Efficient EK-means: Extended K-means Clustering for Categorical data with High Processing Speed

Chen Xi* and Peng Shuo

Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, School of computer and communication engineering, Changsha University of Science & Technology, No. 960, Wanjiali South Road, Hunan Province, China

*E-mail: chenxi_cscu@csust.edu.cn

Abstract. The typical representative of the hard clustering algorithm, K-means, is one of the fastest processing algorithms with good scalability. However, it cannot deal with categorical attributes, which is one of the important indicators to measure the pros and cons. Due to the lack of processing capabilities on categorical attributes, k-means has a large limit on data processing capabilities. This paper proposes a clustering algorithm extends from K-means. This algorithm introduces the concept of a Pseudo-mean distance calculation formula and a counting table so that categorical attributes can be processed while reducing the time cost as much as possible. Experimental results illustrate the proposed Pseudo-means can extend the processing range of k-means to category-type data, and the counting table also effectively reduces the time cost.

1. Introduction

The categorical attribute has a strong internal relationship while disorder and no numerical analysis. Values of a categorical attribute can often be divided into several types which are far less than the amount of data. How to complete the analysis of such valuable data has been widely concerned by scholars all over the world. One of the solutions is clustering.

Clustering is an irreplaceable method in data analysis, especially for categorical attributes. The primary goals are to group similar patterns into the same cluster and discovering the meaningful structure of the data[1].

There are many traditional clustering algorithms: K-modes, k-means, etc. With the advantages of low time complexity, high accuracy, and high readability, k-means has become one of the most representative traditional clustering algorithms. Unfortunately, it limits numeric-only.

The rest of the paper is organized as follows. In Section 2, we present related work of categorical data respectively. In Section 3, we describe the details of Ek-means and its shortcomings. In Section 4, we propose a solution to the defect of Ek-means and the experimental results for UCI datasets. In Section 5, we give conclusions.

2. Related work

For categorical data, the number is huge, the connotation is strong, and the application is wide[2-4]. Therefore research on how to achieve efficient processing of categorical attributes has been ongoing since the 1990s, like BIRCH[5], CLARANS[6], and other methods[7-10]. In this section, we review
the common study on categorical attributes clustering.

Ralambondrainy proposed a conceptual version of the K-means algorithm in which all multi-category attributes are converted to binary attributes and treated as numbers containing only 0 and 1[11].

Huang advanced the Hemingway distance to address the problem of k-means fails to handle categorical data[12]. The concept of the prototype was introduced soon later in [13], which realized the hybrid data clustering problem of numerical and categorical data.

Sudupto Guha presented a novel method that measured similarity between a pair of objects, using the concept of links instead of clusters, to avoid errors in the calculation of distances caused by boolean and categorical data[14].

Yangdong Ye proposed an sIB algorithm that can be used for Category attribute processing[15]. Through the expansion and binary processing of data attributes, the joint distribution calculation of Ye’s method is carried out based on the appearance of attribute values.

K-modes is an extension of K-means. Since proposed, it has been widely recognized for efficient computing speed and good processing ability on category attributes. Due to the sensitive problem of initialization, a lot of research on initial clustering problems has been continued on k-modes.

Zhou Caiying proposed a method, in which the efficiency of fuzzy K.prototype operations improved by changing the initial value and the maximum number of iterations[16].

Liang Bai presented a new categorical attribute initialization method in [17]. It overcomes the shortcomings of the categorical attribute initialization method by combining the distance and density to select the initial center.

Li Tao-Ying advanced a new distance based on RD (Relationship Degree) to obtain a good initial clustering center and a better effect of clustering on indirectly related data[18].

S. Saranya introduced a new fuzzy K-pattern clustering improved model in [19]. It ensures that K initial cluster centers are not selected from the same cluster by adopting improved initial distance and initial entropy initialization methods.

Besides, there are many other types of research on categorical attribute clustering, such as reducing time costs through big data platforms[20].

Although there is a lot of research, most of them are focused on the improvement of the initial center and a little on distance. The most effective algorithm, K-Modes, ignores the integrity of clusters also. Hence, it is very necessary to propose a new method for categorical data which overcomes the shortcomings of distance calculation.

3. EK-means

As we know, the structural data are stored in a table, where each row (tuple) represents facts about an object. More formally, a categorical data table is defined as a quadruple \( IS = (U, A, V, f) \), can be formatted as follows:

- \( U \) is the nonempty set of objects, called a universe, \( U = \{X_1, \ldots, X_n\} \).
- \( A \) is the nonempty set of attributes, \( A = \{A_1, \ldots, A_m\} \).
- \( V \) is the union of attribute domains, i.e., \( V = \bigcup_{A_i \in A} A_i \), where \( V_{A_j} \) is the value domain of attribute \( A_j \) and it is finite and unordered.
- \( f: U \times A \rightarrow V \) is an information function such that for any \( A_i \in A \) and \( X \in U \), \( X \in U \).

The objective of Ek-means is to cluster \( U \) into \( k \) clusters by minimizing the function

\[
P(W, Q) = \sum_{i=1}^{k} \sum_{q \in Q} w_q d(X_i, Q_q)
\]  

Subject to
\[ \sum_{i=1}^{k} w_{il} = 1, \quad 1 \leq i \leq n \]
\[ w_{il} \in \{0,1\}, \quad 1 \leq i \leq n, 1 \leq l \leq k \]

Where \( k \leq n \) is a known number of clusters. \( W = [w_{il}] \) is a \( k \)-by-\( n \) matrix and \( Q_i \) is the vector of the \( i \)-th cluster. The process of EK-means can be described as Figure 1.

**Algorithm: EK-means**

**Input:**
- \( k \): the number of Cluster
- \( U \): the Data set

**Output:** Sets of \( K \) Clusters

**Step:**
1. Select \( K \) initial center points;
2. Allocate an object to a center point whose distance is smallest according to (3) to form initial clusters;
3. Allocate an object to the cluster whose distance is smallest according to (6);
4. Repeat 3 until no object has changed clusters after a full cycle test of the whole data set

**Figure 1. Process of EK-means.**

### 3.1. Initial Clustering

Since the cluster is calculated as a whole in the distance calculation of EK-means, it is necessary to use Hemingway distance to form initial clusters. We select \( k \) objects as the initial center points and then use (3) to calculate the distance between all objects in the data set and each initial center point. Through it, the objects to be clustered can be allocated to the closest center point as the initial cluster.

\[ d(X,Y) = \sum_{i=1}^{n} \delta(X_i,Y) \]

Where

\[ \delta(X_i,Y) = \begin{cases} 0 & X_i \neq Y_j \\ 1 & X_i = Y_j \end{cases} \]

Here, \( X_j \) stands for the value of the \( j \)-th attribute of \( X \) while \( Y_j \) stands for object \( Y \).

### 3.2. Pseudo-mean distance

Supposing a numeric object \( X \) including \( m \) categorical attributes, and the cluster \( C_i = \{Y_{i1},...,Y_{ij},...,Y_{ik}\} \) including \( s \) data like \( X \). According to the definition of mean, the mean of differences is equal to the differences of the mean. Hence, the average distance between \( X \) and \( C_i \) can be shown as formula(5).

\[ d = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{X_{ij} - Y_{ij}}{s} = \sum_{i=1}^{k} \frac{\sum_{j=1}^{m} Y_{ij}}{s} \]

Promote the concept of mean, "Pseudo-mean distance" in the category attribute can be indicated as:

\[ D(X,Q) = \sum_{j=1}^{m} \left( 1 - \frac{b_j}{a_j} \right) \]

Here, \( a_j \) stands for the number of objects having the \( j \)-th attribute, \( b_j \) indicates the number of objects having the same value of \( X \). \( Q \) stands for the vector of the current cluster.
Thus, in this paper, we can find the distance between object $X$ and the entire cluster rather than “center” or some particular points.

$$Q = \begin{pmatrix} q_1 \\ \vdots \\ q_n \\ \vdots \\ q_m \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} x_{i1}, x_{i2}, \ldots, x_{in} \\ \vdots \\ x_{i1}, x_{i2}, \ldots, x_{in} \\ \vdots \\ x_{i1}, x_{i2}, \ldots, x_{in} \end{pmatrix}$$  \hspace{1cm} (7)

The cost function can now be expressed as:

$$P(W, Q) = \sum_{i=1}^{c} \sum_{j=1}^{n} w_i \left( 1 - \frac{b_i}{a_j} \right)$$  \hspace{1cm} (8)

where $w_i \in W$.

### 3.3. The advance of the distance method

Currently, k-modes is the most effective and widely used algorithm of those developed from k-means. The dissimilarity measure of K-modes can be easily understood and avoid complex calculations, but there is still a big flaw. An example will be given below.

As shown in Table 1, we assume that there is a clustering table including categorical attributes—COLOR, TYPE, and AREA. Each record has a class label indicating which class it belongs to.

| Objects | Attributes | COLOR | TYPE | AREA | Class |
|---------|------------|-------|------|------|-------|
| X_0     | r          | C     | N    | D_1  |
| X_1     | r          | T     | B    | D_1  |
| X_2     | r          | T     | N    | D_1  |
| X_3     | b          | C     | N    | D_1  |
| X_4     | b          | T     | B    | D_1  |
| X_5     | b          | T     | N    | D_2  |
| X_6     | w          | C     | B    | D_2  |
| X_7     | w          | T     | B    | D_2  |
| X_8     | w          | C     | N    | D_2  |

It is easy to know that the distance between $X_0$ and $D_1$ is 6/5 and the distance between $X_0$ and $D_2$ is 7/4, then its ownership remains. The same to know is that the distance between $X_5$ and $D_1$ is 8/5, and the distance between $X_5$ and $D_2$ is 7/4, so move $X_5$ to class $D_1$. Repeat the above calculation until no object has changed clusters after a full cycle test of the whole data set. The final result generated by EK-means is $D_1 = \{X_0, X_1, X_2, X_3, X_4, X_5\}$, $D_2 = \{X_6, X_7, X_8\}$ and the number of iterations $t=1$ while the result generated by K-modes is $D_1 = \{X_0, X_1, X_2, X_3, X_5\}$, $D_2 = \{X_4, X_6, X_7, X_8\}$, $t=1$. And there is a case that TYPE=“C” and TYPE=“T” be considered as the same. TYPE loses its value in clustering, i.e., the meaning of the TYPE is ablated.

Through comparative analysis between Pseudo-mean distance and Hemingway distance, advantages of the former can be summarized as:

- Considered as a whole, which is more in line with the definition and target of clustering.
- Invalidation of the attribute can be avoided so that all attributes can be used more effectively.

### 3.4. The defect of EK-means

By introducing the concept of Pseudo-mean distance, EK-means extends the k-means algorithm to
categorical data. However, after preliminary experiments on Mushroom Dataset, the results shown in Figure 2 are obtained.

![Figure 2](image)

**Figure 2.** Computational times for different numbers of objects.

To solve this problem, Efficient EK-means proposed.

4. **Efficient EK-means**

Inspired from SLIQ[21] and SPRINT[22], a solution of EK-means’ defect will be introduced in this section. Also, the experimental analysis will be shown in the last part with the comparison with K-modes.

4.1. **List structure**

| Attribute 1 | Attribute 2 | … | Attribute m |
|-------------|-------------|---|-------------|
| Amount of Value 1 1 | Amount of Value 2 1 | … | Amount of Value m 1 |
| Amount of Value 1 2 | Amount of Value 2 2 | … | Amount of Value m 2 |
| … | … | … | … |
| Amount of Value 1 n | Amount of Value 2 n | … | Amount of Value m n |

**Table 2.** The structure of the count table.

Algorithm: Efficient EK-means

| Input | Algorithm | Output |
|-------|-----------|--------|
| k: the number of Cluster | Efficient EK-means | Sets of K Clusters |
| U: the Data set | | |

**Step:**
1. Select K initial center points;
2. Allocate an object to a center point whose distance is smallest according to (3) to form initial clusters;
3. Create an initial count table for each cluster;
4. Allocate an object to the cluster whose distance is smallest according to (6);
5. Update the count table of the cluster;
6. Repeat 4 and 5 until no object has changed clusters after a full cycle test of the whole data set

![Figure 3](image)

**Figure 3.** Process of Efficient EK-means.
There is no more difference between EK-means and the Efficient one, but the special list structure used to reduce the times of go through the data set. The list represented as table 2 used to record the number of each value of each attribute.

The initial count table is established after completing the statistics on the initial cluster. In a loop, rather than go through all the data, the counter table needs to update each value only. In EK-means, the cluster vector \( \mathbf{Q} \) is expanded into the form of a counter table.

Hence, the process of Efficient EK-means changed to Figure 3. In Step 1-2, we select K center points and use Hemingway distance to form initial clusters. In Step 3, we create initial count tables to record the condition of the attributes’ domain. After entering a loop, we use (6) to calculate the distance and update the counter table.

For a better introduction, the examples in table 1 will be used again. Table 3 and table 4 are the initial count table, in which NoC stands for the number of each value in the domain of COLOR and NoT, NoA stands for it of the corresponding attribute.

| Table 3. Initial count table of D1. | Table 4. Initial count table of D2. |
|------------------------------------|------------------------------------|
| COLOR    | NoC | TYPE | NoT | AREA | NOA |
| r        | 3   | C    | 2   | N    | 3   |
| w        | 0   | T    | 3   | B    | 2   |
| b        | 2   |      |     |      |     |

| COLOR    | NoC | TYPE | NoT | AREA | NOA |
| r        | 0   | C    | 2   | N    | 2   |
| w        | 3   | T    | 2   | B    | 2   |
| b        | 1   |      |     |      |     |

After an iteration of clustering, the count table changed to table 5 and table 6. Rather than go through the dataset, Efficient take a movement only. For instance, move one “r” of COLOR from D2 to D1 with one “C” and one “N”.

| Table 5. Count table of D1. | Table 6. Count table of D2. |
|-----------------------------|-----------------------------|
| COLOR    | NoC | TYPE | NoT | AREA | NOA |
| r        | 3   | C    | 2   | N    | 4   |
| w        | 0   | T    | 3   | B    | 2   |
| b        | 3   |      |     |      |     |

| COLOR    | NoC | TYPE | NoT | AREA | NOA |
| r        | 0   | C    | 1   | N    | 1   |
| w        | 3   | T    | 2   | B    | 2   |
| b        | 0   |      |     |      |     |

It is easy to know that the time complexity of the algorithm is \( O(n) = n^*t(1 + O(e)) \), where \( n \) is the amount of data, and \( t \) is the number of iterations. \( O(e) \) is the time cost of the attribute and value search. In the worst case, all attribute values of \( X \) are the last one recorded in the table; in the best case, all attribute values of \( X \) are the first in the record. Supposing that there are \( m \) attributes and all attributes have at most \( x \) values, the time complexity of search can be expressed as \( O(e) = 0.5*m*x \). Therefore, the time complexity of Efficient EK-means can be expressed as:

\[
O(n) = n^*t \left(1 + \frac{1}{2} * m * x \right)
\]

(9)

4.2. Experimental analysis
There are two types of analysis on data sets—— Mushroom, Car Evaluation, Balance Scale and the Congressional Voting downloaded from UCI.

4.2.1. Performance analysis. For data sets of categories known as the actual types, the accuracy of the clustering algorithm can be calculated using (10), where \( n \) is the size of the data set and \( a_r \) is the number of objects allocated to the right cluster.

\[
\sum_{r=1}^{k} \frac{a_r}{n}
\]

(10)
Since the accuracy of clustering is sensitive to the initial center, different results may be caused by different initial centers. Therefore we use a method of averaging 50 accuracy values and gives the variance to determine the stability in the experiment. All missing attribute values are treated as special values.

Each object of the Mushroom Dataset is a sample of mushrooms. Each record has a class label indicating whether mushrooms are toxic or non-toxic. This data set contains 8124 records, of which the toxic mushroom is 4208 and the non-toxic mushroom is 3916.

| Table 7. The performance on Mushroom Dataset. |
|-----------------|-----------------|-----------------|-----------------|
| Accuracy | Variance | Maximum | Minimum |
| K-modes | 0.68455818 | 0.019379356 | 0.894633 | 0.550951 |
| Efficient EK-means | 0.8516448 | 0.008576154 | 0.897774 | 0.566224 |

The Congressional Voting Dataset is a record of 1984 US Congress votes. Each tuple in the data set represents a member's vote on 16 issues. All attributes are Boolean, and their values are Yes (represented by y) and No (represented by n). Each tuple has a class label: Republican or Democrat. The data set contains 435 tuples, 168 of which are Democratic voters and 267 are Republican voters.

| Table 8. The performance on Congressional Voting Dataset. |
|-----------------|-----------------|-----------------|-----------------|
| Accuracy | Variance | Maximum | Minimum |
| K-modes | 0.858355692 | 5.68442E-05 | 0.866667 | 0.850575 |
| Efficient EK-means | 0.873739846 | 1.28747E-06 | 0.875862 | 0.871264 |

The Car Evaluation Data set was derived from a simple hierarchical decision model. The model evaluates cars according to 6 categorical attributes and divides cars into 4 parts named unacc, acc, good, vgood. This data set contains 1728 records, of which the unacc is 1210, the acc is 384, the good is 69, and the vgood is 65.

| Table 9. The performance on Car Evaluation Data set. |
|-----------------|-----------------|-----------------|-----------------|
| Accuracy | Variance | Maximum | Minimum |
| K-modes | 0.858355692 | 5.68442E-05 | 0.866667 | 0.850575 |
| Efficient EK-means | 0.873739846 | 1.28747E-06 | 0.875862 | 0.871264 |

The Balance Scale Data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. This data set contains 625 records, of which the left is 291, the right is 291 and another is 43.

| Table 10. The performance on Balance Scale Data set. |
|-----------------|-----------------|-----------------|-----------------|
| Accuracy | Variance | Maximum | Minimum |
| K-modes | 0.545846154 | 0.001573054 | 0.5904 | 0.4704 |
| Efficient EK-means | 0.560523077 | 0.001520037 | 0.6304 | 0.4784 |

From Table 7 to Table 10, it’s clear that Efficient EK-means is better than K-modes in both accuracy and stability.

4.2.2. Scalability analysis. To test the scalability of the new algorithm, we choose the Mushroom dataset from UCI. The computational results are performed by using a computer equipped with AMD A6-4400M and 4G RAM. The computational times of the proposed algorithm are plotted with respect to the number of objects, clusters, and m*x (the product of the number of attributes and the number of maximum fields in the attribute).
Make the number of attributes is 22 and the number of clusters is 2, the computation time against the number of objects can be expressed as Figure 4. From Figure 4, it can be known that the time of data processed by Efficient EK-means increases linearly with the increase of the amount of data similar to K-Modes, and much faster than EK-means.

Set the number of clusters equals 2 and the number of objects 8124, the relationship between \( m \times x \) and the computational time shown as Figure 5. From Figure 5, we can know that the time cost increasing with the rise of \( m \times x \) slower and slower.

Make the number of attributes is 22 and the number of objects is 8124. From Figure 6, we can find that the change in the value of K does not cause a significant increase in time.

5. Conclusion
Category attributes worth analyzing on a realistic level even different from numerical attributes, characterized by the inability to perform mathematical operations. K-means is an almost perfect algorithm except dealing with categorical data. Although K-modes creatively provides a way for solving the clustering of category attributes, the existing unreasonable calculation methods still plague researchers. Through theoretical analysis and experimental results, this paper proves that the EfficientEK-means algorithm can deal with the clustering problem of category attributes well in both stability, accuracy and processing speed by changing the distance calculation method and introducing a new table structure.

References
[1] Jiawei Han and Micheline Kamber. DATA MINING CONCEPTS and Techniques. (Peking: China Machine Press), p 223-225
[2] M.Jupri and Riyanarto Samo.Taxpayer Compliance Classification Using C4.5, SVM , KNN, Naive Bayes and MLP. Int. Information and Communications Technology (ICOIACT), 2018.9.10
[3] Wang Yu and Xu Lining.Research on Text Categorization of KNN Based on K-means for Class Imbalanced Problem. Int. Instrumentation & Measurement, Computer, Communication, and Control, 2016
[4] Farideh Sanei, Abbas Harifi, and Shahram Golzari. Improving the Precision of the KNN Classifier using Nonlinear Weighting Method based on the Spline Interpolation. Int. Computer and Knowledge Engineering (ICCKE), 2017.
[5] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 1996.
[6] Raymond T. Ng and Jiawe Han. CLARANS: A Method for Clustering Objects for Spatial Data Ming. IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, 2002, 14(5):1003-1016.

[7] Cao Changhu and Li Yafei. Improved Chameleon Algorithm for Cluster Analysis. SCIENCE TECHNOLOGY AND ENGINEERING, 2010, 10(33):8160-8166.

[8] Luo Min. Novel Graph Partition Based Clustering Algorithm—GAGPBCUK. COMPUTER SCIENCE, 2012, 39(9):162-169.

[9] WenBin Liang, ChangDong Huang, and JianHuang Lai. Weighted Numerical and Categorical Attribute Clustering in Data Streams. Int. Neural Networks (IJCNN), 2017.

[10] Dong Huang, Changdong Wang, and Jianhuang Lai. Locally Weighted Ensemble Clustering. IEEE Transactions on Cybernetics, 2018, 48(5):1460-1473.

[11] H. Ralambondrainy. A conceptual version of the K-means algorithm. Pattern Recognition Letters 16(1995), 1147-1157.

[12] Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Ming and Discovery 2, 283-304.

[13] Huang, Z. Clustering large data sets with mixed numeric and categorical values. Proceedings of the 1st Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore: World Scientific, pp. 21–34.

[14] Sudipto Guha, Ra Rajeev Rastogi, and Kyuseok Shim. A Robust Clustering Algorithm for Categorical Attributes. Information System, 2000, 345-366.

[15] Ye Yang-dong, HE Xi-dian, and JIA Li-min. CD-sIB: A Kind of sIB Algorithm Orient to Categorical Data. ACTA ELECTRONICA SINICA, 2009, 10:2165-2172.

[16] Zhou C and Huang L. The improvement of the initial point selection method for Fuzzy K-Prototype clustering algorithm. Int. Education Technology and Computer. IEEE, 2010:V4-549-V4-552.

[17] Bai L, Liang J and Dang C.A cluster centers initialization method for clustering categorical data. Expert Systems with Applications An International Journal, 2012, 39(9):8022-8029.

[18] Li T Y, Chen Y and Jin Z H. Initialization of k-modes clustering for categorical data. Int. Management Science and Engineering. IEEE, 2013:107-112.

[19] Saranya S and Jayanthi P. A modified initialization method to find an initial center for Fuzzy K-modes clustering. Intelligent Techniques in Control, Optimization and Signal Processing. IEEE, 2018.

[20] Guo T, Ding X, and Li Y. Parallel k-modes algorithm based on MapReduce. 3rd Int. Digital Information, Networking, and Wireless Communications. IEEE, 2015:176-179.

[21] Mehta M, Agrawal R, and Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining. Int. Extending Database Technology: Advances in Database Technology. Springer Verlag, 1996:18-32.

[22] Shafer J C, Agrawal R, and Mehta M. SPRINT: A Scalable Parallel Classifier for Data Mining. Proc Vldb, 1996:544-555