A Framework for Clustering Mixed Attribute Type Datasets

Jongwoo Lim[1], Jongeun Jun[2], Seon Ho Kim[2] and Dennis McLeod[1]
[1]Department of Computer Science, University of Southern California, CA, USA
[2]Integrated Media Systems Center, University of Southern California, CA, USA
{jonglim,jongeunj,seonkim,mcleod}@usc.edu

We propose a clustering framework that supports clustering of datasets with mixed attribute type (numerical, categorical), while minimizing information loss during clustering. Real world datasets such as medical datasets and its ontology have mixed attribute type datasets. However, most conventional clustering algorithms have been designed and applied to datasets containing only single attribute type (either numerical or categorical). Recently, approaches to clustering for mixed attribute type datasets have emerged, but they are mainly based on transforming attributes to straightforwardly utilize conventional algorithms. The problem of such approaches is the possibility of distorted results due to the loss of information because significant portion of attribute values can be removed in the transforming process. This results in a lower accuracy clustering. To address this problem, we propose a clustering framework for mixed attribute type datasets without transforming attributes. We first utilize an entropy based measure of categorical attributes as our criterion function for similarity. Second, based on the results of entropy based similarity, we extract candidate cluster numbers and verify our weighting scheme with pre-clustering results. Finally, we cluster the mixed attribute type datasets with the extracted candidate cluster numbers and the weights. Our experimental results demonstrate that the proposed framework is effective in increasing accuracy.

Key Words: mixed attribute type clustering, entropy based similarity measure, weighting scheme, data mining,

1. INTRODUCTION

Clustering is widely used in data mining applications to find patterns in data. Conventional clustering techniques have been focused on a single type of attributes, either numerical or categorical attributes of datasets. As a criterion function in clustering process, similarity measure has been used as one of the essential steps, i.e., in determining the candidate cluster number. The unique characteristics of categorical attributes are that the values of categorical attributes are not only discontinuous but also disordered while the values of numerical attributes are continuous in computing the distance between two values. Due to the difference of the characteristics between categorical and numerical attributes, similarity measures for categorical attributes or numerical attributes have focused on just their own characteristics, for example, entropy based similarity measure for categorical attributes and distance measure for numerical attributes.

As mixed attribute type datasets are common in real life, clustering techniques for mixed attribute type datasets is required in various informatics fields such as bio informatics, medical informatics, geo informatics, information retrieval, to name a few. These mixed attribute datasets provide challenges in clustering because there exist many attributes in both categorical and numerical forms so mixed attribute type should be considered together for more accurate and meaningful clustering. However, conventional approaches are designed

mainly for a single type attributes they are not appropriate for mixed attribute type datasets [9]. Recently, some approaches to clustering for mixed attribute have been introduced by converting categorical attribute values to numerical ones and applying traditional clustering algorithms with only numerical values [8]. However, those approaches have the possibility of distorting the results by losing characteristics of attributes [8]. Due to the fundamental differences in two data types, the conversion from categorical attribute values to numerical ones may not be perfect and not semantically meaningful. Many times, domain experts should be involved in the conversion process to provide semantic relations among different data types for a better conversion. However, this can be also subjective and incomplete. Thus, the loss of information in the conversion process incurs considerable inaccuracy in clustering. To address this problem, we propose a clustering framework for mixed attribute type datasets without losing information of attributes that works effectively on mixed attribute datasets.

Our framework is based on a couple of observations in mixed attribute type datasets. Categorical domains typically have a smaller number of values than numerical domains [3]. Thus, considering similarity for categorical attributes first and then combining numerical attributes in the same objects might produce a smaller variance in clustering result than the opposite order. Rather than converting categorical values to numerical ones, it can provide useful insights so that we first apply clustering separately to each type (without loss of information) and analyze the results to enhance the overall clustering of mixed attribute type datasets. Then, we can check the degree of balance of each clustering result to figure out which type can have a higher priority in overall clustering process.

The proposed clustering framework consists of three main steps (see Fig. 1). In Step 1, we use an entropy based similarity measure with only categorical attributes and extract candidate cluster numbers by evaluating the difference of values with entropy based similarity measure. We analyze the difference of total entropy among clusters in an exhaustive manner by reducing the number of clusters until all of clusters merge into one cluster and extract candidate cluster numbers by using the difference in entropy values.

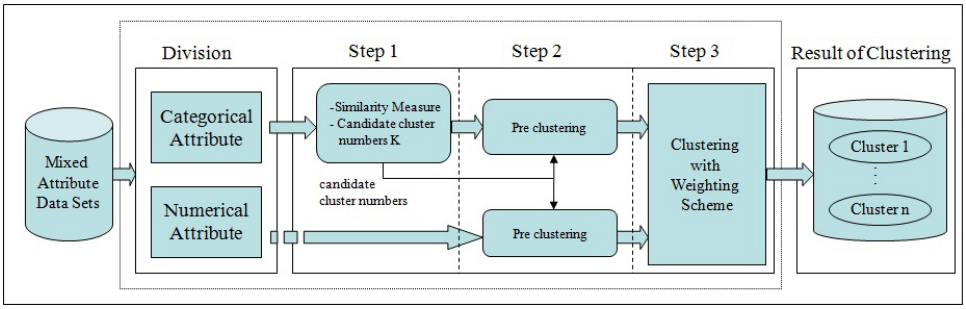

Fig. 1 Overview of the Proposed Clustering Framework

Second, we apply the extracted candidate cluster numbers *K* from Step 1 to cluster the dataset using only numerical attributes (Step 2). Now, we have two clustering results, one by using only categorical attributes and the other by using

numerical ones. Note that the number of clusters is decided solely by categorical attributes.

In Step 3, a weighting scheme is applied using the degree of balance in number of objects in the clusters. After the pre-clustering, we can compare how two clustering results are balanced. The main point of the weighting scheme is to put more weight onto the better-balanced clustering between categorical and numerical one. After determining the weights, the final clustering is processed for the mixed attribute type dataset using the extract candidate cluster numbers from Step 1 and the weights.

The remainder of this paper is organized as follows. We summarize conventional clustering algorithms in section 2. Section 3 describes the proposed clustering framework for mixed attribute type dataset. The experimental results are shown in section 4. Finally, a conclusion follows in Section 5.

2. RELATED WORKS

Based on the properties of attributes in a dataset, clustering algorithms can be classified into three categories such as categorical, numerical and mixed attribute type algorithms.

For categorical attributes, Squeezer algorithm reads each tuple t in sequence over all dataset and determine it using the similarity values between t and clusters [1]. ROCK is an adaptive one of an agglomerative hierarchical clustering algorithm [2] and CACTUS is a fast summarization based algorithm [3]. For numerical attributes, a density based algorithm has been used for large spatial databases [4], and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is for a very large scale database [5].

Although the above conventional algorithms have concentrated on a single characteristic attribute clustering, they can be also applied to mixed attribute datasets by converting or reforming attribute values. Z. Huang [6] presented two algorithms. One is the k-mode algorithm, extending the k-means algorithm, which is used a new distance measure for categorical attributes. The other is the k-prototypes using a weighted sum of Euclidean distance between numerical values. However, inappropriate weights decided by *a priori* parameters may result in unexpected clusters. C. Li and G. Biswas [7] proposed the SBAC algorithm which is based on a similarity measure with weights and uses an agglomerative algorithm. It is not appropriate for large scale datasets due to the increasing complexity of the SBAC algorithm. Yosr Naija, et al. [8] proposed an extension of partitional clustering methods devoted to mixed attribute type datasets. Zengyou, et al. [9] proposed a cluster ensemble approach method for mixed attribute data. J. Suguna and M. Arul Selvi [19] also proposed ensemble fuzzy clustering for mixed numeric and categorical data. They all convert categorical attribute values into numerical ones before applying clustering.

Amir Ahmad and Lipika Dey [10] proposed a k-mean clustering algorithm for mixed numeric and categorical data. Ming-Yi Shih, et al. [18] proposed a two-step method for clustering mixed categorical and numerical data. It first constructs similarity or relationships among categorical attributes based on their co-occurrence and then those

categorical attributes are converted into numeric data. Finally, the hierarchical and partitioning clustering algorithms used for clustering the data including converted into numeric data.

For the similarity measure, entropy concept has been used for categorical data in the literature. As an element of information theory, entropy is also a measure of the uncertainty with a random variable. The total entropy value is created using a classical entropy theory, Shannon Entropy [12]. Entropy based clustering is a method that finds similar objects in clusters based on their total entropy values and determines the number of clusters and identifies the location of the cluster center. The basic idea of entropy based clustering is that the lowest entropy value between two objects represents the highest similarity among objects [15].

3. PROPOSED FRAMEWORK

In this section, we introduce our three-step clustering framework for mixed attribute type datasets using entropy based similarity measure. We present an overview of our proposed clustering framework in Fig 1.

The proposed framework begins with dividing mixed attribute type datasets into categorical and numerical attributes sub dataset. In Step 1, we measure the similarity of categorical attribute sub dataset by utilizing entropy based similarity measure using an agglomerative process. Based on the results of the similarity measure, we analyze the changes in total entropy total entropy value while building clusters in agglomerative way and extracting candidate cluster numbers, K (i.e., a list of desirable cluster numbers), for mixed attribute type dataset clustering. In Step 2, using the candidate cluster numbers K, we pre-cluster each type attributes to determine appropriate weights based on the resulting structure of pre-clustering. In Step 3, we cluster mixed attribute type datasets by using the candidates cluster numbers and weighting each type of attributes with different values

We have two hypotheses in designing our proposed framework:

1) The candidate cluster numbers from categorical attributes can be also candidate cluster numbers of mixed attribute type datasets in the given dataset. In case, measuring total similarity for categorical attributes first and then combining numerical attributes in same objects, the variance of its clustering result is less than the opposite case. Since one of characteristics of categorical attribute is not continuous and ordered, it is not appropriate to use a classical distance measure for similarity in a categorical attribute dataset. Reforming numerical attributes into categorical can be the cause of possibility of distorting candidate cluster numbers. So, we utilize entropy based similarity measure focused on categorical attribute dataset.

2) As one of the critical conditions for effective clustering is the degree of the balance of the number of objects in clusters. We determine that the better balanced attribute between categorical attribute and numerical one will receive the higher weight in a mixed attribute type datasets.

There are three essential processes in the proposed algorithm as described above. We will explain the details of each process in following three subsections.

3.1 Calculating entropy-based similarity measure

As a criterion function, measuring similarity between objects is one of the primary steps in clustering process. There are many well known methods for measuring distance between objects for the purpose of clustering, but these methods are known that they have pros and cons in some level. As one of elements of information theory, entropy can be used to measure the uncertainty of random variables. On that point, we utilized it as a similarity measure for the categorical attributes in mixed attribute type datasets.

Distance functions such as Euclidean distance are used as similarity measure for numerical attribute since they well represent the inherent distance meaning between numerical attributes but they are not for categorical attribute. It is difficult for categorical attribute to measure similarity in that its values cannot be directly compared each other because they are not ordered nor continuous, whereas numerical attributes are ordered and continuous.

On account of the problem for categorical attribute, we developed an entropy based similarity measure which can be an effective and practical similarity measure for categorical clustering [15] to our proposed framework. We first give the notations of a classical entropy definition, which is the Shannon's entropy definition [12] for entropy based similarity measure. The entropy $H(X)$ is simply defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

where $p(x)$ is the probability mass function of the random variable x. X is the set of possible outcomes of x. We consider that a dataset $X = SC + SN$ (Where SC is a subset of categorical attributes and SN is a subset of numerical attributes) in the presence of R objects. Let $m = cn + nn$ be the total number of attributes in a given dataset, where cn is the number of categorical attributes, and nn is the number of numerical attributes. Then, $SC = \{D_1, D_2, ..., D_{cn}\}$, where D_i is the i^{th} categorical attribute, and $SN = \{N_1, N_2, ..., N_{nn}\}$, where N_i is i^{th} numerical attribute. At_i is the set of distinct values in i^{th} categorical attribute (D_i).

The definition of total entropy in the given dataset can be redefined as [17]:

$$H(SC) = -\sum_{i=1}^{cn} \sum_{v \in At_i} p(v) \log p(v)$$

where $p(v)$ is the probability of occurrence of value v in ith categorical attribute(D_i). We calculate the entropy value of sub-datasets then extract candidate cluster numbers for mixed attribute, which will be used for the next preclustering process.

In order to extract , we first assume that a sub-dataset SC can be partitioned into K clusters. It can be represented as follows:

$C^K = \{C_k\}$ for $1 \le k \le K$, where $1 \le K \le R$ and C_k is a cluster having $n_k$records, $1 \le n_k \le R-(K-1)$ in the categorical attribute sub-dataset. By maximizing the entropy criterion [15], the classical entropy based clustering attempts to find the optimal candidate C^K. The entropy criterion for optimal candidate cluster C^K as follows:

$$OC(C^K) = \frac{1}{cn}[(H(SC) - \frac{1}{K}\sum_{k=1}^{K} H(C_k)] \qquad (1)$$

where $H(SC)$ is the total entropy in the given dataset, $AH(C^K) = \frac{1}{K}\sum_{k=1}^{K} H(C_k)$ is the average entropy of C^K. And it is supposed to be minimized in order to maximize $OC(C^k)$. We notate the average entropy of partition C^K as $AH(C^K)$ in the following sections.

3.2 Extracting Candidate Cluster Numbers

As mentioned in the first hypothesis above, to prevent distorted result from converting attribute values, we utilize only categorical attribute sub-dataset to measure similarity by using entropy function. Since the difference between $AH(C^K)$ and $AH(C^{K+1})$ is closely correlated with the similarity between the clusters, we can extract candidate cluster numbers for clustering by exploring the difference of each cluster's average entropy while clusters are merged in an agglomerative way.

We assume that each object in a dataset is initially regarded as a singleton cluster. By merging clusters based on the entropy criterion, the difference between $AH(C^K)$ and $AH(C^{K+1})$ varies in each merging step because the probability distribution of values in clusters changes in uncertain ways when two clusters are merged. If two highly similar clusters are merged into one, then the variance of average entropy will not change much. However, it significantly varies when two very different clusters are merged.

When two clusters are merged into one, the resulting cluster's entropy increases. For the proof of increasing entropy, we provide another notations that $C_a \bigcup C_b$ is the mergence of two clusters C_a and C_b, and C_a has n_a records and C_b has n_b records. The following relation based on the expected entropy was proved in [17].

$$(n_a + n_b)H(C_a \bigcup C_b) \ge n_a H(C_a) + n_b H(C_b) \qquad (2)$$

This relation shows that the expected entropy is always identical or increased by merging clusters, and the average entropy also has a provable relation as shown above. We notate the difference of average entropy of clusters ($Diff_{ent}$) as follow

$$Diff_{ent}(C_a, C_b) = H(C_a \cup C_b) - \frac{1}{2}[H(C_a) + H(C_b)] \ge 0 \quad (3)$$

$$Diff_{ent}(C_a, C_b) = 0 \text{ Where } C_a \text{ is identical with } C_b \qquad (4)$$

Using Eq. (3), our algorithm recursively merges clusters. Initially, each object is considered as a singleton cluster and the initial entropy $Diff_{ent}$ for all possible cluster pairs in the dataset (Fig. 2). So, when n objects exist in the dataset, $n^2/2$ $Diff_{ent}$ values will be calculated initially.

The merge process consists of the following steps:

1) In the presence of n clusters, calculate $Diff_{ent}$ for all possible pairs of clusters. Calculate the average entropy of partition C^n and store it.
2) Find the pair having the minimum entropy, say $Diff_{ent}(C_i, C_j)$. Consequently, the cluster C_i and C_j are being merged. This is because the minimum change in entropy implies potentially better clustering.
3) Next step is to determine which cluster will be deleted or updated. Between i and j, the higher numbered cluster will be merged in to the lower numbered cluster. That is, the lower numbered cluster will be updated and the other will be deleted. See Fig. 3.
4) Next, the $Diff_{ent}$ table will be updated by recalculating $Diff_{ent}$ for all possible pairs of remained clusters. Note that $n = n-1$ now. Calculate the average entropy of partition C^n and store it.
5) Steps 2-4 will be iterated until the number of clusters becomes one whole cluster.

	N1	N2	N3	...	Nn
N1	Diff(N1, N1)	Diff(N1, N2)	Diff(N1,N3)	...	Diff(N1, Nn)
N2		Diff(N2, N2)	Diff(N2, N3)	...	Diff(N2, Nn)
N3			Diff(N3, N3)	...	Diff(N3,Nn)
.				...	...
					...
.					...
Nn					Diff(Nn, Nn)

Fig 2 Initializing $Diff_{ent}$

	N1	...	Ni	...	Nj	...	Nn
N1	0	...	Updated	...	Deleted	...	.
...		...	Updated	...	Deleted	...	.
Ni			0	Updated	**Merged**	Updated	Updated
...				...	Deleted	...	.
Nj					0	Deleted	Deleted
...						...	.
Nn							0

Fig 3 Snapshot of Merging Cluster

Using the results of the above merging process, specifically $\{AH(C^i)\}$ where $1 \le i \le R$, we determine the optimal candidate cluster numbers which will be used in the final clustering. The main point is to monitor the changes of average entropy during the merging process. As shown in Eq. (2), the average entropy of

resulting clusters increases after merging two different clusters. By comparing $AH(C^K)$ to $AH(C^{K+1})$, we can identify sudden changes in entropy.

Let D^K be the difference of entropy between $AH(C^{K-1})$ and $AH(C^K)$. The algorithm computes the set of entropy values, $D=\{D^K\}$ for all $2 \le K \le R$ until the number of cluster being one. By monitoring the value changes in D^K, the algorithm determines the candidate cluster numbers as follows: 1) find a subset of D, i.e., D_S, with all D^K which satisfies $D^{K-1} < D^K$ and $D^K > D^{K+1}$. 2) select P (P is an input parameter) greatest values of D^K values in D_S then the set of K values becomes the candidate cluster numbers. See Fig. 5.

3.3 Weighting Scheme

Based on the algorithm described in subsection 3.2, the candidate cluster numbers are decided. Using the numbers, the given dataset is clustered with only categorical and numerical attributes respectively (Step2: preclustering in Fig. 1). Then, we analyze how each clustering result is balanced. In general, well structured clustering shows that the numbers of objects in clusters are balanced. By comparing the balance of clustering of one result with categorical attributes to that with numerical ones, our approach sets a priority on one type of attribute over the other. So as we mentioned in the second hypothesis above, we consider the weight of each attribute type in a dataset before finally clustering mixed attribute type datasets. We first give more weight for the better balanced attribute type between categorical attribute and numerical one to improve the results of the final clustering. However, it is hard to formulate generally the weighting scheme for mixed attribute type datasets due to the difficulty in extracting correlation between attribute types. Our weighting scheme only focused on a given dataset. The weight condition of our mixed attribute clustering is defined as $\omega_t = \omega_c + \omega_n$ where ω_c is weight for categorical attribute and ω_n is weight for numerical attribute and $\omega_t = 1 \; and \; 0 \le \omega_c \le 1, 0 \le \omega_n \le 1$.

Once the weight are determined, the final clustering is performed based on the following similarity measure using the weights:

$$S_M = \omega_C S_C + \omega_N S_N \qquad (5)$$

where S_M is the similarity of mixed attribute type datasets (i.e., total) and S_C, S_N is the similarity of categorical and numerical attributes, respectively. With S_M values, our algorithm utilizes the agglomerative hierarchical clustering method for the final result.

4. EXPERIMENTAL EVALUATION

In our experiment, we use a heart disease dataset from UCI Data Repository [20]. The dataset has 13 attributes in total, 7 of them are categorical and the others are numerical. The dataset has 270 objects. The dataset has the groundtruth per object about the confirmed diagnosis of heart disease, i.e., positive or negative diagnosis of heart disease.

First, Fig 4 illustrates that the increase of total entropy, $AH(C^K)$, after merging clusters (showing the property in Eq. (2)). Note that, although the dataset has 270 objects, we disregard the case when the number of clusters is greater than 20 since the entropy increase is negligible.

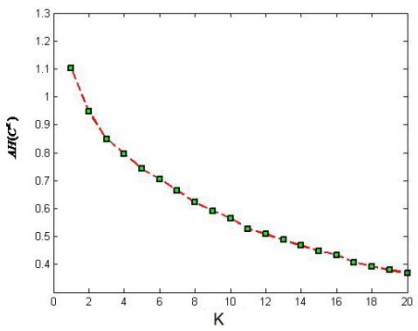

Fig 4 Average Entropy in each cluster

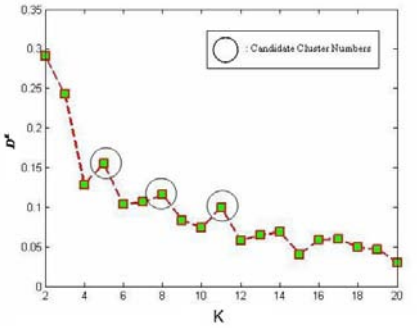

Fig 5 The difference of Average Entropy

Fig 5 demonstrate how the difference of the average entropy, i.e., D^K, varies while merging clusters. For example, D^5 means the difference of the average entropies between when $K=5$ and $K=4$. When 6 clusters are merged into 5 clusters, we can find that the difference of average entropy is relatively small. However, when 5 clusters are merged into 4 clusters, the difference of average entropy is increased highly. When 4 clusters are merged into 3 clusters, it becomes relatively small again. Abrupt change in entropy means that dissimilar clusters are merged, so this merging is not desirable. Thus, $K=5$ becomes a candidate cluster number. In the same way, we identify cluster number 8 and 11. Finally, the algorithm generates a subset, D_S, by choosing top three values (i.e., $P=3$ is the input parameter used in the experiment). So, D_S ={5, 8, 11}, which represents the candidate cluster numbers.

After extracting the candidate cluster numbers, we need to determine the weight values for the given dataset by checking the balance of clusters. We cluster the given dataset using each type of attribute for each candidate cluster number. For examples, Table 1 compares the balance of clusters in the result using only categorical attributes (Categorical) and using only numerical attributes (Numerical), respectively. It shows that the Categorical is far better balanced than Numerical while the number of clusters is changed (e.g., $K=5$, $K=6$) in the given dataset. Thus, a higher weight is assigned to Categorical. Finally, the final clustering is done based on the new similarity measure using Eq. (5) with mixed attribute type dataset (i.e., all attributes).

[number of clusters = 5]

	C1	C2	C3	C4	C5
Categorical	44	34	56	85	51
Numerical	21	1	245	2	1

[number of clusters = 6]

	C1	C2	C3	C4	C5	C6
Categorical	44	34	56	43	51	42
Numerical	21	1	244	2	1	1

Table 1 Comparison of pre clustering results

To verify the applicability of our algorithm, we evaluated the accuracy of the result of clustering by comparing them with the ground truth. The accuracy of clustering measures the extent how well the resulting clusters group similar objects, which can be either negative class or positive class in the given dataset. We define the clustering accuracy as follows.

$$Accuracy = \frac{\sum_{i=1}^{k} O_i}{R},$$

where R is the number of all objects in the given dataset, and $O_i = MAX[count(positive), count(negative)]$ in cluster i.

	0.2	0.3	0.4	0.5	0.6	0.7	0.8
5	0.7781	0.7525	0.8530	0.7772	0.8271	0.8217	0.7649
6	0.7826	0.7978	0.8069	0.7727	0.8350	0.8183	0.7458
7	0.7722	0.8077	0.8055	0.7683	0.8561	0.8152	0.7579
8	0.8048	0.8137	0.8106	0.7639	0.8588	0.8204	0.7641
9	0.7824	0.7848	0.7780	0.7797	0.8454	0.8136	0.7820
10	0.7699	0.7748	0.7861	0.7984	0.8431	0.8081	0.7604
11	0.7653	0.7778	0.7715	0.7832	0.8372	0.8110	0.7707

Table 2 Accuracy of clustering result

Table 2 shows the accuracy of clustering result while varying ω_C. It presents that the higher weights into categorical attributes, the higher the accuracy. We show the candidate cluster numbers from 5 to 11 for a focused discussion. We can determine that the highest accuracy of the clustering is when cluster number K is 8 and the weight on Categorical is 0.6.

The experimental result shows that the candidate cluster numbers extracted from our approach can be the candidate cluster numbers for mixed attribute dataset and our weighting scheme can provide a higher accuracy of mixed attribute type clustering result, as we expected. The results of our experiments on heart disease dataset is verified with accuracy ground truth and our proposed framework works well with mixed attribute dataset.

Fig. 6 shows that the comparison using the average accuracy between our approach and a conventional one. The conventional approach is transforming categorical data to numerical data without background knowledge of the categorical data set and using Euclidean distance as a similarity measure. Then, it uses *k*-means algorithm for clustering. The result of our approach is better than the conventional one. For example, when K=8, our result shows the highest accuracy while the conventional one provides the worst case.

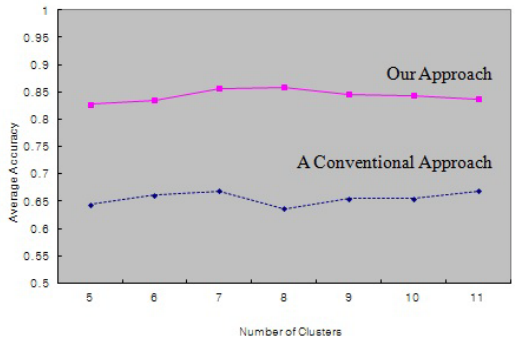

Fig 6 Comparison between our approach and a conventional one

5. CONCLUSION

In this paper, we proposed a clustering framework for mixed attribute type dataset. Conventional clustering algorithms have focused on a single attribute type such as either numerical or categorical attribute. There exist approaches to clustering mixed attribute type datasets by transforming one type into the other. One of challenges in such approaches is the loss of information during the conversion due to the difference between two data types.

Without transforming data, our proposed framework uses a pre-clustering process focused on categorical attributes to better understand which type of attributes can be more influential in clustering mixed attribute type datasets. It divides dataset into categorical attribute and numerical attribute sub datasets. Based on the expected entropy as a similarity measure, it evaluates the average entropy between different numbers of clusters and then extracts candidate cluster numbers with the results of evaluating the difference. After pre-clustering, the balance of clustering is analyzed and used to determine weight values of each attribute type. Finally, clustering process is performed with the extracted candidate cluster numbers and weight values. Our experimental results show that the candidate cluster number extracted from only categorical attributes can be used as the candidate cluster number for mixed attribute type dataset in the given dataset and the proposed weighting scheme based on the degree of balance of clustering can improve the accuracy of clustering.

As future work, we will research subspace clustering algorithms for large scale dataset with mixed attribute types, investigate feature selection techniques to detect correlation between different type attributes, and investigate other alternative weighting scheme algorithms to improve the proposed framework.

REFERENCES

[1] Z. He, X. Xu, S. Deng, “Squeezer: An Efficient Algorithm for Clustering Categorical Data,” *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611-625, 2002.

[2] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes”, *Int. Conf. Data Engineering*, pp.512-521, Sydney, Australia, Mar 1999.

[3] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, “CACTUS- Clustering Categorical Data Using Summaries”, *Int. Conf. Knowledge Discovery and Data Mining*, pp. 73-83, 1999.
[4] M. Ester, H. P. Kriegel, J. Sander, X. Xu, “A Density-based Algorithm for Discovering Clusters in Large Spatial Databases”, *Int. Conf. Knowledge Discovery and Data Mining(KDD ’96),* pp. 226-231, Aug 1996.
[5] T. Zhang, R. Ramakishnan, M. Livny, “An Efficient Data Clustering Method for Very Large Databases”, *Proc. of the ACM_SIGMOD Int’l Conf. Management of Data*, pp. 73-84, 1999
[6] Z. Huang, “Entensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values”, *Data Mining and Knowledge Discovery*, pp. 283-304, 1998.
[7] C. Li, G. Biswas, “Unsupervised Learning with Mixed Numeric and Nominal Data”, IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 4, 2002.
[8] Yosr Naija, Salem Chakhar, Kaouther Blibech, Riadh Robbana, “Extension of Partitional Clustering Methods for Handling Mixed Data“, ICDMW, pp. 257-266 IEEE Intl. Conf on Data Mining Workshops, 2008.
[9] He, Z, Xu, X. and Deng, S. “Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach”, ARXiv Computer Science e-prints, 2008.
[10] Amir Ahmad, Lipika Dey, “A k-mean clustering algorithm for mixed numeric and categorical data”, Data & Engineering, Vol 63, Issue 2, pp. 503-527, 2007
[11] Han, J. Kamber, M., “Data Mining: Concepts and Techniques”, Morgan Kauffman, ISBN 1-55860-489-8, CA, USA
[12] C.E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, 1948.
[13] Yao, J., Dah, M., S.T. & Liu, H., “Entropy-based Fuzzy Clustering and Fuzzy Modeling”, Fuzzy Sets and Systems (Elsevier), Vol. 113, pp. 381-288, ISSN 0165-0114, 2000.
[14] Barbara, D., Couto, J., Li, Y., “COOLCAT: an entropy-based algorithm for categorical clustering”, *Proceedings of the Eleventh ACM CIKM Conference*, pp. 582-589, 2002
[15] Tao Li, Sheng Ma, Mitsunori Ogihara, “Entropy-Based Criterion in Categorical Clustering”, *Proceeding of the twenty-first international conference on Machine Learning*, pp. 68, Canada, 2004
[16] C.H. Cheng, A. W.C. Fu, and Y. Zhang. *Entropy-based subspace clustering for mining numerical data. Proc. of ACM SIGKDD Conference,* 1999.
[17] Keke Chen., Ling Liu.,”The Best K for Entroy-based Categorical Data Clustering”, SSDBM 2005: 235-262
[18] Ming-Yi Shih, Jar-Wen Jheng, and Lien-Fu Lai., "A Two-Step Method for Clustering Mixed Categroical and Numeric Data", *Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11_19 (2010)*
[19] J. Suguna and M. Arul Selvi., " ensemble fuzzy clustering for mixed numeric and categorical data", International Journal of Computer Application, Vol 42-No3(Mar 2012)
[20] UCI Data Repository : http://archive.ics.uci.edu/m