An Extenics-based Criteria Clustering Method

Xingsen Li, Haolan Zhang
Research Center on Intelligent Computing and Data Management
Ningbo Institute of Technology, Zhejiang University,
Ningbo, China
e-mail: lixs(haolan.zhang)@nit.zju.edu.cn

Wei Deng
College of Information Science & Technology
University of Nebraska
Omaha NE 68182, USA
e-mail: wdeng@unomaha.edu

***Abstract*—Clustering has been applied in many field of management for better decision making with a lot of algorithms such as K-means. Based on Extenics, we found that most of algorithms calculate the similarity of elements in a certain set by distance to each other; they focus on the position of each element and neglect their criteria. However, in the real world, there are usually exist criteria to score the elements. Therefore, we present a new clustering method. In our method, we use distance in Extenics for similarity calculating based on criteria, and compared a simple case with traditional K-means algorithm. The results show that our method is more practical and has much potential value for data mining and knowledge management.**

Keywords- clustering; Extension distance; criteria clustering; Extenics; K-means

I. INTRODUCTION

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other clusters [1]. It is a main task of exploratory data mining, and a common technique for statistical data analysis widely used in many fields, including knowledge management, machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics et al. Cluster analysis is also an iterative process of knowledge discovery or interactive multi-objective optimization.

There are many clustering algorithms. K-means clustering [2] is one of the most popular methods for cluster analysis aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

The k-means problem [2-3] is to find cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center. Although finding an exact solution to the k-means problem for arbitrary input is NP-hard, the standard approach to finding an approximate solution is used widely and frequently finds reasonable solutions quickly. However, most current clustering algorithms including the k-means algorithm has at least two major theoretic shortcomings:

First, Clustering algorithm score the similarity of each other by the distance between each other, a density threshold or the number of expected clusters depend on the individual data set and neglect their criteria or the business goal. For example, it's qualified for math test if score >= 60, 59.5 and 60 is quite similar but totally different according to qualification. This often leads to misunderstandings between researchers coming from varies of fields, since they use the same terms and often the same algorithms, but have different goals.

Second, a kind of criteria usually is a interval. But in real variable function, all the distance is 0 if a point is in the interval. This can not make a distinction between A and B both in the interval by their distance.

To overcome such shortages, the purpose of this paper is to propose a new clustering method that would support criteria oriented cluster both theoretically and practically. The rest of the paper is organized as follows. Section 2 present a new definition of extension distance based on Exteincs. Section 3 put forward a theory framework and process of criteria clustering. Section 4 introduces a case study and compares our method with traditional k-means method, followed by a brief summary and some future research scopes in Section 5.

II. EXTENSION DISTANCE IN EXTENICS

A. Distance in classical mathematics

The distance between point x and point y on a real number axis is: $\rho(x,y)=|x-y|$

The distance between point x on a real number axis and finite interval $X=<a,b>$ is:

$$d(x,X)=\begin{cases}0, & x\in X\\ \inf\limits_{y\in X}\rho(x,y), & x\notin X\end{cases}$$

For convenient, the expression method of interval $<a, b>$ in this paper can indicate an open interval, a closed interval, or a half-open and half-closed interval.

The distance between point $x=(x_1,...,x_n)$ and $y=(y_1,...,y_n)$ is:

$$d(x,y)=\sqrt[2]{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

It's specified by the basic concept of "distance" that "the distance between any point in interval and the interval is zero". Therefore, it materially specifies the qualitative description that cannot express the "quantitative change" and "qualitative change" of a point in the interval.

B. *Definition of extension distance in Extenics in 1D*

Extenics is a science initiated by Professor Cai Wen in *1983*[4]. It is at the intersection of mathematics, philosophy, and engineering [4, 14]. Extenics focuses on solving contradictory problems. It is based on modeling and remodeling, on transforming and retransforming until getting a reasonable solution to apparently an unreasonable problem [5-6]. Everything is dynamic; so we have dynamic structure, dynamic classification, and dynamic change. Appling these ideas to data mining [8], knowledge management [9] and innovation [10-11], we structured a lots of new models and methods [12, 13].

In order to describe the difference of points in the same interval, we difine the distance between point x and interval X=<a,b> is specified [6,7]:

Definition 1. Suppose x is any point in real axis, and X =<a,b> is any interval in real field, we define

$$\rho(x, X) = \left| x - \frac{a+b}{2} \right| - \frac{b-a}{2} \tag{1}$$

as the distance between point x and interval X, where <a,b> can be an open interval, a closed interval, or a half-open and a half-closed interval.

As to any point x_0 on real axis, we have:

$$\rho(x_0, X) = \left| x_0 - \frac{a+b}{2} \right| - \frac{b-a}{2} = \begin{cases} a - x_0, & x_0 \le \dfrac{a+b}{2} \\ x_0 - b, & x_0 \ge \dfrac{a+b}{2} \end{cases}$$

The association between point-interval distance $\rho(x,X)$ in Extenics and "point-interval distance" $d(x,X)$ in classical mathematics is [6]:

① when $x \notin X$ or x= a, b, $\rho(x, X)$= $d(x, X) \ge 0$;

② when $x \in X$ and $x \neq a$, b, $\rho(x, X) < 0$, $d(x, X) = 0$.

The introduced concept of extension distance can precisely score the locational relation between a point and an interval in quantitative form. When the point is in the interval, it's considered by classical mathematics that the distance between any point and the interval is 0; while in Extenics, the difference of locations of the point in the interval can be described according to different values of distance.

C. *Extension Distance in Extenics in 2D*

Professor Florentin Smarandache presented the Extension *2D*-Distance formula as following [5] and as shown in fig 1:

$$\rho((x_0, y_0), AMBM) = d(P(x_0,y_0), A(a_1,a_2)MB(b_1,b_2)N)$$
$$= |PO| - |P'O|$$
$$= \sqrt{(x_0 - \frac{a_1+b_1}{2})^2 + (y_0 - \frac{a_2+b_2}{2})^2} - \sqrt{(\frac{a_1-b_1}{2})^2 + (y_{P'} - \frac{a_2+b_2}{2})^2}$$
$$= \pm |PP'|$$
$$= \pm \sqrt{(a_1 - x_0)^2 + (y_{P'} - y_0)^2}$$

where $$y_{P'} = y_0 + \frac{a_2 + b_2 - 2y_0}{a_1 + b_1 - 2x_0}(a_1 - x_0)$$

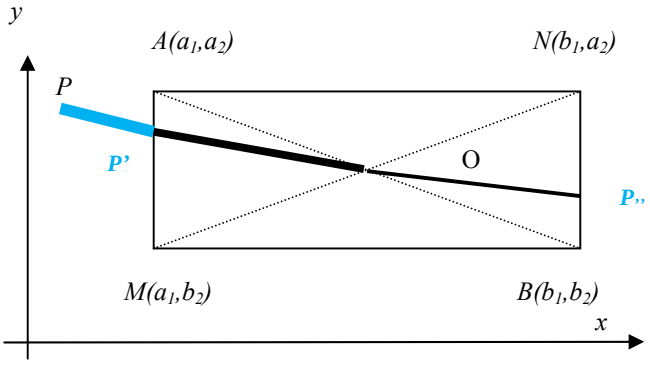

Figure 1. Extension Distance in 2D

III. CRITERIA CLUSTERING METHOD BASED ON EXTENSION DISTANCE

A. *Basic Ideas of Criteria Clustering*

Data is a kind of description of matter, actions or relations of the world with certain criteria. For example, we can descript blood status with Blood pressure and Hemoglobin levels (HB), for male adults, normal Blood pressure is between 90~140 mmHg for systolic pressure, 60~90 mmHg for diastolic pressure and normal HB is between 120~160 g/L. under criteria of systolic pressure, 85 is quite different from 90 although their distance is only 5(90-85 = 5), while 90 is similar to 138 although their distance is 48(138-90 = 48).

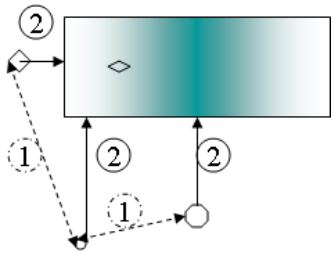

Figure 2. Two kinds of distances for clustering

So there are two kinds of distances as shown in fig 2. ① is distance between each other; ② is the distance to the criteria. **Extenics-based criteria clustering** is the task of grouping a set of objects in such a way that objects in the same group are more similar in extension distance to the criteria than others to the criteria.

B. *Procedure and algorithm of Criteria Clustering*

A criterion clustering is suitable for the field of application in which the data has certain meanings and has an interval of criteria. The procedures are as following:

1) Define the interval of criteria: The interval can be satisfied interval and best fit interval acording to the difinition in Extenics[10].

2) Calculate each extension distance to the criteria and store the distance in D[n].

3) Select K data as center point from D[n] by random. k<=n.

4) Calculate each distance of the rest point to the center and classify them to nearest center in D[1…k].

5) Re-calculate each group's center and store the new center in C[k].

6) Repeat step 4 and 5, until the new center keep the same or the change of value less than a small data given in the beginning.

In step 4 to 6, besides distance, we also can use density to score the value of the center. The exact algorithm is as follows:

- For each data point x, compute extension distance D(x).
- Choose *k* centers uniformly at random from among the data points.
- Use D(x) as the data to compute the distance between x and the chosen centers.
- Now that the initial centers have been chosen, precede using standard k-means clustering.
- Repeat Steps 2 to 4 until k centers keep the same or less than certain value.

IV. EXAMPLE AND COMPARATIONS

In order to explain our method, we give a simple example and compare the clustering result with traditional K-means. The normal criteria of blood pressure is (90,140), the normal criteria of blood pressure is (15, 20). According to formula (1), we compute the extension distance D(x) separately as shown in table 1.

TABLE I. TESTING DATA FOR CRITERIA CLUSTERING

No	Blood pressure	Urine density	D(x) of B	D(x) of U
1	115	20	-25	0
2	137	14	-3	1
3	128	13	-12	2
4	148	16	8	-1
5	152	22	12	2
6	120	17	-20	-2
7	91	19	-1	-1
8	102	18	-12	-2

K-means clustering on original data based on distance with 3 clusters of sizes 2, 3, 3 and 4 clusters of sizes 2, 2, 2, 2 is shown in table 2.

TABLE II. CLUSTERING WITH TRANDITIONAL K-MEANS

No	Blood pressure	Urine density	k=3	k=4
1	115	20	3	2
2	137	14	2	3
3	128	13	3	3
4	148	16	2	1
5	152	22	2	1
6	120	17	3	2
7	91	19	1	4
8	102	18	1	4

The figure of original data is shown in fig 3 and the figure of extension distance is shown in fig 4.

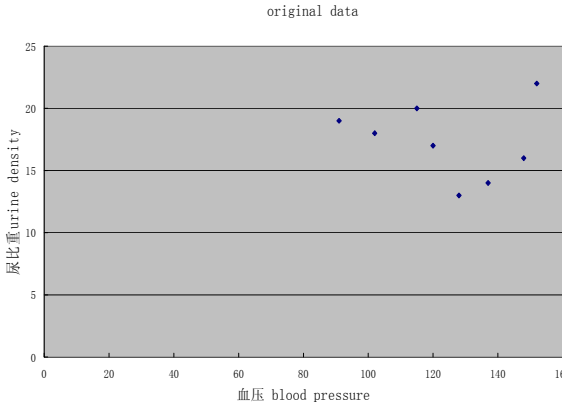

Figure 3. figure of original data for clustering

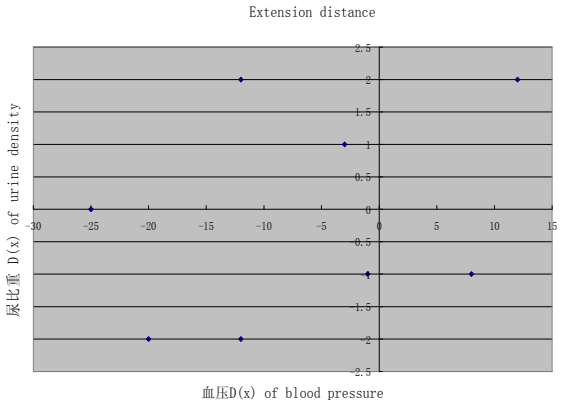

Figure 4. figure of extension distance for clustering

K-means clustering based on extension distance with 3 clusters of sizes 2, 3, 3 and 4 clusters of sizes 2, 2, 2, 2 is shown in table 3.

TABLE III. CRITERIA CLUSTERING WITH K-MEANS ON EXTENSION DISTANCE

No	D(x) of B	D(x) of U	k=3	k=4
1	-25	0	1	3
2	-3	1	3	1
3	-12	2	1	4
4	8	-1	2	2
5	12	2	2	2
6	-20	-2	1	3
7	-1	-1	3	1
8	-12	-2	1	4

Experts' judgments show that the result of criteria clustering is more practical with treatment of health care. This cluster can guide how to transform those who not qualified to be qualified.

V. CONCLUSIONS AND FUTURE WORK

In this paper we analyze the shortcomings of clustering by distance, then by giving a short description of Extenics and its extension distance, we present a new clustering method based on extension distance related to criteria. Further more, we present a framework of its process and algorithm in 1*D* distance. To explain our method, we select 8 records as an example and make comparations by K-means on distance and extension distance. The result shows that criteria clustering method is more suitable to guide transformation from unqualified to be qualified.

However, in this paper, we do criteria clustering in testing use extension distance separately in 1*D*, actually, 2*D* distance formula is more suitable for the criteria clustering. This is also one of the future research works. We will compute the value of a point and its Dependent Function of a point from *1D* to *2-D*, then to *3D* and n-*D* spaces in the future.

In summary, our research is just begin and the method is in primary stage, how to consider criteria well clustering needs further research, for many kinds of data has its normal intervals and abnormal intervals. The final aim of our research is to promote our life better, from unhappy to be happy, unqualified to be qualified and failure to success. More theories will be applied in our methods, such as basic element theory, extension set theory and extension logic.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (#71271191, #70871111), Scientific research project (#2014SCG204), Zhejiang Research Institute of Education Science, Zhejiang Soft Science Research Program (#2013C35085) and the Scientific Research Project (#JG2013300, #Y201122111), Education Department of Zhejiang Province.

REFERENCES

[1] A.K. Jain, M.N. Murty and P.J. Flynn: Data Clustering: A Review, ACM Computing Surveys, vol.31, no.3, pp.264-323, Sep.1999.

[2] D. Arthur and Vassilvitskii, S., How slow is the k-means method?, "Proceedings of the twenty-second annual symposium on Computational geometry", ACM New York, NY, USA: pp.144–153, 2006

[3] Y.YANG, XIANG Chang-cheng, WEI Dai-jun.Data cluster based on extension K nearest neighbor algorithm. Computer Engineering and Applications, vol.46, no.21, 2010, pp.156-159.

[4] C. Yang, Cai Wen. Extension Engineering. Beijing. Science press, July 2006

[5] Florentin Smarandache, Generalizations of the Distance and Dependent Function in Extenics to 2D, 3D, and n-D, "Global Journal of Science Frontier Research (GJSFR)" [USA, U.K., India], vol. 12, Issue 8, 2012, pp. 47-60

[6] C.Yang, W. Cai, *Extenics: Theory, Method and Application*, Science Press, Beijing. 2013

[7] Q. Li and X. Li, The Method to Construct Elementary Dependent Function on Single Interval, Key Engineering Materials vols. 474-476 2011, pp. 651-654

[8] W. Cai, C. Yang, W. Chen and X. Li, *Extensible set and extensible data mining*. Science Press, Beijing. Aug. 2008

[9] X. Li, Y. Shi, L. Zhang, From the information explosion to intelligent knowledge management, Beijing: Science Press, April 2010

[10] C. Yang, X. Li, Research Progress in Extension Innovation Method and its Applications, *Industrial Engineering Journal*, vol.15, no.1, 2012, pp.131-137

[11] Z. Zhou, X. Li, Research on Extenics-based innovation model construction and application of enterprise independent innovation, *Studies In Science of Science*, vol. 28,no.5, 2010, pp.769-776

[12] X. Li, H. Zhang, Z. Zhu, Z. Xiang, Z. Chen, Y. Shi, An Intelligent Transformation Knowledge Mining Method based on Extenics, *Journal of Internet Technology*, vol.14, no.2, 2013,pp.315-325

[13] X. Li, Y. Tian, F. Smarandache and R. Alex, An Extension Collaborative Innovation Model in the Context of Big Data, *International Journal of Information Technology & Decision Making*, Online Ready : pp. 1-23, Vol. 13 ,2014, DOI: 10.1142/S0219622014500266

[14] W. Cai, C. Yang, F. Smarandache, L. Vladareanu, Q. Li, G. Zou, Y. Zhao and X. Li, Extenics and Innovation Methods, communications in cybernetics , systems science and engineering, CRC Press, Balkema, Aug. 2013