = \frac{\sum_{j=1}^{r} u_{1j}, u_{2j}, \cdots, u_{nj}}{|u|} H(u_{1j}, u_{2j}, \cdots, u_{nj})
\]  

(6)

In formula (6), \(u_{ij}\) is the row of data for attribute \(a\) with the same value \(a_j\). \(u_j\) contains data objects of \(D_i\) with the number of \(u_{ij}\).

3.2.2 Information Gain
The weight value of attribute determined by rough set alone will cause information loss. It can be corrected by information gain in the information entropy theory. Information gain can truly reflect importance degree of attribute. The higher the information gain of attribute is, the greater the influence of attribute on final class is, and the greater the importance of attribute to information system is.

Definition 10: information gain is effective reduction of information entropy. Information gain of attribute \(a\) is as formula (7).

\[
\text{Gain}(a) = H(D) - H(a)
\]  

(7)

Definition 11: importance degree \(\mu(a)\) of information gain is as formula (8).

\[
\mu(a) = \frac{\text{Gain}(a)}{\sum_{a \in C} \text{Gain}(a)}
\]  

(8)
4 K-modes Clustering Algorithm Based on Rough Set and Information Entropy

Rough set theory and information entropy theory are introduced in this paper to solve the problem of redundant attributes and weight of each attribute. In the process of clustering, this paper firstly removes redundant attributes by attribute reduction algorithm of rough set, obtains reduction set of training set, and gets attribute importance degree of each reduction set. Then, information gain of each attribute in the data set is calculated, and the weight of each attribute value is determined through combining with rough set. Finally, it can achieve better clustering analysis results.

4.1 Dissimilarity between Samples

Definition 12: given any two samples \( x_i \) and \( x_j \), according to rough set and information entropy, the dissimilarity between these two samples is the distance between their attributes. The dissimilarity between samples can be indicated as formula (9).

\[
d^{\text{new}}(x_i, x_j) = \sum_{r=1}^{t} \left( \frac{s_{pq(a)} + \mu(a)}{2} \right) \delta(x_{ir}, x_{jr})
\]

(9)

\[
\delta(x_{ir}, x_{jr}) = \begin{cases} 
0, & x_{ir} = x_{jr} \\
1, & x_{ir} \neq x_{jr} 
\end{cases}
\]

4.2 Description of Algorithm Improvement

The thought of algorithm improvement is described as below. Firstly, it utilizes rough set to make data preprocessing and applies attribute reduction algorithm of rough set to eliminate redundant attributes and get a reduction set of training set. Then it calculates importance degree of attribute in each reduction set. On the basis of importance degree, it combines with information gain to determine the weight value of each attribute and finally realizes effective clustering.

The specific process of improved K-modes algorithm based on rough set and information entropy is given as below.

For the input, the number of sample objects is \( n \) and the number of classes is \( k \). The output is clustering result.

Step (1): randomly choose \( k \) objects from dataset \( X \) as initial clustering centers.

Step (2): calculate distance \( d^{\text{new}}(x_i, Q_j) \), between all objects in dataset \( X \) and \( k \) initial centers according to formula (9), and assign each object to the class with the smallest distance to it, then obtain \( k \) classes as \( Q_1, Q_2, \ldots, Q_k \).

Step (3): in each attribute of each class, select the attribute value with the maximum frequency as the attribute value of the mode of this class. This value is chosen as the new clustering center.

Step (4): repeat step (2) and step (3) until the cluster center no longer changes.

5 Simulation Experiment and Result Analysis

5.1 Evaluation Index

In order to evaluate clustering results, this paper adopts two kinds of evaluation indexes: classification precision (PR) and classification accuracy (AC). PR and AC are respectively defined as

\[
PR = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i}, \quad AC = \frac{\sum_{i=1}^{k} a_i}{n}
\]

where \( a_i \) represents the number of objects correctly classify to class \( i \), \( b_i \) refers to the number of objects incorrectly classify to other classes, \( n \) is the number of all objects in dataset \( X \), and \( k \) is the number of classes through clustering.

5.2 Experimental Analysis

In this paper, MATLAB R2015b was utilized to compile program to analyze effectiveness of algorithm. The experimental datasets were all taken from UCI machine learning library, including Soybean-Small
dataset, Zoo dataset, Vote dataset, Breast-Cancer dataset and Mushroom dataset. The description of these datasets is shown in Table 1.

| Dataset            | Samples' number | Attributes' number | classes' number |
|--------------------|-----------------|--------------------|-----------------|
| Soybean-Small      | 47              | 35                 | 4               |
| Vote               | 435             | 16                 | 2               |
| Breast-Cancer      | 699             | 9                  | 2               |
| Mushroom           | 8124            | 22                 | 2               |

Compared with K-modes clustering algorithm based on rough set and information entropy proposed in this paper, Huang’s K-modes algorithm is referred to as traditional K-modes algorithm and selected to make a comparative experiment. Before analysis, data in each dataset need to be normalized to eliminate interference caused by different value range for each data attribute.

Samples’ number \( n \) and classes’ number \( k \) should be set according to actual situation in Table 1. The final classification precision (PR) and classification accuracy (AC) of Huang’s K-modes method and this paper’s K-modes method are shown from Table 2 to Table 5.

From calculation results of Table 2 to Table 6, it can be concluded that the improved algorithm in this paper has a larger increase on the performance of five datasets with PR raised 0.0593 on average and AC increased an average value of 0.0727. Thus, the improved K-modes method based on rough set and information entropy in this paper has higher classification accuracy compared with traditional K-modes algorithm, and it is indicated that this improved method is both effective and feasible. Qualitatively, its effectiveness is derived from two aspects:
(1) K-modes algorithm based on rough set and information entropy introduces concept of rough set, so it has ability to deal with non-deterministic problems. When background knowledge is uncertain, incomplete or has noise, this improved algorithm can also make relatively correct analysis and judgment without bringing in any prior knowledge.

(2) When calculating attribute distance, K-modes algorithm based on rough set and information entropy takes advantage of information entropy to determine weight in the calculation process, so as to ensure that the weight of attributes having a key impact on effective clustering result is larger. This is more conducive to precision and accuracy of clustering analysis and this clustering result is more consistent with displaying attribute features of things.

6. Conclusion
This paper proposes new K-modes clustering algorithm based on rough set and information entropy on the basis of traditional K-modes algorithm. The combination of information entropy and rough set theory is applied to K-modes clustering algorithm can improve efficiency and accuracy of traditional K-modes algorithm. Experimental results indicate that the proposed K-modes algorithm based on rough set and information entropy in this paper is superior to traditional k-modes algorithm in terms of classification precision and classification accuracy.

Acknowledgments
This project was supported by the State Key Research Development Program of China (Grant No.2018YFC0808303) National Natural Science Foundation of China, General Project, 51974236, Multi-field information superposition determination of coal spontaneous combustion hazard in goaf area and fire source location inversion identification method, 2020-01 to 2023-12,

References
[1] Huang, Z.X. (1998) Extensions to the K-means algorithm for clustering large data sets with categorical values. J. Data Mining and Knowledge Discovery., 2 (3): 283-304.
[2] Chu, L. lu., Jiang, F. (2014) K-modes clustering algorithm based on weighted overlap distance. J. Computer CD Software and Application., (01):140-140.
[3] Hong, J., Yuu-ming, C. (2014) A new distance for unsupervised learning of categorical data. C. In Proceedings of International Joint Conference on Neural Networks.
[4] Li, W., Feng, Y., Li, D. (2016) Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. J. Automatic Control and Computer Sciences, 50 (4) : 271-277.
[5] Wang, Y.S., Zheng, X.F. (2016) An integrated classifier based on rough set attribute measurement. J. Journal of Engineering Sciences, 038(003): 425-431.
[6] Zhao, L., Liu, J.H., Zhang, Z.Z. (2017) K-modes clustering algorithm based on Bayes distance. J. Computer Engineering and Science, 39(1): 188-193.
[7] Wen, W., Li, P.Q. (2018) KNN Classification Algorithm Based on K Center Point and Rough Set .[J]. Computer Engineering and Design , 39(11): 3930-3934.
[8] Yuan, F.(2020) Research on hybrid data clustering algorithm based on K-modes algorithm . D.
[9] Lu, R.Q., Ma, F, Zhang, T.F. (2018) Rough K-means clustering algorithm based on interval 2-type fuzzy metrics J. Pattern Recognition and Artificial Intelligence, 31(003):265-274.
[10] Pawlak, Z. Rough Sets, Int. J. Comput. Informat. Sci, 11(5): 341-356.
[11] Pawlak, Z., Grzymala, B. J., Slowinski, R. (1995) Rough Sets, Communication of the ACM,38(11): 89-95[28].
[12] Shannon, C. E. (1948) The mathematical theory of communication. M. Bell System Technical Journal,(3-4):373-423.