A Unified Metric for Categorical and Numerical Attributes in Data Clustering

Yiu-ming Cheung and Hong Jia

Department of Computer Science and Institute of Computational and Theoretical Studies
Hong Kong Baptist University, Hong Kong SAR, China

Outline

Outline

Clustering and Attribute

Clustering:

- A widely utilized technique in variant scientific areas;
- The main task is to discover the natural group structure of objects represented by numerical or categorical attributes (*Michalski et al., 1998*).

Attribute:

- An attribute is a property or characteristic of an object;
- Each object is described by a collection of attributes;
- There exists two different types of attributes:
  - *Numerical attributes:* can be ordered by numbers;
  - *Categorical attributes:* cannot be ordered by their values, but can be separated into groups.

An Example: Diagnostic Records of Patients

UCI Heart Disease Data set: contains 8 categorical and 5 numerical attributes.

Attribute	Descriptor	Property	Type
Age		continuous	numerical
Sex	{F, M}	discrete	categorical
Chest pain type	{typical angina, atypical angina, ...}	discrete	categorical
Resting blood pressure		continuous	numerical
Serum cholestoral		continuous	numerical
Fasting blood sugar	$\{> 120mg/dl, \leq 120mg/dl\}$	discrete	categorical
Resting electrocardiographic	{type I, type II, type III}	discrete	categorical
Maximum heart rate		continuous	numerical
Exercise induced angina	{yes, no}	discrete	categorical
ST depression		continuous	numerical
Slope of ST segment	{upsloping, flat, downsloping}	discrete	categorical
CA		continuous	numerical
THAL	{normal, fixed defect, reversable defect}	discrete	categorical

Problem

- Traditional clustering methods often concentrate on purely numerical data only.
- There exists an awkward gap between the similarity metrics for categorical and numerical data.
- Transforming the categorical values into numerical ones will ignore the similarity information embedded in the categorical values and cannot faithfully reveal the similarity structure of the data sets (*Hsu, TNN'2006*).

It is desirable to solve this problem by finding a unified similarity metric for categorical and numerical attributes.

Previous Work

Roughly, the existing approaches dealing with categorical attributes in clustering analysis can be summarized into the four categories:

- Methods based on the perspective of similarity
  - *Similarity Based Agglomerative Clustering (SBAC) algorithm (Li and Biswas, TKDE'02)*
- Methods based on graph partitioning
  - *CLICKS algorithm (Zaki and Peters, ICDE'2005)*
- Entropy-based methods
  - *COOLCAT algorithm (Barbara et al., CIKM'2002)*
- Approaches that attempt to give a distance metric for categorical values
  - *K-prototype algorithm (Huang, PAKDD'97)*

Objective

- Give a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes;
- Design an efficient clustering algorithm which is applicable to the three types of data: numerical, categorical, and mixed data.

Outline

Clustering Task

Clustering a set of N objects, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, into k different clusters, denoted as C_1, C_2, ..., C_k, can be formulated to find the optimal $\mathbf{Q}^*$ via

$$\mathbf{Q}^* = \arg\max_{\mathbf{Q}} F(\mathbf{Q}) = \arg\max_{\mathbf{Q}} [\sum_{j=1}^{k} \sum_{i=1}^{N} q_{ij} s(\mathbf{x}_i, C_j)], \quad (1)$$

where $s(\mathbf{x}_i, C_j)$ is the similarity between object $\mathbf{x}_i$ and Cluster C_j, and $\mathbf{Q} = (q_{ij})$ is an $N \times k$ partition matrix satisfying

$$\sum_{j=1}^{k} q_{ij} = 1,\ 0 < \sum_{i=1}^{N} q_{ij} < N,\ \text{and } q_{ij} \in [0, 1]. \quad (2)$$

Evidently, the desired clusters can be obtained as long as the metric of object-cluster similarity is determined.

Representation of Mixed Data

Suppose the mixed data $\mathbf{x}_i$ with d different attributes consists of d_c categorical attributes and d_u numerical attributes ($d_c + d_u = d$).

$\mathbf{x}_i$ can be denoted as $[\mathbf{x}_i^{cT}, \mathbf{x}_i^{uT}]^T$ with $\mathbf{x}_i^c = (x_{i1}^c, x_{i2}^c, \ldots, x_{id_c}^c)^T$ and $\mathbf{x}_i^u = (x_{i1}^u, x_{i2}^u, \ldots, x_{id_u}^u)^T$.

Here, we have:

- x_{ir}^u $(r = 1, 2, \ldots, d_u)$ belonging to $\mathbf{R}$;
- x_{ir}^c $(r = 1, 2, \ldots, d_c)$ belonging to $dom(A_r)$, where $dom(A_r)$ contains all possible values that can be chosen by categorical attribute A_r.
- Specially, $dom(A_r)$ with m_r elements can be represented with $dom(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$.

Definition of $s(\mathbf{x}_i, C_j)$ (I)

Observations: In clustering analysis, numerical attributes are usually treated as a whole vector while the categorical attributes are investigated individually.

Definition: Let the object-cluster similarity $s(\mathbf{x}_i, C_j)$ be the average of the similarity calculated based on each attribute, we will then have

$$\begin{aligned} s(\mathbf{x}_i, C_j) =& \frac{1}{d} s(x_{i1}^c, C_j) + \frac{1}{d} s(x_{i2}^c, C_j) + \ldots + \frac{1}{d} s(x_{id_c}^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j) \\ =& \frac{1}{d} \sum_{r=1}^{d_c} s(x_{ir}^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j). \end{aligned} \tag{3}$$

Here, the similarity between each numerical attribute and the cluster C_j is replaced with the similarity between the cluster and the whole numerical vector $\mathbf{x}_i^u$.

Definition of $s(\mathbf{x}_i, C_j)$ (II)

If we denote the similarity between $\mathbf{x}_i^c$ and C_j as $s(\mathbf{x}_i^c, C_j)$, we can get

$$s(\mathbf{x}_i^c, C_j) = \frac{1}{d_c}\sum_{r=1}^{d_c} s(x_{ir}^c, C_j) = \sum_{r=1}^{d_c}\frac{1}{d_c} s(x_{ir}^c, C_j). \tag{4}$$

Then, previous Eq. (3) can be further rewritten as

$$s(\mathbf{x}_i, C_j) = \frac{d_c}{d} s(\mathbf{x}_i^c, C_j) + \frac{d_u}{d} s(\mathbf{x}_i^u, C_j), \tag{5}$$

Subsequently, the object-cluster similarity metric can be obtained based on the definitions of $s(\mathbf{x}_i^c, C_j)$ and $s(\mathbf{x}_i^u, C_j)$.

Similarity Metric for Categorical Attributes (I)

Taking into account the unequal importance of different categorical attributes for clustering analysis, the computation of $s(\mathbf{x}_i^c, C_j)$ should be further modified with

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j), \quad (6)$$

where w_r is the weight of categorical attribute A_r satisfying $0 \leq w_r \leq 1$ and $\sum_{r=1}^{d_c} w_r = 1$.

That is, the object-cluster similarity for categorical part is the *weighted summation of the similarity between the cluster and each attribute value*.

Similarity Metric for Categorical Attributes (II)

Definition 1

The similarity between a categorical attribute value x_{ir}^c and cluster C_j is defined as:

$$s(x_{ir}^c, C_j) = \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}, \quad (7)$$

where $\sigma_{A_r = x_{ir}^c}(C_j)$ counts the number of objects in cluster C_j that have the value x_{ir}^c for attribute A_r, $NULL$ refers to empty.

Therefore, the object-cluster similarity for categorical part is calculated by

$$s(\mathbf{x}_i^c, C_j) = \sum_{r=1}^{d_c} w_r s(x_{ir}^c, C_j) = \sum_{r=1}^{d_c} w_r \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)}. \quad (8)$$

Calculation of Categorical Attribute Weights

From the view point of information theory, the **importance** of any categorical attribute A_r can be estimated by

$$H_{A_r} = -\frac{1}{m_r}\sum_{t=1}^{m_r} p(a_{rt})\log p(a_{rt}) \text{ with } p(a_{rt}) = \frac{\sigma_{A_r=a_{rt}}(X)}{\sigma_{A_r \neq NULL}(X)}, \quad (9)$$

where $a_{rt} \in dom(A_r)$, X is the whole data set and m_r is the number of values can be chosen by A_r.

The **weight** of each attribute is then computed as

$$w_r = H_{A_r} / \sum_{t=1}^{d_c} H_{A_t}. \quad (10)$$

Similarity Metric for Numerical Attributes (I)

- It is a universal law that the distance and perceived similarity between numerical vectors are related via an exponential function as follows:

$$s(\mathbf{x}_A, \mathbf{x}_B) = \exp(-Dis(\mathbf{x}_A, \mathbf{x}_B)), \quad (11)$$

  where Dis stands for a distance measure.

- Moreover, to avoid the influence of different magnitudes of distances, we can further use proportional distance instead of absolute distance.

Similarity Metric for Numerical Attributes (II)

Definition 2

The object-cluster similarity between numerical vector $\mathbf{x}_i^u$ and cluster C_j is given by

$$s(\mathbf{x}_i^u, C_j) = \exp\left(-\frac{Dis(\mathbf{x}_i^u, \mathbf{c}_j)}{\sum_{t=1}^{k} Dis(\mathbf{x}_i^u, \mathbf{c}_t)}\right), \quad (12)$$

where $\mathbf{c}_j$ is the center of all numerical vectors in cluster C_j.

In practice, different distance metrics can be utilized to calculate $Dis(\mathbf{x}_i^u, \mathbf{c}_j)$.

Calculation of Object-cluster Similarity

According to previous descriptions, the object-cluster similarity metric for mixed data is given by

$$s(\mathbf{x}_i, C_j) = \frac{d_c}{d} \sum_{r=1}^{d_c} \left(\frac{H_{A_r}}{\sum_{t=1}^{d_c} H_{A_t}} \cdot \frac{\sigma_{A_r = x_{ir}^c}(C_j)}{\sigma_{A_r \neq NULL}(C_j)} \right) + \frac{d_u}{d} \exp\left(- \frac{Dis(\mathbf{x}_i^u, \mathbf{c}_j)}{\sum_{t=1}^{k} Dis(\mathbf{x}_i^u, \mathbf{c}_t)} \right), \quad (13)$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, k$.

Outline

Clustering Criterion

- We concentrate on hard partition only, i.e., $q_{ij} \in \{0, 1\}$.
- Given a set of N objects, the optimal $\mathbf{Q}^* = \{q^*_{ij}\}$ in Eq. (1) can be given by

$$q^*_{ij} = \begin{cases} 1, & \text{if } s(\mathbf{x}_i, C_j) \geq s(\mathbf{x}_i, C_r), 1 \leq r \leq k, \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

- Similar to the learning procedure of k-means, an iterative algorithm can be conducted to implement the clustering analysis.

OCIL Algorithm

Iterative clustering learning based on object-cluster similarity metric:

Require: data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, number of clusters k
Ensure: cluster label $Y = \{y_1, y_2, \ldots, y_N\}$

Calculate the importance of each categorical attribute if applicable
Set $Y = \{0, 0, \ldots, 0\}$ and randomly select k initial objects, one for each cluster
repeat
  Initialize $noChange = true$
  for $i = 1$ **to** N **do**
    $y_i^{(new)} = \arg \max_{j \in \{1,\ldots,k\}} [s(\mathbf{x}_i, C_j)]$
    if $y_i^{(new)} \neq y_i^{(old)}$ **then**
      $noChange = false$
      Update the information of clusters $C_{y_i^{(new)}}$ and $C_{y_i^{(old)}}$, including the frequency of each categorical value and the centroid of numerical vectors
    end if
  end for
until $noChange$ is $true$
return Y

Outline

Evaluation Criteria

- *Clustering Accuracy (ACC)*:

$$ACC = \frac{\sum_{i=1}^{N} \delta(c_i, map(r_i))}{N},$$

where $map(r_i)$ maps the obtained cluster label r_i to the equivalent label from the data corpus by using the Kuhn-Munkres algorithm.

- *Clustering Error Rate*:

$$e = 1 - ACC$$

.

Mixed Data Sets

Table 1 : Statistics of mixed data sets

Data set	Instance	Attribute ($d_c + d_u$)	Class
Statlog Heart	270	$7+6$	2
Heart Disease	303	$7+6$	2
Credit Approval	653	$9+6$	2
German Credit	1000	$13+7$	2
Dermatology	366	$33+1$	6
Adult	30162	$8+6$	2

Clustering Errors on Mixed Data Sets

Table 2 : Clustering errors of OCIL on mixed data sets in comparison with k-prototype and k-means

Data set	K-means	K-prototype	OCIL
Statlog	0.4047±0.0071	0.2306±0.0821	**0.1716±0.0065**
Heart	0.4224±0.0131	0.2280±0.0903	**0.1644±0.0030**
Credit	0.4487±**0.0016**	0.2619±0.0976	**0.2519**±0.0966
German	0.3290±0.0014	0.3289±**0.0006**	**0.3057**±0.0007
Dermatology	0.7006±**0.0216**	0.6903±0.0255	**0.3051**±0.0896
Adult	0.3869±**0.0067**	0.3855±0.0143	**0.3079**±0.0305

Comparison of Convergence Rate

Table 3 : Comparison of average convergent time and iterations between k-prototype and OCIL

Data set	Time		Iterations	
	K-prototype	OCIL	K-prototype	OCIL
Statlog	0.0519s	**0.0516**s	3.09	**3.07**
Heart	0.0639s	**0.0576**s	3.54	**3.02**
Credit	**0.1323**s	0.1625s	**3.18**	4.26
German	0.2999s	**0.2023**s	5.29	**3.15**
Dermatol	0.3674s	**0.1888**s	7.27	**4.32**
Adult	15.2795s	**9.6774**s	10.93	**6.78**

Categorical Data Sets

Table 4 : Statistics of categorical data sets

Data set	Instance	Attribute	Class
Soybean	47	35	4
Breast	699	9	2
Vote	435	16	2
Zoo	101	16	7

Clustering Errors on Categorical Data Sets

Table 5 : Comparison of clustering errors obtained by three different methods on categorical data sets

Data set	H's k-modes	N's k-modes	OCIL
Soybean	0.1691±0.1521	**0.0964**±0.1404	0.1017±**0.1380**
Breast	0.1655±0.1528	0.1356±0.0016	**0.0934**±**0.0009**
Vote	0.1387±0.0066	0.1345±0.0031	**0.1213**±**0.0010**
Zoo	0.2873±0.1083	0.2730±**0.0818**	**0.2681**±0.0906

H's k-modes: original k-modes algorithm (Huang, SIGMOD'97);
N's k-modes: k-modes algorithm with Ng's dissimilarity metric (Ng et al., TPAMI'07);

Outline

Conclusion

- A general clustering framework based on object-cluster similarity has been proposed.
- A unified similarity metric for both categorical and numerical attributes has been presented.
- An iterative algorithm which is applicable to clustering analysis on various data types has been introduced.
- The advantages of the proposed method have been experimentally demonstrated in comparison with the existing counterparts

Outline

Acknowledgment

- Collaborative Graduate Program in Design, Kyoto University;
- Department of Computer Science, Hong Kong Baptist University.

References

1. Michalski, R.S., Bratko, I., Kubat, M.: Machine learning and data mining: methods and applications. Wiley, New York (1998)
2. Hsu, C.C.: Generalizing self-organizing map for categorical data. IEEE Transactions on Neural Networks **17**(2) (March 2006) 294–304
3. Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering **14**(4)(July/August 2002) 673–690
4. Zaki, M.J., Peters, M.: Click: Mining subspace clusters in categorical data via k-partite maximal cliques. In: Proceedings of the 21st International Conference on Data Engineering. (2005) 355–356
5. Barbara, D., Couto, J., Li, Y.: Coolcat: An entropy-based algorithm for categorical clustering. In: Proceedings of the 11th ACM Conference on Information and Knowledge Management. (2002) 582–589
6. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining. (1997) 21–24
7. Huang, Z.: A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Know ledge Discovery. (1997) 1–8
8. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(3) (2007) 503–507

Thank You!