A COMPUTATIONAL GEOMETRY APPROACH TO CLUSTERING PROBLEMS

F.Dehne H.Noltemeier

Lehrstuhl fuer Informatik I , Univ. of Wuerzburg
Am Hubland, 8700 Wuerzburg, W.-Germany

Abstract:
This paper deals with the relationship between cluster analysis and computational geometry describing clustering strategies using a Voronoi diagram approach in general and a line separation approach to improve the efficiency in a special case. We state the following theorems :

1. The set of all centralized 2-clusterings (S_1,S_2) of a planar point set S with $|S_1|=a$ and $|S_2|=b$ is exactly the set of all pairs of labels of opposite Voronoi polygons $v_a(S_1,S)$ and $v_b(S_2,S)$ of $V_a(S)$ and $V_b(S)$ respectively.
2. An optimal centralized 2-clustering [centralized divisive hierarchical 2-clustering] can be constructed in $O(n\ n^{1/2}\ \log^2 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ [$O(n\ n^{1/2}\ \log^3 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ respectively] steps with $P_F(n)$ and $U_F(n)$ being the time complexity to compute and update a given clustering measure f.

1. Intoduction

Given a set S of n points $x_1,\ldots,x_n \in R^d$ (this paper will deal only with planar point sets - d=2 - and Euclidean metric), a partition of S into C disjoint "natural groupings" $S_1,\ldots,S_C$ is called a "C-clustering" of S.
There are several ways to specify "natural groupings". You can ask for minimization (maximization) of some "clustering measure" $f:(S_1,\ldots,S_C)\text{-->}r\in R$ (e.g. minimize the maximum diameter) or you give an algorithmic specification.
Most of the proposed strategies in clustering literature can be classified according to fig.1 .

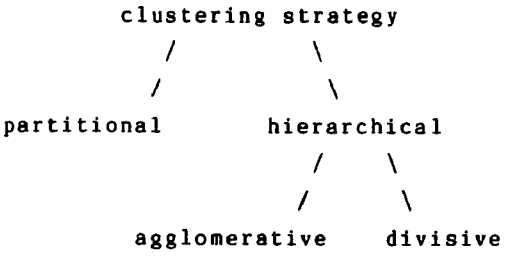


fig.1

Agglomerative hierarchical (divisive hierarchical) algorithms produce a sequence of nested partitions with decreasing (increasing) number of clusters hoping to approach the given goal. Partitional strategies divide S into C clusters at once trying mostly to improve this partitioning in some postprocessing steps (keeping the number of clusters constant)- refer to [DE], [DJ], [M], [P], [R].
This paper will deal with the relationships between cluster analysis and computational geometry describing two divisive hierarchical clustering strategies using computational geometry methods.

2. Cluster centers and Voronoi Diagrams

2.1. Basic definitions and properties

Several clustering methodologies (e.g. FORGY/ISODATA, see [DJ]) select C cluster centers from S assigning the remaining n-C points to their nearest cluster center (consult [DJ] for more details).
We extend this to the following

Definition 1:
(a)
A cluster $S_i \subseteq S$ is called "centralized", if there exists a center $x \in R^2$ with S_i being the set of s_i nearest neighbors of x with respect to S. (Let $s_i := |S_i|$ for the remaining of this paper.)
(b)
A C-clustering $(S_1,\ldots,S_C)$ of S is called centralized, if all S_i $(1 \le i \le C)$ are centralized.
(c)
A C-clustering $(S_1,\ldots,S_C)$ of S is called "balanced", if for all $1 \le i < j \le C$: $|s_i - s_j| \le 1$ (This is the most interesting case in practice).

Let $v_k(S_i,S)$ be the order k Voronoi polygon of some $S_i \subseteq S$ $(k = s_i)$ and $V_k(S)$ be the order k Voronoi diagram of S (see [SH] and [L]). We shall call S_i the "label" of the Voronoi polygon $v_k(S_i,S)$. Using the notations of [SH], [L] and [D] it is easy to prove the following

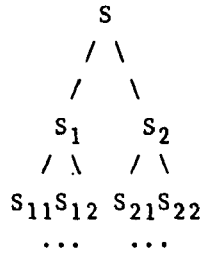
Lemma 1:
1.1
$S_i \subseteq S$ is a centralized cluster if and only if S_i is the label of some Voronoi polygon $v_k(S_i,S) \neq \{\}$.
1.2
$(S_1,\ldots,S_C)$ is a centralized C-clustering if and only if all S_i $(1 \le i \le C)$ are labels of some Voronoi polygon of some Voronoi diagram $V_k(S)$ and S is the disjoint union of $S_1,\ldots,S_k$.
1.3
$(S_1,\ldots,S_C)$ is a balanced centralized C-clustering of S if and only if all S_i $(1 \le i \le C)$ are labels of some Voronoi polygon of $V_{\lfloor n/C \rfloor}(S)$ or $V_{\lceil n/C \rceil}(S)$ and S is the disjoint union of $S_1,\ldots,S_C$.

The proof of Lemma 1 follows immediatly from the Definition of $V_k(S)$ and Def.1 .

Thus, a centralized C-clustering is a selection of disjoint labels of Voronoi polygons. This leads to the idea. to use the geometric properties of Voronoi diagrams for the design of clustering methodologies.

2.2. Applications to divisive hierarchical clustering

Using our above definitions a (C-nested) divisive hierarchical clustering is a nested sequence of C-clusterings (which we will call clustering steps) successively decomposing S into smaller subsets as demonstrated in fig.2 .



$(S_1,S_2),(S_{11},S_{12}),(S_{21},S_{22})$ is a 2-clustering of S , S_1,S_2 respectively

fig.2

We shall call a divisive hierarchical clustering centralized (balanced), if all clustering steps are centralized (balanced).

This chapter will demonstrate the relationships between order k Voronoi diagrams and 2-nested centralized divisive hierarchical clustering.

Definition 2:

Two disjoint Voronoi polygons vp_1 and vp_2 are "opposite" to each other, if there are two nonparallel straight lines g and g' each containing two disjoint rays r_1,r_2 and r_1',r_2' , respectively , with $r_1,r_1' \subseteq vp_1$ and $r_2,r_2' \subseteq vp_2$. (Note that opposite Voronoi polygons are always open.)

With this definition we prove the following lemmata:

Lemma 2:

Let a,b be two positive integers with $a+b \leq n$, $a=|S_1|$, $b=|S_2|$ and $v_a(S_1,S)$, $v_b(S_2,S)$ two nonempty Voronoi polygons which are opposite, then S_1 and S_2 are disjoint.

Proof:

Assume conversely that there is a point p in $S_1 \cap S_2$, and S_1 and S_2 are opposite with $|S_1|+|S_2| \leq n$ (see fig.3). Since it follows that $|S_1|+|S_2|<n$, there must be a point q in $S-(S_1 \cup S_2)$. Now consider the bisector B(p,q) of p and q and let h(p,q) denote the halfplane of all points closer to p than to q. $v_a(S_1,S)$ [$v_b(S_2,S)$] is defined to be the intersection of all h(x,y) with $x \in S_1$ and $y \in S-S_1$ [$x \in S_2$ and $y \in S-S_2$], thus we have $v_a(S_1,S) \subseteq h(p,q)$ and $v_b(S_2,S) \subseteq h(p,q)$. Since S_1 and S_2 are opposite, and B(p,q) must intersect at least one of the lines g and g' this leads to a contradiction. ∎

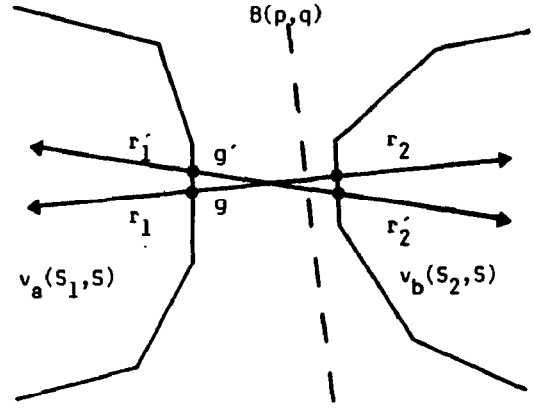


fig.3

Lemma 3:

Let a,b be two positive integers with $a+b=n$, $a=|S_1|$, $b=|S_2|$ and $v_a(S_1,S)$, $v_b(S_2,S)$ two Voronoi polygons with S being the disjoint union of S_1 and S_2, then $v_a(S_1,S)$ and $v_b(S_2,S)$ are open and opposite.

Proof:

Select two distinct points p_1 and p_1' [p_2 and p_2'] from the interior of $v_a(S_1,S)$ [$v_b(S_2,S)$], such that the two lines g and g' defined by (p_1,p_2) and (p_1',p_2') are not parallel to each other. Let r_1 [r_2] be the maximum distance between p_1 [p_2] and all points of S_1 [S_2], and c_1 [c_2] denote the circle with center p_1 [p_2] and radius r_1 [r_2]. Since $c_1 \cap c_2 \cap S = \{\}$, and S is a finite set, S_1 and S_2 are always separable by some line l perpendicular to g . With this it is easy to prove that there is some point x_1 [x_2] on g, such that the points of g to the left [right] of x_1 [x_2] are closer to all points of S_1 [S_2] than to S_2 [S_1], thus being points of $v_a(S_1,S)$ [$v_b(S_2,S)$]. Since the same holds for g', S_1 and S_2 are open and opposite. ∎

Summarizing this, we have

Theorem 1:

The set of all centralized 2-clusterings (S_1,S_2) of S with $|S_1|=a$ and $|S_2|=b$ is exactly the set of all pairs of labels of opposite Voronoi polygons $v_a(S_1,S)$ and $v_b(S_2,S)$ of $V_a(S)$ and $V_b(S)$ respectively.

Because every $S_1 \subseteq S$ has exactly one complement $S_2=S-S_1$, it follows immediatly, that every open order k Voronoi polygon $v_k(S_1,S)$ has exactly one opposite order n-k Voronoi polygon , thus the four bounding rays of these two polygons having pairwise exactly opposite direction.

This is an interesting property of order k Voronoi diagrams, which appears to be new.

Consider the problem of constructing an optimal centralized 2-clustering (S_1,S_2) of S with respect to some clustering measure $f(S_1,S_2) \in R$ and $|S_1|=k$, $|S_2|=n-k$. We assume a given algorithm F, which is able to compute $f(S_1,S_2)$ in time $P_F(n)$ and exchange exactly one element of S_1 and S_2, respectively, in $U_F(n)$ steps (eventually using heraditary properties). The following steps are appropriate to solve the problem:

(1)

Compute all open order k (and n-k) Voronoi polygons sorted by the angle of their bounding rays (respectively). (There are $O(n\,k^{1/2})$ such polygons; see Theorem 1, Lemma 4 and [EW1])

(2)

Follow exactly one revolution of a rotating line pointing at the current pair of opposite Voronoi polygons and select the optimal one with respect to f computing $O(n\,k^{1/2})$ updates using F.

From the aspect of computational complexity step (1) is the most expensive one. Lee [L] has proposed an algorithm to construct an order k diagram in $O(k^2 n \log n)$ steps. With k being of order n in most cases of 2-clustering this would normally lead to an $O(n^3 \log n)$ algorithm, but [ERS] describe some methods to construct all Voronoi diagrams in $O(n^3)$. So current state of the art (as known by the

authors) in constructing Voronoi diagrams leads to an $O(n^3+n\ n^{1/2}\ U_F(n) + P_F(n))$ algorithm to compute an optimal centralized 2-clustering.

A centralized divisive hierarchical clustering will be obtained by a successive application of this algorithm to the current partition of S. This leads to the same asymptotic time complexity.

Note, that we compute much more information than we actually need, leaving us with the problem to look for some better algorithm to construct an order k Voronoi diagram or all of its open polygons, respectively. This will significantly improve the complexity of our algorithm.
For the special case of 2-clustering we will give a more efficient solution in the following chapter.

3. An $O(n\ n^{1/2}\ \log^2 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ algorithm to construct an optimal centralized 2-clustering

To construct an optimal centralized 2-clustering (S_1,S_2) of S with $|S_1|=k$ and $|S_2|=n-k$ we state the following

Lemma 4:

(S_1,S_2) is a centralized 2-clustering of S if and only if S_1 and S_2 are separable and S is the disjoint union of S_1 and S_2.

Proof: very similar to the proof of lemma 3.

After constructing the k-belt of T(S) (see [EW2]) in $O(n\ k^{1/2}\ \log^2 n)$ steps we search along its upper and lower border, respectively, update the clustering value $f(S_1,S_2)$ and select an optimal partition. From [EW1] and [EW2] we know, that our dynamic updating procedure F will be executed $O(n\ k^{1/2})$ times, leading to an $O(n\ n^{1/2}\ \log^2 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ algorithm .

By a successive application of this procedure as described in 2.2. we obtain a centralized divisive hierarchical clustering in $O(n\ n^{1/2}\ \log^3 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ steps.
So we have

Theorem 2:

An optimal centralized 2-clustering [centralized divisive hierarchical 2-clustering] can be constructed in $O(n\ n^{1/2}\ \log^2 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ [$O(n\ n^{1/2}\ \log^3 n + U_F(n)\ n\ n^{1/2} + P_F(n))$ respectively] steps.

4. Remarks

Allowing cluster centers to be points of R^2 gives us the possibility to apply the geometric structure of order k Voronoi diagrams as an interesting tool for solving clustering problems. The described Voronoi diagram approach has the additional advantage of apparently being extendible to centralized C-clustering (in contrast to chapter 3).This is subject of further research.

References

[DE] Day, Edelsbrunner: EFFICIENT ALGORITHMS FOR AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS, Report F122, Institut fuer Informations- verarbeitung, TU Graz, Graz, Austria

[DJ] Dubes,Jain: CLUSTERING METHODOLOGIES IN EXPLORATORY DATA ANALYSIS, in M.C.Yovits (Ed.): Advances in Computers, Vol.19,1980

[D] Dehne: AN $O(N^4)$ ALGORITHM TO CONSTRUCT ALL VORONOI DIAGRAMS FOR K NEAREST NEIGHBOR SEARCHING, Proc. 10th Colloquium on Automata, Languages and Programming , 1983

[ERS] Edelsbrunner, O'Rouke, Seidel: CONSTRUCTING ARRANGEMENTS OF LINES AND HYPERPLANES WITH APPLICATIONS, Report F123 (see [DE]), 1983

[EW1] Edelsbrunner,Welzl: ON THE NUMBER OF LINE- SEPARATIONS OF A FINITE SET IN THE PLANE, Report F97 (see [DE]), 1982

[EW2] Edelsbrunner,Welzl: HALFPLANAR RANGE ESTIMATION Report F98 (see [DE]), 1982

[EW3] Edelsbrunner,Welzl: HALFPLANAR RANGE SEARCH IN LINEAR SPACE AND $O(n^{0.695})$ QUERY TIME, Report F111 (see [DE]), 1983

[L] D.T.Lee: AN APPROACH TO FINDING THE K-NEAREST NEIGHBORS IN THE EUCLIDEAN PLANE, Report , Department of Electrical Engineering and Computer Science, Northwestern Univ., Evanston, Il. 60201, USA, 1981

[M] Murtagh: EXPECTED-TIME COMPLEXITY RESULTS FOR HIERARCHIC CLUSTERING ALGORITHMS WHICH USE CLUSTER CENTERS, Information Processing Letters 16, 1983

[OL] Overmars,Van Leeuwen: MAINTENANCE OF CONFIGURATIONS IN THE PLANE, Journal of Computer and System Science, Vol.23, No.2, 1981

[P] Page: A MINIMUM SPANNING TREE CLUSTERING METHOD Communications of the ACM, Vol.17, No.6, 1974

[R] Rohlf: HIERARCHICAL CLUSTERING USING THE MINIMUM SPANNING TREE, The Computer Journal, Vol.16, No.1, 1973

[S] Schrader: APPROXIMATIONS OF CLUSTERING AND SUBGRAPH PROBLEMS ON TREES, Discrete Applied Math. 6, 1983

[SH] Shamos, Hoey: CLOSEST POINT PROBLEMS, Proc. 16th Ann. IEEE Symp. on Found. of Comp. Sci., 1975

[W] Willard: POLYGON RETRIEVAL, SIAM J.Comput., Vol.11, No.1, 1982

[Y] F.F.Yao: A 3-SPACE PARTITION AND ITS APPLICATION (Extended Abstract), Proc. 15th ACM Symp. on Theory of Comp., 1983