AUG 15 2000

SANDIA REPORT

SAND2000-1766
Unlimited Release
Printed July 2000

RECEIVED
SEP 01 2000
OSTI

Some Provable Properties of VERI Clustering

G. C. Osbourn

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

RECEIVED
SEP 01 2000
OSTI

SAND2000-1766
Unlimited Release
Printed July 2000

Some Provable Properties of VERI Clustering

G. C. Osbourn
Laser, Optics, and Remote Sensing
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1423

Abstract

We present mathematical proofs for two useful properties of the clusters generated by the visual empirical region of influence (VERI) shape. The first proof shows that, for any d-dimensional vector set with more than one distinct vector, that there exists a bounded spherical volume about each vector v which contains all of the vectors that can VERI cluster with v, and that the radius of this d-dimensional volume scales linearly with the nearest neighbor distance to v. We then prove, using only each vector's nearest neighbor as an inhibitor, that there is a single upper bound on the number of VERI clusterings for each vector in any d-dimensional vector set, provided that there are no duplicate vectors. These proofs guarantee significant improvement in VERI algorithm runtimes over the brute force $O(N^3)$ implementation required for general d-dimensional region of influence implementations and indicate a method for improving approximate O(NlogN) VERI implementations. We also present a related region of influence shape called the VERI bow tie that has been recently used in certain swarm intelligence algorithms. We prove that the VERI bow tie produces connected graphs for arbitrary d-dimensional data sets (if the bow tie boundary line is *not* included in the region of influence). We then prove that the VERI bow tie also produces a bounded number of clusterings for each vector in any d-dimensional vector set, provided that there are no duplicate vectors (and the bow tie boundary line *is* included in the region of influence).

I Introduction

The visual empirical region of influence (VERI) is a specific version [1,2] of the more general region of influence (ROI) concept from graph theory [3]. ROIs are two-dimensional (2-D) shapes that are defined with respect to pairs of vectors. The size of the ROI for each pair of vectors is scaled with the separation of the points, and the orientation of the ROI is determined by alignment with the line segment that would connect the two points. ROIs determine pairwise clusterings through a seemingly simple exclusionary rule -- any pair of vectors in a data set are clustered together iff no third vector in the data set is inside the ROI defined by the pair. The ROI is applied to all pairs of vectors in the data set. Many simple mathematical ROI shapes have been explored in the literature, and the mathematical properties of the resulting graphs have been examined. The VERI ROI shape resulted from our novel hypothesis that human perceptual grouping of vectors might be modeled using the ROI concept. In this approach, the VERI shape was regarded as an empirical entity to be fitted to data on human cluster judgments. The VERI ROI shape was first discovered by Osbourn and Martinez through a set of psychophysical studies of human cluster perception [1,2]. That work has confirmed that VERI provides a reasonable model of visual cluster judgments obtained from a consensus of human subjects with normal visual perception.

A fast implementation of the VERI algorithm is essential for the many real world applications that it can address [2,4,5]. However, a naïve implementation of the general ROI graph calculation requires an $O(N^3)$ computation, where N is the number of vectors in the data set. This follows from the apparent need to examine every pair of points for potential clustering, and the apparent need to consider every remaining point in the data set as a potential inhibitor of each pair. We have previously described [1,6] an $O(N^2)$ cluster implementation and an $O(N_{train}*N_{test})$ VERI pattern recognition algorithm that exist if the total number of pairwise VERI clusterings scales as O(N) when only the first nearest neighbors of each vector are considered as inhibitors. However, no proof of this condition has been presented previously. Here we prove a stronger result (which implies this condition), namely that there is a single upper bound on the number of clusterings for each vector in any data set (with no identical vectors) when only the first nearest neighbors of the pairs of vector are considered as inhibitors. Thus, the runtime performances are now guaranteed to be $O(N^2)$ for the VERI cluster implementation and $O(N_{train}*N_{test})$ for the VERI pattern recognition algorithm for all data sets. We also present a proof that specifies an upper bound on the distance from any vector v that must be searched to find potential VERI clusterings with v. The upper bound is proportional to the distance to the nearest neighbor of v. This property allows constant factor runtime improvements in the VERI implementation by directly eliminating unnecessary consideration of many pairs of vectors that can not cluster together. We conclude our consideration of the VERI ROI with a discussion of the implications of these proofs for certain approximate O(NlogN) VERI implementations that have been proposed previously [6].

In the second half of this work we consider a ROI that has not been published previously. This "VERI bow tie" ROI is shown in Fig. 1, and has proved useful in unpublished work on swarm intelligent agent communication protocols. This ROI is made up of the union of two wedge-shaped regions. The radius of each wedge is equal in length to the line segment that would join the pair of vectors (for the rest of this proof we will call this line segment the pair line). Each wedge area is obtained by sweeping each radius in an arc, fixed at one of the pairs of vectors, with equal sweep areas on either side of the pair line. The VERI bow tie ROI can have a variable total angle of sweep (centered about the pair line) up to 120 degrees. Such bow tie ROIs contain vectors that are no farther from both of the pair vectors than the separation of the pair itself. We prove that the VERI bow tie produces connected graphs for arbitrary d-dimensional data sets (if the bow tie boundary line is *not* included in the region of influence). We then prove that the VERI bow tie also produces a bounded number of clusterings for each vector in any d-dimensional vector set, provided that there are no duplicate vectors (and the bow tie boundary line *is* included in the region of influence). These properties have implications for the robustness of the intelligent agent applications that we will discuss in a separate publication.

II Proof: For d-dimensional data sets with more than one unique vector, a finite spherical volume about each vector v contains all of the vectors that can VERI cluster with v, and the spherical radius scales linearly with the first nearest (nonidentical) neighbor distance to v

The key requirement for the proof is that the ROI completely enclose the two data vector positions in the ROI such that the two data vector positions are not on the ROI boundary. Any ROI that contains such regions, e.g. the VERI ROI, will satisfy this property. We assume, without loss of generality, that the ROI regions around each of the pair of points contain a circular area of nonzero radius r less than one in length (relative to a unity pair spacing) which we illustrate in Fig. 2. By definition, any pair of vectors contained by a distance R will not have a VERI grouping if a third vector exists which is within one of the circular areas. This condition occurs when the third vector has a distance to one of the vectors, call this RR, such that

$$RR <= r*R \qquad \text{Eq. 1.}$$

Let v be any vector in a d-dimensional data set S that contains two or more unique vectors. Consider the nearest (nonidentical) neighbor of v, which we call v1, for the data set S. Let the nonzero distance between v and v1 be R1. Now consider any vector v2 in S with distance R2 to vector v. Vector v1 will inhibit VERI clustering of v and v2 when Eq. 1 is satisfied, i.e. when

$$R2 => R1/r. \qquad \text{Eq. 2}$$

(obtained by dividing both sides of Eq. 1 by the positive quantity r). Thus, all VERI clusterings with arbitrary vector v must occur with vectors that are a distance less than R1/r (proportional to the first nearest nonidentical neighbor spacing R1).

III Discussion of proof consequences

This proof provides a mechanism for aborting the search for potential cluster partners in certain cases. In particular, suppose that one has obtained a list of the j nearest neighbors of each vector in preparation for executing the $O(N^2)$ VERI implementation [1,6]. If the first nearest neighbor distance D and the jth nearest neighbor distance Dj of a vector satisfy the equation

$$Dj \Rightarrow D/r, \qquad \text{Eq. 3}$$

then only the j-1 list of nearest neighbors of this vector are candidates for clustering, and there is no need to consider the entire data set when computing clusterings with this vector.

The proof also provides guidance for making an approximate O(NlogN) "divide and conquer" implementation of the VERI clustering [6] perform well. A divide and conquer technique can efficiently partition the entire d-dimensional data set into M rectangular boxes that contain at most a constant number of vectors n. The partitioning is an O(NlogN) computation. This approach allows M VERI computations to take place "independently" in the M boxes, and the computation times for each box take (constant) n^2 time. The results of this computation can be further improved by considering possible clusterings that occur across box boundaries. The proof provides the conditions under which potential clusterings from a vector in one box and vectors in certain distant boxes may be ignored.

IV Proof: A single bound on the number of VERI cluster neighbors exists for all vectors in an d-dimensional data set that contains no identical vectors when only the first nearest neighbors of the vector pairs are considered as inhibitors.

We define a VERI cluster neighbor of a vector v in a data set S as: any vector w in S that is not identical to v and is clustered by VERI to v. Now we prove that VERI produces a single upper bound on the number of cluster neighbors for all data vectors in any d-dimensional data set when only the nearest neighbors of all vectors are considered as inhibitors. The key requirement is again that the ROI completely enclose the two data point positions in the ROI such that the two data point positions are not on the ROI boundary. We again assume, without loss of generality, that the ROI regions around each of the pair of points contain a circular area of nonzero radius r (relative to a unity pair spacing). The value of r must also be less than one to be appropriate for the VERI ROI, so we also assume this condition.

We start by assuming that there is a vector v, in a data set S which contains no identical vectors, that exhibits VERI cluster neighbors that increase in number without bound as new vectors are added to the set S. We then show that this leads to a contradiction, so that the number of VERI cluster neighbors for each vector is bounded. The proof concludes with the construction of a single upper bound for the number of cluster neighbors that depends only on the dimensionality of the data set and the size of r.

Let Sv be the set of cluster neighbors of v in S. For any fixed number N of vectors in set S that is large enough so that the set Sv is nonempty, we must be able to add arbitrarily many new vectors w to S, each of which: is also a cluster neighbor of v, i.e. is also in Sv; does not eliminate any of the other vectors in Sv by acting both as a nearest neighbor and as an inhibitor, i.e. the increase in the number of vectors in Sv by adding w is not offset by eliminating any of the existing cluster neighbors of v.

Consider the nearest neighbor of v, which we call v1, for the data set S of size N. Let the nonzero distance between v and v1 be R1. As proved above, all groupings to v must occur with vectors inside the boundary of radius R2 given by

$$R2 = R1/r. \qquad \text{Eq. 4}$$

Thus, any additional vector w that is added to S must be within the finite volume of the d-dimensional sphere of radius R2 around vector v.

Similarly, a new vector w will inhibit *all* of the existing VERI clusterings if it is closer to v than distance R0 given by

$$R1 = R0/r. \qquad \text{Eq. 5}$$

Since r is less than unity, such a new vector would be closer to v than the previous nearest neighbor and would become the new (inhibitor) nearest neighbor of v. R1 is the previous nearest neighbor distance to v, so that all prior members of Sv must be at least that far from v. Thus, any additional vector w that is added to Sv must be more than a fixed, lower bound distance R0 away from v to avoid eliminating existing cluster neighbors from Sv.

Finally, note that any new cluster neighbor w, by definition, is surrounded by a volume that contains no nearest neighbor inhibitor to prevent the clustering of w with v. Since all such w must be at least distance R0 away from v, there must be a vector-free volume around each w that has a radius Rw which satisfies

$$Rw > r*R0. \qquad \text{Eq. 6a}$$

or

$$Rw > r^2 * R1. \qquad \text{Eq. 6b}$$

We have now established that each additional cluster neighbor w must simultaneously satisfy two properties:
(1) it must be located in a finite volume spherical d-dimensional shell centered around v.
(2) it must be located at the center of a spherical d-dimensional volume, with radius greater than a fixed, nonzero lower bound, that contains no other vectors.

Property 2 guarantees that the volume occupied by the vectors in Sv increases without bound as the number of such vectors increases, which contradicts property 1. Thus, every vector has a bounded number of cluster neighbors.

Properties 1 and 2 can be used to construct a *single* upper bound that applies to any vector. The total spherical volume available to cluster neighbors around any vector is given by

$$C1*(R1/r)^d, \qquad \text{Eq. 7}$$

where d is the dimensionality of the data set, C1 is a constant that depends only on d and R1 is the distance from the vector to the first nearest neighbor. Each cluster neighbor must occupy the center of a spherical volume

$$C1*(R1*r^2)^d \qquad \text{Eq. 8}$$

that contains no other vectors. The upper bound on the number of cluster neighbors is determined by how many of the d-dimensional spherical volumes given in Eq. 8 can be packed into the volume given in Eq. 7 such that the center of each spherical volume does not occupy any other spherical volume. Alternatively, the bound can be determined from the packing of spheres with a radius reduced by half (i.e. equal to $R1*r^2/2$) such that the sphere volumes do not overlap. The packing value is proportional to the ratio of the volumes in Eqs. 7 and 8. Thus, the number of cluster neighbors m for any vector must satisfy

$$m < C2/r^{3d}, \qquad \text{Eq. 9}$$

where C2 is a constant (associated with the packing efficiency of smaller d-dimensional spheres into a larger d-dimensional sphere) that depends only on the dimensionality d. This single upper bound depends only on the dimensionality of the data set and the r value of the ROI, and thus holds for all vectors.

V Proof: (Edgeless) VERI bow tie ROI produces a single connected graph for any d-dimensional data set

The "edgeless" VERI bow tie ROI does not include the boundary line of the shape, i.e. it does not inhibit pairwise groupings if there are vectors exactly on the ROI boundary line.

We assume that there is a d-dimensional data set A that yields more than one connected subgraph via the edgeless bow tie ROI, and show that this leads to a contradiction. For each pair of distinct connected subgraphs produced for data set A, there exists (at least) one "interpair" of points (i.e. one of the two points is in each of the two subgraphs) such that the distance between this point pair is a greatest lower bound on the distances between all such interpair distances of this pair of subgraphs. Call this the closest point interpair for the pair of subgraphs. There further exists (at least) one of these closest point interpairs such that the distance between this pair of points is a greatest lower bound on the set of distances of the closest point interpairs for all pairwise combinations of distinct subgraphs. Call the points of this particular pair p1 and p2, and call the subgraphs that contain these points A1 and A2, respectively. Since p1 and p2, by definition, come from distinct connected graphs, there cannot be a grouping between p1 and p2. Now consider the edgeless circular bow tie applied to p1 and p2. For the grouping to be inhibited, there must be another point p3 that is in the ROI area. By the definition of the edgeless VERI bow tie shape (with sweep angles limited to less than 120 degrees), this point p3 is closer to both p1 and p2 than the distance between p1 and p2. Point p3 belongs to one of the distinct connected subgraphs of A. Assume p3 belongs to the subgraph A1. Then the subgraphs A1 and A2 have an interpair of points (p2,p3) with distance less than the (p1,p2) distance. But this contradicts (p1,p2) being the closest interpair for A1 and A2. Similarly, assuming the p3 belongs to subgraph A2 leads to a contradiction that (p1,p2) is the closest interpair for A1, A2. The only possibility remaining is that p3 belongs to another distinct subgraph, call this A3, that is different from A1 and A2. Subgraph A1 and A3 have an interpair (p1,p3) with smaller distance than the (p1,p2) distance. This means that the closest interpair distance for A1 and A3 must also be less than the closest interpair distance for A1 and A2. This contradicts (p1,p2) being the greatest lower bound on all distinct subgraph closest interpair distances. Thus our original assumption leads to contradiction, so there does not exist an N-D data set that yields more than one connected subgraph via the edgeless VERI bow tie.

We note that this property for planar graphs (i.e. 2-D data) can also be obtained by recognizing that the edgeless VERI bow tie is contained within the well-known LUNE ROI[3]. The LUNE ROI has been proven to produce connected planar graphs.

VI Proof: An upper bound on the number of VERI bow tie cluster neighbors exists for all vectors in an d-dimensional data set that contains no identical vectors

We define a VERI bow tie cluster neighbor of a vector v in a data set S as: any vector w in S that is not identical to v and is clustered by the VERI bow tie ROI to v. For this proof we consider the boundary line of the bow tie to be included in the ROI. We prove that the VERI bow tie produces an upper bound on the number of cluster neighbors for all data vectors in any d-dimensional data set (without identical vectors).

This proof is conceptually similar to the proof in Sec. IV. Solid angles in d-dimensional space play the same role here that spherical volumes played in the previous proof. The present proof can be extended in a straightforward manner to show that a single upper bound exists for all vectors. We start the proof by assuming that there is a vector v, in a data set S which contains no identical vectors, that exhibits VERI bow tie cluster neighbors that increase in number without bound as new vectors are added to the set S. We then show that this leads to a contradiction, so that the number of VERI bow tie cluster neighbors for each vector is bounded.

Let A be the nonzero sweep angle of the VERI bow tie ROI. Let Sv be the set of bow tie cluster neighbors of v in S. For any fixed number N of vectors in set S that is large enough so that the set Sv is nonempty, we must be able to add arbitrarily many new vectors w to S, each of which: is also a bow tie cluster neighbor of v, i.e. is also in Sv; does not eliminate any of the other vectors in Sv by acting as an inhibitor, i.e. the increase in the number of vectors in Sv by adding w is not offset by eliminating any of the existing cluster neighbors of v.

Consider the nearest cluster neighbor of v, which we call v1, for the data set S of size N. Let the nonzero distance between v and v1 be R1. Any additional vector w that is added to S and Sv cannot occupy a d-dimensional cone of sweep angle A that radiates from v and is centered about the pair line associated with v and v1. This is true because the vector w will either be inhibited from clustering with v (if the distance from w to v is greater than R1) or it will inhibit the clustering of v and v1 (if the distance from w to v is less than or equal to R1). Each additional member of Sv that does not eliminate existing members of Sv must be placed in an infinite cone of nonzero solid angle that contains no other vectors. Thus, the d-dimensional solid angle around v required for new members of Sv that don't eliminate existing members must increase without bound. But this is impossible for a vector space of finite dimensionality, since such spaces contain only a finite solid angle around each vector in the space.

Acknowledgements

This work was sponsored by the U.S. Department of Energy under Contract DE-AC04-94AL85000. Sandia is a multiprogram laboratory operated by Sandia Corp., a Lockheed Martin Co., for the U.S Department of Energy.

References

1. G. C. Osbourn and R. F. Martinez, EMPIRICALLY DEFINED REGIONS OF INFLUENCE FOR CLUSTERING ANALYSIS, Pattern Recognition, vol. 28, no. 11, 1793-1806 (1995).

2. G. C. Osbourn, J. W. Bartholomew, R. F. Martinez and A. M. Bouchard, VERI: A USER'S GUIDE,
on the website http://www.sandia.gov/imrl/XVisionScience/X1155home.htm.
3. G. T. Toussaint, THE RELATIVE NEIGHBORHOOD GRAPH OF A FINITE PLANAR SET, Pattern Recognition, vol. 12, 261-268 (1980).
4. G. C. Osbourn, J. W. Bartholomew, R. F. Martinez and G. C. Frye, VISUAL-EMPIRICAL REGION-OF-INFLUENCE PATTERN RECOGNITION APPLIED TO CHEMICAL MICROSENSOR ARRAY SELECTION AND CHEMICAL ANALYSIS, Accounts of Chemical Research, vol. 31, no. 5, 297-305 (1998).
5. K. M. Horn, B. S. Swartzentruber, G. C. Osbourn, A. Bouchard and J. W. Bartholomew, ELECTRONIC STRUCTURE CLASSIFICATIONS USING SCANNING TUNNELING MICROSCOPY CONDUCTANCE IMAGING, Journal of Applied Physics, vol. 84, no. 5, 2487-2496 (1998).
6. R. F. Martinez and G. C. Osbourn, EXTENDING APPLICABILITY OF CLUSTER BASED PATTERN RECOGNITION WITH EFFICIENT APPROXIMATION TECHNIQUES, SAND97-0589 (1997).

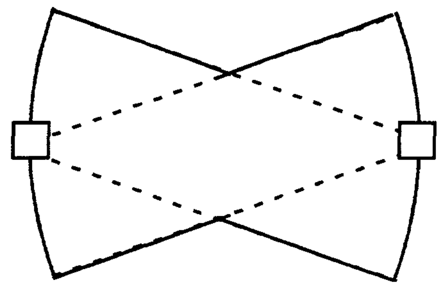

Fig. 1 VERI bow tie region of influence. The squares indicate the positions of the pair of vectors that are tested for clustering. The dashed lines indicate the wedge-shaped regions discussed in the text.

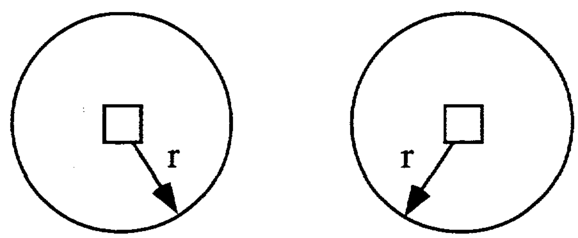

Fig. 2 The circular subset of the VERI region of influence used in the proofs. The squares indicate the positions of the pair of vectors that are tested for grouping. The radii r are identical and are less than one in length, where the separation of the squares is taken to be unity spacing.

DISTRIBTUTION:

1	MS0188	LDRD Office, 4001
10	MS1423	G. C. Osbourn, 1118
1	MS0612	Review and Approval Desk, 9612
2	MS0899	Technical Library, 9616
1	MS9018	Central Technical Files, 8940-2