[20] M. Yamamoto, "Direct estimation of three-dimensional motion parameters from image sequence and depth, *Trans. Inst. Electron. Commun. Eng. Japan*, vol. J68-D, no. 4, pp. 562–569, 1985.
[21] ——, "A general aperture problem for direct estimation of 3D motion parameters," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 11, no. 5. pp. 528–536, 1989.
[22] M. Yamamoto, P. Boulanger, J. -A. Beraldin, M. Rioux, and J. Domey, "Direct estimation of deformable motion parameters from range image sequence," in *Proc. 3rd Int. Conf. Comput. Vision* (Osaka), 1990, pp. 460–464.

Threshold Validity for Mutual Neighborhood Clustering

Stephen P. Smith

Abstract— **Clustering algorithms have the annoying habit of finding clusters in random data. This note presents a theoretical analysis of the threshold of the mutual neighborhood clustering algorithm (MNCA) [1] under the hypothesis of random data. This yields a theoretical minimum value of this threshold below which even unclustered data is broken into separate clusters. To derive the threshold, a theorem about mutual near neighbors in a Poisson process is stated and proved. Simple experiments demonstrate the usefulness of the theoretical thresholds.**

Index Terms—**Clustering, Possion point processes, random graphs.**

I. Introduction

Clustering algorithms have the annoying habit of finding clusters in random data [2]. The mutual neighborhood cluster algorithm (MNCA) [1], [3] is no exception. It employs a user-specified threshold to determine the clusters. Small values of this threshold yield a large number of clusters, whereas large values yield a small number of clusters. The purpose of this correspondence is to outline a theoretical method to determine a reasonable lower bound for this threshold.

II. Approach

We define random data as data from a uniform distribution over some compact convex subset of K-dimensional space [5]. If a clustering algorithm is applied to random data, it should, with high probability, yield only a single cluster containing all the points. This is because a random data set should have no meaningful subsets that are clusters.

Thus, if we were able to determine the MNCA threshold at which $(1-\alpha)$ of all random data sets were formed into one cluster by the MNCA, we would, with probability $(1-\alpha)$, be assured that the MNCA, when employed with a larger threshold, would find no clusters in random data.

In this correspondence, we make certain simplifying assumptions that allow us to derive an estimate of the needed threshold.

Manuscript received December 2, 1990; revised June 7, 1991. Recommended for acceptance by Editor-in-Chief A. K. Jain.

The author was with the Northrop Research and Technology Center, Palos Verdes Peninsula, CA 90274. He is now with Intel Corporation, Chandler, AZ 85226.

IEEE Log Number 9205819.

III. The MNCA in Graph Theoretic Terms

It will be advantageous to view the MNCA in graph theoretic terms. Each point (pattern) in an N-point data set is viewed as a node in a complete, weighted graph. The weight on the edge between two nodes X and Y is derived from near neighbor information of X and Y. Specifically, if Y is the nth nearest neighbor of X and X is the mth nearest neighbor of Y, then the weight on edge (X, Y) is $n+m$.

The MNCA operates by taking a specified integer edge weight threshold, say d, and producing a threshold graph G_d from the original graph. A threshold graph is a subset of the original complete graph with the same nodes but with all edges whose weight is greater than the threshold deleted. For the MNCA, clusters are defined as connected components of the G_d, that is, points X and Y are in the same cluster if and only if there is a path from node X to node Y in G_d. Thus, if a G_d of a data set is connected, then the MNCA with a d threshold will produce one cluster for the data.

For computation reasons, in actual implementations of the MNCA, large weights in the complete graph are estimated and not computed [3]. Only the Lth nearest neighbors of each point are actually used, where $L<N$. The weight of an edge between two nodes, such that at least one of them is not in the set of the other's Lth nearest neighbors, is set to an arbitrary large value greater than $2L$.

IV. Assumptions

We would like to find the probability that G_d is connected for random data. This is very difficult to derive. We make the following simplifying assumptions:

- A1 We only consider the probability that a point that is the nth nearest neighbor of point X has point X as its mth nearest neighbor.
- A2 This conditional distribution of mutual near neighbors for random data can be approximated by the corresponding distribution of mutual near neighbors from a Poisson point process [2]. A Poisson point process is a stationary, isotropic process that scatters points such that the number of points in any finite Borel subset of K-dimensional space is a random variable following the Poisson distribution. Further, the points are independent of one another, and the number of points in disjoint subsets of space are independent.
- A3 The probability that a threshold graph with V edges generated from random data is connected is approximately the same as the probability that a random graph with V edges is connected [4]. A random graph is a graph chosen at random from the set of all undirected graphs with N (labeled) nodes and V edges.

Assumption A1 ignores the metric structure of mutual near-neighbor distance, i.e., the dependence of mutual near-neighbor distances among points in a set. Assumption A2 ignores edge effects, which increase with increasing K. Assumption A3 ignores the metric structure of Euclidean space, which gives quite severe constraints due to the triangle inequality in two dimensions, although these lessen for large K.

V. Derivation

Let e_d denote the number of edges in G_d. We have

$$e_d = (1/2) \sum_{n=1}^{\min(L,d)} \sum_{i=1}^{N}$$

TABLE I
THRESHOLD VALUE FOR d ABOVE WHICH $100(1-\alpha)$ PERCENT OF ALL SIZE N RANDOM DATA SETS IN K DIMENSIONS ARE CLUSTERED INTO ONE CLUSTER BY THE MNCA

		$(1-\alpha)$		
K	N	0.8	0.95	0.99
2	20	8	11	13
2	50	11	14	17
2	100	13	16	19
2	180	15	18	21
4	20	9	11	13
4	50	11	14	17
4	100	13	16	19
4	180	15	18	21
8	20	9	11	13
8	50	12	14	17
8	100	13	16	19
8	180	15	18	21

$$I\left[\begin{array}{l}\text{Point } i \text{ is at least the } (d-n)\text{th nearest neighbor of the}\\ \text{point which is point } i\text{'s } n\text{th nearest neighbor}\end{array}\right]$$

where $I[.]$ is the indicator function. Taking the expected value of both sides yields

$$E[e_d] = (1/2) \sum_{n=1}^{\min(L,d)} \sum_{i=1}^{N} \sum_{m=1}^{d-n}$$
$$P\left[\begin{array}{l}\text{Point } i \text{ is the } m\text{th nearest neighbor of the point}\\ \text{which is point } i\text{'s } n\text{th nearest neighbor}\end{array}\right]$$

Let X and Y be arbitrary points in a Poisson point process, and let $P_n(m)$ be the conditional probability that point X is the mth nearest neighbor of point Y given that point Y is the nth nearest neighbor of point X. Using assumptions A1 and A2, we obtain

$$E[e_d] = N/2 \sum_{n=1}^{\min(L,d)} \sum_{m=1}^{d-n} P_n(m). \tag{1}$$

Theorem 1, which is stated and proved in the next section, gives the formula for $P_n(m)$. Substituting this result into (1) and using assumption A3, we can numerically solve for the minimum d such that $E[e_d] \geq V_\alpha$ since the results of [4] give V_α such that

$$P[\text{A random graph with } N \text{ nodes and } V_\alpha \text{ edges is connected}] = 1-\alpha.$$

Thus, given a value of α, we can determine a threshold value for d above which the MNCA is expected to cluster random data sets into a single cluster about $100(1-\alpha)\%$ of the time. Table I shows numerical values of d for various values of K, N, and α when $L = N-1$. Note that these values turn out to be insensitive to changes in K.

VI. MUTUAL NEAR NEIGHBORS IN A POISSON PROCESS

The following theorem states mutual near neighbor probabilities for Poisson point processes.

Theorem 1: Let X and Y be arbitrary points in a Poisson point process of intensity λ in K-dimensional space. Let $P_n(m)$ be the conditional probability that point X is the mth nearest neighbor of point Y given that point Y is the nth nearest neighbor of point X. Then

$$P_n(m) = \frac{(1-\beta_K)^{M-2}}{(2-\beta_K)^{M-1}} \sum_{i=0}^{m-1} \binom{n-1}{i} \binom{M-i-2}{m-i-1} ((1-\beta_K)^{-2}-1)^i$$

where $M = n+m$, and $\beta_K \equiv I_{\frac{3}{4}}[\frac{K+1}{2}, \frac{1}{2}]$ is a value of the incomplete beta ratio

$$I_x[p,q] = \int_0^x t^{p-1}(1-t)^{q-1}dt \Big/ \int_0^1 t^{p-1}(1-t)^{q-1}dt.$$

Let $S(X,r)$ denote a hypersphere of radius r centered on point X, let $\mu(.)$ denote K-dimensional Lebesgue measure, and let $S_K \equiv \mu(S(X,1))$ denote the volume of the unit hypersphere in K dimensions.

We proceed to find $P_n(m)$ by conditioning on the distance from X to Y. Let r_n be the random distance from a point to its nth nearest neighbor. From the definition of a Poisson process, the spherical volume $\mu(S(X,r_n))$ has a gamma distribution with parameters $1/\lambda$ and n. Therefore, the density function for r_n is

$$f_{r_n}(r) = \frac{K(\lambda S_K)^n}{(n-1)!} r^{Kn-1} \exp(-\lambda S_K r^K).$$

Let $P_n(m|r)$ be the conditional probability that X is the mth nearest neighbor of Y given that the distance from X to its nth Y is in the interval $(r, r+dr)$. By the definition of conditional probability

$$P_n(m) = \int_0^\infty P_n(m|r) f_{r_n}(r) dr. \tag{2}$$

Given that the distance from X to Y is in the interval $(r, r+dr)$, then X is the mth nearest neighbor of Y if and only if $|S(Y,r)| = m-1$, where $|A|$ is the number of points from the Poisson process in set A.

Let $B = S(X,r) \cap S(Y,r)$ and $C = S(Y,r) - B$. Since Y is the nth nearest neighbor of X

$$\begin{aligned} P_n(m|r) &= P[\ |S(Y,r)| = m-1\ ] \\ &= P[\ |B \cup C| = m-1\ ] \\ &= \sum_{i=0}^{m-1} P[\ |B| = i\ ]\ P[\ |C| = m-i-1\ ]. \end{aligned}$$

Since points are distributed according to a Poisson point process with intensity λ, $|C|$ has the Poisson distribution $p(i|\lambda\mu(C))$. Now, $|S(X,r)| = n-1$ and $B \subset S(X,r)$; therefore, $|B|$ has the binomial distribution $b(i|n-1, \mu(B)/\mu(S(X,r)))$. Therefore

$$P_n(m|r) = \sum_{i=0}^{m-1} b(i|n-1, \frac{\mu(B)}{\mu(S(X,r))})\ p(m-i-1|\lambda\mu(C)).$$

Substituting into this equation, the required p.d.f's and

$$\begin{aligned} M &\equiv m+n \\ S_K &= \frac{\pi^{\frac{K}{2}}}{\Gamma(\frac{K}{2}+1)} \\ \mu(B) &= \frac{(r\sqrt{\pi})^K}{\Gamma(\frac{K}{2}+1)}\ \beta_K \\ \frac{\mu(B)}{\mu(S(X,r))} &= \beta_K \\ \mu(C) &= S_K r^K (1-\beta_K) \\ \beta_K &\equiv I_{\frac{3}{4}}[\frac{K+1}{2}, \frac{1}{2}] \\ I_x[p,q] &\equiv \int_0^x t^{p-1}(1-t)^{q-1}dt \Big/ \int_0^1 t^{p-1}(1-t)^{q-1}dt \end{aligned}$$

yields

$$P_n(m|r) = \sum_{i=0}^{m-1} \binom{n-1}{i} \beta_K^i (1-\beta_K)^{M-2i-2} \frac{(\lambda S_K r^K)^{m-i-1}}{(m-i-1)!} \exp(-\lambda S_K r^K (1-\beta_K)).$$

TABLE II
THRESHOLD VALUE FOR d ABOVE WHICH $100(1-\alpha)$ PERCENT OF ALL 500 SIZE N RANDOM DATA SETS GENERATED IN A UNIT HYPERSQUARE IN K DIMENSIONS WERE CLUSTERED INTO ONE CLUSTER BY THE MNCA. (The numbers in parentheses show the 90% confidence interval for these quantiles.)

K	N	$(1-\alpha)$ 0.8	0.95	0.99
2	20	9 (9,10)	11 (11,11)	13 (13,16)
2	50	10 (10,10)	12 (12,13)	15 (14,17)
2	100	11 (10,11)	13 (12,13)	14 (14,19)
2	180	11 (11,11)	13 (13,13)	15 (15,16)
4	20	9 (8, 9)	10 (10,11)	13 (12,13)
4	50	10 (10,10)	13 (12,13)	15 (14,16)
4	100	11 (10,11)	13 (13,14)	17 (17,19)
4	180	11 (11,11)	14 (14,15)	18 (18,22)
8	20	8 (8, 9)	11 (10,11)	12 (12,14)
8	50	11 (11,11)	14 (14,15)	18 (17,22)
8	100	13 (13,14)	18 (17,19)	22 (20,28)
8	180	15 (15,16)	20 (19,22)	24 (23,29)

Substituting this into (2), interchanging the order of integration and summation, and noting that

$$\int_0^\infty r^{K(M-i-1)-1}\exp(-\lambda S_K r^K(2-\beta_K))dr = \frac{\Gamma(M-i-1)}{K(\lambda S_K(2-\beta_K))^{M-i-1}}$$

we obtain

$$P_n(m) = \sum_{i=0}^{m-1} \cdot \binom{n-1}{i}\frac{\beta_K^i(1-\beta_K)^{M-2i-2}}{(2-\beta_K)^{M-i-1}}\frac{(M-i-2)!}{(n-1)!(m-i-1)!} = \frac{(1-\beta_K)^{M-2}}{(2-\beta_K)^{M-1}}\sum_{i=0}^{m-1}\binom{n-1}{i}\binom{M-i-2}{m-i-1}((1-\beta_K)^{-2}-1)^i.$$

VII. EXPERIMENTS

Given the number and severity of assumptions made in deriving this note's main result, it is prudent to verify the usefulness of that result.

A. Monte Carlo Validation

Table II presents the order statistics on the mutual near-neighbor threshold d needed to connect a Monte Carlo sample of random data into one cluster using the MNCA. For each Monte Carlo sample, N points were generated uniformly in a K-dimension unit hypersquare, and the MNCA was applied. This process was repeated 500 times and the observed order statistics tabulated.

Comparing Table I with Table II shows that our theoretical analysis holds fairly well in the $K = 8$ case. For k=2 and 4, Table I significantly overestimates the true quantiles for $N \geq 100$ and thus presents conservative minimums. This implies that our assumption A3 is the most suspect, as increasing dimension appears to be required to reduce triangle inequality constraints. However, this could also be due to assumption A2 as our estimates might overshoot their true mark until edge effects compensate.

B. Example: Well-Separated Clusters

Fig. 1 gives an illustation of a sample 2-D data set. It contains two well-separated clusters, where each is composed of 90 points and distributed uniformly in a square. For this data, when $9 \leq d < 20$, G_d is composed of two connected components, where each represents one of the clusters. Table I implies that with significance between 0.95 and 0.99, these two clusters are not an artifact of applying MNCA to random data. Setting the MNCA threshold lower that 9 and, thus, having the MNCA break these two cluster further is not justified by our results.

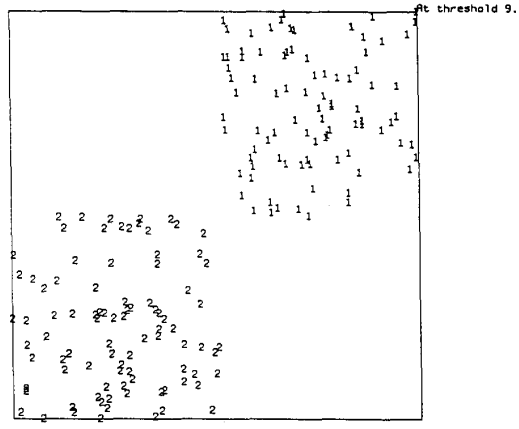

Fig. 1. Example 2-D data set with two well-separated clusters, each containing 90 uniformly distributed points

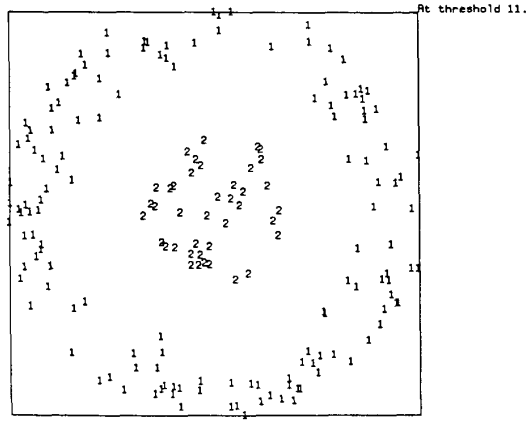

Fig. 2. Example 2-D data set with a unit radius circle containing 40 uniformly distributed points inside a annulus of unit width containing 140 uniformly distributed points. The inner radius of the annulus is 1.75 units, and the number of points is chosen so that the point density in each cluster is constant.

Other similarly distributed data sets, with a larger geometric separation between clusters than that shown in Fig. 1, have a larger d threshold needed to connect them. Our results show that it is justified to consider these data sets to be composed of the two *a priori* clusters.

C. Example: Concentric Clusters

Fig. 2 shows another type of data set that is particularly appropriate for the MNCA since it has the cluster topology of a circle inside an annulus. When $11 \leq d < 35$, G_d for this 180-point data set is composed of two connected components, where each represents one of the clusters. Table I implies that with significance greater than 0.99, these two clusters are not an artifact of applying MCNA to random data. Having the MNCA break these two clusters further by lowering d below 11 cannot be justified by our results.

Fifteen trials on data sets identically distributed to that in Fig. 2 shows that the threshold of d suggested by Table I is appropriate. The

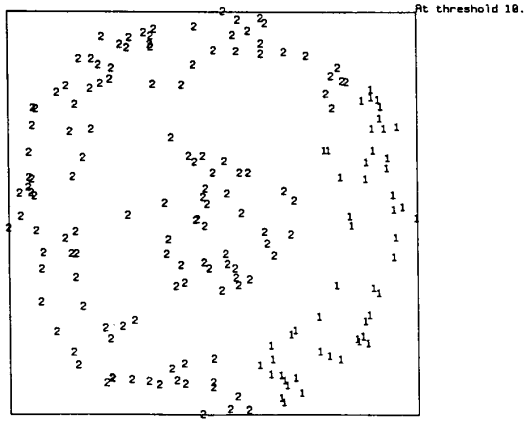

Fig. 3. Example 2-D data set similar to that of Fig. 2 but with the inner radius of the annulus of 1.4 units. Note the incorrect cluster labelings.

minimum value for the 15 trials of the smallest value of d at which G_d was connected was 24. For the 15 trials, the maximum value of d at which G_d was composed of more than two clusters was 14. In all cases, the two clusters found by the MNCA correspond to the two *a priori* clusters. Table I suggests that these two cluster partitions are unusual in random data and that further partitioning is not justfied.

Only when the distance between the inner radius of the annulus and the radius of the circle was reduced to less than 0.5 units did the value of d below which G_d was disconnected fall below the significance thresholds of Table I.

For instance, in 15 trails with 180 points and a separation distance of 0.4 units, G_d became disconnected in the range $10 \leq d \leq 19$. In eight of these cases, the disconnected G_d produced the correct *a priori* clustering; in the other seven cases, the MNCA would produce an incorrect clustering. If we set a threshold on d of 14 as suggested by the liberal reading of Table I, we reject six of the seven incorrect clusterings while rejecting only three of the eight correct clusterings. Fig. 3 shows the incorrect clustering on an example data set from this experiment when $10 \leq d < 11$.

D. Example: The 80X data

As a final test of the suggested threshold values of Table I, we examined the 80X data [2]. This data consists of 45 points in eight dimensions. Each point corresponds to eight features measured on a handprinted character. The data contains 15 examples of each the characters "8," "0," and "X." It is well known that the structure of this data is such that, except for a few outlier points, most of the "X" points separate out from a main mixture of the "8" and "0" points.

For the 80X data, G_d becomes disconnected at $d = 25$ into two components, where one consists of an outlier "0" point, and the second is composed of all the other points. Similarly, two "8"s separate from the other points at $d = 22$ and $d = 19$. G_d does not change between $12 < d < 19$. At $d = 12$, the large component of G_d is broken into two subcomponents, where one is composed of all 15 "X" points, and the other contains the remaining "8" and "0" points. Table I suggests that this clustering has some modest significance but that lowering the threshold any further is not justified. Thus, our results show that there is strong evidence of three outlier points in the 80X data and modest evidence of a separate cluster of "X" points.

VIII. Conclusions

This correspondence has presented a theoretical analysis of the threshold of the mutual neighborhood clustering algorithm under the hypothesis of random data. In order to derive an estimate of this threshold, a general theorem about the distribution of mutual near neighbors in a Poisson process was stated and proved.

Our analysis yielded a theoretical minimum value of the clustering threshold below which even unclustered data is broken into separate clusters. A simple Monte Carlo experiment validated the thresholds produced, and examples showing the use of these thresholds were given.

Clustering algorithms have the annoying habit of finding clusters in random data. This note provides a small step at alleviating this problem for the mutual neighborhood clustering algorithm.

References

[1] K. C. Gowda and G. Krishna, "Disaggregative clustering using the concept of mutual nearest neighborhood," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-8, no. 12, pp. 888–895, Dec. 1978.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[3] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," *IEEE Trans. Comput.*, vol. C-22, pp. 1025–1034, 1973.

[4] R. F. Ling and G. G. Killough, "Probability tables for cluster analysis based on a theory of random graphs," *J. Amer. Stat. Assoc.*, vol. 71, no. 354, pp. 293–300, June 1976.

[5] S. P. Smith and A. K. Jain, "Testing for uniformity in multidimensional data," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-6, no. 1, pp. 73–81, Jan. 1984.