CREATIVE MATH. & INF.
17 (2008), 98 - 105

The convergence of some clustering techniques for elements ideally grouped in clusters

DANA AVRAM LUPŞA

ABSTRACT. Depending on the characteristics of the clusters we want to determine, different clustering techniques are employed. Data characterization is usually not perfect, the model suffers and the clustering results are not always what the user expected. This paper argues that, for elements ideally grouped in clusters, clustering techniques converges. We propose a characterization of elements ideally grouped in clusters and prove the uniqueness of the optimum clusters for some different clustering criteria.

1. INTRODUCTION

Unsupervised classification, or clustering, is a method that infers groups based on inter-object similarity. It tends to be an unsupervised learning technique.

A vast collection of clustering algorithms is available [2, 4]. Depending on the characteristics of the groups we want to determine, different clustering techniques are employed. Usually, different clustering techniques rely on different data characterization. The multitude of existing clustering techniques as well as the much discussed problem of ideal clusters are indicators of the fact that data characterization is not perfect. The model suffers and the clustering results are not always what the user expected.

When clustering, we can choose to permit (or not to permit) for an object to be member of two clusters. This choice determines the existence of two classes of clustering methods. Soft clustering methods determine the degree of membership of each object in each cluster. Hard clustering algorithms assign each object to exactly one cluster. In this paper, we limit our discussion to hard clustering problem.

This paper argues that, in case of elements ideally grouped in clusters, some clustering algorithms obtain the ideal clusters. That is, the algorithms converge to the same solution.

This paper is organized as follows: Section 2 presents the fundamental problem of clustering. Section 3 describes two traditional clustering techniques. Section 4 defines the notion of elements ideally grouped in clusters. The properties of clustering in case of elements ideally grouped in clusters are presented in Sections 5 and 6. This paper ends with some conclusions.

2. THE PROBLEM

Unsupervised classification algorithms partition a set of objects in groups. The fundamental problem of clustering [3, 6] can be stated as follows:

Received: 11.09.2006. In revised form: 5.10.2007.
2000 *Mathematics Subject Classification.* 03C13, 03C45, 62H30, 91C20.
Key words and phrases. *Clustering, cluster analisys.*

Given:

- $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ - a set of elements
- $simi : \mathcal{X} \times \mathcal{X} \to \Re^+$ - a similarity function between elements
- k , $1 \leq k \leq m$ - a pre-determined number of the clusters

Results: classes of elements $C_1, C_2, \ldots, C_k$ with the next properties:

- $C_1, C_2, \ldots, C_k \in \mathcal{P}art(\mathcal{X})$ (form a partition of $\mathcal{X}$), that is:
  $C_i \cap C_j = \phi, \forall i \neq j$
  $\bigcup_{i=1}^{k} C_i = \mathcal{X}$
- the elements from the same class to be as much similar as possible (function $simi$), and the objects from different classes to be as dissimilar as possible.

3. CLUSTERING TECHNIQUES

Traditionally, clustering techniques are broadly divided in hierarchical and partitional.

3.1. **Hierarchical clustering.** Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters. Hierarchical clustering methods are categorized into agglomerative and divisive. An agglomerative clustering starts with one point clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion is achieved. To merge or split subsets of elements, the similarity between individual elements has to be generalized to the similarity between subsets.

In this paper we are going to study an agglomerative clustering, with the stopping criterion constructed on the request that a fixed number of clusters must be achieved. Single link will be considered as the inter-cluster similarity measure.

3.2. **Partitional clustering.** Partitional algorithm divide data into several subsets. Elements are iteratively re-assigned to clusters and clusters are gradually improved.

One approach to data partitioning is to start with the definition of objective function depending on a partition. It is considered that to compute inter- and intra- cluster similarity measure, based on elements pair-wise similarities, would be too expensive. Such methods use cluster representatives to compute the objective function. Using unique cluster representatives make the objective function to become linear. Depending on how representatives are constructed, the partitional algorithms are subdivided into K-means and K-medoids methods. In K-mean, a cluster is represented by its centroid, which is a mean of elements (viewed as points) within a cluster. K-medoids use the most appropriate element within a cluster to represent it.

In this paper we study properties of some objective functions used in case of K-medoids methods.

4. Elements Ideally Grouped in Clusters

To establish some clusters means to approximate the similarities among the elements from the same cluster with the same value. In an ideal case, we don't have to do those approximations; the similarities among elements in a cluster have the same values. Based on this observation, the definitions for elements ideally grouped in clusters is built [1].

Definition 4.1. Consider $\mathcal{X}$ a set with m elements and $simi : \mathcal{X} \times \mathcal{X} \to \Re^+$ a similarity function between elements.
The elements of the set $\mathcal{X}$ are ideally grouped according with the similarity function $simi$, into the set of ideal clusters $\{C_1, C_2, \dots C_k\}$ if:

(1) $\exists \alpha$ so that $\forall x, y \in C_i : \ simi(x, y) = \alpha, \forall i \in \{1, 2, \dots k\}$
(the similarity between any two elements from a cluster has the value α);

(2) $\exists \beta < \alpha$ so that $\forall x \in C_i, y \in C_j, i \neq j : \ simi(x, y) = \beta, \forall i, j \in \{1, 2, \dots k\}$
(the similarity between any two elements that are not in the same cluster has the same value, smaller than α).

The Definition 4.1 of the elements ideally grouped in clusters guarantees that the similarities between any two elements from the same cluster are equal, but they are greater than the similarities between elements from different clusters. In the conditions of Definition 4.1 we say that elements are ideally grouped in clusters, and that $\{C_1, C_2, \dots C_k\}$ are ideal clusters.

5. Properties of Hierarchical Clustering When Elements Are Ideally Grouped in Clusters

5.1. **The Algorithm.** We built our discussion on a version based on the agglomerative hierarchical clustering algorithm presented in [5], which build hierarchical clusters until a stop condition is met, and which stores all the levels of clusters. The algorithm is presented in Table 1.

The stop condition can be $|C^{<step>}| = 1$, when we want to build all levels of clusters, or $|C^{<step>}| = k$, when we know that we need k clusters, and $k \geq 2$.

The similarity $Sim(C_u, C_v)$ can be a function depending on the similarity between elements in the clusters C_u and C_v. We will consider the *single-link* similarity:

$$Sim_{SL}(C_u, C_v) = \max_{x_i \in C_u, y_j \in C_v} simi(x_i, y_j) \tag{5.1}$$

In what follows, we are going to prove that, if we have objects ideal grouped in the ideal clusters $C_1, C_2, \dots C_k$, the agglomerative hierarchical clustering algorithm, with the stop condition : the number of the cluster on the last level is k, will obtain the ideal clusters, when the similarity function between clusters is SL.

5.2. **When Elements Are Ideally Grouped in Clusters.** The algorithm (see Table 1) obtains the ideal clusters when the elements are ideally grouped in clusters. This is stated by Proposition 5.1.

Proposition 5.1. *If elements are ideally grouped in clusters $C_1, C_2, \dots C_k$, according with the similarity function between elements $simi : \mathcal{X} \to \mathcal{X}$, the agglomerative hierarchical clustering algorithm (see Table 1), with:*

Input

The set $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ of m objects to be clustered,
the similarity function $Sim : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \Re$.

Output

The set of hierarchical clusters $C = \{C_1^0, C_2^0, \ldots, C_j^{<step>}\}$

BEGIN
FOR i = 1 TO m DO $C_i^0 \leftarrow \{x_i\}$ ENDFOR
$step = 0$
$C = \{C_1^0, C_2^0, \ldots, C_m^0\}$
(A) ————
WHILE *not* $stopcondition(C)$ DO
(B) ————
$(C_{u^*}^{<step>}, C_{v^*}^{<step>}) = argmax_{C_u^{<step>}, C_v^{<step>}} Sim(C_u^{<step>}, C_v^{<step>}), u <> v$
$C_*^{<step>} = C_{u^*}^{<step>} \cup C_{v^*}^{<step>}$
$C^{<step>+1} = C^{<step>} - \{C_{u^*}^{<step>}, C_{v^*}^{<step>}\} \cup C_*^{<step>}$
$C = C \cup C^{<step>+1}$
$step = step + 1$
(C) ————
ENDWHILE
(D) ————
END

TABLE 1. Hierarchical agglomerative clustering algorithm

- *stop condition : the number of clusters on the current level is k,*
- *similarity function between clusters: $Sim = Sim_{SL}$ (equation 5.1)*

will obtain, on the last level, the ideal clusters $C_1, C_2, \ldots C_k$.

Proof.

- From Definition 4.1: $simi(e_i, e_j) \leq \alpha, \forall e_i \in \mathcal{X}, e_j \in \mathcal{X}$ (5.1.1)
- If, at a certain moment, there are more than k clusters on the level $< step >$, and k is the number of ideal clusters $\Rightarrow$ there are at least two clusters C_u, C_v which have elements from the same ideal cluster C_l. That is: $\exists\, C_l$ - ideal cluster, $e_u \in C_u^{<step>}$ and $e_v \in C_v^{<step>}$ so that $e_u \in C_l$ and $e_v \in C_l \Rightarrow simi(e_u, e_v) = \alpha$.
  By using (5.1.1) $\Rightarrow Sim_{SL}(C_u, C_v) = \alpha$. (5.1.2)
- At every step, at point B in the algorithm, $stopcondition(C)$ is not fulfilled. That means that there are more than k clusters on the level $< step >$ at point B. (5.1.3)
- From (5.1.2) and (5.1.3) $\Rightarrow$ at each step, in point B, there are at least two clusters C_u, C_v so that $Sim_{SL}(C_u, C_v) = \alpha$. Then the similarity Sim_{SL} between clusters that are unified is α. (5.1.4)
- We are going to prove that $\forall C_r^{<step>}$, $\exists$ ideal cluster C_l so that $C_r^{<step>} \subseteq C_l$, in each cut point: (A), (B), (C) and (D). We prove that the condition is true in point (A) and then we prove that the condition holds for all the possible paths between (A), (B), (C) and (D). (5.1.5)
  At point (A) in the algorithm
  - $C_i^0 = \{e_i\} \Rightarrow \forall C_i^0$ $\exists C_l$ - ideal cluster so that $C_i^0 \subseteq C_l$

  The path: (A) $\rightarrow$ (B)
  no variable value changes: condition true in point A $\Rightarrow$ true in B
  The path: (B) $\rightarrow$ (C)

suppose condition true in point (B)
(because, if we get here, we come from a cut point where condition is true)
$\forall C_i^{<step>} \exists C_l$ - ideal cluster so that: $C_i^{<step>} \subseteq C_l$ (5.1.5.1)
– $C_{u^*}^{<step>}$ and $C_{v^*}^{<step>}$ are two clusters (at point B)
From (5.1.5.1) $\Rightarrow \exists\, C_{lu}$ and C_{lv} - ideal clusters so that:
$C_{u^*}^{<step>} \subseteq C_{lu}$ and $C_{v^*}^{<step>} \subseteq C_{lv}$ (5.1.5.2)
– cluster changes consist on unification of $C_{u^*}^{<step>}$ and $C_{v^*}^{<step>}$;
from (5.1.4): $Sim_{SL}(C_{u^*}^{<step>}, C_{v^*}^{<step>}) = \alpha$
$\Rightarrow \exists x_i \in C_{u^*}^{<step>}$ and $\exists x_j \in C_{v^*}^{<step>}$ so that: $simi(x_i, x_j) = \alpha$,
(from Definition 4.1) $\Rightarrow \exists C_{l*}$ - ideal cluster so that: $x_i, x_j \in C_{l*}$ (5.1.5.3)
– From (5.1.5.2), (5.1.5.3),
$\Rightarrow x_i \in C_{u^*}^{<step>} \subseteq C_{lu}$ and $x_i \in C_{l*} \Rightarrow x_i \in C_{l*} \cap C_{lu}$
$\Rightarrow C_{l*} \cap C_{lu} \neq \phi$, where C_{l*} , C_{lu} - ideal clusters $\Rightarrow C_{l*} = C_{lu}$
$\Rightarrow C_{u^*}^{<step>} \subseteq C_{l*}$ (5.1.5.4)
– From (5.1.5.2), (5.1.5.3) results, in a similar manner, that $C_{v^*}^{<step>} \subseteq C_{l*}$ (5.1.5.5)
– From (5.1.5.4), (5.1.5.5) $\Rightarrow C_{u^*}^{<step>} \cup C_{v^*}^{<step>} \subseteq C_{l*}$

The path: (C) $\rightarrow$ (B)
no variable value changes: condition true in point C $\Rightarrow$ true in B
The path: (C) $\rightarrow$ (D)
no variable value changes: condition true in point C $\Rightarrow$ true in D
The path: (A) $\rightarrow$ (D)
no variable value changes: condition true in point A $\Rightarrow$ true in D

- From (5.1.5): $\forall C_r^D \exists C_l$ ideal cluster so that: $C_r^{<step>} \subseteq C_l$ (5.1.6)
- Each level of clusters is also a partition of the given set $\mathcal{X}$.
The last level of clusters is get at point D. Because we can get in point D only if $stopcondition$ is true, the number of clusters at last level is k. Let us note them: $C_1^D, C_2^D, \ldots C_k^D$
On the other hand, the ideal clusters are also k and form a partition. (5.1.7)
- From (5.1.6), (5.1.7)
$\Rightarrow \quad \{C_1^D, C_2^D, ..., C_k^D\} = \{C_1, C_2, ..., C_k\}$.

□

6. Properties of Partitional Clustering When Elements Are Ideally Grouped in Clusters

Some classes of partitional algorithms divide data into subsets that optimize objective functions (or criterion function) [7]. It is considered that those functions fully characterize the clusters needed for some specific problems. Algorithms based on criterion functions build clusters that reach the optimum (maximum or minimum) of the criterion function.

The Definition 4.1 introduces some restrictions on pair-wise similarity among elements. We study some criterion functions that can be fully described by similarities between elements. We suppose that the similarities between elements satisfy the conditions from Definition 4.1. We also suppose that we know that we want to obtain k clusters. We are going to study if there is a relation between the clusters for which the optimum of criterion function is reached and the ideal clusters. We prove that the optimum is achieved only for the k ideal clusters for the two criterion functions that are studied.

6.1. **Property of I_1 Criterion.**

I_1 criterion function maximizes the sum of the average pairwise similarities between the elements assigned to each cluster, weighted

according to the size of each cluster. I_1 function must be maximized and its formula is:

$$I_1 = \sum_{r=1}^{k} \frac{1}{n_r} \left(\sum_{x_i, x_j \in S_r} simi(x_i, x_j) \right) \tag{6.2}$$

Note that this equation includes the pairwise similarities involving the same pairs of elements, as it is given in [7].

Proposition 6.1. *When elements are ideally grouped in k clusters, the maximum of I_1 criterion function (equation 6.2) is reached when and only when elements are grouped in the ideal clusters.*

Proof.

- In the following, for a cluster C_r, we denote by n_r the number of elements in C_r : $n_r = card(C_r)$. Then:
$$\sum_{r=1}^{k} n_r = \sum_{r=1}^{k} card(C_r) = m$$
- Let us consider that the ideal k clusters are $C_1, C_2, ..., C_k$ and the set of clusters that maximizes the formula 6.2 are $S_1, S_2, ..., S_k$. We will show that the partition $S_1, S_2, ..., S_k$ is the same with the partition $C_1, C_2, ..., C_k$.
- We compute a maximum for I_1 criteria, in conditions of Definition 4.1. From the definition, we use the inequality: $simi(x_i, x_j) \leq \alpha, \forall x_i, x_j \in \mathcal{X}$

$$\begin{aligned} I_1 &= \sum_{r=1}^{k} \frac{1}{n_r} \left(\sum_{x_i, x_j \in S_r} simi(x_i, x_j) \right) \\ &\leq \sum_{r=1}^{k} \frac{1}{n_r} \left(\sum_{x_i, x_j \in S_r} \alpha \right) = \sum_{r=1}^{k} \frac{1}{n_r} \left(n_r^2 \times \alpha \right) \\ &= \alpha \times \sum_{r=1}^{k} n_r = \alpha \times m \end{aligned}$$

- We prove that the maximum of I_1 is reached when clusters $S_1, S_2, \ldots, S_k$ are the ideal clusters $C_1, C_2, \ldots, C_k$.
The similarity among elements from an ideal cluster is α
$\Rightarrow \sum_{x_i, x_j \in C_r} simi(x_i, x_j) = \sum_{x_i, x_j \in C_r} \alpha$

$$\begin{aligned} I_1 &= \sum_{r=1}^{k} \frac{1}{n_r} \left(\sum_{x_i, x_j \in S_r} simi(x_i, x_j) \right) \\ &= \sum_{r=1}^{k} \frac{1}{n_r} \left(\sum_{x_i, x_j \in C_r} \alpha \right) = \alpha \times m \end{aligned}$$

- We prove that, if the maximum is reached in case of a set of k clusters, then those clusters are the ideal clusters.
Suppose that there are a set of k clusters, $S_1, S_2, ..., S_k$, different than $C_1, C_2, ..., C_k$ and with the property that I_1 is maximum: $I_1 = \alpha \times m$.
  - From the condition that I_1 reaches maximum, we have that:
  $simi(x_i, x_j) = \alpha, \forall x_i, x_j \in S_r$
  From Definition 4.1 $\Rightarrow \exists C_l \; \forall x_i, x_j \in S_r$ so that $x_i, x_j \in C_l$
  $\Rightarrow \forall S_r \; \exists C_l$ so that $S_r \subseteq C_l$ (6.1.1)

– On the other hand, $S_1, S_2, ..., S_k$ and $C_1, C_2, ..., C_k$ are clusters (partitions of $\mathcal{X}$, formed by k subsets of $\mathcal{X}$) (6.1.2)
– From (6.1.1) and (6.1.2)
$\Rightarrow$ the partition $S_1, S_2, ..., S_k$ is the same with the partition $C_1, C_2, ..., C_k$.

□

6.2. **Property of I_2 Criterion.** I_2 criterion function is used by the popular vector-space variant of the K-means algorithm. In this algorithm each cluster is represented by its centroid ($centr_r$) and the goal is to find the clustering solution that maximizes the similarity between each element and the centroid of the cluster that is assigned to. I_2 function must be maximized and its formula is:

$$I_2 = \sum_{r=1}^{k} \sum_{x_i \in S_r} simi(x_i, centr_r) \tag{6.3}$$

I_2 criterion is specific to K-means algorithm. It uses a computed centroid of each cluster. Such a task is used in cases when each element is a vector of attribute values, that is, each element can be seen as a point in a multidimensional space. In order to avoid this, we look at the centroid in a K-medoid manner and consider as centroid of a cluster the most representative element from that cluster. In this case, $simi(x_i, centr_r)$ is the similarity between an ordinary element x_i of the cluster and the most representative element of the cluster which is $centr_r$. In the ideal case, we consider that the centroid of a cluster can be any element of the cluster.

Note that this equation includes the self-similarities between centroid of each cluster, as it is defined in [7].

Proposition 6.2. *When elements are ideally grouped in k clusters, the maximum of I_2 criterion function (equation 6.3), in a K-medoid method of determining the centroid $centr_r$ of a cluster, is reached when and only when elements are grouped in the ideal clusters.*

Proof.

- We know that: $centr_r \in S_r \subseteq \mathcal{X} \Rightarrow centr_r \in \mathcal{X}$
- We compute a maximum for I_2 criteria, in conditions of Definition 4.1. From the definition, we use the inequality: $simi(x_i, x_j) \leq \alpha, \forall x_i, x_j \in \mathcal{X}$

$$\begin{aligned} I_2 &= \sum_{r=1}^{k} \sum_{x_i \in S_r} simi(x_i, centr_r) \\ &\leq \sum_{r=1}^{k} \sum_{x_i \in S_r} \alpha = \sum_{r=1}^{k} n_r \times \alpha = \alpha \times \sum_{r=1}^{k} n_r \\ &= \alpha \times m \end{aligned}$$

- We prove that the maximum of I_2 is reached when clusters $S_1, S_2, \ldots, S_k$ are the ideal clusters $C_1, C_2, \ldots, C_k$.
The similarity between elements from an ideal cluster is α. For any representative of the cluster $centr_r$, which is an element of the cluster C_r, the next relation holds: $simi(x_i, centr_r) = \alpha, \forall x_i \in C_r$. That implies:

$$\begin{aligned} I_2 &= \sum_{r=1}^{k} \sum_{x_i \in C_r} simi(x_i, centr_r) \\ &= \sum_{r=1}^{k} \sum_{x_i \in C_r} \alpha = \sum_{r=1}^{k} n_r \times \alpha = \alpha \times m \end{aligned}$$

- We prove that, if the maximum is reached in case of a set of k clusters, then those clusters are the ideal clusters.
  Suppose that there are a set of k clusters, $S_1, S_2, ..., S_k$, different than $C_1, C_2, ..., C_k$ and with the property that I_2 is maximum: $I_2 = m \times \alpha$.
  - From the condition that I_2 is maximum, we have that:
    $simi(x_i, centr_r) = \alpha, \forall x_i \in S_r$
    From Definition 4.1, the similarity between two elements is α only if they are in the same ideal cluster: $\forall x_i \in S_r \ \exists C_{il}$ - ideal cluster so that: : ($x_i \in C_{il}$ and $centr_r \in C_{il}$)
    $\Rightarrow centr_r \in C_{il}$, $\forall x_i \in S_r \Rightarrow \forall C_{il} \ \exists C_l$ - ideal cluster so that: $C_l = C_{il}$
    $\Rightarrow \exists C_l$ so that ($\forall x_i \in S_r$: $x_i \in C_l$) and $centr_r \in C_l$
    $\Rightarrow \exists C_l$ so that $S_r \subseteq C_l$ (6.2.1)
  - On the other hand, $S_1, S_2, ..., S_k$ and $C_1, C_2, ..., C_k$ are clusters (partitions of $\mathcal{X}$, containing k subsets of $\mathcal{X}$). (6.2.2)
  - From (6.2.1) and (6.2.2)
    $\Rightarrow$ the partition $S_1, S_2, ..., S_k$ is the same with the partition $C_1, C_2, ..., C_k$

□

7. CONCLUSIONS

In this paper we have studied the properties of some clustering techniques in case of elements ideally grouped in clusters. We have proved that some clustering techniques converges in case of elements ideally grouped in clusters, in the sense that they all obtains the ideal clusters. We have proved that the agglomerative hierarchical clustering algorithm obtains the ideal clusters when elements are ideally grouped in clusters. We have also studied clustering techniques that are based on criterion functions and that obtain the clusters that maximizes that criterion. We have proved that the criterion functions I_1 and I_2 reach their optimum when and only when the elements are grouped in the ideal clusters, in case when elements are ideally grouped in clusters.

REFERENCES

[1] Avram-Lupşa, D., *Extraction of Semantic Information from Texts Using Unsupervised Classification*, PhD. Thesis, Babes-Bolyai University Cluj-Napoca, 2006 (in Romanian)
[2] Berkin, P., *Survey of Clustering Data Mining Techniques*, 2002
[3] Dumitrescu, D., *Mathematical Foundations of the Classification Theory*, Romanian Academy Press, 1995
[4] Jain, A.K. and Murty, M.N., *Data Clustering: A Review, ACM Computing Surveys*, Vol. 31, No.3, September 1999
[5] Manning, C. and Schutze, H., *Foundation of statistical natural language processing*, MIT, 1999
[6] Pop, H.F., *Intelligent Systems in Classification Problems*, Mediamira, 2004 (in Romanian)
[7] Zhao, Y. and Karypis, G., *Criterion Functions for Document Clustering. Experiments and Analysis*, Technical Report #01-40, University of Minnesota, Department of Computer Science, 2002

BABEŞ-BOLYAI UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
1, M. KOGĂLNICEANU STREET
400084 CLUJ-NAPOCA, ROMANIA
E-mail address: dana@cs.ubbcluj.ro