Clustering Large Datasets of Mixed Units

Simona Korenjak-Černe, Vladimir Batagelj
University of Ljubljana, Faculty of Mathematics and Physics, and
Institute of Mathematics, Physics and Mechanics, Dept. of TCS
Jadranska 19, 1 000 Ljubljana, Slovenia

Summary: In the paper we propose an approach for clustering large datasets of mixed units based on representation of clusters by distributions of values of variables over a cluster – histograms, that are compatible with merging of clusters. The proposed representation can be used also for clustering symbolic data. On the basis of this representation the adapted versions of leaders method and adding method were implemented. The proposed approach was successfully applied to several large datasets.

Key words: large datasets, clustering, mixed units, distribution description compatible with merging of clusters, leaders method, adding method.

1. Introduction

In the paper we propose an approach for clustering large datasets of *mixed* (nonhomogenous) units – units described by variables measured in different types of scales (numerical, ordinal, nominal).

Let E be a finite *set of units*. Its nonempty subset $C \subseteq E$ is called a *cluster*. A set of clusters $\mathbf{C} = \{C_i\}$ forms a *clustering*. In this paper we shall require that every clustering $\mathbf{C}$ is a partition of E.

The clustering problem can be formulated as an optimization problem:

Determine the clustering $\mathbf{C}^* \in \Phi$, for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where Φ is a set of *feasible clusterings* and $P : \Phi \rightarrow \mathbb{R}_0^+$ is a *criterion function*.

Most approaches to the clustering are based explicitly or implicitly on some kind of criterion function that measures the deviation of units from selected description or *representative* of its cluster. Usually $P(\mathbf{C})$ takes the form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, R_C)$$

where R_C is a representative of cluster C and d a dissimilarity.

The cluster representatives usually consist of variable-wise resumes of variable values over the cluster, ranging from central values (mean, median, mode), (min, max) intervals (symbolic objects; Diday 1997), Tukey's (1977) box-and-whiskers plots, to detailed distribution (histogram or fitted curve).

In this paper we investigate a description taken from the middle of this range satisfying two requirements:

- it should require a fixed space per variable;
- it should be *compatible* with merging of clusters – knowing the description of two clusters we can, without additional information, produce the description of their union.

Note that only some of the cluster descriptions are compatible with merging: mean (as sum and number of units) for numerical variables and (min, max) intervals for ordinal variables.

2. A cluster representative compatible with merging

In our approach a cluster representative is composed from representatives of each variable. They are formed, depending on the type of the scale in which a variable is measured, in the following way. Let $\{V_i, i = 1, \ldots, k\}$ be a partition of the range of values of variable V. Then we define for a cluster C the sets

$$Q(i, C; V) = \{X \in C : V(X) \in V_i\},\ i = 1, \ldots, k$$

where $V(X)$ denotes the value of variable V on unit X.

In the case of ordinal variable V (numerical scales are a special case of ordinal scales) the partition $\{V_i, i = 1, \ldots, k\}$ usually consists of intervals determined by selected threshold values $t_0 < t_1 < t_2 < t_3 < \cdots < t_{k-1} < t_k$, $t_0 = \inf V$, $t_k = \sup V$.

For nominal variables we can obtain the partition, for example, by selecting $k-1$ values $t_1, t_2, t_3, \ldots, t_{k-1}$ from the range of variable V (usually the most frequent values on E) and setting $V_i = \{t_i\},\ i = 1, \ldots, k-1$; and putting all the remaining values in class V_k.

Using these sets we can introduce the *concentrations*

$$q(i, C; V) = \operatorname{card} Q(i, C; V)$$

and *relative concentrations*

$$p(i, C; V) = \frac{q(i, C; V)}{\operatorname{card} C}$$

It holds

$$\sum_{i=1}^{k} p(i, C; V) = 1$$

The description of variable V over C is the vector of concentrations of V.

Note that in the special case $C = \{X\}$ we have

$$p(i, C; V) = \begin{cases} 1 & X \in Q(i, C; V) \\ 0 & \text{otherwise} \end{cases}$$

Similarly we can consider a missing value on V for unit X by setting $p(i, \{X\}; V) = \frac{1}{k}$, $i = 1, \ldots, k$. Using concentrations we can describe also symbolic data.

It is easy to see that for two clusters C_1 and C_2, $C_1 \cap C_2 = \emptyset$ we have

$$Q(i, C_1 \cup C_2; V) = Q(i, C_1; V) \cup Q(i, C_2; V)$$

$$q(i, C_1 \cup C_2; V) = q(i, C_1; V) + q(i, C_2; V)$$

The description is compatible with merging.

The threshold values are usually determined in a way that for the given set of units E (or the space of units $\mathcal{E}$) it holds $p(i, E; V) \approx \frac{1}{k}$, $i = 1, \ldots, k$.

As a compatible description of nominal variable over a cluster C also its range $V(C)$ can be used, since we have

$$V(C_1 \cup C_2) = V(C_1) \cup V(C_2)$$

3. Dissimilarity

Most clustering methods are based on some dissimilarity between clusters or a unit and a cluster. For our descriptions we define

$$d(C_1, C_2; V) = \frac{1}{2} \sum_{i=1}^{k} |p(i, C_1; V) - p(i, C_2; V)|$$

and in the case of nominal variable described by set of values

$$d(C_1, C_2; V) = \frac{\text{card}\,(V(C_1) \oplus V(C_2))}{\text{card}\,(V(C_1) \cup V(C_2))}$$

We shall use the abbreviation $d(X, C; V) = d(\{X\}, C; V)$. In both cases it can be shown that

- $d(C_1, C_2; V)$ is a semidistance on clusters; i.e.
  1. $d(C_1, C_2; V) \geq 0$
  2. $d(C, C; V) = 0$
  3. $d(C_1, C_2; V) + d(C_2, C_3; V) \geq d(C_1, C_3; V)$
- $d(C_1, C_2; V) \in [0, 1]$

and for the vector representation also

$$X \in Q(i, E; V) \Rightarrow d(X, C; V) = 1 - p(i, C; V)$$

The semidistances on variables can be combined into a semidistance on units by

$$d(C_1, C_2) = \sum_{j=1}^{m} \alpha_j d(C_1, C_2; V_j), \qquad \sum_{j=1}^{m} \alpha_j = 1$$

where m is the number of variables and α_j weights (Batagelj and Bren, 1995); often $\alpha_j = \frac{1}{m}$.

4. Clustering procedures

The proposed approach allows us to first recode the original nonhomogenous data to a uniform representation by integers – indices of intervals. For the recoded data efficient clustering procedures can be built by adapting leaders method (Hartigan, 1975) or adding clustering method (Batagelj and Mandelj, 1993). In this paper, because of limited space, we shall describe only the procedure based on the dynamic clusters method (a generalization of the leader method).

To describe the dynamic clusters method for solving the clustering problem let us denote (Diday, 1979; Batagelj, 1985): Λ a set of *representatives*; $\mathbf{L} \subseteq \Lambda$ a *representation*; Ψ a set of *feasible representations*; $W : \Phi \times \Psi \to \mathbb{R}_0^+$ *extended criterion function*; $G : \Phi \times \Psi \to \Psi$ *representation function*; $F : \Phi \times \Psi \to \Phi$ *clustering function* and suppose that the following conditions are satisfied:

W0. $P(\mathbf{C}) = \min_{\mathbf{L}\in\Psi} W(\mathbf{C}, \mathbf{L})$

the functions G and F tend to improve (diminish) the value of the extended criterion function W:

W1. $W(\mathbf{C}, G(\mathbf{C}, \mathbf{L})) \leq W(\mathbf{C}, \mathbf{L})$
W2. $W(F(\mathbf{C}, \mathbf{L}), \mathbf{L}) \leq W(\mathbf{C}, \mathbf{L})$

then the dynamic clusters method can be described by the scheme:

$\mathbf{C} := \mathbf{C}_0; \mathbf{L} := \mathbf{L}_0;$
repeat
    $\mathbf{C} := F(\mathbf{C}, \mathbf{L});$
    $\mathbf{L} := G(\mathbf{C}, \mathbf{L})$
until the goal is attained

To this scheme corresponds the sequence $v_n = (\mathbf{C}_n, \mathbf{L}_n), n \in \mathbb{N}$ determined by relations

$$\mathbf{C}_{n+1} = F(\mathbf{C}_n, \mathbf{L}_n) \qquad \text{and} \qquad \mathbf{L}_{n+1} = G(\mathbf{C}_{n+1}, \mathbf{L}_n)$$

and the sequence of values of the extended criterion function $u_n = W(\mathbf{C}_n, \mathbf{L}_n)$.

Let us assume the following model $\mathbf{C} = \{C_i\}_{i\in I}$, $\mathbf{L} = \{L_i\}_{i\in I}$, $\mathbf{L}(X) = \{L_i : X \in C_i\}$ and further $L = [L(V_1), \ldots, L(V_m)]$, $L(V) = [s(1, L; V), \ldots, s(k, L; V)]$, $\sum_{j=1}^{k} s(j, L; V) = 1$ and

$$d(C, L; V) = \frac{1}{2} \sum_{j=1}^{k} |p(j, C; V) - s(j, L; V)|$$

For

$$W(\mathbf{C}, \mathbf{L}) = \sum_{X\in E} d(X, \mathbf{L}(X)) = \sum_{i\in I} p(C_i, L_i)$$

where

$$p(C, L) = \sum_{X\in C} d(X, L)$$

we define $F(\mathbf{L}) = \{C'_i\}$ by

$$X \in C'_i : i = \min \operatorname*{Argmin}_j \{d(X, L_j) : L_j \in \mathbf{L}\}$$

each unit is assigned to the nearest leader; and we define $G(\mathbf{C}) = \{L'_i\}$ by

$$L'_i = \operatorname*{argmin}_{L \in \Psi} p(C, L)$$

To solve the last optimization problem we consider

$$p(C, L; V) = \sum_{X \in C} d(X, L; V) = \operatorname{card}(C) - \sum_{j=1}^{k} q(j, C; V) s(j, L; V)$$

This expression has a minimum value iff the sum has a maximum value. The unique symmetric optimal solution is

$$s(i, L'; V) = \begin{cases} \frac{1}{t} & j \in M \\ 0 & \text{otherwise} \end{cases}$$

where $M = \{j : q(j, C; V) = \max_i q(i, C; V)\}$ and $t = \operatorname{card} M$.

Evidently W0, W1 and W2 hold. It holds also $P(\mathbf{C}) = W(\mathbf{C}, G(\mathbf{C}))$.

Usually $t = 1$ – we include in the representative of a cluster the most frequent range of values of variable on this cluster. This provides us with very simple interpretations of clustering results.

For example, in a clustering of types of cars we obtained among 14 clusters:

Cluster 5: (42 units) coupes
doors = 2, height(mm) ∈ (– , 1388], power(KW) ∈ (125, –), acceleration time(s) ∈ (– , 9.1], max speed(km/h) ∈ (215 , –), price(1000 SIT) ∈ (6550, –).

Cluster 8: (57 units) minivans
doors = 5, passengers = 7, length(mm) ∈ (4555, 4761], height(mm) ∈ (1490, –), fuel tank capacity(l) ∈ (66, 76], weight(kg) ∈ (1540, –), wheelbase(mm) ∈ (2730, –).

Cluster 9: (193 units) small-cars
doors = 4, passengers = 5, length(mm) ∈ (– , 4010], width(mm) ∈ (– , 1680], luggage capacity(l) ∈ (–, 279], fuel tank capacity(l) ∈ (–, 48], weight(kg) ∈ (– , 940], power(KW) ∈ (– , 55], torque ∈ (– , 124], brakes = disc/drums, max speed(km/h) ∈ (– , 163], price(1000 SIT) ∈ (– , 2100].

Cluster 13: (62 units) sport-utility
doors = 5, height(mm) $\in$ (1490, –), fuel tank capacity(l) $\in$ (76, –), weight(kg) $\in$ (1540, –), cargo capacity(l) $\in$ (560, –), torque $\in$ (270, –).

For another criterion function

$$W(\mathbf{C}, \mathbf{L}) = \alpha \sum_{i \in I} d(C_i, L_i) + \sum_{i \in I} p(C_i, L_i)$$

where α is a large constant, the first term can be set to 0 by setting $s(j, L_i; V) = p(j, C_i; V)$, $j = 1, \ldots, k$. If this defines the function G, and F is defined as in the previous case, we obtain a procedure which works very well on real data. We conjecture that in this case it holds in general that $W(\mathbf{C}_{i+1}, \mathbf{L}_{i+1}) \leq W(\mathbf{C}_i, \mathbf{L}_i)$ and $W(\mathbf{C}_{i+1}, \mathbf{L}_i) \leq W(\mathbf{C}_i, \mathbf{L}_{i-1})$. Note also that for the obtained (local) minimum $(\mathbf{C}^*, \mathbf{L}^*)$ it holds

$$P(\mathbf{C}^*) = \min_{\mathbf{L} \in \Psi} W(\mathbf{C}^*, \mathbf{L}) = W(\mathbf{C}^*, G(\mathbf{C}^*)) = \sum_{C \in \mathbf{C}^*} p(C, G(C))$$

5. Conclusion

We successfully applied the proposed approach to the dataset of types of cars (1349 units) and also to some large datasets from AI collection

`http://www.ics.uci.edu/~mlearn/MLRepository.html`

An extended version of this paper is available at

`http://vlado.fmf.uni-lj.si/pub/cluster/`

References

Batagelj, V. (1985). Notes on the dynamic clusters method, in: *IV conference on applied mathematics*, Split, May 28-30, 1984. University of Split, Split, 139-146.

Batagelj, V. & Bren, M. (1995). Comparing Resemblance Measures, *Journal of Classification*, **12**, 1, 73-90.

Batagelj, V. & Mandelj, M. (1993). Adding Clustering Algorithm Based on L-W-J Formula, Paper presented at: *IFCS 93*, Paris, 31.aug-4.sep 1993.

Diday, E. (1979). *Optimisation en classification automatique*, Tome 1.,2.. INRIA, Rocquencourt, (in French).

Diday, E. (1997). Extracting Information from Extensive Data sets by Symbolic Data Analysis, in: *Indo-French Workshop on Symbolic Data Analysis and its Applications*, Paris, 23-24. September 1997, Paris IX, Dauphine, 3-12.

Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New York.

Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.