RESEARCH ARTICLE OPEN ACCESS

Selection of Initial Seed Values for K-Means Algorithm Using Taguchi Method as an Optimization Technique

Aparna K*, Mydhili K Nair**
*Associate Professor, Dept of MCA, BMS Institute of Technology, Bangalore, Karnataka, INDIA
**Associate Professor, Dept. Of ISE, M S Ramaiah Institute of Technology, Bangalore, Karnataka, INDIA

ABSTRACT

This paper proposes an enhancement of the performance of the traditional K-Means algorithm of Partitional clustering by using Taguchi method as an optimization technique. K-Means algorithm requires the desired number of clusters to be known in priori. Given the desired number of clusters, the initial seed values are selected randomly. The K-means algorithm does not have any specific mechanism to choose the appropriate initial seeds and selecting different initial seeds may generate huge differences in terms of the clustering results. This is true especially when the target sample contains many outliers. In addition, random selection of initial seeds often makes the clustering to fall into local optimization. Therefore, it is very important to select appropriate initial seeds when using the traditional clustering method. Hence, to select the appropriate initial seeds, we propose the use of Taguchi method in this paper as a tool. Using the Taguchi method, the initial seed values are selected based on the signal-to-noise ratios and with these values as the initial input, the K-Means algorithm can be continued to form the clusters for the given data.

Keywords: K-Means algorithm, Taguchi method, Clustering Technique, Partitional Clustering, Data Mining, Orthogonal Array

I. INTRODUCTION

Clustering high dimensional data is an observable fact in real-world data mining applications. Developing efficient clustering techniques for high dimensional dataset is a challenging role because of the curse of dimensionality [6]. The aim of clustering is to find structure in data and is therefore exploratory in nature. Clustering has a long and rich history in a variety of scientific fields. One of the most popular and simple clustering algorithms is K-Means. K-Means validity measure is based on the intra-cluster and inter-cluster distance measures which allows the number of clusters to be determined automatically. The main disadvantage of the K-Means algorithm is that the number of clusters ‘K’ must be supplied as a parameter [4]. Moreover, to start with the K-Means algorithm, initial seed values based on the number of clusters required are selected randomly. There is no specific mechanism to select these initial seed values and hence the results can vary for every iteration or for every set of initial seed values. The objective of this paper is to make use of Taguchi technique to select the initial seed values. Once the seed values are selected, the K-Means algorithm is continued until the desired number of clusters are obtained. The initial seed values are identified by their rank based on signal-to-noise ratios using the Taguchi method [7]. The rest of the paper is organized as follows. The rest of the paper is organized as follows: Section II gives the overview of K-Means algorithm and Taguchi method. Section III illustrates the implementation of the work. The performance and results are shown in section IV and the conclusion is given in section V.

II. OVERVIEW OF TECHNIQUES USED

A. K-Means Algorithm

The K-Means algorithm is composed of the following steps

Step 1: Place K points into the space represented by the objects that are being points represent initial group centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the K centroids.

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

K-Means clustering algorithm partitions the given data set into k number of clusters of points. That is, the K-Means algorithm clusters all the data points in the data set such that each point falls in one and only one of the k partitions. Points with the same cluster similarity are in the same cluster, while points with different cluster difference fall in different clusters. In clustering algorithms, points are grouped by some notion of “closeness” or “similarity.” In K-Means, the default measure of closeness is the

Euclidean distance. The Euclidean distance formula is

EUCLIDEAN DISTANCE(X, Y)

$$= (|X_1 - Y_1|^2 + |X_2 - Y_2|^{2+\cdots+} |X_{N-1} - Y_{N-1}|^{2+} |X_N - Y_N|^2)^{1/2}$$

B. Taguchi Method

Design of Experiments (DOE) is a collection of statistical methods for studying the relationships between input variables (independent variables) and their interactions on a response variable (dependent variable) [2]. Taguchi method provides simple, efficient and systematic approach to optimize designs for performance, quality and cost. This method combines the engineering and statistical knowledge to optimize design and manufacturing processes to achieve high quality at low cost and time [1]. Taguchi design is an important tool for robust design. It helps to identify the process parameters and their levels to put quality characteristic on target and to limit number of experiments for optimization [3]. In this paper, Taguchi Design of Experiments (DOE) approach has been used to analyze the selection of initial seed values of Step 1 in the above K-Means algorithm.

III. EXPERIMENTAL DESIGN

The experiment using Taguchi method was considered to calculate the initial seed values from a huge data set in order to predict the value of relative humidity. Relative Humidity is calculated based on the values of Temperature, Dew Point temperature and two constant values based on August-Roche-Magnus Approximation [5] (taken from Clausius-Clapeyron Relation). Given these values the Relative Humidity is calculated using the formula

RELATIVE HUMIDITY:

$$RH = 100 \frac{exp(\frac{aT_d}{b+T_d})}{exp(\frac{aT}{b+T})}$$

Where T is the Temperature in °C, Td is Dew point temperature in °C. The range of values considered valid based on August-Roche-Magnus Approximation is

$0°C < T < 60°C$
$0°C < T_d < 50°C$

The K-Means algorithm was then used to form clusters based on the similar values of the four factors considered. Table I below shows four factors and three levels used in the experiment. If three levels were assigned to each of these factors and a factorial experimental design was employed using each of these values, number permutations would be 3^4. The fractional factorial design reduced the number of designs to nine. The orthogonal array of L_9 type was used and is represented in Table II below.

TABLE I
LEVEL OF EXPERIMENTAL PARAMETERS

Symbol	Factor	Level		
		1	2	3
A	Temperature	23	28	33
B	Dew Point	20	25	27
C	Constant a	17.212	17.368	17.625
D	Constant b	235.48	238.88	243.04

TABLE II
TAGUCHI'S L_9 (3^4) ORTHOGONAL ARRAY

Std. order	Factors			
	A	B	C	D
1	23	20	17.212	235.48
2	23	25	17.368	238.88
3	23	27	17.625	243.04
4	28	20	17.368	243.04
5	28	25	17.625	235.48
6	28	27	17.212	238.88
7	33	20	17.625	238.33
8	33	25	17.212	243.04
9	33	27	17.368	235.48

IV. PERFORMANCE AND RESULTS

Nine experiments were conducted based on the above values and the complete response table for these data is shown in Table III below.

TABLE III
EXPERIMENTAL DATA AND SAMPLE STATISTICS

Expt. No.	Observed Response values for RH			Mean	Standard Deviation	Logarithmic value of Std. Deviation	S/N Ratio
1	83.183	82.123	82.781	82.696	0.53505	-0.62540	43.7818
2	112.758	110.678	111.345	111.594	1.06211	0.06026	40.4294
3	126.935	127.987	125.123	126.682	1.44867	0.37065	38.8348

4	62.272	63.989	61.712	62.658	1.18645	0.17097	34.4545
5	83.409	83.001	82.990	83.133	0.23856	-1.43312	50.8434
6	94.370	95.190	94.559	94.706	0.42929	-0.84562	46.8726
7	45.883	44.761	45.083	45.242	0.57777	-0.54858	37.8758
8	63.616	64.109	63.923	63.883	0.24882	-1.39104	48.1901
9	70.595	70.231	71.092	70.639	0.43222	-0.83882	44.2668

In this paper, optimal humidity is taken as the indicator of better initial value. Therefore, the nominal is best is selected for obtaining initial values.

The S/N Ratio for nominal-is-best is calculated using the formula:

S/N Ratio = (10*Log10(Ybar**2/s**2))

The plot of factor effects on S/N Ratio is shown below in fig1.

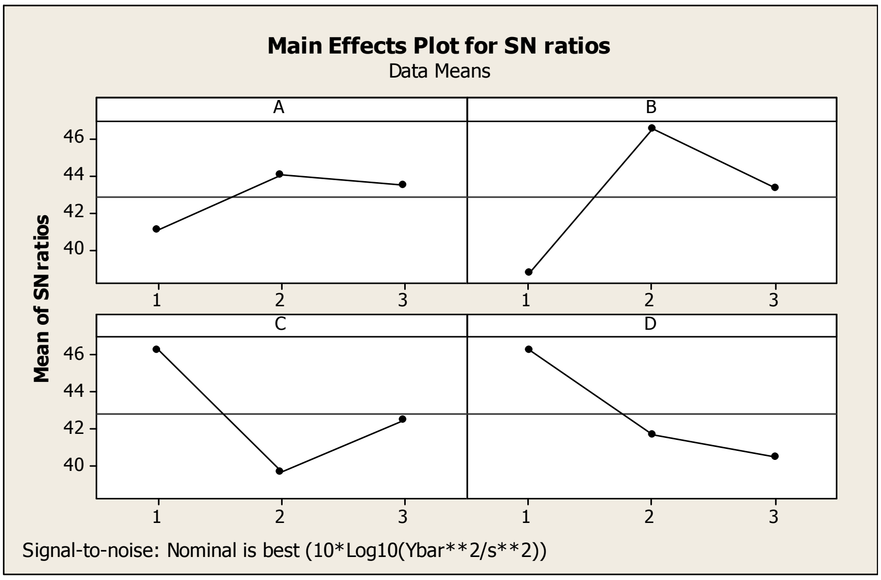

Fig 1. Plot of factor effects on S/N Ratio

The Taguchi analysis for prediction of optimal values is shown in the response tables below.

Response Table for Signal to Noise Ratios
Nominal is best (10*Log10(Ybar**2/s**2))

Level	A	B	C	D
1	41.02	38.70	46.28	46.30
2	44.06	46.49	39.72	41.73
3	43.44	43.32	42.52	40.49
Delta	3.04	7.78	6.56	5.80
Rank	4	1	2	3

Response Table for Means

Level	A	B	C	D
1	106.99	63.53	80.43	78.82
2	80.17	86.20	81.63	83.85
3	59.92	97.34	85.02	84.41
Delta	47.07	33.81	4.59	5.58
Rank	1	2	4	3

Response Table for Standard Deviations

Level	A	B	C	D
1	1.0153	0.7664	0.4044	0.4019
2	0.6181	0.5165	0.8936	0.6897
3	0.4196	0.7701	0.7550	0.9613
Delta	0.5957	0.2536	0.4892	0.5594
Rank	1	4	3	2

Predicted values

S/N Ratio	Mean	StDev	Ln (StDev)
35.4470	92.3112	1.17344	0.312732

Factor levels for predictions

A	B	C	D
1	1	3	2

Hence the predicted values by Taguchi method are level 1 for factors A and B, level 3 for factor C and level 2 for factor D. These values can be taken as the initial seed value for performing K-Means algorithm.

V. CONCLUSION

Taguchi's prediction gives the optimal results. Hence we can use these predicted values as the initial seed values for the K-Means algorithm. This gives better quality of clusters rather than selecting the initial seed values randomly. This also results in a standard mechanism before the final results, rather the clusters, are obtained.

REFERENCES

[1]. W. H. Yang, Y.S. Tang, Design optimization of cutting parameters for tuning operations based on Taguchi method, *Journal of Materials processing technology,* 84, 1998, pp 122-129

[2]. Howard S Gitlow, Alan J Oppenheim, Rosa Oppenheim, David M Levine, Quality Management, McGraw Hill, 2009, pp 427-429

[3]. Jaharah A Ghani, et al., Philosophy of Taguchi Approach and method in Design of Experiment, *Asian Journal of Scientific Research,* 6(1), pp 27 – 37, 2013.

[4]. Joydeep Ghosh and Alexander Liu, *Journal of Machine Learning Research (JMLR)* Vol.6, pp. 1705-1749. 2005.-"K-Means-type Algorithms A Generalized Convergence Theorem and characterized Local Optimality"

[5]. Clausius-Clapeyron Equation – *Wikipedia* – Redirected from August-Roche-Magnus Approximation

[6]. *Data Mining: Concepts and Techniques*, Jiawei Han and Micheline Kambar, Elsevier, 2nd Edition

[7]. Mustafa Zaidi, Bushra A Saeed, I. Amin, Nukman Yusoff, "Experimental Data Mining Techniques", International Journal of Computer Science Issues, Vol 9, Issue 3, No.3, May 2012

[8]. Huei Chun Wang, Chih-Chou Chiu, Chao-Ton Su, "Data Classification using the Mahalanobis – Taguchi System", Journal of the Chinese Institute of Industrial Engineers", Vol 21, No, 6, pp. 608-618, 2004.