Improvement of *K*-means clustering algorithm based on MIP optimization

IOP Conf. Series: Journal of Physics: Conf. Series **1053** (2018) 012100 doi:10.1088/1742-6596/1053/1/012100

Improvement of *K*-means clustering algorithm based on MIP optimization

Wenbing Chang, Xinglong Yuan and Shenghan Zhou[1]

School of reliability and systems engineering, Beihang University, China

[1]Email: zhoush@buaa.edu.cn

Abstract. The *k*-means algorithm is a widely used partition clustering algorithm. The traditional *k*-means algorithm has two problems: it is easy to fall into the local optimal solution; it is very sensitive to the initial solution. In this paper, a *k*-means algorithm model based on mixed integer linear programming is established. The experiment shows that the effect of the new algorithm is better than the traditional *k*-means algorithm, and the above two problems are solved well.

1. Introduction

1967, MacQueen [1] put forward the *k*-means algorithm. He summarized the research achievements of Cox, Fisher and sebertyen, and summarized *k*-means algorithm in detail. *K*-means clustering algorithm has a strong interpretative and simple expression, so the efficiency of processing data is very high. *K*-means clustering has been widely used in various fields because of its advantages. However, with the study of *k*-means clustering algorithm, some inherent defects of the algorithms are gradually emerging.

From the process of *k*-means algorithm, we can see that the whole process is a partial optimization process, not the strategy of global optimization. In 2012, Wang Qian et al. made a summary of the current *k*-means clustering algorithms, found that the *k*-means algorithm can be considered as a NP-hard problem, that is, it will always fall into the local optimal solution [2]. Bighnaraj N et al. proposed a particle swarm optimization based clustering algorithm (PSOBC) to avoid local optimal cluster center in cluster analysis [3]. AJM Rani and L Parthipan used Improved Particle Swarm Optimization (IPSO) +*K*-means document clustering algorithm, the proposed solution generates more accurate, robust and better clustering results when compared with *K*-means [4].

K-means clustering algorithm can't produce an accurate and unique result. The main reason is that when we are doing arithmetic operations, the initial centers chosen are randomly generated, so the final result will lead to different clustering results. Davies (2002) [5]points out that such a result is not credible. Y. Chen et al. proposed a GEP automatic clustering algorithm with dynamic penalty factors, this algorithm combines penalty factors and GEP clustering algorithm, and doesn't rely on any priori knowledge of the data set [6]. Xuemei W introduced Artificial Bee Colony algorithm based on *K*-means, then put forward an improved Artificial Bee Colony algorithm combined with *k*-means clustering algorithm at the same time, the experiments showed that the method has solved algorithm stability of *k*-means clustering algorithm well [7].

H Pirim et al. think that integer programming models for clustering have applications in diverse fields addressing many problems such as market segmentation and location of facilities, which are flexible in expressing objectives subject to some special constraints of the clustering problem [8]. Burcu S et al. presents a mixed-integer programming based clustering approach with the objective of

minimizing the maximum cluster diameter among all clusters [9]. In 2015, Zhang Jie, Dong Jianrui and Xiao Yiyong improved the *k*-means clustering algorithm by using method of P-partition, a new method, which proved efficient and globally optimal [10]. Yanchi L et al. presented a detailed study of 11 widely used internal clustering validation measures for crisp clustering [11]. The purpose of this paper is to propose a mathematical programming model based on MIP (mix integer linear programming model)for the classical clustering partition method k- means, in order to solve the problem of the sensitivity of the initial solution and the problem of easily falling into the local optimal solution. In this paper, we use the sum of distance from each point to its cluster center as a criterion for judging the effect of clustering.

2. Model

Using different distance computing methods, the *k*-means clustering algorithm is essentially constant. This paper modeling by Manhattan distance. Manhattan distance: Calculation of distance between two data objects using Manhattan distance. Represented by data of two dimensional attributes:

$$d_{ij} = \sqrt{\left(d_{ij}^{x}\right)^2 + \left(d_{ij}^{y}\right)^2} \tag{1}$$

d_{ij}: distance between two data objects; d_{ij}^{x}, d_{ij}^{y}: difference between *x, y* attributes.

2.1. Traditional k-means algorithm model

We set up the *k*-means algorithm as a mathematical programming model. The *k*-means algorithm can be summed up as the optimization problem of the intra group distance, that is, the minimum inner distance of the optimized group. The grouping of data objects and class centers can be represented by decision variables w_{ik}, Q_k respectively. The distance between data objects for class centers can be expressed as $d(X_i, Q_k)$. We use the symbols shown in table 1 to describe the *k*-means clustering problem. The mathematical programming model of Manhattan distance can be described as follows:
Minimize:

$$\min F(W,Q) = \sum_{k=1}^{l}\sum_{i=1}^{n} w_{ik} d(X_i, Q_k) \tag{2}$$

Subject to:

$$\sum_{k\in K} w_{ik} = 1\,, \forall i \in N \tag{3}$$

$$d(X_i, Q_k) = \sum_{j=1}^{m} \delta\left(x_{ij}, q_{kj}\right), \forall i \in N, j \in M, k \in K \tag{4}$$

$$\delta\left(x_{ij}, q_{kj}\right) = \left|x_{ij} - q_{kj}\right|, \forall i \in N, j \in M, k \in K \tag{5}$$

The target function (2) express the sum of the distance between all the data objects and the center of their class. w_{ik} is the decision variable, while $d(X_i, Q_k)$ is determined by the decision variable Q_k, so the objective function is composed of two decision variable products, not a linear expression. At the same time, only the decision variable w_{ik}=1 is satisfied (that is, the data object *i* is subordinate to the *k* class), distance $d(X_i, Q_k)$ (that is, the distance between the data object *i* and the class center of class *k*) will be added to the target function, which ensures that the target function is optimized by the intra cluster distance. The optimal direction of the objective function is to minimize the sum of the intra group distance. Constraint condition (3) is the data object *i* that can only and must be divided into a cluster / class. Constraint condition (4) is the sum of the distance between the data object points of the data set and the class center. Constraint condition (5) represent the distance calculation between various attributes.

In order to solve such an integer programming model, the method of partial optimization is generally adopted to solve the problem. Because there are two decision variables in the objective function, the method is used to fix one of the variables by rotation and then the whole integer programming model is solved. The whole process can be summarized as follows:

(1) set the initial l initial class center.

(2) fix the value of Q, then solve the integer programming model, and find the most suitable W to minimize the target function $F(W,Q)$.

(3) fix the value of W, then solve the integer programming model, and find the most suitable Q to minimize the target function $F(W,Q)$.

(4) repeated iteration steps (2) and step (3) until the target function $F(W,Q)$ are no longer optimized.

Table 1. The symbol definition of the traditional K-means clustering algorithm.

symbol	explanation
N	A collection of data objects that need to be clustered
M	Set of numeric attributes
K	A set of clusters (classes)
n	he number of data objects in the set of the required cluster data objects
m	The number of attributes in a set of numeric attributes
l	The number of clusters in a cluster
i	The index of the collection of data objects
j	Index of a set of numeric attributes
k	Index of cluster sets $k \in K$
X_i	X_i represents the i data object point in the data object set and is described by m numeric attributes, $X_i=\{x_{i1}, x_{i2},\ldots,x_{im}\}$.
Q_k	The class center of the k cluster (class) is also described by the m numeric attributes, $Q_i=\{q_{k1}, q_{k2},\ldots,q_{km}\}$
W	Class member decision variables set
w_{ik}	Class member decision variables, if the w_{ik}=1, data object X_i belongs to class k, and vice versa
$d(X_i, Q_k)$	The distance between the class center of the class k and the data object i.

2.2. k-means algorithm model based on MIP optimization

A decision variable d_{ikj} is used instead of the product of the two decision variables of the original objective function (2). That is to say, the new decision variable d_{ikj} should contain the membership information of the data object of decision variable w_{ik} and the information of data object to the class center distance of decision variable $d(X_i, Q_k)$. This can ensure that the optimized target function is the sum of the intra group distance. The optimal k-means clustering model based on MIP is described with the symbols in table 2.

It can be seen that only when the data object i is subordinate to the class k, the Manhattan distance between the data object i and the class k is calculated, and the d_{ikj} value for the rest of the case is 0. This restriction ensures that the sum of the decision variables d_{ikj} is the sum of the intra group distance that is expected to be optimized. So the foundation of the k-means clustering model based on MIP optimization is established in Manhattan. Therefore, the optimal k-means algorithm model based on MIP in Manhattan is established as follows:

Minimiz：

$$\min F(d) \sum_{k=1}^{l} \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ikj} \tag{6}$$

Subject to:

$$d_{ijk} \geq \left| x_{ij} - u_{kj} \right| - T(1 - w_{ik}), \forall i \in N, k \in K, j \in M \tag{7}$$

$$d_{ijk} \geq 0, \forall i \in N, k \in K, j \in M \quad (8)$$

$$\sum_{k=1}^{l} w_{ik} = 1, \forall i \in N \quad (9)$$

$$\sum_{i=1}^{n} w_{ik} \geq 1, \forall k \in K \quad (10)$$

In the above constraints, Constraint conditionals (7) and (8) define the distance between each data object and its subordinate class center in each attribute. Formula (7) in $T(1\text{-}w_{ik})$ to ensure that only w_{ik}=1 (data i belonging to the class k), the variable d_{ikj} to $|x_{ijt}\text{-}u_{kit}|$. If the w_{ik}=0 (data object i does not belong to the class k), then the d_{ikj} can only be valued at 0. So this ensures that the objective function (6) is the sum of the intra group distance. Constraints (9) guarantee that each data object must belong to one category and can only belong to one category. Constraints (10) guarantee that there must be at least one data object point (class member) within each category.

In this way, by objective function (6) and conditional (7) - (10), we can build the k-means clustering algorithm in Manhattan to the mixed integer linear programming model (MIP). It can be seen that all expressions are linear, so the model can be solved directly by the branch and bound method.

Table 2. The symbol definition of K-means model based on MIP optimization.

symbol	explanation
N	A collection of data objects that need to be clustered
n	Set the number of data objects in N to meet n=card (N)
i,	The coefficient of data objects, meet $i \in N$
M	A collection of data object attributes
m	The number of attributes in the attribute set M, which satisfies the m=card(M)
j	Properties of the coefficient, meet $j \in M$
A_j	The range of the property J
K	Cluster / class set
l	The number of classes required to be gathered in a class set, l=card (K)
k	Coefficients in class sets, $k \in K$
T	A large number
w_{ik}	The 0/1 decision variable, if w_{ik} =1, indicates that the data object i belongs to the k class
u_{kj}	The value of the attribute J of the class k class of the decision variable on the value of the range A_j.
x_{ij}	The value of the attribute j of the data object i on the range A_j.
d_{ikj}	The distance between the center of the data object i and the class center of the k class it belongs to in the attribute j is defined. The distance from the data object can be based on the following rules. $d_{ikj} = \begin{cases} \lvert x_{ij} - u_{ik} \rvert, & if\ w_{ik} = 1 \\ 0, & otherwise \end{cases}$

3. Experiment

3.1. Experimental environment

The programming language AMPL is used to compile the mathematical programming model, and the widely used MIP solver CPLEX is called to use, and the CPLEX version of the call is 12.6.1.0. AMPL software reads the model and data file, and calls the solver CPLEX to solve the problem

3.2. Experimental method and data

Using the total distance between all points and the cluster centers that they belong to as criterion for judging the effect of clustering. The smaller the total distance is, the better the clustering effect is. In order to verify that our proposed algorithm solves two disadvantages of the traditional *k*-means algorithm, we use the *k*-means algorithm model based on linear integer programming and the traditional *k*-means algorithm model respectively to cluster the same group of data ten times. Because the models currently built are based on two-dimensional numerical data, so we use the dataset called 'Syn. data' in two-dimensional plane as an example. There are 10 data objects on the plane, which need to be grouped into 3 types. The coordinates of each data point are shown in table 3.

Table 3. 'Syn. data' coordinate value.

No.	*X*	*Y*
1	0.303333	0.368724
2	0.24815	0.167396
3	0.46615	0.712839
4	0.18097	0.451399
5	0.80228	0.794894
6	0.97858	0.744839
7	0.61091	0.253925
8	0.62399	0.066260
9	0.55559	0.824103
10	0.92293	0.515927

3.3. Results

Table 4 is the clustering results of the two models. The objective function mean value (the distance between all cluster centers and the points that they belong to) that is solved by the improved model is 1.61633, and the result is constant. The objective function mean value of the traditional *k*-means clustering method is 2.40676, variance is 0.71019. So, the effect of the *k*-means algorithm based on mixed integer programming is better than the traditional *k*-means algorithm. The result of our new model do not change as the initial cluster centers change.

Table 4. The sum of distance from each point to its cluster center.

No.	MIP	traditional
1	1.61633	2.76666
2	1.61633	2.32581
3	1.61633	3.81044
4	1.61633	1.70775
5	1.61633	1.90053
6	1.61633	2.15732
7	1.61633	2.01180
8	1.61633	4.09500
9	1.61633	1.75835
10	1.61633	1.53395
mean value	1.61633	2.40676
variance	0	0.71019

4. Conclusions

There are two problems in the traditional *k*-means algorithm: (1) it is easy to fall into the local optimal solution. (2) it is very sensitive to the initial solution. In this paper, a *k*-means algorithm model based on mixed integer programming is established, and the two problems mentioned above are solved well. The experimental results show that the new model can obtain the optimal solution and is not sensitive to the initial solution.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No.71501007 & 71672006). The study is also sponsored by the Aviation Science Foundation of China and the Graduate Student Education & Development Foundation of Beihang University.

References

[1] J Macqueen 1967 Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Berkeley: University of California Press) **1** 281-297
[2] Q Wang, C Wang, ZY Feng and YE Jin-Feng 2012 *China. Electronic Design Engineering* **7** 21-24
[3] Bighnaraj N, Subhra S, Dayal K B, Sarita M and Bharat K P 2013 *Conf. on Computing and Communication Systems* Durgapur, India
[4] Rani A J M and Parthipan L 2014 *Conf. on Sustainable Energy and Intelligent Systems* Tiruchengode, India
[5] Davies 2002 Pattern recognition and image preprocessing (Boca Raton :CRC Press) **57** 8-9
[6] Chen Y, Li K and Yang L 2016 *Journal of System Simulation*
[7] Xuemei W and Jinbo W 2014 *Applied Mechanics and Material* **556-562** 3852-3855
[8] H Pirim, B Eksioglu and FW Glover 2018 Journal of Optimization Theory and Applications. **2** 1-17
[9] Burcu S, F. Sibel S, Serpil S and Metin T 2006 *European Journal of Operational Research.* **3** 866-879
[10] J Zhang, J Dong and Y Xiao 2015 *Conf. on The 27th Chinese Control and Decision Conference* Qingdao
[11] Yanchi L, Zhongmou L, Hui X, Xuedong G and Junjie W 2010 *Conf. on IEEE International Conference on Data Mining* Sydney, Australia