# Adapted Methods For Clustering Large Datasets Of Mixed Units

Simona Korenjak-Černe
IMFM Ljubljana, Dept. of TCS,
Jadranska 19, 1 000 Ljubljana, Slovenia
E-mail: simona.korenjak@fmf.uni-lj.si

**Keywords:** clustering, large datasets, mixed units, hierarchical clustering, cluster description compatible with merging of clusters, leaders method, adding clustering method

**Edited by:** Cene Bavec and Matjaž Gams

**Received:** October 17, 1999 **Revised:** October 30, 1999 **Accepted:** December 11, 1999

*The proposed clustering methods are based on the recoding of the original mixed units and their clusters into a uniform representation. The description of a cluster consists for each variable of the frequencies of the variable values over its range partition. The proposed representation can be used also for clustering symbolic data. On the basis of this representation the adapted version of the leaders method and adding clustering method were implemented. We describe both approaches, which were successfully applied on several large datasets.*

## 1 Introduction

Abstraction is the main tool to deal with large amounts of data. The first step is to identify groups of similar units - clusters. In data analysis this is a task of clustering methods. The most popular are hierarchical clustering methods. Because they usually use a similarity/dissimilarity matrix they are appropriate only for clustering datasets of a moderate size (some hundreds of units). On the other hand well known nonhierarchical methods are mostly implemented for datasets of variables measured in the same scale type (such as for example 'k-means method'). Because of these limits we are searching for new clustering methods or at least trying to adapt known methods to be appropriate for clustering **large** datasets of **mixed** units, where variables (properties) of the units are measured in different scales.

Let $E$ be a finite *set of units*. A 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 $\Phi$ is a set of *feasible clusterings* and $P : \Phi \to \mathbb{R}_0^+$ is a *criterion function*.

In many clustering methods the criterion function measures the deviation of units from *representatives* (*leaders*) of corresponding clusters. In our method we select the criterion function in one of the most frequent 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 $d$ is a dissimilarity.

The cluster representatives usually consist of variable-wise summaries of variable values over the cluster. For homogeneous units with only numerical variables their means are usually selected as representatives of clusters. For mixed (nonhomogeneous) units a new description has to be selected.

In this paper we investigate a description satisfying two additional requirements:

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

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

## 2 A description of a cluster

For our adaptation of clustering methods to be appropriate for clustering large datasets of mixed units, we choose a cluster description based on frequencies. For this purpose, the ranges of the variables are partitioned into selected number of classes. Let $\{V_i, i = 1, \ldots, k(V)\}$ be a partition of the range of values of variable $V$ (the number of classes $k(V)$ depends on variable). Then we can define for a cluster $C$ the sets

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

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

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

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

Units are not necessarily represented with single value for each variable, but they can also be represented with frequencies over the classes of variables ranges.

Using classes of ranges we get *frequencies*

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

and *relative frequencies*

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

Note that

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

When only a single unit is in the cluster $C$ we get

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

We can add, for each variable, a new class for a missing value and treat it as a special value, or we can also consider a missing value on $V$ for a unit $X$ by setting $p(i, \{X\}; V) = \frac{1}{k(V)}$, $i = 1, \ldots, k(V)$ (or by some other distribution).

It is easy to see that such a description is compatible with merging, because for two disjoint clusters $C_1$ and $C_2$ 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 threshold values are usually determined in such a way that, for the given set of units $E$ (or the space of units $\mathcal{E}$), it holds that $p(i, E; V) \approx \frac{1}{k(V)}$, $i = 1, \ldots, k(V)$.

As a compatible description of nominal variable over a cluster $C$ also its range $V(C)$ can be used, since we have $V(C_1 \cup C_2) = V(C_1) \cup V(C_2)$.

**Example:** Recoding of flags dataset

Original data are taken from the address
`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/flags`
(Flags from Collins Gem Guide to Flags, donated by Richard S. Forsyth.)

Let us consider the following three variables:
- *population* (in round millions),
- *mainhue* (predominant color in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)),
- *text* (1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise).

The range of the variable *population* is divided into 5 classes with approximately the same number of units in each of them. The ranges of the others variables are so small that we put for discretization of them each possible value in a separate class:

| var= *population* | var= *mainhue* | var= *text* |
|---|---|---|
| map | map | map |
| 1 = {0} | 1 = {red} | 1 = {0} |
| 2 = (0, 4] | 2 = {green} | 2 = {1} |
| 3 = (4, 18] | 3 = {blue} | |
| 4 = (18, 158] | 4 = {gold} | |
| 5 = (158, 1100] | 5 = {white} | |
| | 6 = {black} | |
| | 7 = {orange} | |

ORIGINAL DATA

| unit ID | population | mainhue | text |
|---|---|---|---|
| Austria | 8 | red | 0 |
| New-Zealand | 2 | blue | 0 |
| Saudi-Arabia | 9 | green | 1 |
| Switzerland | 6 | red | 0 |
| USA | 231 | white | 0 |

RECODED DATA

| | | | |
|---|---|---|---|
| Austria | 3 | 1 | 1 |
| New-Zealand | 2 | 3 | 1 |
| Saudi-Arabia | 3 | 2 | 2 |
| Switzerland | 3 | 1 | 1 |
| USA | 5 | 5 | 1 |

In our case for each variable a unit is represented with index of the appropriate class.

The description of a cluster $C_6$ (only for considered variables) obtained with the leaders method is

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $q(C_6; population)$ | 8 | 1 | 1 | 0 | 0 | | |
| $q(C_6; mainhue)$ | 1 | 0 | 9 | 0 | 0 | 0 | 0 |
| $q(C_6; text)$ | 6 | 4 | | | | | |

>From this description we can see that in eight countries the population is less than a million, in one country is between 1 and 4 millions and in one country population is between 4 and 18 millions. This cluster is one of the seven clusters obtained with the adapted version of the leaders method with maximal allowed dissimilarity between a unit and its nearest leader 0.5. In one of the countries flags red is a dominant color and all of the remaining units have blue

mainhue. In six units some text is presented and four units inside cluster $C_6$ have no text in their description. For better understanding, cluster $C_6$ consists of: Bermuda, Brit. Virg. Isles, Cayman Islands, Falklands Malvi, Fiji, Hong-Kong, Montserrat, St. Helena, Turks Cocos Islands and Tuvalu.

# 3 Dissimilarity between clusters

Let us return to our approach to clustering problem as an optimization problem. After deciding to use the uniform representation of units and clusters, we have to define a measure of dissimilarity between clusters (a unit is a special case of a cluster with only one element). First the dissimilarity between clusters for individual variable $V$ is defined as

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

We shall use the abbreviation

$$d(X, C; V) = d(\{X\}, C; V).$$

In both cases it can be shown that

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

and for the representation of a single unit also

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

The semidistances on clusters for individual variable can be combined into a semidistance on clusters for complete descriptions by

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

where $m$ is the number of variables and $\alpha_j$ are weights ($\alpha_j \geq 0$ and $\sum_{j=1}^{m} \alpha_j = 1$); often $\alpha_j = \frac{1}{m}$. We can use weights to consider dependencies among variables or to tune the dissimilarity to a given learning set in AI applications.

# 4 Clustering procedures

In the proposed approach the original nonhomogeneous data are first recoded to a uniform representation. For the recoded data efficient clustering procedures can be built by adapting leaders method (Hartigan, 1975) or adding clustering method (Zupan 1982, Jambu and Lebeaux 1983, Batagelj and Mandelj 1993).

## 4.1 The adapted version of the leaders method

The adapted version of the leaders method is a variant of a dynamic clustering method (Diday 1979, Batagelj 1985). To describe the dynamic clustering method for solving the clustering problem let us denote: $\Lambda$ a set of *representatives*; $\mathbf{L} \subseteq \Lambda$ a *representation*; $\Psi$ a set of *feasible representations*; $P : \Phi \to \mathbb{R}_0^+$ *criterion function*; $G : \Phi \to \Psi$ *a representation function*; $F : \Psi \to \Phi$ *a clustering function* and suppose that the functions $G$ and $F$ tend to improve (diminish) the value of the criterion function $P$. Then a simple version of the dynamic clustering method can be described by the scheme:

$\mathbf{L} := \mathbf{L}_0$;
**repeat**
  $\mathbf{C} := F(\mathbf{L})$
  $\mathbf{L} := G(\mathbf{C})$
**until** the leaders stabilize

We begin with the initial representation and then repeat to assign each unit to the nearest leader and after that select leaders for each (new) cluster until we reach the minimum of the criterion function or until the leaders don't change any more (local minimum).

Let us assume the following model $\mathbf{C} = \{C_i\}_{i \in I}$, $\mathbf{L} = \{L_i\}_{i \in I}$, $\mathbf{L}(X) = L_i : X \in C_i$ (the nearest leader to the unit $X$), $L = [L(V_1), \ldots, L(V_m)]$, $L(V) = [s(1, L; V), \ldots, s(k(V), L; V)]$, $\sum_{j=1}^{k(V)} s(j, L; V) = 1$ (the description of a leader has the same form as the description of a cluster) and

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

For selected criterion function

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

where

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

we define $F(\mathbf{L}) = \{C_i'\}$ with

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

This means that each unit is assigned to the (first) nearest leader.
We define $G(\mathbf{C}) = \{L_i'\}$ with

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

The unique symmetric optimal solution of this optimization problem is

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

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

The representative (leader) of a cluster is obtained from the most frequent range(s) of values of variables on this cluster.

**Example:** Leader of a cluster

For the description of a cluster $C_6$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $q(C_6; population)$ | 8 | 1 | 1 | 0 | 0 | | |
| $q(C_6; mainhue)$ | 1 | 0 | 9 | 0 | 0 | 0 | 0 |
| $q(C_6; text)$ | 6 | 4 | | | | | |

the optimal leader $L_6$ is

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $q(C_6; population)$ | 1 | 0 | 0 | 0 | 0 | | |
| $q(C_6; mainhue)$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $q(C_6; text)$ | 1 | 0 | | | | | |

The characteristics of the cluster are

| | |
|---|---|
| population = less than a million | 80 % |
| mainhue in the flag = blue | 90 % |
| text in the flag = no | 60 % |

For example, 80% of all countries in the cluster $C_6$ have less than a million inhabitants, 90% of all countries flags have blue mainhue and 60% of the flags in the cluster have no text in their descriptions.

**Properties of the leaders method**

The main properties of the adapted version of the leaders method are:

1. Selection of the leaders and formation of new clusters diminish the value of the criterion function.
2. The program always stops (converges). The number of iterations is usually less than 10.
3. The program is suitable for clustering (very) large datasets.
4. The leaders descriptions provide us with simple interpretations of clustering results.

## 4.2 The adapted adding method

The adding clustering method is a hierarchical clustering method in which a new unit is added in a clustering tree. Each vertex corresponds to a cluster. For large datasets usually only the upper part of the hierarchy is maintained, the lower levels subtrees are replaced by 'bags' containing all units from a subtree.

We shall use the same description of a cluster (vertex) and the same definition of a dissimilarity as in the leaders method. Every time we add a unit in a cluster (vertex) the frequencies are recalculated. There are two possible ways how to add a new unit:

a) To maximize the dissimilarity between clusters (sons) of the current vertex or,

b) To minimize the dissimilarity from clusters (sons) of the current vertex.

In the first case (see Figure 1) the dissimilarities between both sons of a current vertex are calculated. Because of greedy approach the case with the biggest dissimilarity is chosen: $\max\{d(C_p \cup \{X\}, C_q), d(C_p, C_q \cup \{X\}), d(C, \{X\})\}$.

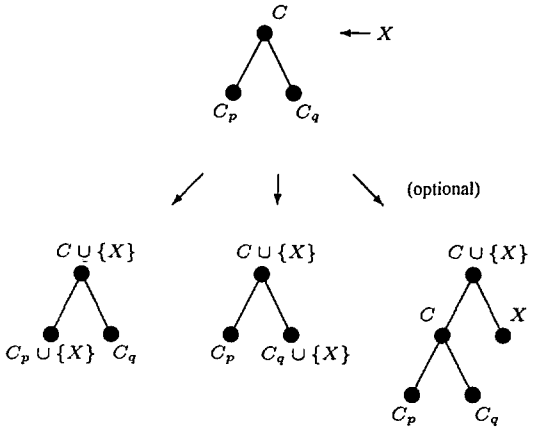


Figure 1: Maximize the dissimilarity between clusters

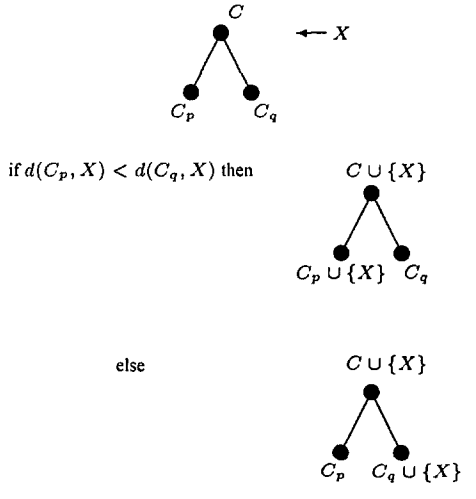


Figure 2: Minimize the dissimilarity from clusters

In the second case (see Figure 2) the dissimilarities from each of the sons of current vertex are calculated and the unit is added to the nearest one: $\min\{d(C_p, \{X\}), d(C_q, \{X\})\}$.

The proposed approaches can also be extended on non-binary trees.

The adding clustering method has some advantages:

1. Presentation of the result with a tree.

2. It can be used for classification.

3. Speed up - if the tree has many (hundreds of) leaves which represent the leaders, it is more efficient adding unit into the tree with this method than to calculate the dissimilarities to each of the leaders.

A drawback of the adding method is that the result strongly depends on the ordering of the input sequence of units. A possible way to avoid this problem is to select a 'good' initial tree. We are suggesting to built the initial tree with some agglomerative hierarchical clustering method on leaders obtained with the leaders method. The other possibility is to include balancing of the tree in the process of adding new unit. Both possibilities are still under the development.

# 5 Conclusion

We successfully applied the proposed approach on the dataset of types of cars (1 349 units, 26 variables), on the ISSP data (45 784 units, 21 variables) and also on some large datasets from AI collection

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

The first version of the program ClaMix (based on the adapted version of the leaders method) and some of the results are available at

`http://www.educa.fmf.uni-lj.si/datana/`

**Acknowledgment:** This work was supported by the Ministry of Science and Technology of Slovenia, Project J1-8532.

# References

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

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

[3] 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.

[4] Brucker, P. (1978) On the complexity of clustering problems. *Lecture Notes in Economics and Mathematical Systems 175*, in: *Optimization and Operations Research, Proceedings*, Bonn. Henn,R., Korte,B., Oettli,W. (Eds.), Springer-Verlag, Berlin 1978.

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

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

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

[8] Jambu, M. & Lebeaux, M.O. (1983) *Cluster Analysis and Data Analysis*. North-Holland Publishing Company.

[9] Korenjak-Černe, S. & Batagelj, V. (1998). Clustering large datasets of mixed units. *Advances in Data Science and Classification*. Rizzi, A., Vichi, M. and Bock, H.-H. (Eds.), Springer, Berlin, 1998, p. 43-48.

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

[11] Zupan, J. (1982) *Clustering of Large Data Sets*. Research Studies Press, John Wiley & Sons LTD.

[12] Flags from Collins Gem Guide to Flags. Collins Publishers (1986). Donated by Richard S. Forsyth.
`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/flags`