Homogeneity Evaluation and Seed Selection in Clustering Graph-Connected Spatial Data

Annalisa Appice, Antonietta Lanza, Donato Malerba, and Antonio Varlaro

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
{appice,lanza,malerba,varlaro}@di.uniba.it

Abstract. CORSO is a spatial clustering algorithm which works on structured observations representing some areal units. Observations may include a variable number of related (spatial) objects of different type and are represented by conjunctions of ground atoms. Dependencies among observations are possible, in which case they are described by a directed graph, called spatial discrete structure, which is considered by CORSO when it cluster observations. In particular, CORSO partitions the graph structure so that each partition corresponds to a cluster of possibly homogeneous (i.e., similar) observations. Similarity between two observations is based on a flexible matching function which is defined for relational representations. In this work we present some issues occurring in the original proposal of CORSO and we propose some solutions to these issues. Experiments on two data sets illustrate the effect of the proposed solutions.

1 Introduction

In the Spatial Data Mining task investigated in this work, observations are areal units which may include multiple objects of different types, such as roads, rivers, and cultivations. The definition of relations between such objects (e.g., a road crosses a river) makes the representation of each single observation inherently structural. In addition, observations cannot be considered independent due to the spatial continuity of events in the space. The dependence between observations is expressed by means of a directed graph where nodes correspond to observations and edges represent spatial relations between observations (e.g., adjacency). Henceforth, such a directed graph is called *discrete spatial structure*. In this work we investigate spatial clustering of (possibly structured) observations related by means of a discrete spatial structure. Spatial clustering corresponds to partitioning the directed graph.

CORSO [2] is a data mining method for clustering relational observations whose dependencies are expressed by means of a discrete spatial structure. Clusters are built by merging partially overlapping homogeneous neighborhoods. A neighborhood is built starting from a seed node and including all nodes directly connected to the seed in the graph and not yet assigned to any cluster. Similarity among areal units is computed as the degree of matching of each areal unit with respect to a common generalization. A neighborhood is joined to a cluster only if it is homogeneous with respect to the cluster generalization (or model). This homogeneity evaluation poses several issues. First, the homogeneity evaluation is performed at the neighborhood (and not cluster) level, hence it cannot guarantee that the entire cluster is homogeneous with respect to its final description (or model). Second, the cluster model is built incrementally as a set of generalizations (i.e., logic theories), one for each neighborhood to be joined to the cluster. A generalization of a neighborhood is learned independently from the model currently associated with the cluster to be expanded. Hence, the same generalization can be learned for separate neighborhoods by introducing duplicates in the evaluation of homogeneity. Third, the clustering expansion operates at the level of an entire neighborhood, which means that it is not possible to include in a cluster only a portion of the neighborhood.

In this work, we propose a different homogeneity evaluation strategy in order to overcome the issues of the original proposal. In addition, we investigate some criteria that exploit the distribution of both nodes and edges in the graph structure in order to "guide" the choice of the seed objects. The validity of our solutions is confirmed by empirical results.

2 Clustering Graph-Connected Spatial Data

Theoretically, the problem solved by CORSO is formulated as follows: *Given* a set of structured objects O, a discrete spatial structure defined by a directed graph $G = \langle O, E \rangle$, and a background knowledge BK;

Find a partition of G into a set of connected subgraphs $G_i = \langle C_i, E_i \rangle$ such that $O = \bigcup_i C_i$ and each cluster C_i is homogeneous with respect to the given BK. Objects of singleton clusters are labeled as *noise*. An object $o_j \in O$ is described by means of a conjunction of ground atoms, while the background knowledge BK is expressed by a set of definite clauses. Each cluster C_i of two ore more objects is intensionally described by a logical theory T_{C_i}.

The cluster construction starts from an arbitrary object o (*seed object*), that is, the first object accessed in O not yet assigned to any cluster, and then it continues by checking whether the neighborhood N_o of o in G is a homogeneous cluster C. N_o is built by including all objects $o_j \in O$ such that there is an arc from o to o_j in G (oRo_j) and o_j is not assigned to any cluster. The logical theory T_C associated to the cluster C is obtained by generalizing on all objects in C. CORSO iteratively expands C by joining neighborhoods N_{o_i} ($o_i \in C$) which are partially "overlapping" with C (neighborhood expansion) and result in a homogeneous set w.r.t the cluster theory T_{C_i} ($C_i = C \cup N_{o_i}$). The neighborhood homogeneity is estimated as follows:

$$h(N_{o_i}; T_{C_i}) = \frac{1}{\#N_{o_i}} \sum_{o_j \in N_{o_i}} h(o_j; T_{C_i}) = \frac{1}{\#N_{o_i}} \sum_{o_j \in N_{o_i}} \frac{1}{\#T_{C_i}} \sum_{H_k \in T_{C_i}} fm(o_j, H_k) \geq h_{min} \tag{1}$$

where $\#N_{o_i}$ is the cardinality of the neighborhood set N_{o_i}, while the new cluster theory T_{C_i} is a set of first-order clauses H_k. T_{C_i} includes both $\{T_1; \ldots; T_w\}$, i.e., the model of C, and T_{w+1}, i.e., the model of N_{o_i}. The model T_i of a neighborhood N_{o_i} is built as a set of first-order clauses, $T_i : \{cluster(X) = c \leftarrow H_{i1}; \ldots; cluster(X) = c \leftarrow H_{iz}\}$, where each H_{ij} is a conjunctive formula to describe a sub-structure (properties and relations) which is shared by one or more objects falling in N_{o_i}. Each object in N_{o_i} can be explained by the model $T_i \cup BK$ ($\forall o \in N_{o_i}$, $BK \cup T_i \models o$). T_i is generated by means of the ILP system ATRE [1], which works only on positive examples (observations in N_{o_i}). Since the generalization of a neighborhood is built independently from the (partially learned) cluster theory, duplicate clauses may be learned for separate neighborhoods and then introduced in the cluster theory. The function $fm(S, R)$ (flexible matching) returns a number in $[0, 1]$, which corresponds to the the probability of precisely matching the *subject* S against the *referent* R, provided that some change is possibly made in the description of S. In CORSO S (R) is a conjunctive ground (non-ground) first-order formula, and $fm(S, R)$ is defined as follows:

$$fm(S, R) = \max_{\theta} \prod_{i=1,\ldots,k} fm_\theta(S, r_i). \tag{2}$$

where θ is a substitution which binds variables in R to constants in S, while fm_θ is the flexible matching computed between S and an atom of $R\theta$. Each atom r_i of R has the form $f_{r_i}(x_1, \ldots, x_n) = v_{r_i}$ where x_k's are variables and f_{r_i} denotes a function (called *descriptor*) with either numerical or categorical range. In the former case the value v_{r_i} is an interval ($v_{r_i} \equiv [a, b]$), while in the latter case v_{r_i} is a subset of values ($v_{r_i} \equiv \{v_1, \ldots, v_M\}$) from the range of f_{r_i}. Atoms of the subject S have the form $f_{s_j}(c_1, \ldots, c_n) = w_{s_j}$, where c_k's denote constants and w_{s_j} corresponds to a single value of the range of f_{s_j}. The flexible matching $fm_\theta(S, r_i)$ evaluates the degree of similarity $fm(s_j, r_i\theta)$ between $r_i\theta$ and the corresponding selector s_j in S such that both r_i and s_j have the same descriptor ($f_{r_i} = f_{s_j}$) and $x_k\theta = c_k$ for each pair of corresponding arguments. More precisely, $fm(s_j, r_i\theta) = fm(w_{s_j}, v_{r_i}) = \max_{v \in v_{r_i}} P(equal(w_{s_j}, v))$. Details are provided in [2].

3 Homogeneity Evaluation

In this section we describe a novel homogeneity evaluation strategy for CORSO. It aims at both improving the intra-cluster homogeneity and avoiding the generation of duplicate clauses in cluster models and permitting the expansion of a cluster with single neighbors (in alternative to the entire neighborhood).

Hence, our proposal provides a solution to three different issue. The first one concerns the evaluation of the cluster homogeneity at the *cluster level* and not at the neighborhood level as in the original proposal. The neighborhood N_{o_i} is merged with the cluster C if and only if the homogeneity of the candidate cluster $C_i = C \cup N_{o_i}$ is greater than a user defined threshold h_{min}. Homogeneity of C_i is evaluated as follows:

$$h(C_i, T_{C_i}) = \frac{1}{\#C_i} \sum_{o_j \in C_i} h(o_j, T_{C_i}) \tag{3}$$

where $T_{C_i} = \{T_C, T_{N_{o_i}}\}$. Obviously, the evaluation of homogeneity at cluster level is more complex than the evaluation at neighborhood level. Some caching techniques are applied to improve the efficiency of the algorithm (details are omitted due to space constraints).

The second issue, namely redundancy in cluster models, is due to the fact that in the original formulation of the clustering method each neighborhood theory $T_{N_{o_i}}$ was learned independently from the current cluster model T_C. The generation of duplicate clauses also affected the evaluation of cluster homogeneity, since a cluster is 'more homogeneous' when a subset of objects is covered by several clauses of the cluster model. We face this problem by generating a neighborhood theory only for those objects in N_{o_i} not already covered by T_C. Besides having a better evaluation of cluster homogeneity, this way we generally improve both the efficiency of homogeneity evaluation and the interpretability of cluster models.

The third issue is due to the fact that in the original proposal the clustering expansion operates at neighborhood level. This prevents the expansion of a cluster by adding only a "portion" of the neighborhood under analysis. To overcome this limitation, CORSO learning strategy has been modified in order to evaluate a single-neighbor based expansion when the neighborhood-based one fails.

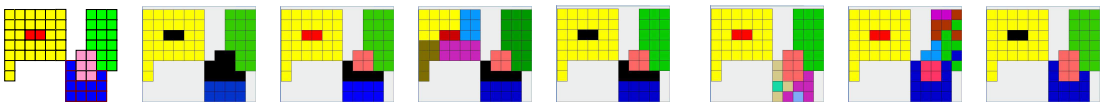
4 Seed Selection Criteria

The cluster shape depends on the object that CORSO chooses at each step as seed of the neighborhood to be processed. CORSO adopts a sequential strategy. In the case a new cluster has to be discovered, the seed is the first object accessed in O not yet assigned to any cluster. In the case an existing cluster has to be expanded, the seed is the first element of a list of cluster objects not yet considered for the expansion step. The sequential seed selection (SEQ) is efficient, but the quality of clustering clearly depends on the "order" in which objects are stored and accessed, which might not be the optimal one for cluster formation. Alternatively, we have empirically investigated some heuristics that take advantage from the graph structure in the seed selection step. In particular, we base the choice of the candidate seed on the concept of density (cardinality) of a neighborhood. For each candidate seed, the candidate "best" seed is the object whose neighborhood in the graph has the highest (DESC) or lowest (ASC) density. In the first case (DESC), we follow the intuition coming from the density-based framework where dense areas are labeled as clusters. Our suggestion is to estimate density in terms of the number of connections in the graph from the candidate seed to the objects not yet assigned to any cluster. Alternatively, a dense area in the graph may correspond to an area covered by contiguous different clusters, hence, discovery should preferentially start from peripheral objects (ASC).

5 Experimental Results

We applied CORSO to both artificial data on the arrangement of computers in a LAN and real data extracted from a topographic map of Apulia (Italy). In LAN data set, each spatial object corresponds to a PC described in terms of operating system available, shared option, share space size, type of net connection and type of IP assignment, while the discrete spatial structure is defined by the adjacency relation. Results obtained with different parameter configurations are shown in Figure 1.

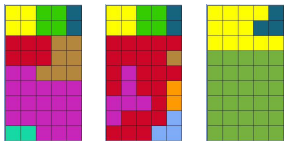
Results confirm that, except for the ASC criterion, the new cluster-based homogeneity evaluation allows CORSO to discover clusters whose shape better fits the original one reported in Figure 1.a. At the same time the number of objects labeled as noise (black objects) decreases when resorting to the cluster-based homogeneity evaluation. The clusters descriptions discovered with cluster-based homogeneity are simpler than those discovered with the neighborhood-based homogeneity evaluation. For example, the description of the yellow cluster in Figure 1.b is composed by 16 similar (or at worst identical) clauses, while the same description includes only one clause in the SEQ configuration (Figure 1.c). This depends on the fact that the cluster-based homogeneity evaluation re-uses the current cluster description when building the generalization of a candidate neighborhood, thus preventing the introduction of redundant clauses. Experimental results confirm that the cluster expansion performed at single-neighbor level allows CORSO to discover irregularly shaped clusters. Although the ASC seed selection criterion generated several fragmented clusters, no object was labeled as noise. On the contrary, the DESC seed selection perfectly detects the actual clustering configuration except for the objects in the original red cluster which are labeled as noise during the expansion of contiguous clusters.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 1. Clusters in the original LAN dataset (a) are compared with clusters discovered with the original version of CORSO (b) and clusters discovered with the cluster-based homogeneity evaluation in different configurations: SEQ (c), ASC (d) and DESC (e) order for seed selection. Results with SEQ (f), ASC (g) and DESC (h) order for seed selection and single-neighbor level in cluster expansion are shown. $h_{min} = 0.95$.

Topographic map data concern adjacent cells of a map described in terms of both geometrical and topological features (e.g., area, distance) of geographical objects (e.g., roads, cultivations) contained in each cell. The discrete spatial structure is a regular grid defined by the relation of adjacency between cells. Details on this dataset are found in [2]. Some results are depicted in Figure 2.



(a) (b) (c)

Fig. 2. Comparison of clustering performed by CORSO with the neighborhood-based evaluation function (a) and the cluster-based homogeneity function combined with the single-neighbor check and the SEQ (b) or DESC (c) order of connectivity in seed selection. $h_{min} = 0.99$

CORSO with the cluster-based homogeneity function and SEQ selection (Figure 2.b) is able to detect irregularly shaped clusters. This is due to the check at level of single-neighbor. The models associated with clusters in Figures 2.b,c are more compact and simpler to read. For instance, the descriptions of the red and brown clusters in Fig. 2.a include the clauses:

red(X):- grapevines(X) ∈ [2..19], hasStreet(X,Y), extension(Y) ∈ [11..1179],
street2parcel(Y,Z)=adjacent, area(Z) ∈ [262..249975]
brown(X):- grapevines(X) ∈ [6..27], hasStreet(X,Y), extension(Y) ∈ [11..1115]
brown(X):- grapevines(X) ∈ [6..24], hasStreet(X,Y), extension(Y) ∈ [22..1115]

while approximatively a super set of the same area is modeled in Fig. 2.b by the red cluster that is described by the single clause:

red(X):- grapevines(X) ∈ [2..15], hasStreet(X,Y), extension(Y) ∈ [11..1023],
street2parcel(Y,Z)=adjacent, area(Z) ∈ [262..249975]

Finally, Figure 2.c shows that the DESC criterion decreases the number of clusters. Indeed, starting from highly connected objects, CORSO generates cluster models representative of a higher number of objects.

References

1. D. Malerba. Learning recursive theories in the normal ilp setting. *Fundamenta Informaticae*, 57(1):39–77, 2003.
2. D. Malerba, A. Appice, A. Varlaro, and A. Lanza. Spatial clustering of structured objects. In *ILP*, pages 227–245. Springer-Verlag, 2005.