Hierarchical Density-Based Clustering of Uncertain Data

Hans-Peter Kriegel, Martin Pfeifle

Institute for Computer Science
University of Munich, Germany
{*kriegel*, *pfeifle*}*@ifi.dbs.lmu.de*

Abstract

The hierarchical density-based clustering algorithm OPTICS has proven to help the user to get an overview over large data sets. When using OPTICS for analyzing uncertain data which naturally occur in many emerging application areas, e.g. location based services, or sensor databases, the similarity between uncertain objects has to be expressed by one numerical distance value. Based on such single-valued distance functions OPTICS, like other standard data mining algorithms, can work without any changes. In this paper, we propose to express the similarity between two fuzzy objects by distance probability functions which assign a probability value to each possible distance value. Contrary to the traditional approach, we do not extract aggregated values from the fuzzy distance functions but enhance OPTICS so that it can exploit the full information provided by these functions. The resulting algorithm FOPTICS helps the user to get an overview over a large set of fuzzy objects.

1. Introduction

The standard OPTICS algorithm [1], like other data mining algorithms, can only analyze uncertain data if we measure the distance (similarity) between uncertain objects by one numerical value. These single-valued distance functions do not reflect the uncertainty of the compared objects, as all possible distances are aggregated by only one distance value. Obviously, aggregation goes hand in hand with information loss. For instance, we have no information about the degree of uncertainty of such a single distance value. Even if we had one, it would be of no use because traditional algorithms, e.g. join algorithms or clustering algorithms, cannot handle this additional information.

In this paper, we propose to use fuzzy distance functions to measure the similarity between uncertain object representations. Contrary to the traditional approaches, we do not extract aggregated values from these fuzzy distance functions but propose to enhance data mining algorithms so that they can exploit the full information provided by these functions. As hierarchical density-based clustering is very useful for analyzing complex data sets, we exemplarily demonstrate in this paper how fuzzy distance functions can be integrated into the hierarchical density-based clustering algorithm OPTICS. We call the resulting clustering algorithm FOPTICS indicating that it is based on fuzzy distance functions measuring the similarity between uncertain data.

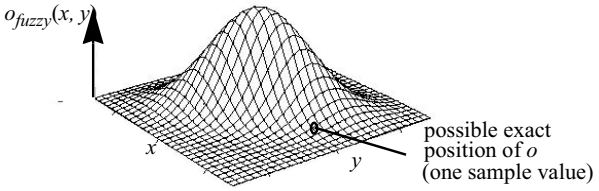


Figure 1: Fuzzy object representation.

2. FOPTICS

In this section, we will present the algorithm FOPTICS suitable for clustering fuzzy object representations. First, in Section 2.1, we present the theoretical foundations underlying the FOPTICS approach. Then, in Section 2.2, we put these theoretical concepts into practice.

2.1 Theoretical Foundations

Fuzzy object representations, also denoted simply as fuzzy objects, occur in many different application ranges. They can be modelled by a probability density function indicating the likelihood that an object is located at a certain position (cf. Figure 1).

Definition 1 *Fuzzy Object Representation*

Let $o \in D \subseteq IR^d$ be an object from a database. A *fuzzy object representation* is a function o_{fuzzy}: $IR^d \rightarrow IR_0^+ \cup \infty$, for which the following condition holds:

$$\int_{IR^d} o_{fuzzy}(v) dv = 1$$

The distance between two fuzzy objects can be described by the following two fuzzy distance functions.

Definition 2 *Distance Density Function*

Let d: $D \times D \rightarrow IR_0^+$ be a distance function, and let $P(a \le d(o, o') \le b)$ denote the probability that $d(o,o')$ is between a and b. Then a probability density function p_d: $D \times D \rightarrow (IR_0^+ \rightarrow IR_0^+ \cup \infty)$ is called a *distance density function* if the following condition holds:

$$P(a \le d(o, o') \le b) = \int_a^b p_d(o, o')(x) dx$$

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

Definition 3 *Distance Distribution Function*
Let d: $D \times D \rightarrow IR_0^+$ be a distance function, and let $P(d(o, o') \leq b)$ denote the probability that $d(o, o')$ is smaller than b. Then a probability distribution function P_d: $O \times O \rightarrow (IR_0^+ \rightarrow [0..1])$ is called a *distance distribution function* if the following condition holds:

$$P_d(o, o')(b) = P(d(o, o') \leq b)$$

Based on these two fuzzy distance functions, we can extend the fundamental definitions of hierarchical density-based clustering as presented in [1] so that they can be used for clustering fuzzy object representations.

Definition 4 *Fuzzy Core Distance*
Let D be a database, and let P_d: $D \times D \rightarrow (IR_0^+ \rightarrow [0..1])$ be a fuzzy distance distribution function. Then, the fuzzy core distance $core-dist^{fuzzy}_{\varepsilon, \mu, d, D} : D \rightarrow (IR_0^+ \rightarrow IR_0^+ \cup \infty)$ is defined as follows:

$$core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)(x) =$$

$$\left(\sum_{\substack{A \subseteq D \\ |A| \geq \mu}} \prod_{p \in A} P_d(p, o)(x) \prod_{p' \in D \setminus A} (1 - P_d(p', o)(x)) \right) / dx$$

Lemma 1. The function $core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)(x)$ is equal to the density probability function reflecting the likelihood that the core distance of o (according to [1]) is equal to x.

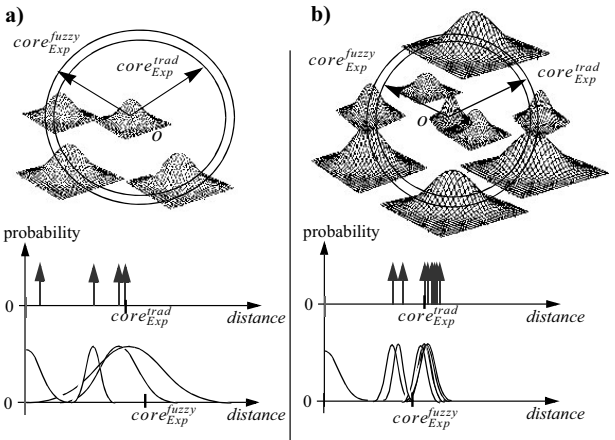
Proof. First, we determine for each subset A of D having a cardinality higher than μ the probability that only the points of A are within an x-range of o but no points of $D \setminus A$. The sum of all these probability values is equal to the probability distribution function $C(o)(x) = P(core\text{-}dist_{\varepsilon, \mu, d, D}(o) \leq x)$ which assigns to each distance value x a probability value that the core distance of o is smaller than x. The function $core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)(x)$ corresponds to the derivation of $C(o)(x)$. □

Note that the traditional core-distance value $core\text{-}dist_{\varepsilon, \mu, d, D}(o)$ of an object o can also be regarded as a fuzzy core distance:

$$core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)(x) = \delta(x - core\text{-}dist_{\varepsilon, \mu, d, D}(o)),$$

where δ denotes the dirac-delta function.

Figure 2 shows that the expectation value $core^{fuzzy}_{Exp}$ of the fuzzy core-distance function $core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)(x)$ differs from $core^{trad}_{Exp}$ which is the μth smallest expectation value of the distance density functions $p_d(p, o)$. In the traditional approach, the distance density functions $p_d(p, o)$ are mapped onto one value, i.e. their expectation value, and based on this aggregated information the expected core distance $core^{trad}_{Exp}$ is computed. In the fuzzy approach, the fuzzy core distance is computed by exploiting the full information provided by the distance density functions. Figure 2 shows that the expectation value of the fuzzy core-distance function represents the exact core-distance much better than the traditional approach. For instance, in Figure 2a, the exact value $core\text{-}dist_{\varepsilon, \mu, d, D}(o)$ seems to be higher than the value $core^{trad}_{Exp}$ as the area around o is not very dense. On the other hand, in Figure 2b, the exact value $core\text{-}dist_{\varepsilon, \mu, d, D}(o)$ seems to be smaller than the value $core^{trad}_{Exp}$ as the area around o is very dense. In both cases, it is very likely that the value $core^{fuzzy}_{Exp}$ better reflects the exact value $core\text{-}dist_{\varepsilon, \mu, d, D}(o)$ than the traditional value $core^{trad}_{Exp}$.



the traditional core-distance value $core^{trad}_{Exp}$ is smaller than the expectation value $core^{fuzzy}_{Exp}$ of the fuzzy core distance.

the traditional core-distance value $core^{trad}_{Exp}$ is greater than the expectation value $core^{fuzzy}_{Exp}$ of the fuzzy core distance.

Figure 2: Fuzzy core distance ($\mu = 4$).

The reachability distance $reach\text{-}dist_{\varepsilon, \mu, d, D}(p, o)$ is defined as $max(core\text{-}dist_{\varepsilon, \mu, d, D}(o), d(p, o))$ [1]. In our case, both of the two input parameters are fuzzy distance functions. The following definition combines these two fuzzy input functions to a fuzzy reachability distance which again is a generalization of the traditional definition.

Definition 5 *Fuzzy Reachability-Distance*
Let $p_d(p,o)$ be the fuzzy distance function between object p and o, and let $core-dist^{fuzzy}_{\varepsilon, \mu, d, D}(o)$ be the fuzzy core distance of o. Then, the fuzzy reachability distance $reach-dist^{fuzzy}_{\varepsilon, \mu, d, D}$: $D \times D \rightarrow (IR_0^+ \rightarrow IR_0^+ \cup \infty)$ is defined as follows:

$$reach-dist^{fuzzy}_{\varepsilon, \mu, d, D}(p, o)(x) =$$

$$p_d(p, o)(x) \cdot \int_{-\infty}^{x} core-dist^{fuzzy}_{\varepsilon, \mu-1, d, D \setminus \{p\}}(o)(t)dt +$$

$$core-dist^{fuzzy}_{\varepsilon, \mu-1, d, D \setminus \{p\}}(o)(x) \cdot \int_{-\infty}^{x} p_d(p, o)(t)dt$$

Lemma 2. The function $reach-dist^{fuzzy}_{\varepsilon, \mu, d, D}(p, o)(x)$ is equal to the probability density function reflecting the likelihood that the reachability-distance of p w.r.t. o (according to [1]) is equal to x.

Proof. As the two density functions $core\text{-}dist^{fuzzy}_{\varepsilon, \mu-1, d, D \setminus \{p\}}(o)$ and $p_d(p, o)$ are independent from each other, we can argue as follows. Let $C(o)(x) = \int_{-\infty}^{x} core-dist^{fuzzy}_{\varepsilon, \mu-1, d, D \setminus \{p\}}(o)(t)dt$ denote the probability that $\mu-1$ objects of $D \setminus \{p\}$ are located in $N_x(o)$. Furthermore, $P_d(p, o)(x) = \int_{-\infty}^{x} p_d(p, o)(t)dt$ denotes the proba-

bility that the distance between o and p is smaller than x. As the two events are independent from each other, the value $C(o)(x) \cdot P_d(p, o)(x)$ denotes the probability that the core distance of o is smaller than x, and that p is located within $N_x(o)$. Finally, $reach-dist^{fuzzy}_{\varepsilon,\mu,d,D}(p, o)(x) = \frac{C(o)(x) \cdot P_d(p, o)(x)}{dx}$ holds due to the product rule for differentiation. □

The traditional reachability distance value $reach\text{-}dist_{\varepsilon,\mu,d,D}(p,o)$ [1] can also be regarded as a fuzzy reachability distance:

$$reach-dist^{fuzzy}_{\varepsilon,\mu,d,D}(p, o)(x) = \delta(x - reach\text{-}dist_{\varepsilon,\mu,d,D}(p, o)),$$

where again δ denotes the dirac-delta function.

The traditional OPTICS algorithm repeatedly updates the reachability values of all elements in the seedlist. These reachability values determine the ordering of the elements within the seedlist. The reachability value of an object p within the seedlist denotes the smallest reachability distance to all objects o which have already been processed.

In our approach, we assign to each object in the seedlist not a single reachability value but a fuzzy object reachability function. More specifically, our approach assigns to each distance value t, a density probability value which reflects the likelihood that the exact object reachability distance is equal to t (cf. Figure 3).

Definition 6 *Fuzzy Object Reachability-Distance*
Let $Pred \subseteq D$ denote the set of all objects which have already been processed during an ($\mathcal{F}$)OPTICS run, and let $p \in D \setminus Pred$ denote an element of the seedlist. Then, $\forall x \in IR_0^+$ the fuzzy object reachability distance $obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\mu,d,D} : D \rightarrow (IR_0^+ \rightarrow IR_0^+ \cup \infty)$ is defined as follows:

$$\int_{-\infty}^{x} obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\mu,d,D}(p)(t)dt = P(min\{reach-dist_{\varepsilon,\mu,d,D}(p, o) | o \in Pred\} \leq x)$$

Note that we cannot order the elements of the seedlist according to a function, but we need an enumeration type. Therefore, we determine the expectation value of the fuzzy object reachability distances, and order the elements of the seedlist according to these values. It is quite a difference whether we form the expectation value for each fuzzy reachability distance and order the objects according to the minimum of these values, or whether we form the expectation value of the fuzzy object reachability distance and order the elements according to this value (cf. Figure 3). Again, the later we aggregate, the more accurate the results are.

2.2 Computational Aspects

Algorithm. The $\mathcal{F}$OPTICS algorithm basically works like the OPTICS algorithm. We always take the first element from the seedlist, add it to the result set, and carry out a range query. As outlined above, the ordering of the seedlist and the final object reachability values are based on the expectation values of the object reachability functions $obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\mu,d,D}$. The remaining question is how to compute these values efficiently and effectively. In this paper, we propose to compute the reachability

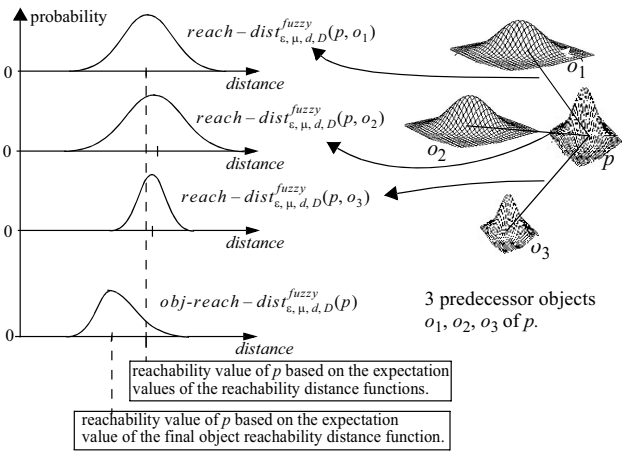


Figure 3: Fuzzy object reachability distance.

plot generated by the $\mathcal{F}$OPTICS algorithm based on monte-carlo sampling. In many applications the fuzzy objects might already be described by a discrete probability density function, i.e. we have the sample set already. If the fuzzy object is described by a continuos probability density function, we can easily sample according to this function and derive thus a sequence of samples. In the following, we assume that each object x is represented by a sequence of s sample points, i.e. x is represented by s different representations $<x_1, ..., x_s>$. Thus, the distance between two fuzzy objects could be described by s^2 many sample distances. If we modelled the distance distribution function between two objects by these s^2 distances, the minimum runtime of the $\mathcal{F}$OPTICS approach would be $O(s^2 \cdot n^2)$, as we have to compute the distance between all $O(n^2)$ many object pairs. Note that index structures cannot help to reduce the runtime as we have to use very high ε-values in order to detect the complete clustering hierarchy.

We aim at an $\mathcal{F}$OPTICS algorithm which effectively and efficiently outperforms a standard OPTICS run based on the distance expectation values which has a runtime of $O(s^2 \cdot n^2)$. The main idea of our implementation is to generate meaningful samples which represent the functions introduced in Definition 4, 5 and 6. In order to guarantee a runtime complexity of $O(s \cdot n^2)$, we perform only O(s) distance computations on the samples to compute a distance distribution function between two fuzzy objects. The fuzzy distance functions of Definition 4, 5 and 6 are then also represented by s sample values (cf. Figure 4).

In the following, we describe the $\mathcal{F}$OPTICS approach. We compute for each query object o, i.e. the first object in the ordered seedlist (cf. Figure 4a), an array $A(o)$ by using one range scan (cf. Figure 4b). The s values a_i of $A(o)$ reflect the core-distance values of o_i w.r.t. instance database $D_i = \{x_i | <x_1, ..., x_i, ..., x_s> \in D\}$, *i.e.* $a_i = core\text{-}dist_{\varepsilon,\mu,d,D_i}(o_i)$. Then, for all objects $p \in D$ we compute the s values $reach\text{-}dist_{\varepsilon,\mu,d,D_i}(p_i, o_i) = max(core\text{-}dist_{\varepsilon,\mu,d,D_i}(o_i), d(p_i,o_i))$ and organize these values $reach_{p_i,o_i}$ in a list $L(p,o)$ which can be regarded as a meaningful representation of the function $reach-dist^{fuzzy}_{\varepsilon,\mu,d,D}(p, o)$ introduced in

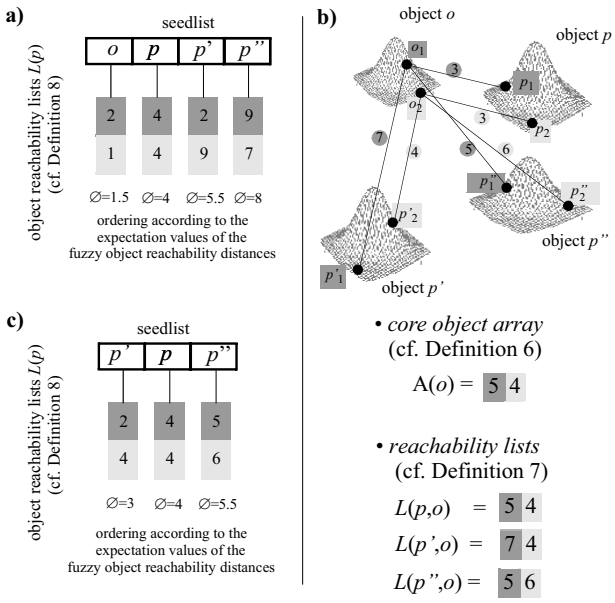


Figure 4: One cycle of the $\mathcal{F}$OPTICS approach (s = 2, μ = 2).
a) seedlist before object o is processed
b) core object array A(o), and reachability lists $L(p,o)$, $L(p',o)$, $L(p'',o)$
c) seedlist after object o is processed

Definition 5 (cf. Figure 4b). In order to compute the fuzzy object reachability distance $obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\,\mu,\,d,\,D}(p)$ according to Definition 6, we merge for a given object p all the lists $L(p, o)$ of all objects o that have already been processed throughout the $\mathcal{F}$OPTICS run. In this merging process, we create the final object reachability list $L(p)$ (cf. Figure 4a and 4c) containing s entries $R(p_1)$, ..., $R(p_s)$ where $R(p_i)$ reflects the reachability value of p in the instance database D_i w.r.t. all objects which have already been processed throughout the $\mathcal{F}$OPTICS run.

Let us now assume that n objects $o^{(1)}$ to $o^{(n)}$ have already been processed. Furthermore, we assume that $o_i^{(1)}$ to $o_i^{(n)}$ are the corresponding instances of D_i. Then, we can easily compute the reachability value $R(p_i)$ of p for D_i by $R(p_i) = min\ \{reach_{p_i,o_i^{(1)}}, ..., reach_{p_i,o_i^{(n)}}\}$. As the $\mathcal{F}$OPTICS algorithm requires only the object function $obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\,\mu,\,d,\,D}(p)$ and not all $reach-dist^{fuzzy}_{\varepsilon,\,\mu,\,d,\,D}(p, o)$ functions of all predecessor objects o, it is enough to assign the list $L(p)$ to each object p stored in the seedlist (cf. Figure 4a and 4c). Note that we can compute the list $L(p)$ without materializing all the other lists $L(p, o^{(1)})$, ..., $L(p, o^{(n)})$ due to the following reasoning. Let $L^{(n)}(p) = (R^{(n)}(p_1), ..., R^{(n)}(p_s))$ be the list representing the object reachability distance $obj\text{-}reach\text{-}dist^{fuzzy}_{\varepsilon,\,\mu,\,d,\,D}(p)$ after n objects have been processed. Furthermore, let s be the used sample rate, and let $o^{(n+1)}$ be the next object to be processed. Then, the list $L^{(n+1)}(p) = (R^{(n+1)}(p_1), ..., R^{(n+1)}(p_s))$ representing the new object reachability distance can be computed as follows:

$$\forall i \in 1 \ldots s : R^{(n+1)}(p_i) = min\ (reach_{p_i,o_i^{(n+1)}}, R^{(n)}(p_i)).$$

Runtime Complexity. From an efficiency point of view, our $\mathcal{F}$OPTICS approach basically carries out s parallel OPTICS runs on simple object samples, i.e. feature vectors. Thereby all of the s OPTICS runs process the database objects in the same order. Therefore, for each database object o we only need $O(1)$ many range scans for the determination of the core object array $A(o)$ (cf. Definition 4), and for the determination of the reachability lists $L(p,o)$ (cf. Definition 5). Furthermore, in each cycle of our $\mathcal{F}$OPTICS approach, we require only $O(s \cdot n)$ many distance computations between d-dimensional feature vectors. The merging of the reachability lists $L(p,o)$ to object reachability lists $L(p)$ can be done without any further distance computations between d-dimensional feature vectors.

To sum up, if we assume n database objects and a sample rate of s, we can summarize the characteristics of the $\mathcal{F}$OPTICS implementation as follows:

- We need $O(n)$ range scans.
- We need $O(s \cdot n^2)$ many distance computations between d-dimensional feature vectors.

Note that the number of distance computations required by OPTICS based on the distance expectation values is equal to $O(s^2 \cdot n^2)$.

3. Conclusion

In this paper, we demonstrated how we can generate a hierarchical density-based clustering on vague and uncertain information which often occurs in modern application ranges like sensor databases, spatio-temporal applications, and biometric information systems. Besides lying the theoretical foundations for hierarchical density-based clustering of uncertain data, we showed how to put these concepts into practice. The resulting algorithm $\mathcal{F}$OPTICS can be used to cluster uncertain data, e.g. moving objects, effectively and efficiently. The algorithm follows the general paradigm of integrating fuzzy distance functions directly into data mining algorithms instead of working on lossy aggregated information. According to our experimental evaluation (which is left out due to space limitations), the newly introduced clustering algorithm $\mathcal{F}$OPTICS achieves much more accurate results than state-of-the-art comparison partners without sacrificing efficiency.

In our future work, we will show that also other data mining algorithms working on vague information can benefit from a direct integration of fuzzy distance functions.

References

[1] Ankerst M., Breunig M., Kriegel H.-P., Sander J.: *OPTICS: Ordering Points To Identify the Clustering Structure*. SIGMOD' 99.