A Multi-Relational Hierarchical Clustering Method for DATALOG Knowledge Bases

Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito

Dipartimento di Informatica – Università degli Studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{fanizzi|claudia.damato|esposito}@di.uniba.it

Abstract. A clustering method is presented which can be applied to relational knowledge bases (e.g. DATALOG deductive databases). It can be used to discover interesting groups of resources through their (semantic) annotations expressed in the standard logic programming languages. The method exploits an effective and language-independent semi-distance measure for individuals., that is based on the resource semantics w.r.t. a number of dimensions corresponding to a committee of features represented by a group of concept descriptions (discriminating features). The algorithm is a fusion of the classic BISECTING K-MEANS with approaches based on medoids that are typically applied to relational representations. We discuss its complexity and potential applications to several tasks.

1 Unsupervised Learning with Complex Data

In this work, we investigate on unsupervised learning for knowledge bases (KBs) expressed in relational languages. In particular, we focus on the problem of conceptual clustering of semantically annotated resources. The benefits of *conceptual clustering* [10] in such a context are manifold: $1)$ *concept formation*: clustering annotated resources enables the definition of new emerging concepts on the grounds of the primitive concepts asserted in a KB; $2)$ *evolution*: supervised methods can exploit these clusters to induce new concept definitions or to refining existing ones; $3)$ *search and ranking*: intensionally defined groupings may speed-up the task of search upon queries; a hierarchical clustering also suggests criteria for ranking the retrieved resources.

Essentially, many existing clustering methods are based on the application of similarity (or density) measures defined over a fixed set of attributes of the domain objects. Classes of objects are taken as collections that exhibit low interclass similarity (density) and high intraclass similarity (density). Often these methods cannot take into account *background knowledge* that could characterize object configurations by means of global concepts and semantic relationship. As pointed out in related surveys [11], initially, most of the proposed similarity measures for concept descriptions focus on the similarity of atomic concepts within simple concept hierarchies or are strongly based on the structure of the terms for specific FOL fragments [4]. Alternative approaches are based on the notions of *feature* similarity or *information content*. In the perspective of exploiting similarity measures in inductive (instance-based) tasks like those mentioned above, the need for a definition of a semantic similarity measure for *instances* arises [2, 8].

Early conceptual clustering methods aimed at defining groups of objects using conjunctive descriptions based on selected attributes [10]. Anyway, in the conceptual clustering perspective, the expressiveness of the language adopted for describing objects and clusters (concepts) is equally important. Alternative approaches, suitable to concept languages, have pursued a different way for attacking the problem, devising logic-based methods [3]. However, these methods may suffer from noise in the data. This motivates our investigation on similarity-based clustering methods which can be more noise-tolerant. We propose a multi-relational extension of effective clustering techniques. It is intended for grouping similar resources w.r.t. a semantic dissimilarity measure in order to discover new concepts. Our relational method derives from the *Bisecting k-means* algorithm [5], a well-known partitional clustering method. Specifically, we recur to the notion of *medoids* (like in algorithm *PAM* [6]) as central individual in a cluster, rather than to the notion of means characterizing the algorithms descending from *k-means* and *EM* [5] developed for numeric (or ordinal) features. Upgrading existing algorithms to work on multi-relational representations such as clausal languages, requires novel similarity measures that are suitable for such representations. Moreover, rather than fix a given number k of clusters of interest (that may be hard when scarce domain knowledge is available), a partitional method may be employed up to reaching a minimal threshold value for cluster *quality* [6, 5] which makes any further bisections useless.

In the next section the dissimilarity measure adopted in the algorithm is defined. The clustering algorithm is presented in Sect. 3. Possible developments are examined in Sect. 4.

2 A Family of Metrics for Instances

In the following, we assume that objects (instances), concepts and relationships among them are defined in terms of a function-free (yet not constant-free) clausal language such as DATALOG, endowed with the standard semantics (see [7]). A *knowledge base* is defined as $\mathcal{K} = \langle \mathcal{P}, \mathcal{D} \rangle$, where $\mathcal{P}$ is a logic program representing the *schema*, with concepts (entities) and relationships defined through definite clauses, *database* $\mathcal{D}$ is a set of ground facts concerning the world state. Without loss of generality, we will consider concepts as described by unary atoms. *Primitive* concepts are defined in $\mathcal{D}$ extensionally by means of ground facts only, whereas *defined* concepts will be defined in $\mathcal{P}$ by means of clauses. The set of the objects occurring in $\mathcal{K}$ is denoted with $\mathsf{const}(\mathcal{D})$. As regards the necessary inference services, our measures will require performing *instance-checking*, which amounts to determining whether an object belongs (is an instance) of a concept in a certain interpretation.

Instances lack a syntactic structure that may be exploited for a comparison. However, on a semantic level, similar objects should *behave* similarly w.r.t. the same concepts, i.e. similar assertions (facts) should be shared. Conversely, dissimilar instances should likely instantiate disjoint concepts. Therefore, we introduce novel dissimilarity measures for objects, whose rationale is the comparison of their semantics w.r.t. a fixed number of dimensions represented by concept descriptions (predicate definitions). Instances are compared on the grounds of their behavior w.r.t. a reduced (yet not necessarily disjoint) committee of features (concept descriptions) $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, expressed in the language taken into account, acting as discriminating *features*. We will consider unary predicates which have a definition in the KB. Following [9], a family of totally semantic distance measures for objects can be defined for clausal representations. In its simplest formulation, inspired by Minkowski's metrics, it is defined as:

Definition 2.1 (family of measures). *Let $\mathcal{K}$ be a KB. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, a family $\{d_p^{\mathsf{F}}\}_{p \in \mathbb{N}}$ of functions $d_p^{\mathsf{F}} : \mathsf{const}(\mathcal{D}) \times \mathsf{const}(\mathcal{D}) \mapsto [0,1]$ is defined as follows*

$$\forall a, b \in \mathsf{const}(\mathcal{D}) \qquad d_p^{\mathsf{F}}(a,b) := \frac{1}{m} \left[\sum_{i=1}^{m} (\delta_i(a,b))^p \right]^{1/p}$$

where $\forall i \in \{1, \ldots, m\}$ the i-th dissimilarity function δ_i *is defined:*

$$\forall a, b \in \mathsf{const}(\mathcal{D}) \qquad \delta_i(a,b) = \begin{cases} 0 & \mathcal{K} \vdash F_i(a) \textit{ iff } \mathcal{K} \vdash F_i(b) \\ 1 & \textit{otherwise} \end{cases}$$

The superscript F will be omitted when the set of features is fixed.

These functions are semi-distances (or pseudo-metrics) [1], namely, it cannot be proved that if $d_p(a,b) = 0$ then $a = b$. However, if the *unique names assumption* is made for the constant names, then a distance can be obtained by using a further feature set F_0 based on the equality: $\delta_0(a,b) = 1$ if $a = b$; $\delta_0(a,b) = 0$ otherwise.

Here, we make the assumption that the feature-set F represents a sufficient number of (possibly redundant) features that are able to discriminate really different objects. In [1], we propose a method for performing a randomized search of optimal feature sets.

Compared to other proposed distance (or dissimilarity) measures, the presented functions are not based on structural (syntactical) criteria.

The definition above might be further refined and extended by recurring to model theory. The set of Herbrand models $\mathcal{M}_{\mathcal{K}} \subseteq 2^{|\mathcal{B}_{\mathcal{K}}|}$ of the KB may be considered, where $\mathcal{B}_{\mathcal{K}}$ stands for its Herbrand base. Given two instances a and b to be compared w.r.t. a certain feature F_i, $i = 1, \ldots, m$, we might check if they can be distinguished in the world represented by a Herbrand interpretation $\mathcal{I} \in \mathcal{M}_{\mathcal{K}}$: $\mathcal{I} \models F_i(a)$ and $\mathcal{I} \models F_i(b)$. Hence, a distance measure should count the cases of disagreement, varying the Herbrand models of the KB. The resulting measure definition will be in this case:

$$\forall a, b \in \mathsf{const}(\mathcal{D}) \quad d_p^{\mathsf{F}}(a,b) := \frac{1}{m \cdot |\mathcal{M}_{\mathcal{K}}|} \left[\sum_{\mathcal{I} \in \mathcal{M}_{\mathcal{K}}} \sum_{i=1}^{m} \mid \delta_i^{\mathcal{I}}(a,b) \mid^p \right]^{1/p}$$

where the dissimilarity functions $\delta_i^{\mathcal{I}}$ are computed for a specific world state as encoded by a Herbrand interpretation $\mathcal{I}$:

$$\forall a \in \mathsf{const}(\mathcal{D}) \quad \delta_i^{\mathcal{I}}(a,b) = \begin{cases} 1 & F_i(a) \in \mathcal{I} \text{ and } F_i(b) \in \mathcal{I} \\ 0 & \textit{otherwise} \end{cases}$$

3 Grouping Objects through Hierarchical Clustering

The conceptual clustering procedure implemented in our method works top-down, starting with one universal cluster grouping all instances. Then it iteratively finds two clusters bisecting an existing one up to the desired number of clusters is reached. Our algorithm can be ascribed to the category of the heuristic partitioning algorithms such as K-MEANS and EM [5]. Each cluster is represented by the center of the cluster. In our setting we consider the medoid [6] as a notion of cluster center. In particular our algorithm can be seen as a hierarchical extension of the PAM algorithm (*Partition Around Medoids* [6]): each cluster is represented by one of the individuals in the cluster, the medoid, that is, in our case, the one with the lowest average distance w.r.t. all the others individuals in the cluster. The bi-partition is repeated level-wise producing a dendrogram. In the following, a sketch of the algorithm is reported.

```
HBAM(allIndividuals, $k$, maxIterations): clusterVector;
input   allIndividuals: set of individuals; $k$: number of clusters;
        maxIterations: max number of inner iterations;
output  clusterVector: array $[1..k]$ of sets of clusters

level := 0;   clusterVector[1] := allIndividuals;
repeat
        ++level;
        cluster2split := selectWorstCluster(clusterVector[level]);
        iterCount := 0;
        stableConfiguration := false;
        (newMedoid1,newMedoid2) := selectMostDissimilar(cluster2split);
        repeat
            ++iterCount;
            // E step
            (medoid1,medoid2) := (newMedoid1,newMedoid2);
            (cluster1,cluster2) := distribute(cluster2split,medoid1,medoid2);
            // M step
            newMedoid1 := medoid(cluster1);
            newMedoid2 := medoid(cluster2);
            stableConfiguration := (medoid1 = newMedoid1) $\wedge$ (medoid2 = newMedoid2);
        until stableConfiguration $\vee$ (iterCount = maxIterations);
        clusterVector[level+1] := replace(cluster2split,cluster1,cluster2,clusterVector[level]);
until (level $= k$);
```

The algorithm essentially consists of two nested loops: the outer one computes a new level of the resulting dendrogram and it is repeated until the desired number of clusters is obtained; the inner loop consists of a run of the PAM algorithm at the current level. Per each level, the next worst cluster is selected (*selectWorstCluster*() function) on the grounds of its quality, e.g. the one endowed with the least average inner similarity (or cohesiveness [10]). This cluster is candidate to being parted in two. The partition is constructed around two medoids initially chosen (*selectMostDissimilar*() function) and then iteratively adjusted in the inner loop. In the end, the candidate cluster is replaced by the newly found parts at the next level of the dendrogram. The inner loop basically resembles to a 2-means (or EM) algorithm, where medoids are considered instead of means, which can hardly be defined in symbolic computations. Then, the classical two steps are performed in an iteration: **E step:** given the current medoids, the first distributes the other individuals in one of the two partitions under construction on the grounds of their similarity w.r.t. either medoid; **M step:** given the bipartition obtained by *distribute*(), this second step computes the new medoids for either cluster. These tend to change on each iteration until eventually they converge to a stable couple (or when a maximum number of iteration have been performed). The medoid of a group of individuals is the individual that has the lowest distance w.r.t. the others. Formally. given a cluster $C = \{a_1, a_2, \ldots, a_n\}$, the medoid is defined: $m = \mathrm{medoid}(C) = \mathrm{argmin}_{a \in C} \sum_{j=1}^{n} d(a, a_j)$. The representation of centers by means of medoids has two advantages. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers.

Each node of the tree (a cluster) may be labeled with an intensional concept definition which characterizes the individuals in the given cluster while discriminating those in the twin cluster at the same level. Labeling the tree-nodes with concepts can be regarded as a number of supervised learning problems in the specific multi-relational representation targeted in our setting. A straightforward solution may be given by the computation of the *least general generalization* (lgg) operator [7] and (an approximation of) the *most specific concept* (msc) operator, which amounts to building a new ground clause whose body is made up of the ground literals in the set of derivable facts (from $\mathcal{K}$) that are linked through their constants:

$$\begin{aligned}\mathsf{msc}_{\mathcal{K}}(a) := \{L : \text{literal}, \mathcal{K} \models L \mid \exists a \in \text{args}(L) \; or \; \exists L' \in \mathsf{msc}(a) \; s.t.& \\ \exists b \in \text{args}(L) \cap \text{args}(L') \; and \; \exists c \in \text{args}(L')\}&\end{aligned}$$

This solution involves the following steps:

- **input:** clusters of individuals C_j
- **output:** new clause

```
1. for each individual $a_{ij} \in C_j$ do
   (a) compute $M_{ij} \leftarrow \mathsf{msc}_{\mathcal{K}}(a_{ij})$
   (b) let $\mathsf{Clause}_{ij} \leftarrow (\mathsf{newConcept}_j(a_{ij}) \text{ :- } M_{ij})$;
2. return $\mathsf{lgg}(\mathsf{Clause}_{ij})$
```

As an alternative, algorithms for learning concept descriptions expressed in DATALOG may be employed.

4 Conclusions and Future Work

This work has presented a clustering method for DATALOG knowledge bases. The method exploits a novel dissimilarity measure, that is based on the resource semantics w.r.t. a number of dimensions corresponding to a committee of features represented by a group of concept descriptions (discriminating features). The algorithm is an adaptation of the classic bisecting k-means to complex LP representations. We have discussed its complexity and the potential applications to a variety of important tasks.

Ongoing work concerns the feature selection task. Namely, we aim at inducing an optimal set of concepts for the distance measure by means of randomized algorithms based on genetic programming and simulated annealing. Furthermore, also the clustering process itself may be carried out by means of a randomized method based on the same approaches. We are also exploiting the outcome of the clustering algorithm for performing similarity search grounded on a lazy-learning procedure and specifically based on the weighted k-nearest neighbor approach, exploiting the distance measures presented in this work. Further applications regards the tasks specified in Sec. 1.

References

[1] C. d'Amato, N. Fanizzi, and F.Esposito. Induction of optimal semantic semi-distances for clausal knowledge bases. In *Proceedings of the 17th International Conference on Inductive Logic Programming, ILP2007*, LNAI. Springer, 2007. (to appear).

[2] W. Emde and D. Wettschereck. Relational instance-based learning. In L. Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning, ICML96*, pages 122–130. Morgan Kaufmann, 1996.

[3] N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro. Concept formation in expressive description logics. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Proceedings of the 15th European Conference on Machine Learning, ECML2004*, volume 3201 of *LNAI*, pages 99–113. Springer, 2004.

[4] A. Hutchinson. Metrics on terms and clauses. In M. van Someren and G. Widmer, editors, *Proceedings of the 9th European Conference on Machine Learning, ECML97*, volume 1224 of *LNAI*, pages 138–145. Springer, 1997.

[5] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[6] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[7] S. Nienhuys-Cheng and R. de Wolf. *Foundations of Inductive Logic Programming*, volume 1228 of *LNAI*. Springer, 1997.

[8] J. Ramon and M. Bruynooghe. A framework for defining distances between first-order logic objects. Technical Report CW 263, Department of Computer Science, Katholieke Universiteit Leuven, 1998.

[9] M. Sebag. Distance induction in first order logic. In S. Džeroski and N. Lavrač, editors, *Proceedings of the 7th International Workshop on Inductive Logic Programming, ILP97*, volume 1297 of *LNAI*, pages 264–272. Springer, 1997.

[10] R. E. Stepp and R. S. Michalski. Conceptual clustering of structured objects: A goal-oriented approach. *Artificial Intelligence*, 28(1):43–69, Feb. 1986.

[11] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach*. Springer, 2007.