Some Characterizations on Clustering Using Equivalent Graphs

Silviu Bârzã
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.
Keywords: *isomorphic graphs, clustering.*
ACM/AMS Classification: 05C30, 05C60, 05C75, 91G20

1. Introduction

Firstly we include here some necessary definitions and results as they appear in [Bârzã, Morogan, 2008].

Definition 1.1 *A graph G is a pair (V, E) where V is a finite set specifying the vertexes (generally considered as $V = \{1, 2, \ldots, n\}$) and E is the set of unordered pairs of numbers from V (generally presented as subsets of two values from V) named edges.*

Definition 1.2 *Let $G = (V, E)$ and $H = (W, F)$ be two graphs. G and H are named isomorphic if and only if there exists a function $f : V \to W$ so that f is bijective and $a = \{x, y\} \in E$ if and only if $f(a) = \{f(x), f(y)\} \in F$.*

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 $G \simeq_i H$.*

Definition 1.4 *Let $G = (V, E)$ be a graph. The graph $H = (V, F)$ with F subset of E is called* ***partial graph of*** *G*

Now, let us remember the basis elements that allow clustering using equivalence graphs, given in [Bârzã, 2012].

We consider a collection of observations $E_1, E_2, \ldots, E_k$, where for any i, $1 \leq i \leq k$, E_i is a set of information formed by individual date $d_{i,1}, d_{i,2}, \ldots, d_{i,m_i}$

and is known that there exists connection between $d_{i,x}$ and $d_{i,y}$ for some $x, y \in \{1, 2, \ldots, m_i\}$.

With this considerations, for any i, $1 \leq i \leq k$, we consider the points $p_{i,1}, p_{i,2}, \ldots, p_{i,m_i}$ as vertexes of a graph $G_i = (V_i, F_i)$ and so

$$V_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,m_i}\}.$$

To define F_i we consider that if there exists a connection between $d_{i,x}$ and $d_{i,y}$, then we have an edge between $p_{i,x}$ and $p_{i,y}$, so $\{p_{i,x}, p_{i,y}\} \in F_i$.

With the above process we associate to the collection of sets of information $E_1, E_2, \ldots, E_k$, a collection of graphs $G_1, G_2, \ldots, G_k$ using $G \simeq_i H$ relationship for clustering.

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 (so $|V| = n$ and $|F| = m$).

The proposed process to make clusters using equivalence relationship on graphs, [Bârzã, 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 $G_k \cong G$ 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 $G_{p+1} = G(V, F)$; replace p with $p + 1$ and stop.

Step 6. Place E in cluster G_k and stop.

2. **Maximum number of sets of information in a cluster**

Considering only the elements of graph theory, especially the definition of isomorphic graphs, the answer related to maximum number of different sets of information seems to be very simple because we must consider the identification of vertexes and so we can have an infinite number of different isomorphic graphs.

If we restrict the graphs to those in which we have the same set of vertexes for all graphs, then the answer results from:

Proposition 2.1 *Let $G = (V, E)$ and $H = (V, F)$ be two graphs with the same sets of vertexes. Then $G \cong H$ if and only if it exists a permutation sigms defined on V so that $H = \sigma(G)$, with $\sigma(G) = (\sigma(V), \sigma(E)) = (V, \sigma(E))$, where $\sigma(E) = \{\{\sigma(x), \sigma(y)\} \,|(x, y) \in E\}$.*

If we give a graph G, then for any permutation σ on V, from proposition 2.1, we have $G \cong \sigma(G)$ and it follows that the number of isomorphic graphs with G is equal with the number of permutation defined on V and so is equal $n!$.

Here we consider that the only graph isomorphic with G and identical with G is obtained only for identity permutation, means the permutation in which for any $x \in V$, $\sigma(x) = x$.

In the case of clustering, we consider that the facts presented above are not applicable. The clustering is a process related to data analysis and the last task is a part of statistics. In statistics we work with samples which are representative for a population and the individuals from a sample is randomly choose and the results of analysis do not depend on individuals.

So, the identification used for the vertexes in graph associated to a set of information is not relevant and we consider that is a mistake to use for the graphs only the set representation to say that two graphs are not identical.

The way in which a graph is represented do not depend by the identifications of its vertexes is the algebraic representation using adjacencies matrix as appears in the next definition.

Definition 2.1 *Let $G = (V, E)$ be a graph in which we may consider that $V = \{x_1, x_2, \ldots, x_n\}$. The matrix $A_G = (a_{ij})_{i,j=1,2,\ldots,n}$ defined by $a_{ij} = 1$ if and only if $\{x_i, x_j\} \in E$ and $a_{ij} = 0$ otherwise is called the **adjacencies matrix of graph** G.*

Using the definition 2.1 we may say that two graphs G and H are identical if and only if $A_G = A_H$.

From this point of view, the value $n!$ for the maximum number of different sets of information in a cluster do not seem to be a reasonable one. To show this affirmation let us consider an extreme case of G, namely if G is a complete graph with n vertexes.

Strictly by the point of view of graph theory, any two complete graphs with n vertexes are isomorphic. As regards the adjacencies matrix, their matrices are equals and so we may consider that any two complete graphs are identical and this is the reason to designate complete graphs with n vertexes as K_n.

On this consideration we may conclude that the maximum number of distinct graphs associated with sets of information, which are complete graphs with n vertexes is equals with 1.

For exact determination of the maximum number of distinct graphs associated with sets of information we must considerate the rows (or columns) of the adjacencies matrix of graph G We will consider that:

Definition 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 $\sigma_{ij} = (i, j)$, we have $A_G = A_{\sigma_{ij}(G)}$. If rows i and j are not dependent, we call the **independent**.*

Using definition 2.2 we may realize a ***dependences based partitioning*** of the set $\{1, 2, \ldots, n\}$ with the sets $I_1, I_2, \ldots, I_k$ so that:

1. for any $j \in I_k$ and any $i \neq j$, $1 \leq i \leq n$, rows i and j are independent;
2. for any $s \in \{1, 2, \ldots, k\}$ for any $i, j \in I_s$ the rows i and j are dependent and for any $i \in I_s$ and for any $j \in \{1, 2, \ldots, n\} / I_s$ rows i and j are independent.

Now, the problem of determination of the maximum number of different graphs associated with sets of information is reduced to the problem of the number of matrix that can be built keeping de dependences on rows with a given matrix with n rows.

If we consider firstly the rows with indexes in I_1 and we fix those rows in the matrix, the remaining problem is like initial one but for $k-1$ sets in the dependences based partition and with $n - |I_1|$ as numbers of rows in the matrix. So we may say that

$$maxA_G(n) = C_n^{|I_1|} \times maxA_G(n - |I_1|)$$

because to the choose of $|I_1|$ value from a set of n value represent a combinatorial problem where the order of values is not important.

Repeating the fixing operation $k-1$ times we have

$$maxA_G(n) = C_n^{|I_1|} \times C_{n-|I_1|}^{|I_2|} \times \ldots \times C_{n-|I_1|-|I_2|-\ldots-|I_{k-2}|}^{|I_{k-1}|} maxA_G(|I_K|)$$

because $n - |I_1| - |I_2| - \ldots - |I_{k-1}| = |I_k|$. Finally we obtain the following results.

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 $I_1, I_2, \ldots, I_k$ 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:*

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

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:

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

Because in considering two graphs as distinct we use the adjacencies matrix we propose that with every set of information E from cluster to store too the adjacencies matrix of the graph associated with E, which will de designated by A_E and so, instead of E we store (E, A_E). In agreement with the above result and final observation we must modify proposed algorithm for clustering to include the new form of characteristic and the new form of stored data.

Now, at the end of this section we may note that, in this point we are able to identify the probability that a new set of information placed in a cluster to be identical in structure with a set of information already stored. This probability is equal with:

$$\frac{|C|}{maxA_G}$$

where by $|C|$ we indicate the number of sets of information already stored in the cluster.

3. **Distance proposal**

In this section we want to define a distance on the space of clusters.
We begin by considering the space of clusters

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

so that for any $i = 1, 2, \ldots, t$ 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

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

Because in this definition d is defined as a sum of modules, it follows that $d(C_i, C_j) \geq 0$ for any $i, j = 1, 2, \ldots, t$. Also, if $d(C_i, C_j) = 0$ and d is defined as sum of non-negative values, it follows that $n_i = n_j$, $m_i = m_j$ and $maxA_{g_i} = maxA_{G_j}$. In the same time, because $maxA_{G_i}$ and $maxA_{G_j}$ are strictly related to G_i and G_j, from $maxA_{G_i} = maxA_{G_j}$ it follows that $G_i \cong G_j$ and so $C_i = C_j$. This facts show that $d(C_i, C_j) = 0$ if and only if $C_i = C_j$.

Now, if we consider three clusters C_i, C_j and C_k, we may write that

$$\begin{aligned} D(C_i, C_j) &= |n_i - n_j| + |m_i - m_j| + |maxA_{G_i} - maxA_{G_j}| = \\ &= |n_i - n_k + n_k - n_j| + |m_i - m_k + m_k - m_j| + \\ &+ |maxA_{G_i} - maxA_{G_k} + maxA_{G_k} - maxA_{G_j}| \leq \\ &\leq |n_i - n_k| + |n_k - n_j| + |m_i - m_k| + |m_k - m_j| + \\ &+ |maxA_{G_i} - maxA_{G_k}| + |maxA_{G_k} - maxA_{G_j}| = \\ &= d(C_i, C_k) + d(C_k, C_j). \end{aligned}$$

In this way we shown that d define a distance on the space of clusters, but we consider that the distance d is poor in information. This is the reason why in the future we must try to find another distance to be defined on clusters space.

4. **Conclusion**

In this paper we try to give some characterization on clusters formed using equivalence on graphs. For the future we wish to extend this work to clusters formed using strong equivalence on graphs and to continue to study the properties which may be associated with this two clustering process.

References

1. Bamdy, J.A., Murty U.S.R., *Graph Theory* "Springer-Verlag", 2007.
2. Bârzã, S., and Morogan, L.M., *Algoritmica grafurilor*, "Editura Fundatiei Romania de Maine", Bucharest, 2008.
3. Bârzã, 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.
4. Berge, C., *Graph Theory and Application*, romanian edition, "Editura Tehnica", Bucharest, 1971.
5. Popescu, D.R., *Combinatorica si teoria grafurilor*, "Editura Societatii de Stiinte matematice din Romania", Bucharest, 2005.
6. Tomescu, I., *Combinatorica si teoria grafurilor*, "Editura Universitatii Bucuresti", Bucharest, 1990.