Unique Distance Measure Approach for K-means (UDMA-Km) Clustering Algorithm

WK Daniel PUN and ABM Shawkat ALI
Central Queensland University/School of Computing Sciences, Rockhampton, QLD, Australia

Abstract— **Clustering technique in data mining has received a significant amount of attention from machine learning community in the last few years and become one of the fundamental research areas. Among the vast range of clustering algorithms, *K-means* is one of the most popular clustering algorithms. The basic principle of the K-means algorithm is to know how different distance measure is defined. It is a critical issue for K-means users. For example, how can we select a unique distance measure method for an optimum clustering task? Our research provides a statistical based Unique Distance Measure Approach for K-means (UDMA-Km) to this issue. We consider 112 supervised datasets and measure the statistical data characteristics using central tendency measure. Those data characteristics are split using well known entropy method to generate the rules. Finally, the generated rules are used to select the unique distance measure for K-means algorithm. The experiment is conducted within 112 problems and 10 fold cross validation methods. The most significant contribution of our study is that a new algorithm was created and the new algorithm can be used and has been used to solve any clustering tasks very quickly and provide much better optimum performance.**

I. Introduction

In general, data mining can quite often be defined as finding hidden information, even explicit knowledge [1], in a very large, huge database [2] (sometimes referred as to a *data warehouse*). Clustering is one kind of useful data mining techniques [3].

Clustering is a type of categorization inflicted rules on a group of objects. A broad definition of clustering could be "the process of categorizing a finite number of objects into groups where all members in the group are similar in some manner". As a result, a *cluster* is a aggregation of objects. All objects in the same cluster have common properties (e.g. distance) whose are different to the objects laying in other clusters.

A comprehensive collection of data is called a *data warehouse*. After datum have been collected, raw data needs to be screened before they are ready for further analysis. If data has tendency to be clustered, we need to select a suitable clustering algorithm such as *K-means algorithm* to perform cluster analysis. The process of using clustering algorithm to analyze data for patterns and relationships is called *data mining*. Figure 1 on next page shows the process of our unique distance measure approach for K-means (UDMA-Km) model.

In 1967, MacQueen [4] developed *K-means clustering algorithm* for classification and analysis of multivariate observations. Since then, while clustering techniques have been studied extensively in the areas of Statistics, Machine Learning, and Data Mining [5], the K-means algorithm has been applied to many problem domains, including the area of data mining, and has become one of the most used clustering algorithms. Matteucci [6] even said that the K-means algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. In our research, the K-means clustering algorithm has been used as a tool for research experiment. We propose a unique distance measure approach model, named UDMA-Km, for K-means algorithm.

Matteucci [6] has also divided clustering algorithms into four groups that are *exclusive*, *overlapping*, *hierarchical*, and *probabilistic*. K-means is an exclusive clustering algorithm because with K-means clustering, data is clustered in an exclusive way. It is just simply true that if data has been arranged into a certain cluster then it can not be arranged into another cluster [7], [8], [9].

A significant element of a clustering algorithm is to measure the distances between data (object) instances, that is, the distances between data instances is very important in clustering. However, the selection of unique distance measure is not that easy. Although if the elements of the data distance vectors are all in the same physical units and the *Euclidean distance metric* could be applied to clusters similar data instances, the Euclidean distance can sometimes be misleading as different relative scalings can lead to different clusterings [6]. A unique distance measure is used for clustering purposes. Matteucci [6] also points out that different mathematical formulae, which are used to combine the distances between the single elements of the data feature vectors into a unique distance measure, can lead to different clusterings. Hence, we can see the importance and difficulty of the unique distance measure selection.

Firstly we obtain 112 supervised datasets for the experiment from University of California, Irvine (UCI) [15], USA. As we know nothing about the data, we start with an unsupervised classification, which is clustering, in MATLAB® [17] system environment. After the process of normalization based on squared Euclidean metric and City-block (Mahalanobis metric), we have two classes of data. And then four statistical data characteristics measures have applied in order to gain a data characteristics matrix. They are Mean, Standard Deviation, Skewness, and Kurtosis. The last step is to use decision tree rules to get our final results.

We start with a brief background to data mining, data warehouse, clustering, K-means algorithm, and distance measure, and briefly describe the method we propose to the problem. In the next section, we look at K-means algorithm, statistical data characteristics measure, and entropy based decision tree approach. In section 3, we show our experimental outcome

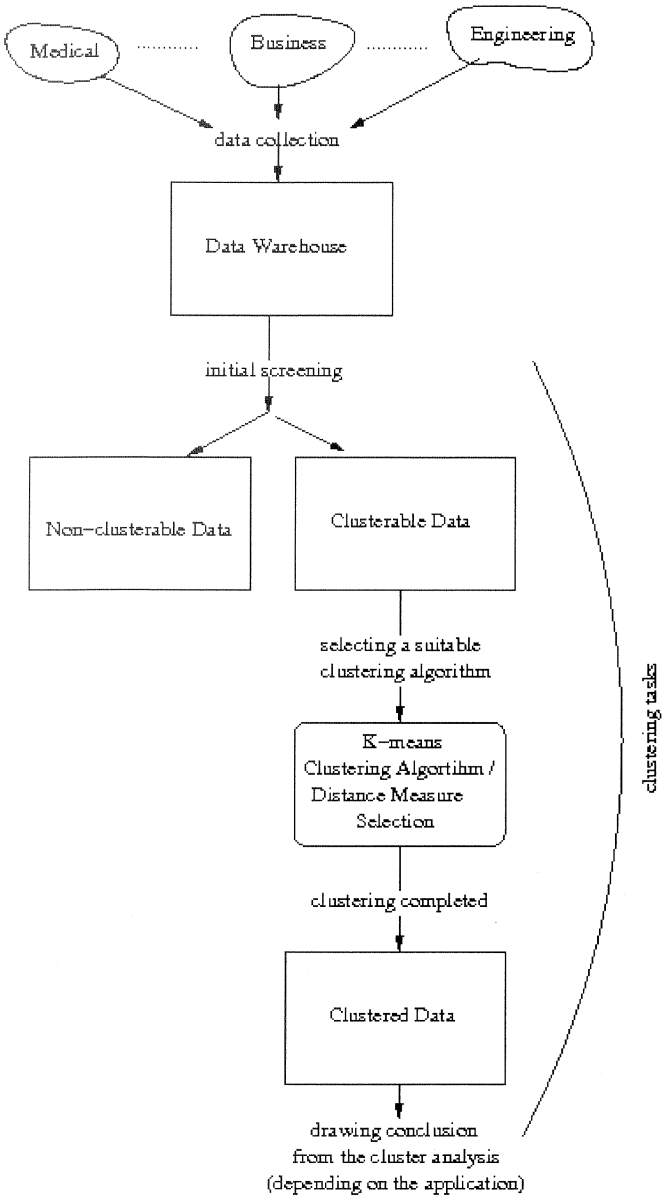

Fig. 1. Process of UDMA-Km model

and the rules for distance method selection. In the last section, we summarize with what we have done, the contribution, and conclude with some future research directions.

II. Algorithm Descriptions

A. K-means Algorithm

In K-means [4] algorithm, the center of a cluster represents the cluster, which is call *centroid*. The algorithm provides an easy way to cluster a given dataset through a number of k clusters. The main point is to define k centroids, one for each cluster. Due to a different location will cause a different result, it is better to put centroids as much as possible far away from each other. The next step we need to do is to take each data (object) in a given dataset and connect the object to the nearest centroid. The K-means algorithm start with giving initial values for means , $m_1, m_2, ..., m_k$, and enter into a loop, giving each item t_i to the cluster that has the closest mean and work out new mean for each cluster, the loop will be stopped if convergence criteria is met [2], [16]. Two things are important for K-means clustering: new center assignment and distance measurement. The common distance measure is squared Euclidean as follows:

$$D_e^2(X_i, m_k) = ||X_i - m_k||^2 \tag{1}$$

where D in the distance parameter and X is the data matrix.

We also adopted City Block distance measure in our experiment as follows:

$$D_c(X_i, m_k) = \sum_{i=1}^{k} |X_i - m_k| \tag{2}$$

It is a critical point, which distance measure is better for our existing clustering tasks. The following section will advise on this issue.

B. Statistical Data Characteristics Measure

Four statistical central tendency measures, *Mean*, *Standard Deviation*, and *Skewness and Kurtosis*, have been used in this research with our datasets to construct a data characteristics matrix. The following paragraph will explain the basic statistical terms.

Mean ($\overline{X}$): The sample mean estimates the population mean, commonly notated as $\overline{X}$. It is a measure of location in the same variable. It also considers all outliers values during the location measure. It may not appear representative of the central region for skewed datasets. It is especially useful as being representative of the whole sample for identify the characteristics of a variable by a few numbers [10].

$$\overline{X} = \frac{1}{n} \sum X_i \tag{3}$$

n is the sequence length.

Standard Deviation (σ): The standard deviation measures the spread of a set of data as a proportion of its mean. The larger the standard deviation indicates the distribution is more widely spread [10]. It is calculated by taking the square root of the variance and is generally symbolized by σ.

$$\sigma = \left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^{\frac{1}{2}} \tag{4}$$

Skewness (s): Skewness is a descriptive statistical measure about the normality of a dataset. When one tail of the distribution is longer than the other, it indicates the dataset is either highly positive or negatively skewed [11]. It can be define as follows:

$$s = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_k - \overline{X})^3}{\sigma^3} \tag{5}$$

Kurtosis (k): Any symmetric distribution could deviate from the normal distribution due to a heavy tail [12]. The deviation is measured by the coefficients of Kurtosis as follows:

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_k - \overline{X})^4}{\sigma^4} \tag{6}$$

Based on these above statistical measure we construct the statistical data characteristics matrix. And then an additional attribute is added towards the end with the data characteristics, which explain the distance method performance ranking.

C. Entropy based Decision Tree Approach

Finally we use entropy measuring method to find out which distance method is the better choice with K-means clustering for a specific dataset.

The concept of entropy [13] and [14] is actually quite simple. It is a measure of how much uncertainty there is in the information. Let us consider three possible classes A, B, and C. If each class has equal probability of occurrence, the entropy is:

$$\begin{aligned} &-p_A \log_2 p_A - p_B \log_2 p_B - p_C \log_2 p_C \\ &= -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} = 1.59 \end{aligned} \tag{7}$$

where p_i is the probability of occurrence of i. Since there are three possible outcomes with equal probability, the probability of each is (1/3). In general, if there are n possibilities with p_i the probability of the event i, the entropy $H(X)$ is given by:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{8}$$

The Greek letter sigma, $\sum$, simply indicates summation of all the values. Equation (8) can be extended to situations where we deal with the conditional probability, for instance, what is the entropy for attribute Y when we already know the outcome of attribute X? Involving two attributes X and Y, we have:

$$H(X) = -\sum p_x \sum p_{Y|X} \log_2 p_{Y|X} \tag{9}$$

where $p_{Y|X}$ is the conditional probability of Y given X. In Information Theory, the information content is maximum when the entropy is minimum. Once we have the entropy, it is a simple matter to construct a rule by figuring out what is the better distance method for K-means algorithm. Based on this entropy outcome we identify the important attribute in the statistical characteristics matrix. Then following the well known decision tree [14] structure we find out the rules to select a distance measure method for K-means algorithm. The rules are summarized in the next section.

III. Experimental Outcome

We observed different distance measures in K-means algorithm and have found a significant impact on choosing methods. For instance, on one hand, K-means performed 71.2% accuracy for UCI 'flare' data with Euclidean distance measure. But it showed 84.3% accuracy with City Block distance. On the other hand, for 'iris' data K-means with Euclidean distance measure showed 96% accuracy, but City Block showed 33.33% accuracy. Therefore, we feel it is very important to choose the right distance measure method for K-means algorithm. We have solved this issue by generating two rules using a statistical approach. The generated rules were verified by ten-fold cross validation and the percentage of accuracy is summarized with the rules. These rules are as follows:

- **Rule 1:** IF $k \leqq 8.4595$ THEN select Square Euclidean distance measure for K-means clustering.
- **Rule 2:** IF $k > 8.4595$ THEN select City Block distance measure for K-means clustering.

The above rules have been driven by setting a threshold value of k using entropy based method. The values remain for unseen datasets.

We set a default rule in our method, if any data does not satisfy the above rule then K-means will select Euclidean distance measure method. Since Euclidean distance method shows a better average performance comparing with City Block in our experiment. The average accuracy of these generated rules is 90.9%. The higher accuracy of this generated rules has made higher acceptance to select a unique distance measure method for K-means algorithm. After adopting these rules inside of the K-means algorithm, it becomes an automatic environment for clustering task.

IV. Conclusions

The main contribution in this study is that we have developed a new algorithm to optimize a distance measure method selection for the most popular clustering algorithm K-means. The analysis of this algorithm used a meta learning approach by constructing a data characteristics matrix. We have figured out that clustering problem is better suitable for which distance measure method. After that, we split the attributes with the help of well known entropy measure to generate the rules. Finally, the generated rules supervised the K-means algorithm to pick up the right distance measure method for the current clustering task. We consider a wide range of problems in our experiment. The generated rules were verified by 10 fold cross validation method and found the higher accuracy of these rules. Therefore, this research contributed for K-means algorithm as a faster algorithm and optimal clustering performance. The effectiveness and efficiency of the new algorithm are demonstrated by our experiments on MATLAB® system environment. We have already made a plan to extend this current research by considering more distance measure methods in across a different types of data from different domains with K-means algorithm as well as other distance based clustering algorithms.

V. Appendix

Dataset	*Dataset name*	*Instances*	*Attributes*	Classes
1	abalone	1253	9	3
2	adp	1351	12	3
3	adult+stret	20	5	2
4	adult-stret	20	5	2
5	allbp	840	7	3

Dataset	*Dataset name*	*Instances*	*Attributes*	Classes
6	ann1	1131	7	3
7	ann2	1028	7	3
8	aph	909	19	2
9	art	1051	13	2
10	australian	690	15	2
11	balance-sca	625	5	3
12	bcw	699	10	2
13	bcw_noise	683	19	2
14	bld	345	7	2
15	bld_noise	345	16	2
16	bos	910	14	3
17	bos_noise	506	26	2
18	breast-canc	286	7	2
19	breast-canc	699	10	2
20	bupa	345	7	2
21	c	1500	16	2
22	cleveland-heart	303	14	5
23	cmc	1473	10	3
24	crx	490	16	2
25	dar	1378	10	5
26	dhp	1500	8	2
27	DNA-n	1275	61	3
28	dna	2000	61	3
29	dna_noise	2000	81	3
30	dph	590	11	2
31	echocardiogram	131	8	2
32	flare	1389	11	2
33	german	1000	25	2
34	glass	214	10	6
35	h-d	303	14	2
36	hayes-roth	132	6	3
37	hea	270	14	2
38	hea_noise	270	21	2
39	heart	270	14	2
40	hepatitis	155	20	2
41	horse-23	368	23	2
42	horse-colic	368	28	2
43	house-votes-84	435	17	2
44	hyp	2847	16	2
45	hypothyroid	1265	26	2
46	iris	150	5	3
47	khan	1063	6	2
48	kr-vs-kp	1279	37	2
49	labor-neg	40	17	2
50	led-noise	1047	10	10
51	lenses	24	6	3
52	letter-a	1334	17	2
53	lung-cancer	32	57	2
54	lymphography	148	19	8
55	mha	1269	9	4
56	monk1	556	7	2
57	monk2	601	7	2
58	monk3	554	7	2
59	mushroom	1137	12	2
60	nettalk_str	1141	8	5
61	page-blocks	1149	11	5
62	pendigits-8	1399	17	2
63	pha	1070	10	5
64	phm	1351	12	3
65	phn	1500	10	2
66	pid	532	8	2
67	Pima	768	9	2
68	poh	527	12	2
69	post-operative	90	9	3
70	primary-tum	339	18	2
71	pro	1257	13	2
72	promoter	106	58	2
73	pvro	590	19	2
74	rph	1093	9	2
75	satimage	1351	11	6
76	shuttle-landing control	15	7	2
77	sick-euthyroid	1582	16	2
78	sma	409	8	4
79	smo	1855	9	2
80	smo_noise	1855	16	2
81	sonar	208	61	2
82	splice	1589	61	3
83	t_series	62	3	2
84	tae	151	6	3
85	tae_noise	151	11	2

Dataset	*Dataset name*	*Instances*	*Attributes*	Classes
86	thy	1887	22	3
87	thynoise	1132	11	3
88	tic-tac-toe	958	10	2
89	titanic	2201	4	2
90	tmris	100	4	2
91	tqr	1107	12	2
92	trains-transformed	10	17	2
93	va-heart	200	9	4
94	veh	846	19	4
95	veh_noise	761	31	4
96	votes_noise	391	31	2
97	waveform	5000	22	2
98	waveform_noise	5000	41	2
99	wdbc	569	31	2
100	wine	178	14	3
101	wpbc	199	34	2
102	xaa	94	19	4
103	xab	94	19	4
104	xac	94	19	4
105	xad	94	19	4
106	xae	94	19	4
107	xaf	94	19	4
108	xag	94	19	4
109	xah	94	19	4
110	xai	94	19	4
111	yha	1601	10	2
112	zoo	101	17	7

REFERENCES

[1] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*, New York: Oxford University, 1995.

[2] M. H. Dunham, *Data Mining Introductory and Advanced Topics*, Upper Saddle River, New Jersey: Pearson Education, Inc., 2003.

[3] H.-J. Mucha and H. Sofyan, "Nonhierarchical clustering," http://www.quantlet.com.mdstat/scripts/xag/html/xaghtmlframe149.html, 18 November 2003, accessed on 27th April, 2007.

[4] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1, pp. 281-297, 1967.

[5] O. R. Zalane, "Principles of knowledge discovery in databases" http://www.cs.ualberta.ca/%Ezaiane/courses/cmput690/slides/, Fall 1999, accessed on 27th April, 2007.

[6] M. Matteucci, "A tutorial on clustering algorithms," http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/, accessed on 27th April, 2007.

[7] A. Moore, "K-means and hierarchical clustering - tutorial slides," http://www-2.cs.cmu.edu/~awm/tutorials/keans.html, 2004, accessed on 27th April, 2007.

[8] B. T. Luke, "K-Means clustering," http://fconyx.ncifcrf.gov/~lukeb/kmeans.html, accessed on 27th April, 2007.

[9] T. Rashid, "Clustering," http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html, accessed on 27th April, 2007.

[10] D. L. Harnett and J. F. Horrell, *Data, Statistics, and Decision Models with Excel*, USA: John Wiley & Sons, Inc., 1998.

[11] J. Shao, *Mathematical Statistics*, New York: Springer-Verlag, 1999.

[12] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press, 1995.

[13] J. R. Quinlan, "Induction of decision trees". *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[14] R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufman Publishers, 1993.

[15] C. Blake and C. J. Merz, UCI Repository of machine learning databases, University of California, Irvine, CA, 2002. http://www.ics.uci.edu/mlearn/mlrepository.html.

[16] R. Zhang and A. I. Rudnicky, "A Large Scale Clustering Scheme for Kernel K-Means," *ICPR Proceedings of the 16th International Conference on Pattern Recognition*, vol. 4, p. 40289, 2002.

[17] MATLAB The Language of Technical Computing, 7.3.0.298 (R2006b), USA: The MathWorks, Inc., www.mathworks.com/patents, 2006.