10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'06), Singapore, 2006

Clustering Multi-Represented Objects Using Combination Trees

Elke Achtert, Hans-Peter Kriegel, Alexey Pryakhin, Matthias Schubert

Institute for Computer Science
University of Munich, Germany
{achtert,kriegel,pryakhin,schubert}@dbs.ifi.lmu.de

Abstract. When clustering complex objects, there often exist various feature transformations and thus multiple object representations. To cluster multi-represented objects, dedicated data mining algorithms have been shown to achieve improved results. In this paper, we will introduce combination trees for describing arbitrary semantic relationships which can be used to extend the hierarchical clustering algorithm OPTICS to handle multi-represented data objects. To back up the usability of our proposed method, we present encouraging results on real world data sets.

1 Introduction

In modern data mining applications, there often exists no universal feature representation that can be used to express similarity between all possible objects in a meaningful way. Thus, recent data mining approaches employ multiple representations to achieve more general results that are based on a variety of aspects. In this paper, we distinguish two types of representations and show how to combine sets of representations containing both types using so-called combination trees. The combination trees are build with respect to domain knowledge and describe multiple semantics. To employ combination trees for clustering, we introduce a multi-represented version of the hierarchical density-based clustering algorithm OPTICS. OPTICS derives so-called cluster orderings and is quite insensitive to the parameter selection. The introduced version of OPTICS is capable to derive meaningful cluster hierarchies with respect to an arbitrary combination tree. The rest of this paper is organized as follows. Section 2 surveys related work. In Section 3, we define combination trees. Section 4 describes a multi-represented version of OPTICS which is based on combination trees. In Section 5, we provide encouraging experimental results.

2 Related Work

In [1] an algorithm for spectral clustering of multi-represented objects is proposed. [2] introduces Expectation Maximization (EM) clustering and agglomerative clustering for multi-represented data. Finally, [3] introduces the framework of reinforcement clustering, which is applicable to multi-represented objects. However, these three approaches do not consider any semantic aspects of the underlying data spaces. In [4], DBSCAN [5] has been adapted to multi-represented

objects distinguishing two possible semantics. However, DBSCAN has several drawbacks leading to the development of OPTICS[6] which is the algorithm the method proposed in this paper is based on.

3 Handling Semantics

In [4], there were two general methods to combine multiple representation for density based clustering, called union and intersection method. The union method states that an object is an union core-object if there are at least k data objects in the union of the local ε-neighborhoods. The intersection method was defined analogously. However, it is not clear which method is better suited to compare an arbitrary set of representations. In [7], the suitability of representations for one or the other combination method is discussed. As a result, two aspects of a data space can be distinguished, the precision space and recall space property. An examples for a good precision space are word vectors because documents containing the same set of words usually describe the same content. An example for a recall space are color histograms because two images having a similar content usually have similar color distributions. Furthermore, we can state that precision spaces should be combined using the union method and recall spaces should be combined using the intersection method. The result of combining recall spaces improves the precision and the result of combining precision spaces improves the recall. Thus, we can successively group representation of both types and construct a so-called combination tree according to the following formalization:

Definition 1 (Combination Tree). *Let $R = \{R_1, \ldots, R_m\}$. A combination tree CT for R is a tree of arbitrary degree fulfilling the following conditions:*

- *CT.root denotes the root of the combination tree CT.*
- *Let n be a node of CT, then n.label denotes the label of n and n.children denotes the children of n.*
- *The leaves are labeled with representations, i.e. for each leaf $n \in CT$: $n.label \in \{R_1, \ldots, R_m\}$.*
- *The inner nodes are labeled with either the union or the intersection operator, i.e. for each inner node $n \in CT$: $n.label \in \{\cup, \cap\}$.*

4 Hierarchical Clustering of Multi-Represented Objects

In order to obtain the comparability of distances, we normalize the distance in representation R_i with respect to the mean value μ_i^{orig} of the original distance d_i^{orig}. The algorithm OPTICS [6] works like an extended DBSCAN algorithm, computing the density-connected clusters w.r.t. all parameters ε_i that are smaller than a generic value of ε. OPTICS does not assign cluster memberships, but stores the order in which the objects have been processed and the information can be used to assign cluster memberships. This information consists of two values for each object, its core distance and its reachability distance. To compute these information during a run of OPTICS on multi-represented objects, we

must adapt the core distance and reachability distance predicates of OPTICS to our multi-represented approach. In the following, we will show how we can use a combination tree CT for a given set of representations R to cluster multi-represented objects. The (global) distance between two objects $o, p \in \mathcal{D}$ w.r.t. a combination tree CT is defined as the combination of the distances of the nodes of CT.

Definition 2 (distance w.r.t. CT).
Let $o, p \in \mathcal{D}$, $R = \{R_1, \ldots, R_m\}$, d_i be the distance function of R_i, CT be a combination tree for R, and let n be a node in CT, i.e. $n.label \in \{\cup, \cap, R_1, \ldots, R_m\}$.

The distance between o and p w.r.t. node $n \in CT$, *denoted by $d^n(o,p)$, is recursively defined by*

$$d^n(o,p) = \begin{cases} \min\limits_{c \in n.children} \{d^c(o,p)\} & \text{if } n.label = \cup \\ \max\limits_{c \in n.children} \{d^c(o,p)\} & \text{if } n.label = \cap \\ d_i(o,p) & \text{if } n.label = R_i \end{cases}$$

The distance between o and p w.r.t. CT, *denoted by $d_{CT}(o,p)$, is defined by*

$$d_{CT}(o,p) = d^{CT.root}(o,p)$$

The (global) ε-neighborhood of an object $o \in \mathcal{D}$ w.r.t. a combination tree CT is defined as the combination of the ε-neighborhoods of the nodes of CT.

Definition 3 (ε-neighborhood w.r.t. CT).
Let $o \in \mathcal{D}$, $\varepsilon \in \mathbb{R}^+$, $R = \{R_1, \ldots, R_m\}$, CT be a combination tree for R, and let n be a node in CT, i.e. $n.label \in \{\cup, \cap, R_1, \ldots, R_m\}$.

The ε-neighborhood of o w.r.t. node $n \in CT$, *denoted by $\mathcal{N}_\varepsilon^n(o)$, is recursively defined by*

$$\mathcal{N}_\varepsilon^n(o) = \begin{cases} \bigcup\limits_{c \in n.children} \mathcal{N}_\varepsilon^c(o) & \text{if } n.label = \cup \\ \bigcap\limits_{c \in n.children} \mathcal{N}_\varepsilon^c(o) & \text{if } n.label = \cap \\ \mathcal{N}_\varepsilon^{R_i}(o) & \text{if } n.label = R_i \end{cases}$$

The ε-neighborhood of o w.r.t. CT, *denoted by $\mathcal{N}_{CT,\varepsilon}(o)$, is defined by*

$$\mathcal{N}_{CT,\varepsilon}(o) = \mathcal{N}_\varepsilon^{CT.root}(o)$$

Since the core distance predicate of OPTICS is based on the concept of k-nearest neighbor (k-NN) distances, we have to redefine the k-nearest neighbor distance of an object o w.r.t. a combination tree CT.

Definition 4 (k-NN distance w.r.t. CT).
Let $o \in \mathcal{D}$, $k \in \mathbb{N}$, $|\mathcal{D}| \geq k$, $R = \{R_1, \ldots, R_m\}$, CT be a combination tree for R, and let n be a node in CT, i.e. $n.label \in \{\cup, \cap, R_1, \ldots, R_m\}$.

The k-nearest neighbors of o w.r.t. CT *is the smallest set* $NN_{CT,k}(o) \subseteq \mathcal{D}$ *that contains (at least)* k *objects and for which the following condition holds:*

$$\forall p \in NN_{CT,k}(o), \forall q \in \mathcal{D} - NN_{CT,k}(o) : d_{CT}(o,p) < d_{CT}(o,q).$$

The k-nearest neighbor distance of o w.r.t. CT, *denoted by* $\text{NN-DIST}_{CT,k}(o)$, *is defined as follows:*

$$\text{NN-DIST}_{CT,k}(o) = \max\{d_{CT}(o,q)\} \mid q \in NN_{CT,k}(o)\}.$$

Now, we can adopt the core distance definition from OPTICS to our combination approach: If the ε-neighborhood w.r.t. CT of an object o contains at least k objects, the core distance of o is defined as the k-nearest neighbor distance of o. Otherwise, the core distance is infinity.

Definition 5 (core distance w.r.t. CT).
Let $o \in \mathcal{D}$, $k \in \mathbb{N}$, $|\mathcal{D}| \geq k$, $R = \{R_1, \ldots, R_m\}$, CT *be a combination tree for* R, *and let* n *be a node in* CT, *i.e.* $n.label \in \{\cup, \cap, R_1, \ldots, R_m\}$.

The core distance of o w.r.t. CT, ε and k, *denoted by* $\text{CORE}_{CT,\varepsilon,k}(o)$, *is defined by*

$$\text{CORE}_{CT,\varepsilon,k}(o) = \begin{cases} \text{NN-DIST}_{CT,k}(o) & \text{if } |\mathcal{N}_{CT,\varepsilon}(o)| \geq k \\ \infty & \text{otherwise.} \end{cases}$$

The reachability distance of an object $p \in \mathcal{D}$ from $o \in \mathcal{D}$ w.r.t. CT is an asymmetric distance measure that is defined as the maximum value of the core distance of o and the distance between p and o.

Definition 6 (reachability distance w.r.t. CT).
Let $o, p \in \mathcal{D}$, $k \in \mathbb{N}$, $|\mathcal{D}| \geq k$, $R = \{R_1, \ldots, R_m\}$, CT *be a combination tree for* R, *and let* n *be a node in* CT, *i.e.* $n.label \in \{\cup, \cap, R_1, \ldots, R_m\}$.

The reachability distance of o to p w.r.t. CT, ε, and k, *denoted by* $\text{REACH}_{CT,\varepsilon,k}(p,o)$, *is defined by*

$$\text{REACH}_{CT,\varepsilon,k}(p,o) = \max\{\text{CORE}_{CT,\varepsilon,k}(p), d_{CT}(o,p)\}$$

5 Performance Evaluation

We implemented the proposed clustering algorithm in Java 1.5 and ran several experiments on a work station with two 1.8 GHz Opteron processors and 8 GB main memory. The experiments were performed on protein data that is described by text descriptions (R_1) and amino-acid sequences (R_2). We employed entries of the Swissprot protein database [1] belonging to 5 functional groups (cf. Table 1). As reference clustering, we employed the classes of Gene Ontology [2]. To evaluate the derived cluster structure C, we extracted flat clusters from OPTICS plots

[1] http://us.expasy.org/sprot/sprot-top.html
[2] www.geneontology.org

and applied the following quality measure for comparing different clusterings w.r.t. the reference clustering K: $Q_K(C) = \sum_{C_i \in C} \frac{|C_i|}{|DB|} \cdot (1 - entropy_K(C_i))$. We employed an combination tree describing the union of both representations. As first comparison partners, we clustered text and sequences separately using only one of the representations. A second approach combines the features of both representations into a common feature space (CFS) and employs the cosine distance to relate the resulting feature vectors. Additionally, we compared reinforcement clustering (RCL) using DBSCAN as underlying cluster algorithm. For reinforcement clustering, we ran 10 iterations and tried several values of the weighting parameter α. The ε-parameters were set sufficiently large and we chose $k = 2$. Table 1 displays the derived quality for our method and the four competitive methods mentioned above. As it can be seen, our method clearly outperforms any of the other algorithms.

Table 1. Description of the protein data sets and results.

	Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 5**
Name	Isomerase	Lyase	Signal Transducer	Oxidoreductase	Transferase
No. of Classes	16	35	39	49	62
No. of Objects	501	1640	2208	3399	4086
$R_1 \cup R_2$	**0.66**	**0.56**	**0.43**	**0.50**	**0.38**
R_1	0.61	0.54	0.32	0.46	0.35
R_2	0.31	0.25	0.36	0.39	0.24
CFS	0.62	0.46	0.28	0.41	0.29
RCL	0.55	0.43	0.25	0.33	0.19

Another, set of experiments were performed on a data set of images being described by 4 representations. The OPTICS clustering based on a 2 level combination trees achieved encouraging results as well. More information about these experiments can be found in [7].

References

1. De Sa, V.R.: Spectral Clustering with two Views. In: Proc. ICML Workshop. (2005)
2. Bickel, S., Scheffer, T.: Multi-View Clustering. In: Proc. ICDM. (2004)
3. Wang, J., Zeng, H., Chen, Z., Lu, H., Tao, L., Ma, W.: ReCoM: Reinforcement clustering of multi-type interrelated data objects. In: Proc. SIGIR. (2003)
4. Kailing, K., Kriegel, H.P., Pryakhin, A., Schubert, M.: Clustering Multi-represented Objects with Noise. In: Proc. PAKDD. (2004)
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proc. KDD. (1996)
6. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: OPTICS: Ordering Points to Identify the Clustering Structure. In: Proc. SIGMOD. (1999)
7. Achtert, E., Kriegel, H.P., Pryakhin, A., Schubert, M.: Hierarchical Density-Based Clustering for Multi-Represented Objects. In: Workshop on Mining Complex Data (MCD 2005)at ICDM05, Houston, TX, USA. (2005)