A NOVEL THREE STAGED CLUSTERING ALGORITHM WITH A NEW SIMILARITY MEASURE

Jamil Al-Shaqsi and Wenjia Wang
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

ABSTRACT

This paper presents a novel three staged clustering algorithm and a new similarity measure. The main objective of the first stage is to create the initial clusters, the second stage is to refine the initial clusters, and the third stage is to refine the initial BASES, if necessary. The novelty of our algorithm originates mainly from three aspects: automatically estimating k value, a new similarity measure and starting the clustering process with a promising BASE. A BASE acts similar to a centroid or a medoid in common clustering method but is determined differently in our method. The new similarity measure is defined particularly to reflect the degree of the relative change between data samples and to accommodate both numerical and categorical variables. Moreover, an additional function has been devised within this algorithm to automatically estimate the most appropriate number of clusters for a given dataset. The proposed algorithm has been tested on 7 benchmark datasets and compared with 11 other commonly used methods including TwoStep, *k*-means, *k*-prototypes and some ensemble based methods including QMI, CSPA, HGPA, and MCLA. The experimental results indicate that our algorithm identified the appropriate number of clusters for the tested datasets and also showed its overall better clustering performance over the compared clustering algorithms.

KEYWORDS

Clustering, similarity measures, automatic cluster detection, centroid selection

1. INTRODUCTION

Clustering is the process of splitting a given dataset into homogenous groups so that elements in one group are much similar to each other than the elements in different groups. Many clustering techniques and algorithms have been developed and used in a variety of applications. Nevertheless, each individual clustering technique has its limits in some areas and none of them can adequately handle all types of clustering problems and produce reliable and meaningful results; thus, clustering is still considered as a challenge and there is still a need for exploring new approaches for clustering.

This paper presents a novel clustering algorithm based on a new similarity definition**.** The novelty of our algorithm comes mainly from three aspects, (1) employing a new similarity measure that we defined to measure the similarity of the relative changes between data samples, (2) being able to estimate the most probable number of the clusters for a given dataset, (3) starting the clustering process with a promising *BASE*. The details of these techniques will be described in Section 2. Section 3 gives the new definition of similarity measure. Section 4 presents the experiments and evaluation of the results. The conclusions highlighting the fundamental issues and the future research are given in the final Section.

2. A NOVEL CLUSTERING ALGORITHM

Based on the literature study, we proposed a three staged clustering algorithm (see Figure 1). The main objective of the first staged is to build up the initial clusters, the second stage is to refine the initial clusters, and the third stage is responsible to refine the initial *BASES*, if necessary. Another important good feature of our algorithm is to have a mechanism to estimate the number of cluster, which is done in preprocess.

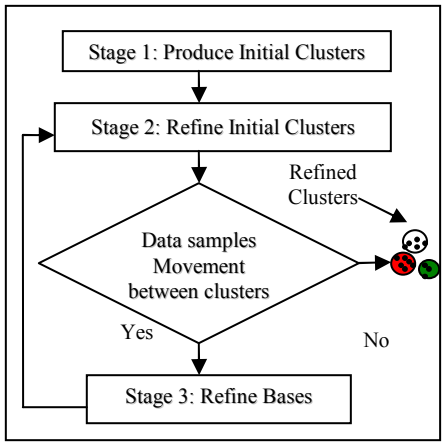

Figure 1. Framework of the proposed algorithm

3. First Stage

The first task in the first stage is to find a *BASE*. A *BASE* is a real sample that acts like a medoid or a centroid. The major steps in the first stage are:

1. Find a *BASE*
   1) Find a mode or a centroid
      - Find a mode (medoid) for each categorical feature. Calculate the frequency of each category for all the categorical features and then take the most frequent category in each feature as its mode.
      - Calculate the average (centroid) for each numerical feature.
   2) Construct a sample with the modes and centriods.
   3) Calculate the similarity between the constructed sample and all the samples in the dataset by using the proposed similarity measure (described in section 4.2).
   4) Select the sample that has the highest similarity value with the constructed sample as a *BASE*.
2. Calculate the similarity between the obtained *BASE* and the remaining samples.
3. Those samples that have similarity value higher than or equal to the set threshold will be assigned to the *BASE's* cluster.
4. If there are any samples that have not been assigned to any clusters then a new *BASE* is required.
5. Repeat steps 1 to 4 until no samples left.

3.1 Second Stage

The second stage commences by selecting the *BASE* of the second obtained cluster and calculates its similarity with all the samples in the first cluster. This is because the second *BASE* has not been used to calculate the similarity with the samples in the first cluster. Therefore, if any record has a greater similarity value than its original cluster, the record has to be moved to the second cluster. This process will go through all the remaining clusters.

3.2 Third Stage

The objective of the third stage is to refine the initial *BASES* to see whether the solution can be further improved. The main steps in this stage include:

1. Calculate the frequency of all categorical features in the first refined clusters.
2. Construct a mode/centroid sample by following the steps mentioned in stage 1.
3. Calculate the similarity between the constructed sample and the cluster's samples.
4. Select the new *BASE* whish is the sample most similar to the constructed mode/centroid sample.
5. Calculate the similarity between the new *BASE* and the cluster's samples.
6. Repeat steps 1 to 5 for the remaining refined clusters.
7. If the obtained *BASES* differ from the original ones, repeat the second stage; otherwise, the clustering process is terminated.
8. Repeat the third stage until no data sample is moved between clusters.

3.3 Automatically Estimating the Appropriate Number of Clusters

Determining the appropriate number of clusters is a critical and challenging task in clustering analysis. Sometimes, for the same dataset there may be possibly different answers depending on the purpose and criterion of the study [1]**.** We devised a mechanism as a component of our proposed algorithm to identify the appropriate number of clusters, *k,* automatically. This is achieved by running the proposed algorithm with a varying similarity threshold (θ) value range from 1% until the interval lengths (L) start getting very small continuously (L < 2%). An interval is the number of times the algorithm produces a constant value of *k* continuously. We then terminate the algorithm and study the first longest interval at which *k* is constant and then stop the algorithm at the θ value that produces better average intra-cluster similarity for all clusters. This approach has been integrated into our clustering method as a preprocess function and tested in the experiments. The results confirmed it works well in most of the cases because the numbers of the clusters it identified are either the same as or very close to the number of true classes.

To better understand this, consider the following example of cancer dataset. In Table 1, columns 1 to 3 present the range of the similarity value threshold (θ), the number of cluster(s), and the interval length,

respectively. Although the longest interval is at $k = 1$, this interval is ignored as nobody is interested at $k = 1$. Therefore, the appropriate number of clusters is 3 as it has the longest interval (see Table 3 and Figure 2).

Table 1. Intervals of cancer dataset

Threshold values (%)	*k*	Interval length, L, (%)
1 – 33.1	1	32.1
33.2 – 47.7	3	12.5
47.8 – 53.4	4	5.6
53.5 – 55.7	5	2.2
55.8 – 56.5	6	0.7
56.6 – 57.4	7	0.8
57.5 – 58.6	8	1.1

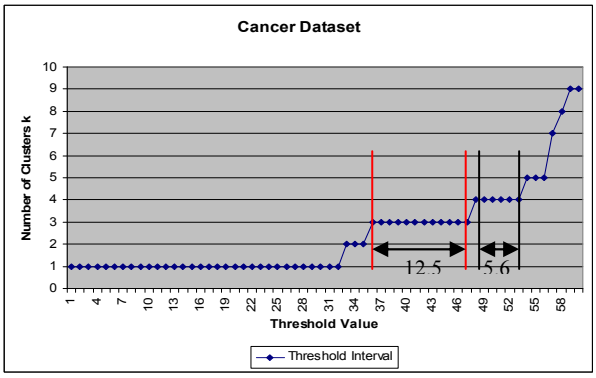

Figure 2. Intervals of cancer dataset

4. SIMILARITY MEASURES

Since measuring the similarity between data samples plays an essential role in all clustering techniques and can determine their performance, after studying the common existing similarity measures and evaluating their weaknesses, we proposed a new similarity measure.

4.1 Existing Similarity Measures

In practice the most similarity measures are defined based on 'distance' between data points and some popular measures are listed in Table 2. The common major weaknesses of these measures are:

1. Unable to handle categorical features
2. Unsuitable for unweighted features. Therefore, one feature with large values might dominate the distance measure.
3. Unable to reflect the degree of change between data samples.

To address theses weaknesses we proposed a novel similarity measure described in the next section.

Table 2. Existing similarity measures

Similarity Measures	Equations
Squared Euclidean distance	$d(x,y)=\sum_{i=1}^{N}(x_i-y_i)^2$
Euclidean distance	$d(x,y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}$
Correlation	$\text{Correlation}(x,y)=\frac{\sum_{i=1}^{N}(z_{xi}z_{yi})^2}{N-1}$
Cosine	$\cos ine(x,y)=\frac{\sum_{i=1}^{N}(x_iy_i)^2}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$
Chebychev (chy)	$chy(x,y)=\max_i\lvert x_i-y_i\rvert$
Manhattan distance	$Block(x,y)=\sum_i\lvert x_i-y_i\rvert$
Minkowski (p)	$p(x,y)=(\sum_i\lvert x_i-y_i\rvert^p)^{1/p}$
Power(p,r)	$Power(x,y)=(\sum_i\lvert x_i-y_i\rvert^p)^{1/r}$

4.2 A Novel Similarity Measure

The new similarity measure, represented by Equation (1), was defined particularly to reflect the degree of the relative change between samples and to cope with both numerical and categorical variables. For numerical variables, Term 1 in Equation (1) is used. For the categorical variable, the similarity between two data samples is the number of the variables that have same categorical values in two considering data samples, and is calculated by Term 2.

$$Sim(x_i, B_k) = 1 - \frac{1}{N}\left[\left(\sum_{j=1,\ \text{for } x_{ij}\in R}^{N}\frac{|x_{ij}-B_{kj}|}{\max\{x_{ij},B_{kj}\}}\right)+\left(\sum_{j=1,\ \text{for } x_{ij}\in Cat}^{N}1 \text{ if } x_{ij}=B_{kj},0 \text{ otherwise}\right)\right] \quad (1)$$

Where ***Sim*** is an abbreviation of similarity, ***N*** is the number of features, ***x*** represents the sample; ***i*** is sample index; ***j*** is feature index; ***B*** the *BASE*, ***k*** the index for clusters and *BASES*. ***R*** and ***Cat*** represent the numerical and categorical features, respectively. In this similarity measure, the similarity value between input $\boldsymbol{x_{ij}}$ and a *BASE* B_{kj} is scaled to [0, 1]. Thus, no one feature can dominate the similarity measure. This definition can be extended to measure similarity between any two samples not only limited to the *BASE*. The more detailed analysis and test on this new definition will be presented in a separate paper later.

5. EXPERIMENTS AND EVALUATION

To evaluate the accuracy of the proposed algorithm and the effectiveness of the new similarity measure we implemented it and conducted the experiments by using the same benchmark datasets that were used by the comparing methods in other papers. The basic strategy of our comparison is to take the most commonly used *k*-means as a baseline, and TwoStep as a competing target because it is generally considered as a more accurate

algorithm. Before presenting the experimental results and carrying out the intended comparison, we give the criteria for measuring clustering accuracy in section 5.1 and the method of using data in section 5.2.

5.1 Measuring the Clustering Accuracy

One of the most important issues in clustering is how to measure and evaluate the clustering performance, usually in terms of accuracy. In unsupervised clustering, there is no absolute criterion of measuring the accuracy of clustering results. However, in some cases where the class labels are available, the quality of a partition can be assessed by measuring how close the clustering results are to the known groupings in the dataset. Thus, the correct clusters should be those clusters that have all the samples with the same labels within their own cluster. It should be noted that, with such a strategy, the class label is not included during the clustering process but just used at the end of the clustering procedure to assess the partition quality.

In practice, the accuracy r is commonly measured by $r = \frac{1}{n}\sum_{i=1}^{k} a_i$ [2], where a_i is the number of majority samples with the same label in cluster i, and n the total number of samples in the dataset. Hence, the clustering error can be obtained by $e = 1 - r$.

5.2 Testing Datasets

Table 3 shows the demographic details of 7 benchmark datasets (obtained from UCI Machine learning Repository [3]) that have been used in our experiment. As commonly strategy in clustering experiments, the whole dataset is used in experiments apart from Credit Approval and Cleve datasets. Regarding the Credit Approval dataset, 24 samples with missing values in numeric features were removed from it as all algorithms lack the ability to handle missing values in numeric features, while the missing value in numeric features in Cleve dataset were replaced with the value of ZERO "0".

Table 3. Details of the benchmark datasets

No.	Datasets	No. of classes	No. of Samples	No. of Features		Missing Values	
				C	N	Total	Removed
1	IRIS	3	150	0	4	0	0
2	Wine	3	178	0	13	0	0
3	Soybean	4	47	0	35	0	0
4	Credit Approval		690	9	6	37	24
5	Cleve		303	8	6	5	0
6	2-Spirals	2	200	0	2	0	0
7	Half-rings	2	400	0	2	0	0

5.3 Results and Evaluation

Tables 4 to 8 give the accuracy of our algorithm and also some other methods for comparison.

In Table, 4 we compared our results with the experimental results given in [4] and TwoStep Algorithm [5]. The best clustering achieved for Iris dataset is by our proposed algorithm at an accuracy of 94%, while SICM (the co-association matrix to select clustering seed [4]) performed the second best at an accuracy 90.45%. For Wine dataset, our proposed algorithm performed nearly 20% better than SICM, SIPR (selecting initial seeds based on pervious results [4]) and *k*-means, but slightly lower than TwoStep. For Soybean dataset, none of the compared algorithms used in [4] manage to get, at least, an accuracy of 80%. On contrast, TwoStep and our proposed algorithm achieved an accuracy of 100%.

The experimental results for Cleve and Credit Approval datasets are presented in Table 5. With respect to the clustering accuracy, it has been illustrated in [6] that the algCEBMC algorithm produces the best clustering results. Hence, this algorithm is chosen for the comparison. For Cleve dataset, the proposed performed the highest accuracy. The algCEBMC and TwoStep achieved the second and third best accuracy, respectively. Concerning the Credit Approval, although the proposed algorithm did not perform the best on this dataset, its accuracy is smaller than that of TwoStep and algCEBMC algorithm and but better than *k*-prototypes algorithm.

Based on the experimental results on Table 6, at k = 5 although other algorithms used high value of H, where H is the number of ensemble components, the best clustering achieved for 2-Spirals dataset is by our proposed algorithm. In this particular dataset, there is a notable decrease in the accuracy of TwoStep algorithm.

On clustering the Half-rings dataset, we compared our algorithm with five other algorithms used in [7] and TwoStep algorithm. As our proposed algorithm estimated 3 clusters, we highlighted the cases where the other algorithms produced 3 clusters and run TwoStep algorithm with k = 3. As shown in Table 7 our algorithm performed the best accuracy among the compared algorithms. TwoStep algorithm performed the second best. EM algorithm had third best accuracy only for small value of H (H = 5). MLCA (Meta-Clustering Algorithm [8]) performed the third best in 4 cases, at H > 10. Regardless the value of H, at k = 2 HGPA (HyperGraph Partitioning Algorithm [9]) always generated the highest error.

Table 8 lists the effectiveness of the proposed algorithms on Iris. As shown that our proposed algorithm is the clear winner by large margins over the other individual methods. CSPA (Cluster-based Similarity Partition Algorithm [10]) achieved the second best results at high value of H (H >= 15). However, in practice, it is not recommend having high value H.

Table 4. Iris, Wine and Soybean datasets

Algorithms	Datasets and Accuracy%		
	Iris, k= 3	Wine, k = 3	Soybean, k= 4
Standard k-means	88	70.78	68.08
SIPR	88.23	72.47	73.06
SICM	90.45	75.28	76.59
TwoStep	86	94.94	100
Jamil & Wang	94	93.3	100

Table 5. Cleve and Credit Approval datasets

Algorithms	Datasets and Accuracy%	
	Cleve, k= 4	Credit Approval, k = 5
Standard k-means	79.20	83.8
k-prototypes	≈79	≈72
algCEBMC	≈83	≈80.9
TwoStep	80.5	83.93
Jamil & Wang	83.3	78.7

Table 6. 2-Spirals dataset

H	k	Algorithms and Accuracy (%)				
		EM	QMI	CSPA	HGPA	MCLA
5	2	56.5	56.4	56.1	50	56.2
5	3	58.9	58.7	60.1	50.5	59.5
5	5	58.8	59	60	57	60
5	7	54.1	54.6	54.6	57.6	56.3
5	10	52.7	54.6	52.3	53.6	56.1
10	2	56.6	56.3	56	50	56.1
10	3	63.1	60	61	50.8	58.3
10	5	61.4	60.6	61.7	59.4	61.1
10	7	53.3	53.3	53.8	**57**	54.3
10	10	53.3	54.4	52.3	52.9	57.6
20	2	56.7	56.4	56.2	50	56.1
20	3	59.3	59.8	62.9	50.7	60
20	5	61.4	60.5	61.8	60	61.9
20	7	54.1	52.4	53.3	55.6	55.8
20	10	51.8	52.8	51.3	52.7	57.8
-	5	Standard k-means: 53				
-	5	TwoStep: 52				
-	5	Jamil & Wang : 71				

Table 7. Half-rings dataset

H	k	Algorithms and Accuracy (%)				
		EM	QMI	CSPA	HGPA	MCLA
5	2	74.6	74.6	74.5	50	74.6
5	3	76	63.2	73.8	51.2	74.9
10	2	73.3	66.8	71.4	50	76.3
10	3	66.5	60.3	75.1	74	75.8
30	2	73.1	59.4	73.8	50	74
30	3	70.7	64.1	73.8	72.5	73.8
50	2	72.8	67.7	70.5	50	78.9
50	3	71.2	64.7	75	75.2	75.4
-	3	Standard k-means: 85.8				
-	3	TwoStep: 87.25				
-	3	Jamil & Wang : 88				

Table 8. Iris dataset

H	k	Algorithms and Accuracy (%)				
		EM	QMI	CSPA	HGPA	MCLA
5	3	89	85.3	88.8	58.6	89.1
10	3	89.2	89.2	88.7	61.8	89.1
15	3	89.1	88.1	90.2	57.2	88.9
20	3	89.1	85.5	90.2	60.9	89.1
30	3	89.1	87.2	92.1	56.6	88.7
40	3	89	87.6	92.3	58.1	88.9
50	3	89.1	86.2	92.1	57.3	88.8
-	3	Standard k-means: 88				
-	3	TwoStep: 86				
-	3	Jamil & Wang : 94				

5.4 Summary of the Results

To sum up, our proposed algorithm achieved the best results in 6 cases, second best in 1 cases and forth best in 1 case, but it never performed the worst (see Table 9). It is worth noting that the accuracy of our algorithm when it scored the second was not really bad. It was comparable and slightly lower than that of TwoStep which performed the best. The proposed algorithm worked best for Soybean, Iris, and Wine datasets as its accuracy were at least 94%. For Iris, Soybean, Cleve, 2-Spirals and Half-rings datasets our algorithm performed the best. For Wine dataset, it preformed the second best. The proposed algorithm performed the forth best in Credit Approval datasets.

Table 9. Summary of the Experimental Results

No.	Datasets	No. of Compared Algorithm		Ranking of our algorithm	
1	Iris	5	7	1	1
2	Wine	5		2	
3	Soybean	5		1	
4	Credit Approval	5		4	
5	Cleve	5		1	
6	2-Spirals	8		1	
7	Half-rings	8		1	

6. CONCLUSIONS

In this paper, we proposed a novel clustering algorithm and a new similarity definition. The proposed algorithm consists of three stages. With respect to the clustering accuracy, the experimental results show that our algorithm has outperformed the individual clustering techniques compared in most of the datasets, and particularly good for Soybean, Iris, and Wine datasets. Our algorithm found the true classes for Soybean dataset. It also performed better than the compared ensembles based methods, although the clustering ensemble supposes to perform better than the individual algorithms. Overall, it achieved the best in 6 cases, second beset in 1 cases and never performed worst. More importantly, our algorithm does not need to specify the number of clusters, *k,* as it is calculated automatically. In addition, it is able to handle both numerical and categorical variables. As the similarity value between features is scaled to [0, 1] all features will have the same weight in calculating the over all similarity value. On the anther hand, our proposed algorithm may have a relatively from high computation complexity. This is because the process of finding and refining the *BASES* is time consuming.

Future work will involve conducting more experiments to refine the proposed algorithm, and investigating the clustering ensemble method.

REFERENCES

[1] Strehl A., "Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining," Austin: University of Texas, 2002, p. 232.

[2] Huang Z., "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values " *Data Mining and Knowledge Discovery,* vol. 2, pp. 283-304, 1998.

[3] Merz C.J. and M. P., "UCI Repository of Machine Learning Databases," Website access at <Http://www.ics.uci.edu/~mlearn/MLRRepository.html>, 1996.

[4] Azimi J., Davoodi S., and Analoui M., "Fast Convergence Clustering Ensemble," in *9th International Multi-conference on Information Society, Data Mining and Data Warehouses (SiKDD 2006)* Ljubljana, Slovenia, 2006.

[5] "TwoStep Cluster Analysis," Website access at <http://www1.uni-hamburg.de/RRZ/Software/SPSS/Algorith.120/twostep_cluster.pdf>, 2007.

[6] He Z., Xu X., and D. S., "Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach," 2005.

[7] Topchy, K. Jain, and W. Punch, "Combining Multiple Weak Clusterings," *Proceedings of the Third IEEE International Conference on Data Mining* pp. 331-338, Nov 2003.

[8] Strehl A. and Ghosh J., "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research,* vol. 3, pp. 583-617, 2003.

[9] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire, "Discovery of Aggregate Usage Profiles for Web Personalization " in *Proceedings of the Workshop on Web Mining for E-Commerce*, 2000.

[10] Kuncheva L.I., Hadjitodorov S.T., and T. L.P., "Experimental Comparison of Cluster Ensemble Methods," in *Information Fusion, 2006 9th International Conference on*, Florence, 2006, pp. 1-7.