Cluster Analysis with Deolistic Graphs

Antonio Juarez Alencar [a,*], Leôncio Teixeira Cruz [a],
Armando Leite Ferreira [b], Eber Assis Schmitz [c],
Priscila M. V. Lima [c], Fernando Silva Pereira Manso [c]

[a] *Institute of Mathematics, Federal University of Rio de Janeiro, P.O. Box 68530, 21941-590 - Rio de Janeiro - RJ, Brazil*

[b] *The COPPEAD School of Business, Federal University of Rio de Janeiro, P.O. Box 68514, 21945-970 - Rio de Janeiro - RJ, Brazil*

[c] *Electronic Computer Center, Federal University of Rio de Janeiro, P.O. Box 2324, 20001-970 - Rio de Janeiro - RJ, Brazil*

Abstract

This work introduces a particular family of acyclic graphs referred to as Deolistic graphs. The formal definition of a Deolistic graph is presented together with a new clustering algorithm that uses logical sentences to reveal hidden structures in data. The sentences yielded by the algorithm can be generated by automatic means, considerably reducing the complexity of data analysis and time spent on the clustering process, not to mention the usefulness of its results. Furthermore, these sentences are rules that can be easily understood, interpreted and disclosed to all interested parties, thereby improving communication and reducing misunderstandings.

Key words: Graphs; Cluster Analysis; Segmentation; Logical Systems; Trees.

1 Introduction

One of the most basic elements of the human learning process is our ability to group together objects that share similar properties. By acknowledging that certain groups of objects have similar properties, one can apply previously acquired knowledge to objects that one encounters for the first time. Without this fundamental ability, which human beings tend to take for granted, any intelligent creature would be in a situation where every object perceived would

* *Corresponding* author. Tel.: +55-21-2598-3310; Fax: +55-21-2598-3301
Email address: juarezalencar@dcc.ufrj.br (Antonio Juarez Alencar).

have to be treated as a unique entity completely dissimilar from anything else in the universe, making learning a much more complex task than it already is [11].

According to the 16^{th} century Swedish naturalist Carl Von Linné [3],

> "All the real knowledge which we possess depends on methods by which we distinguish the similar from the dissimilar. The greater the number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method, and the more necessary."

the relevance of classification for learning has motivated extensive mathematical investigation into the subject. In this respect, cluster analysis is the field of study in which quantitative methods for classification are conceived, analyzed and refined. From a strictly numerical point of view, clustering, i.e. classification is a multi-step interactive process where data on a given set of objects is gathered and used to successively classify these objects into segments, aiming to obtain segments that exhibit both high internal homogeneity and high external (between-segments) heterogeneity. The process comes to an end when an acceptable arrangement of segments of objects is obtained [15].

Once a clustering process has been finalized, two important tasks are to be carried out by the data analyst: identifying what distinguishes the segments yielded by the clustering process from each other and determining what properties the objects placed in the same segment have in common.

It should be noted that these distinctions and properties are ultimately what makes the arrangement of segments useful for practical purposes. Without a clear appreciation of these differences and similarities, learning will not have taken place in its broadest sense. Unfortunately, despite all the developments undergone by the clustering analysis over the last decades, current clustering methods tend to offer little support in this respect. As a result, data analysts must resort to descriptive statistics and hypothesis tests in order to obtain information as required. Everitt *et al.* [15] present a comprehensive discussion of the concepts and methods commonly used in cluster analysis.

All of these aspects tend to turn cluster analysis into a multivariate tool that many professionals find difficult to use, especially those with restrict background in mathematics and statistics. Regrettably, the vast majority of those who are currently employed by organizations of all sizes.

This work introduces a family of acyclic graphs referred to as Deolistic graphs, which uses logical sentences to bring together objects that share similar properties and reveal hidden structures in data. These sentences are rules that can be translated into any natural language and are easily understood, in-

terpreted and disclosed to all interested parties, improving communication, reducing misunderstandings and making cluster analysis a more easily accessible tool for professionals of all areas.

2 Conceptual Framework

Although the Greek philosopher and natural scientist Theophrastus (372-287 B.C.) is credited with the authorship of the first written works on classification due to his investigation into the taxonomy of plants, the term cluster analysis as a quantitative discipline was first introduced by Tryon [4] in the late 1930's. Since then, a multitude of clustering methods has been proposed. See [15] and [10] for a comprehensive account of these methods.

In the increasingly complex world in which we live, cluster analysis has permeated all branches of science and areas of business. For example, cluster analysis serve as a basis to study the chemical composition of planetary nebulae [12], to distinguish different kinds of stars [5], to show similarities and differences among families of methods, techniques tools and substances [8,1], to help the identification of biological and chemical agents responsible for different kinds of diseases and the development of healing therapies [18], to investigate the reasons that lead to psychological depression [7] etc. The list is literally endless.

According to Han and Kamber [10], all available clustering methods can be broadly classified into five major groups, namely

- *Partitioning methods* - given the number of segments one wants objects to be classified into, partitioning methods creates an initial arrangement of such segments. Subsequently, it uses an interactive relocation technique to improve the arrangement of segments by moving objects from one group to another. The process comes to an end when an acceptable arrangement of objects is obtained.
- *Hierarchical methods* - a family of methods that creates a hierarchical decomposition of a given set of objects by either successively splitting the existing segments of objects into new segments (divisive methods) or successively merging the existing segments to form larger segments (agglomerative methods).
- *Density-based methods* - a group of methods that regards the data space as dense regions of objects that are separated from each other by regions of low density. The basic idea resides in enlarging the existing segments of objects until the number of objects in the vicinity of those segments reaches some predetermined threshold.
- *Grid-based methods* - a clustering approach that divides the data space in

a finite number of cells forming a hierarchical grid structure.

- *Model-based methods* - a set of methods that attempts to optimize the fit between objects in the data space and a predetermined mathematical model, such as the hierarchical structure or neurons that serve as a basis for neural network competitive learning and self-organizing maps clustering algorithms.

However, despite the particular family of clustering methods that one finds suitable to use, there are generally nine steps to be followed in the clustering process of a set of objects, as enumerated in Table 1. It should be observed that the last two steps of a clustering process rely on establishing the dissimilarities that exist among the different segments yielded by the process and the similarities shared by objects placed in the same segment. Methods that strongly supports the establishment of these differences and similarities not only make clustering easier, but also help reducing the time and money spent on clustering analysis.

3 Deolistic Graph

According to Narsingh Deo [14], the Millican Chair Professor of Computer Science of the University of Central Florida, each vertice of a binary tree represents a test of some kind with two possible outcomes. Starting at the root, the outcome of the test carried out at that level indicates which of the descents vertices should be taken, where further tests are made. This process goes on until a pendant vertice is reached. It is important to notice that whatever statement can be made about the pendant vertice as a result of the testing process, it reveals a property that is held by that particular vertice. Based on the ideas of Deo, we define Deolistic graphs.

However, before the definition of Deolistic graphs can be presented there are two concepts that have to be defined in formal terms, namely the concept of observation, a collection of measurements of interest on a population of subjects, and the concept of variable, an attribute of a population to be measured in each observation.

Definition 1 *An observation o is a tuple $(a_1, a_2, \ldots, a_n)$, where each tuple component a_i may take value in a different domain. If $o = (a_1, a_2, \ldots, a_n)$ is an observation, then $o[i]$ is the value of the tuple component a_i of o, for $i \in [1\,..\,n]$.*

Definition 2 *Given a set of observations $\boldsymbol{O} = \{\ o_1,\ o_2,\ \ldots,\ o_m\ \}$, such that each $o_i \in \boldsymbol{O}$ is a tuple $(a_1, a_2, \ldots, a_n)$, a variable w_i in $\boldsymbol{O}$ takes value in the set $\{o_1[i],\ o_2[i],\ \ldots,\ o_m[i]\}$, for $i \in [1\,..\,n]$.*

Step	Label	Action
1	*Data gathering*	Select the objects to be classified and gather as much related relevant data as possible.
2	*Variable selection*	Choose the different dimensions (variables) to be considered in the classification process.
3	*Method selection*	Indicate the classification method that should be used to bring together similar objects.
4	*Similarity metric selection*	Elect the metric to be used by the classification method to calculate how similar any two objects are.
5	*Variable transformation*	In accordance with the similarity metric, make the necessary variable transformation so that they can be properly used by the classification method. This includes the standardization of variable values.
6	*Variable elimination*	Some clustering methods and similarity metrics impose strict restrictions on the kind of variable that may submitted to a clustering process. Eliminate the variables that do not comply with those restrictions.
7	*Number of segments*	Select the number of segments to be yielded by the classification process, if this is requested by the classification method.
8	*Classification*	Apply the classification method to the given set of objects.
9	*Evaluation*	Evaluate the arrangement of segments. If it is satisfactory, move on to the next and final step in the process, otherwise consider reviewing all your decisions from the first step.
10	*Labeling*	Label the segments of objects.

Table 1: General steps of a clustering process.

The notion of logical systems used in the definition that follows has been formalized by Goguen and Bustal with the concept of institutions [9].

Definition 3 *Given a logical system* $\boldsymbol{L}$ *and a set of observations* $\boldsymbol{O} = \{ o_1, o_2, \ldots, o_m \}$*, such that each* $o_i \in \boldsymbol{O}$ *is a tuple* $(a_1, a_2, \ldots, a_n)$*, a Deolistic Graph* $\boldsymbol{G} = (\boldsymbol{V},\boldsymbol{E})$ *is a binary tree in which*

(1) each vertice $v_i \in \boldsymbol{V}$ *has an associated set of observations* $\boldsymbol{O}_i \subseteq \boldsymbol{O}$*,*

(2) *each internal vertice* $v_k \in \boldsymbol{V}$ *has an associated well-defined sentence* l_k *in* $\boldsymbol{L}$ *involving the variables in* $\boldsymbol{O}_k$*, such that*
   (a) *the left child of* v_k*, i.e.* $left(v_k)$*, is associated with the observations of* v_k *for which* l_k *holds, and*
   (b) *the right child of* v_k*, i.e.* $right(v_k)$*, is associated with the observations of* v_k *for which* $\neg l_k$ *holds,*

Diagram 1 presents a Deolistic graph built from data presented by Witten *et al.* [6] concerning favorable weather conditions for playing golf. The following properties enjoyed by Deolistic graphs stem directly from its definition.

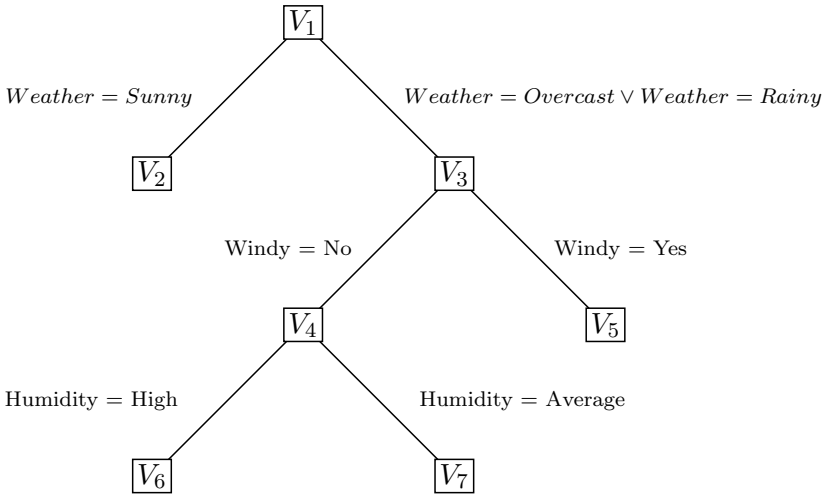

Diagram 1: A Predicate Deolistic Graph

Proposition 4 *Given a Deolistic Graph* $\boldsymbol{G}$ *and a path* P *in* $\boldsymbol{G}$*, such that* P *begins at the root of* $\boldsymbol{G}$*, if* v_i *immediately precedes* v_j *in* P*, then the set of observations associated with* v_i *contains the observations associated with* v_j*, i.e.* $\boldsymbol{O}_i \supseteq \boldsymbol{O}_j$*.*

Proof: Let l_i *be the predicate associated with* v_i*. If* v_i *precedes* v_j *in* P*, then* v_j *is either the left or the right child of* v_i*. If* v_j *is the left child, then, by Definition 3,* v_j *is associated exclusively with the subset of objects in* $\boldsymbol{O}_i$ *for which* l_i *holds. Otherwise, if* v_j *is the right child, then, by Definition 3,* v_j *is associated exclusively with the subset of objects in* $\boldsymbol{O}_i$ *for which* $\neg l_i$ *holds. Either way* $\boldsymbol{O}_i \supseteq \boldsymbol{O}_j$ $\square$

Proposition 5 *Given a Deolistic Graph* $\boldsymbol{G=(V,E)}$ *and path* $P = v_1 \overset{e_1}{\frown} v_2 \overset{e_2}{\frown} v_3 \ldots v_n$*, beginning at the root of* $\boldsymbol{G}$*, such that* $n \geq 2$*, then the set of observations associated with* v_1 *contains the observations associated with* v_2 *which contains the set of observations associated with* v_3 *...,* *i.e.* $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_2 \ldots \boldsymbol{O}_n$*.*

Proof: Let us take a sub-path Q *of* P*, such that* $Q = v_1 \overset{e_1}{\frown} v_2 \ldots v_k$*, where* $1 \leq k \leq n$*. Note that if* $k = 1$*, then, obviously,* $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_1$*. If* $k = 2$*, then* $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_2$*, by Proposition 4.*

Now, let us suppose that $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_2 \ldots \boldsymbol{O}_k$ *holds, for some* $k \geq 2$*. For* $k+1$, $\boldsymbol{O}_k \supseteq \boldsymbol{O}_{k+1}$ *by Proposition 4, because* v_k *immediately antecedes* v_{k+1} *in* Q*. Because* $\supseteq$ *is a transitive relation, it follows that* $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_2 \ldots \boldsymbol{O}_k \supseteq \boldsymbol{O}_{k+1}$*. Therefore, by the principle of finite induction* $\boldsymbol{O}_1 \supseteq \boldsymbol{O}_2 \ldots \boldsymbol{O}_n$ □

Proposition 6 *Given a Deolistic Graph* ***G=(V,E)*** *and path* $P = v_1 \overset{e_1}{\frown} v_2 \overset{e_2}{\frown} v_3 \ldots v_n$*, beginning at the root of* $\boldsymbol{G}$*, such that* $n \geq 2$*, then the conjunction* $q_1 \wedge q_2 \ldots q_{n-1}$ *holds for* $\boldsymbol{O}_n$*, the observations associated with* v_n*, where* q_i *is* l_i *or* q_i *is* $\neg l_i$*, for* $i \in [1 \,..\, n-1]$*, and* $l_1, l_2 \ldots l_{n-1}$ *are the sentences associated with* $v_1, v_2 \ldots v_{n-1}$ *respectively.*

Proof: Let us take a sub-path Q *of* P*, such that* $Q = v_1 \overset{e_1}{\frown} v_2 \ldots v_k$*, where* $k \geq 2$*. Note that if* $k = 2$*, then, by Definition 3, either* l_1 *or* $\neg l_1$ *holds for* $\boldsymbol{O}_2$*. Therefore, there is a* q_1 *that holds for* $\boldsymbol{O}_2$*.*

Now, suppose that for some $k > 2$*,* $q_1 \wedge q_2 \wedge q_3 \ldots q_{k-1}$ *holds for* $\boldsymbol{O}_k$*. By Definition 3, either* l_k *or* $\neg l_k$ *holds for* $\boldsymbol{O}_{k+1}$*. Therefore, there is a* q_k *that holds for* $\boldsymbol{O}_{k+1}$*. Moreover, Definition 3 tells us that,* $\boldsymbol{O}_k \supseteq \boldsymbol{O}_{k+1}$*. If* $q_1 \wedge q_2 \wedge q_3 \ldots q_{k-1}$ *holds for* $\boldsymbol{O}_k$*, it holds for every subset of* $\boldsymbol{O}_k$*, in particular for* $\boldsymbol{O}_{k+1}$*. Consequently, as both* q_k *and* $q_1 \wedge q_2 \wedge q_3 \ldots q_{k-1}$ *hold for* $\boldsymbol{O}_{k+1}$*,* $q_1 \wedge q_2 \wedge q_3 \ldots q_{k-1} \wedge q_k$ *hold for* $\boldsymbol{O}_{k+1}$*. By the principle of finite induction,* $q_1 \wedge q_2 \ldots q_{n-1}$ *hold for* $\boldsymbol{O}_n$ □

3.1 Clustering with Deolistic Graphs

Before a simple and sound clustering algorithm based upon Deolistic graphs can be shown, some basic definitions on similarities of observations need to be presented.

Definition 7 *Given two observations* o_i *and* o_j*, let* $d(o_i, o_j)$ *be a metric that indicates how similar* o_i *and* o_j *are. The matrix of similarities* $\boldsymbol{M}$ *of a set of observations* $\boldsymbol{O} = \{ o_1, o_2, \ldots, o_m \}$ *is an* $m \times m$ *matrix such that* $M(\boldsymbol{O})(i, j) = d(o_i, o_j)$*, for* $i, j \in [1 \,..\, m]$*.*

Over the years many different metrics have been proposed to indicate how similar two given observations are, including the very popular Euclidian distance. See [15] for a comprehensive discussion on other similarity measures.

Definition 8 *For a set of observations* $\boldsymbol{O} = \{ o_1, o_2, \ldots, o_m \}$ *and its corresponding matrix of similarities* $M(\boldsymbol{O})$*, let* $\lambda(M(\boldsymbol{O})) \geq 0$ *be a metric that indicates how diverse the elements of* $M(\boldsymbol{O})$ *are, such that the smaller the value yielded by* $\lambda(M(\boldsymbol{O}))$*, the more similar the objects in* $\boldsymbol{O}$ *are.*

Among the many metrics that can be used to indicate how dissimilar a set of data is, the most widely used are the variance and the standard derivation. However, there are many circumstances in which other measures of variation are preferred. See [17] for a discussion on the pros and cons of these measures.

Algorithm 1 adds exactly one node to a Deolistic graph, Algorithm 2 builds a Deolistic graph from scratch and Algorithm 3 successively trims a Deolistic graph until a certain number of pendents vertices is reached.

Algorithm 1 *Given a vertice v_i of a Deolistic Graph $\boldsymbol{G}$, its associated set of observations $\boldsymbol{O}_i = \{o_1, o_2, \ldots, o_m\}$ and a minimum rate of decrease in diversity ξ*

(1) Find the sentence l that maximizes the equation, β, that follows

$$\lambda(M(\boldsymbol{O}_i)) \times \frac{|\boldsymbol{O}_i|}{|\boldsymbol{O}|} - \lambda(M(\boldsymbol{O}_{left(v_i)})) \times \frac{|\boldsymbol{O}_{left(v_i)}|}{|\boldsymbol{O}|} - \lambda(M(\boldsymbol{O}_{right(v_i)})) \times \frac{|\boldsymbol{O}_{right(v_i)}|}{|\boldsymbol{O}|},$$

where $\boldsymbol{O}_{left(v_i)}$ and $\boldsymbol{O}_{right(v_i)}$ are the set of observations to be associated to the left and right children of v_i respectively, if l is associated to v_i. It should be noticed that β is the weighted difference between the measure of diversity of the objects associated to v_i and the measure of diversity of the objects associated to its children.

(2) If $\beta \geq \xi$ then
  (a) Associate l to v_i,
  (b) Associate β to v_i,
  (c) Create the left child of v_i, i.e. $v_{left(v_i)}$, and connect it to v_i,
  (d) Associate the observations in $\boldsymbol{O}_i$ for which l holds to $v_{left(v_i)}$,
  (e) Create the right child of v_i, i.e. $v_{right(v_i)}$, and connect it to v_i,
  (f) Associate the observations in $\boldsymbol{O}_i$ for which $\neg l$ holds to $v_{right(v_i)}$.

Algorithm 2 *Given a graph $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{E})$, such that $|\boldsymbol{V}| = 1$ and $\boldsymbol{E} = \{\}$, a non-empty set of observations $\boldsymbol{O} = \{o_1, o_2, \ldots, o_n\}$ and a minimum rate of decrease in diversity $\xi > 0$*

(1) Let v be the root vertice of $\boldsymbol{G}$,
(2) Associate $\boldsymbol{O}$ to the root vertice v,
(3) Apply the procedures described in Algorithm 1 to v and recursively to each of its children.

Algorithm 3 *Given a Deolistic graph $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{E})$ and the maximum number of desired segments $\eta > 0$*

(1) Let $\rho(\boldsymbol{G})$ be the set of pendant nodes in G,
(2) While $|\rho(\boldsymbol{G})| > \eta$
  (a) Let $\phi(\boldsymbol{G}) = \{v \in \boldsymbol{V} \wedge v \notin \rho(\boldsymbol{G}) \,|\, left(v) \in \rho(\boldsymbol{G}) \wedge right(v) \in \rho(\boldsymbol{G})\}$,

(b) Let v_i be the node in $\phi(\boldsymbol{G})$ that has the smallest associated β,
(c) Remove the left child of v_i, i.e. $left(v_i)$, from $\boldsymbol{G}$ and the vertice that connects v_i to $left(v_i)$,
(d) Remove the right child of v_i, i.e. $right(v_i)$, from $\boldsymbol{G}$ and the vertice that connects v_i to $right(v_i)$,
(e) Dissociate l_i, the logical sentence associated to v_i, from v_i.

4 Clustering with Predicate Deolistic Graphs

The example that follows uses a particular type of Deolistic graphs, namely restricted predicate Deolistic graph, the construction of which through computational means is made easier due to restrictions imposed on the form that its logical sentences are allowed to take.

Definition 9 *A predicate Deolistic graph is a Deolistic graph that has predicate logic as its logical system. A restricted predicate Deolistic graph is a predicate Deolistic graph where the predicates associated with each vertice take the following form: $\boldsymbol{w_i} \leq \boldsymbol{r}$, if w_i is a continuous variable, or $\boldsymbol{w_i} = \boldsymbol{k}$, if w_i is a categorical variable, where w_i is a variable in $\boldsymbol{O}$, the set of observations, and r and k are constants or variables of the appropriate type.*

In order to build a predicate Deolistic Graph by computational means, it suffices to provide a procedure that searches the domain of each variable in the set of observations to determine the sentence that maximizes the equation β introduced in the first step of Algorithm 1. In order to make the search computationally feasible, the domain of continuous variables has to be made discrete by being divided into an arbitrary number of segments.

4.1 Bumpus' Sparrows

After a severe winter storm that hit the northeast coast of the Unites States on February 1^{st}, 1898, a number of moribund sparrows were taken to Professor Hermon Bumpus' biology laboratory at Brown University, Rhode Island, USA. Subsequently, about half of the birds died. Seeing this as an opportunity to study the effect of Charles Darwin's natural selection principle on birds, Professor Bumpus took several morphological measures of each bird and weighted them [2]. Table 2 presents the results of five of these measures taken from female birds. Birds from 1 to 21 died and the remaining birds survived.

Diagram 2 introduces a restricted predicate Deolistic graph built from these

BId	TL	AE	LBH	LH	LKS	BId	TL	AE	LBH	LH	LKS
1	156	245	31.6	18.5	20.5	26	160	250	31.7	18.8	22.5
2	154	240	30.4	17.9	19.6	27	155	237	31.0	18.5	20.0
3	153	240	31.0	18.4	20.6	28	157	245	32.2	19.5	21.4
4	153	236	30.9	17.7	20.2	29	165	245	33.1	19.8	22.7
5	155	243	31.5	18.6	20.3	30	153	231	30.1	17.3	19.8
6	163	247	32.0	19.0	20.9	31	162	239	30.3	18.0	23.1
7	157	238	30.9	18.4	20.2	32	162	243	31.6	18.8	21.3
8	155	239	32.8	18.6	21.2	33	159	245	31.8	18.5	21.7
9	164	248	32.7	19.1	21.1	34	159	247	30.9	18.1	19.0
10	158	238	31.0	18.8	22.0	35	155	243	30.9	18.5	21.3
11	158	240	31.3	18.6	22.0	36	162	252	31.9	19.1	22.2
12	160	244	31.1	18.6	20.5	37	152	230	30.4	17.3	18.6
13	161	246	32.3	19.3	21.8	38	159	242	30.8	18.2	20.5
14	157	245	32.0	19.1	20.0	39	155	238	31.2	17.9	19.3
15	157	235	31.5	18.1	19.8	40	163	249	33.4	19.5	22.8
16	156	237	30.9	18.0	20.3	41	163	242	31.0	18.1	20.7
17	158	244	31.4	18.5	21.6	42	156	237	31.7	18.2	20.3
18	153	238	30.5	18.2	20.9	43	159	238	31.5	18.4	20.3
19	155	236	30.3	18.5	20.1	44	161	245	32.1	19.1	20.8
20	163	246	32.5	18.6	21.9	45	155	235	30.7	17.7	19.6
21	159	236	31.5	18.0	21.5	46	162	247	31.9	19.1	20.4
22	155	240	31.4	18.0	20.7	47	153	237	30.6	18.6	20.4
23	156	240	31.5	18.2	20.6	48	162	245	32.5	18.5	21.1
24	160	242	32.6	18.8	21.7	49	164	248	32.3	18.8	20.9
25	152	232	30.3	17.2	19.8						

Table 2: Body measurement of female sparrows in millimeters (BId = bird identification number , TL = total length, AE = alar extent, LBH = length of beak and head, LH = length of humerus, LKS = length of keel of sternum).

data with the help of a software we developed to support Deolistic graphs [1] ,

[1] We call this software IDGC, which stands for Intelligent Deolistic Graphs for Clustering.

using the Euclidian distance to calculate how similar any two birds were. It should be noted that the rules generated by the graph clearly indicate the dissimilarities that exist among the different segments and the similarities shared by objects placed in the same segments. As a result, even those with limited knowledge of biology should face little difficulty to understand how the birds are grouped together according to their morphological measures.

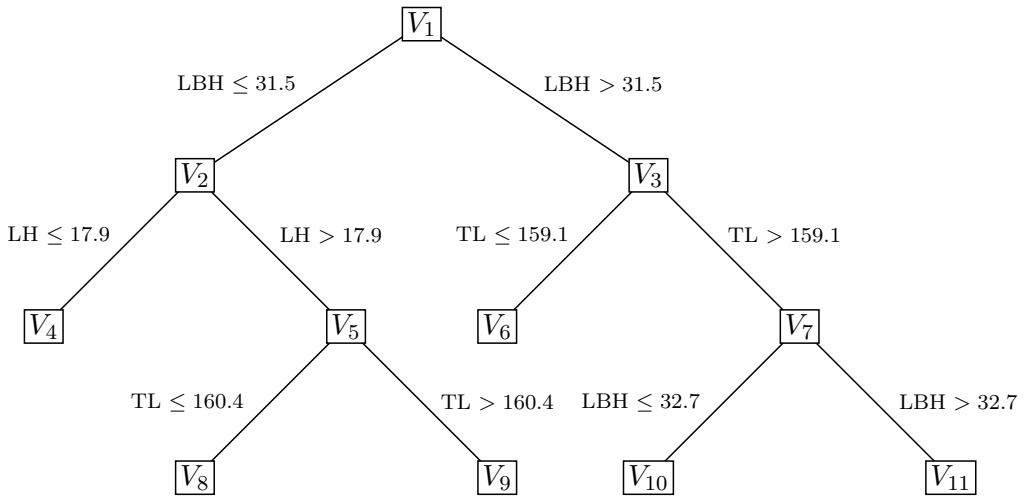

Diagram 2: A Predicate Deolistic Graph

Table 3 shows the quantity and survival rate of female sparrows per pendant node. The average survival rate of the 49 female sparrows is 42.9%. Nodes V_6 and V_8 hold 26 sparrows with survival rate above the average survival rate of the group of birds taken to Prof. Bumpus. The rules that indicate the similarities shared by the sparrows that belong to these nodes are respectively

$$\text{LBH} > 31.5 \wedge \text{TL} \leq 159.1.$$

and

$$\text{LBH} \leq 31.5 \wedge \text{LH} > 17.9 \wedge \text{TL} \leq 160.4$$

Therefore, while node V_6 holds sparrows that have the length of beak and head (LBH) greater than 31.5 mm and the total length (TL) smaller than or equal to 159.1 mm, node V_8 holds sparrows that have the length of beak and head (LBH) smaller than or equal to 31.5 mm, the length of humerus (LH) greater than 17.9 mm and the total length (TL) smaller than or equal to 160.4 mm.

Figure 1 presents the dendrogram resulting from the application the hierarchical agglomerative complete-linkage method to the same data, using the Euclidian distance to measure similarities among objects. The hierarchical clustering methods are among the most commonly used clustering algorithms [13,16]. The reader is encouraged to examine the dendrogram closely and ver-

Node	Quantity	Survival Rate
V_4	7	28.5%
V_6	6	50.0%
V_8	20	60.0%
V_9	2	0.0%
V_{10}	12	33.3%
V_{11}	2	0.0%

Table 3: Quantity and survival rate of sparrows per pendant node.

ify how little information it discloses to data analysts when compared with Deolistic graphs and, as a result, the limited support that it offers to the learning process.

5 Discussion

At the outset of this article we undertook to introduce a family of direct acyclic graphs that uses logical sentences to reveal hidden structures in data. We set forth bellow the answer to some key questions about the development of such a family of graphs and discuss the implication of their existence for data analysis professionals, organizations and the general public.

5.1 Why is it important to provide a clustering algorithm that clearly reveals hidden structures in data?

We live in a world that is increasingly complex in all of its aspects, be they political, economical, social, scientific, educational, ethnical etc. Improvements in telecommunication over the last decades have helped to make it even more complex by allowing people from different places to share concerns and experience events that would otherwise go by almost unnoticed. Moreover, recent enhancements in computer technology have made it possible for both the government and the private sector to gather an enormous volume of information about who we are, where we live and work, how we behave, what products and services we buy, what we believe in and fear most, our main concerns etc.

In such a complex information-driven world, it is of paramount importance to learning that we are able to group together objects that share similar proper-

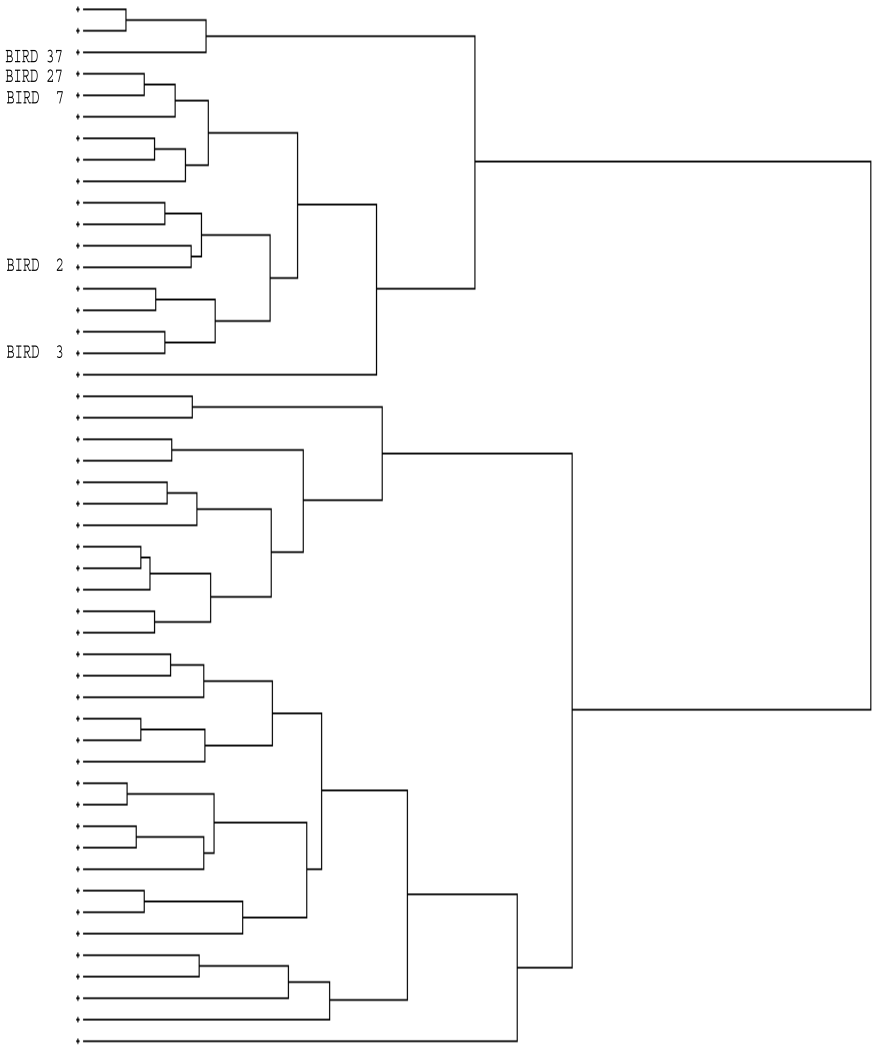

Fig. 1. Dendrogram of Bumpus' sparrows.

ties. Without this ability we would not be able to apply previously acquired knowledge to new objects we encounter, making learning a strenuous task.

However, comprehensive learning only takes place if we are able to identify the properties shared by objects that are placed in the same segment and the dissimilarities that exist among objects that are placed in different segments. Clustering algorithms that clearly support the establishment of these common properties and dissimilarities, such as Deolistic graphs, not only favor learning, but also speed up the clustering process, saving time, effort and money that can be undoubtedly used for other noble purposes.

5.2 Why should one resort to logical sentences, especially to predicate logic sentences, to reveal hidden structures in data?

The logical sentences generated by Deolistic graphs are rules that can be easily understood, interpreted and disclosed to all interested parties, thereby improving communication and reducing misunderstandings.

Predicate logic is one of the simplest and most intuitive existing forms of logic there is. In many countries it is the subject of mathematical studies partly or completely thought in intermediate and high school. As a result, even professionals with little background in mathematics should face only minor difficulties in mastering its syntax and semantics. Moreover, predicate-logic sentences can be easily translated in any written language. Altogether, this makes predicate logic a valuable ally to reveal and communicate hidden structures in data.

5.3 What is the impact of Deolistic graphs on the use of clustering methods?

Because learning is such an important part of modern life and classification is such an intrinsic part of the learning process, there has been an increasing need for the use of clustering methods and tools. Moreover, computer technology has made it possible for organizations to gather large amounts of data about business and everyday life, raising the demand for clustering. Remarkably, most of this demand has been kept held back because current clustering methods require a certain knowledge of mathematics and statistics that, although simple, is not mastered by many people.

By facilitating the use of clustering analysis, the Deolistic graph clustering algorithm helps to ease such a demand, making the whole field more popular and attractive to both market professionals and scholars. The more attention clustering analysis receives, the more likely it is that the development pace of the field will increase with the introduction of new and better clustering methods.

5.4 What is the potential impact of Deolistic graphs on business strategy?

In order to keep an organization competitive in the increasingly dynamic and complex world where we live in, executive management must keep abreast of the markets whose needs are satisfied by the products and services offered by their organization. If it is the case that a large number of products and services is being offered to a large customer base by many different organizations,

than grasping the most recent events in these markets is bound to require that products, services, customers and competing organizations are grouped together into segments of homogeneous objects.

By helping to lower the barriers that prevent the use of clustering methods and at the same time that increasing the quality of the segmentation process, Deolistic graphs contribute to the existence of more competitive organizations and products that are better adjusted to their customers' needs.

5.5 How can society benefit from Deolistic graphs?

The potential impact of Dolistic graphs on all branches of science and areas of business is considerable. For example, by making it easier and faster to understand the hidden structure of the *modus operandi* of criminals, Deolistic graphs favor more effective crime fighting; by clearly revealing differences and similarities among students, Deolistic graphs support the development of better educational policies; by providing organizations with a clearer insight on the profile of their clients, Deolistic graphs show the way to better products and services etc. The list is unequivocally endless.

Therefore, Deolistic graphs encourage the existence of products and services that satisfy their customers' needs more effectively in both the public and private sectors. In the public sector, better products and services increase the return on investment made by taxpayers and save money. In the private sector, better products and services favor the existence of more competitive organizations that tend to remain in business for longer periods, employing people and paying taxes that, if properly applied, will benefit society even further.

6 Conclusions

Although bringing together objects that share similar properties is a human skill of paramount importance for learning and, as a result, for a better understanding of the world in which we live, many professionals face considerable difficulties in making use of quantitative clustering methods. These difficulties stem not only from both the mathematical principles and structures upon which those methods are based, but also from the way clustering analysis results are presented.

By conceiving an algorithm that provides an easy and straightforward way of distinguishing segments of objects and determining the common properties of

objects placed in the same segments, we encourage the use of cluster analysis by professionals of all areas. For data analysis professionals such as statisticians, actuaries, mathematicians, computer specialists etc., Deolistic graphs provide considerable benefits, as it reduces time and effort spent on cluster analysis and makes it easier to understand and communicate the hidden structure of data. Moreover, it improves the quality of the final result yielded by cluster analysis, favoring an interactive process that allows a simple and fast evaluation of the impact caused by the introduction and withdrawal of variables.

References

[1] Charles J. A., Crane F. A. A., and Furness J. A. G. *Selection and Use of Engineering Materials*. Butterworth-Heinemann, 3^{rd} edition, 1996. 352 pages.

[2] Bumpus H. C. The elimination of the unfit as illustrated by the introduced sparrow, passer domesticus. *Biology Lectures: Woods Hole Marine Biological Laboratory*, pages 209–225, 1899.

[3] Linnaeus C. *Genera plantarum eorumque characteres naturales secundum numerum, figuram, situm & proportionem omnium fructificationis partium*. Leiden, 1737. In Latin.

[4] Tryon R. C. *Cluster Analysis*. Edwards Brothers, Ann Arbor, 1939.

[5] Celeux G. and Govaert G. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–322, 1995.

[6] Witten I. H. and Frank E. *Data mining: practical machine learning tools and techniques with java implementations*. Morgan Kaufmann, 1999. 371 pages.

[7] Pilowsky I., Levine S., and Boulton D. M. The classification of depression by numerical taxonomy. *British Journal of Psychiatry*, 115:937–945, 1969.

[8] Bird J. and Ross C. *Mechanical Engineering Principles*. Newnes, 2002. 288 pages.

[9] Gougen J. and Burstal R. Institutions: Abstract model theory for specification and programming. *Journal of the ACM*, 39(1):95–146, January 1992.

[10] Han J. and Kamber M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 1^{st} edition, 2000. 500 pages.

[11] Kaufman L. and Rousseeuw P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1990. 368 pages.

[12] Faúndez-Abans M., Ormeno M. I., and Oliveira-Abans M. Classification of planetary nebulae by cluster analysis and artificial neural network. *Astronomy Astrophysics Supplement*, 116:395–402, 1996.

[13] Kim M. and Compton P. Incremental development of browsing for domain-specific document retrieval systems. In Handschuh S., Dieng R., and Staab S., editors, *Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation*, Victoria, B.C., Canada, October 21 2001.

[14] Deo N. *Graph Theory with Applications to Enginering and Computer Science*. Prentice Hall, 1964. 478 pages.

[15] Everitt B. S., Landau S., and Leese M. *Cluster Analysis*. Arnold, 4^{th} edition, 2001. 237 pages.

[16] Raychaudhuri S., Chang J. T., Imam F., and Altman R. B. The computational analysis of scientic literature to define and recognize gene expression clusters. *Nucleic Acids Research*, 31(15):4553–4560, 2003.

[17] Hogg R. V., Craig A., and McKean J. W. *Introduction to Mathematical Statistics*. Prentice Hall, 6^{th} edition, 2004. 692 pages.

[18] Feuerstein G. Z., Libby P., and Mann D. L. *Inflammation and Cardiac Diseases (Progress in Inflammation Research)*. Birkhauser, Boston, MA, USA, 2003. 416 pages.