Some Characterizations on Clustering Using Equivalent Graphs

Silviu Bârză
Faculty of Mathematics and Informatics
Spiru Haret University
Bucharest, Romania
silviu.barza@gmail.com

Abstract: In this paper I wish to give some characterizations on a clustering method applied to sets information with connectivity links inside each set of information. Because such a structure can be associated with a graph, in a previous paper I consider that clustering is done using so called equivalent graphs and I have proposed a clustering process. The characterization include the maximum number of sets of information in a cluster, the probability that a new set of information placed in cluster to have identical associated graph with a set of information already existent in that cluster. I will introduce too a possible distance between clusters.

Paper sumary

1. Introduction
2. Maximum number of sets of information in a cluster
3. Distance proposal
4. Conclusion

References

Introduction

Definition 1.3 On the set of graphs, we say that two graphs G and H are equivalent if and only if, by definition, $G \cong H$ and we write $G \overline{\sim}_i H$.

We consider that a cluster C is characterized by

$$char(C) = [n, m, G=(V,F)] ;$$

where: n is the number of vertexes in associated graph G (and the number of individual pieces of information in observations included in cluster); m is the number of edges in associated graph G (and the number of known connection between individual pieces of information in observations included in cluster); and G=(V,F) represents the graph associated with observations included in cluster

Introduction

The proposed process to make clusters using equivalence relationship on graphs, [Bârză, 2012], is:

Step 1. It is generated the graph $G=(V,F)$ associated with E.

Step 2. If does not exists a cluster C_i with $n_i=|V|$ and $m_i=|F|$, for $1 \leq i \leq p$, then we go to step 5.

Step 3. Let $C_{i_1}, C_{i_2}, \ldots, C_{i_r}$ be the clusters for which $n_{i_j}=|V|$ and $m_{i_j}=|F|$,

$1 \leq j \leq r$, $i_1, i_2, \ldots, i_r \in \{1,2,\ldots,p\}$.

Step 4. We test if G and G_{i_j} are isomorphic graphs, $1 \leq j \leq r$. If such a test

is true then if $G_k = G$ go to step 6, otherwise continue.

Step 5. It is created a new cluster C_{p+1} with $n_{p+1}=|V|$, $m_{p+1}=|F|$ and

$G_{p+1}=(V,F)$, replace p with p + 1 and stop.

Step 6. Place E in cluster G_k and stop.

Maximum number of sets of information in a cluster

Denition 2.2 Let G=(V,E) be a graph with A_G its adjacencies matrix and n=|V|. The rows i and j from A_G are dependent if for the transposition σ_{ij}=(i,j), we have $A_G = A_{\sigma_{ij}(G)}$. If rows i and j are not dependent, we call the independent.

Proposition 2.2 Let G=(V,E) the graph associated with sets of information from a cluster with |V|=n and A_G the adjacencies matrix for G. If $A_1, A_2, \ldots, A_k$ is the dependences based partitions for the rows of A_G then, the maximum number of different graphs associated for sets of information in cluster is:

$$maxA_G = \frac{n!}{\prod_{i=1}^{k-1} |A_i|!}$$

Maximum number of sets of information in a cluster

Because the value $maxA_G$ is a constant for a cluster we must not calculate it every time we use the cluster and so we may extend the cluster characteristics to include this new information. The form of the cluster characteristic will be now:

$$Char(C)=(n,m,G=(V,F),maxA_G)$$

Distance proposal

We begin by considering the space of clusters

$$\{C_1, C_2, \ldots, C_t\}$$

so that for any i = 1,2,…,t we have

$$Char(C_i) = (n_i, m_i, G_i=(V_i, E_i), maxA_{G_i}):$$

A very easy way to specify a distance between clusters is to consider only the information from $Char(C_i)$ and so we can define a function

$$d(C_i, C_j) = |n_i - n_j| + |m_i - m_j| + |maxA_{G_i} - maxA_{G_j}|.$$

References

Bârză, S, Equivalent and Strong Equivalent Graphs with Application in Clustering, "Analele Universitatii Spiru Haret, Seria Matematica-Informatica", vol.VII, nr. 2, Editura Fundatiei Romania de Maine, Bucharest, 2012.