Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume IV
Pro–Z

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Production:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Typesetter:	Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Copyright © 2009 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

If a library purchased a print copy of this publication, please go to http://www.igi-global.com/agreement for information on activating the library's complimentary electronic access to this publication.

Proximity-Graph-Based Tools for DNA Clustering

Imad Khoury
School of Computer Science, McGill University, Canada

Godfried Toussaint
School of Computer Science, McGill University, Canada

Antonio Ciampi
Epidemiology & Biostatistics, McGill University, Canada

Isadora Antoniano
IIMAS-UNAM, Ciudad de Mexico, Mexico

Carl Murie
McGill University, Canada & McGill University and Genome Quebec Innovation Centre, Canada

Robert Nadon
McGill University, Canada & McGill University and Genome Quebec Innovation Centre, Canada

INTRODUCTION

Clustering is considered the most important aspect of unsupervised learning in data mining. It deals with finding *structure* in a collection of unlabeled data. One simple way of defining clustering is as follows: the process of organizing data elements into groups, called clusters, whose members are similar to each other in some way. Several algorithms for clustering exist (Gan, Ma, & Wu, 2007); proximity-graph-based ones, which are untraditional from the point of view of statisticians, emanate from the field of computational geometry and are powerful and often elegant (Bhattacharya, Mukherjee, & Toussaint, 2005). A proximity graph is a graph formed from a collection of elements, or points, by connecting with an edge those pairs of points that satisfy a particular neighbor relationship with each other. One key aspect of proximity-graph-based clustering techniques is that they may allow for an easy and clear visualization of data clusters, given their geometric nature. Proximity graphs have been shown to improve typical instance-based learning algorithms such as the *k*-nearest neighbor classifiers in the typical nonparametric approach to classification (Bhattacharya, Mukherjee, & Toussaint, 2005). Furthermore, the most powerful and robust methods for clustering turn out to be those based on proximity graphs (Koren, North, & Volinsky, 2006). Many examples have been shown where proximity-graph-based methods perform very well when traditional methods fail miserably (Zahn, 1971; Choo, Jiamthapthaksin, Chen, Celepcikay, Giusti, & Eick, 2007)

The most well-known proximity graphs are the nearest neighbor graph (*NNG*), the minimum spanning tree (*MST*), the relative neighborhood graph (*RNG*), the Urquhart graph (*UG*), the Gabriel graph (*GG*), and the Delaunay triangulation (*DT*) (Jaromczyk, & Toussaint, 1992). The specific order in which they are introduced is an inclusion order, i.e., the first graph is a subgraph of the second one, the second graph is a subgraph of the third and so on. The *NNG* is formed by joining each point by an edge to its nearest neighbor. The *MST* is formed by finding the minimum-length tree that connects all the points. The *RNG* was initially proposed as a tool for extracting the shape of a planar pattern (Jaromczyk, & Toussaint, 1992), and is formed by connecting an edge between all pairs of distinct points if and only if they are relative neighbors. Two points A and B are relative neighbors if for any other point C, the maximum of *d*(A, C), *d*(B, C) is greater than *d*(A, B), where *d* denotes the distance measure. A triangulation of a set of points is a planar graph

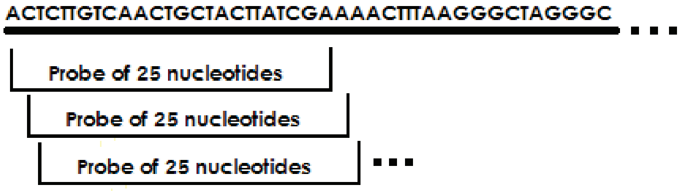
Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

connecting all the points such that all of its faces, except for the outside face, are triangles. The *DT* is a special kind of triangulation where the triangles are as "fat" as possible, i.e., the circumcircle of any triangle does not contain any other point in its interior. The *UG* is obtained by removing the longest edge from each triangle in the *DT*. Finally, the *GG* is formed by connecting an edge between all pairs of distinct points if and only if they are Gabriel neighbors. Two points are Gabriel neighbors if the hyper-sphere that has them as a diameter is empty, i.e., if it does not contain any other point in its interior. Clustering using proximity graphs consists of first building a proximity graph from the data points. Then, edges that are deemed long are removed, according to a certain edge-cutting criterion. Clusters then correspond to the connected components of the resulting graph. One edge-cutting criterion that preserves Gestalt principles of perception was proposed in the context of *MST*s by C. T. Zahn (Zahn, 1971), and consists in breaking those edges *e* that are at least say, twice as long as the average length of the edges incident to the endpoints of *e*. It has been shown that using the *GG* for clustering, or as part of a clustering algorithm, yields the best performance, and is adaptive to the points, in the sense that no manual tweaking of any particular parameters is required when clustering point sets of different spatial distribution and size (Bhattacharya, Mukherjee, & Toussaint, 2005).

The applications of proximity-graph-based clustering, and of clustering in general, are numerous and varied. Possibilities include applications in the fields of marketing, for identifying groups of customers with similar behaviours; image processing, for identifying groups of pixels with similar colors or that form certain patterns; biology, for the classification of plants or animals given their features; and the World Wide Web, for classifying Web pages and finding groups of similar user access patterns (Dong, & Zhuang, 2004). In bioinformatics, scientists are interested in the problem of DNA microarray analysis (Schena, 2003), where clustering is useful as well. Microarrays are ordered sets of DNA fragments fixed to solid surfaces. Their analysis, using other complementary fragments called probes, allows the study of gene expression. Probes that bind to DNA fragments emit fluorescent light, with an intensity that is positively correlated, in some way, to the concentration of the probes. In this type of analysis, the calibration problem is of crucial importance. Using an experimental data set, in which both concentration and intensity are known for a number of different probes, one seeks to learn, in a supervised way, a simple relationship between intensity and concentration so that in future experiments, in which concentration is unknown, one can infer it from intensity. In an appropriate scale, it is reasonable to assume a linear relationship between intensity and concentration. However, some features of the probes can also be expected to have an effect on the calibration equation; this effect may well be non-linear. Arguably, one may reason that if there is a natural clustering of the probes, it would be desirable to fit a distinct calibration equation for each cluster, in the hope that this would be sufficient to take into account the impact of the probes on calibration. This hope justifies a systematic application of unsupervised learning techniques to features of the probes in order to discover, such a clustering, if it exists.

The main concern remains whether one is able to discover the absence or presence of any real clustering of the probes. Traditionally, clustering of microarray probes has been based on standard statistical approaches, which were used to validate an empirically found clustering structure; however, they were usually complex and depended on specific assumptions (Johnson, & Wichern, 2007). An alternative approach

Figure 1.Probes of 25 nucleotides to be clustered. Shown is a gene sequence and a probe window sliding by one nucleotide.


ACTCTTGTCAACTGCTACTTATCGAAAACTTTAAGGGCTAGGGC ...
Probe of 25 nucleotides
Probe of 25 nucleotides
Probe of 25 nucleotides ...

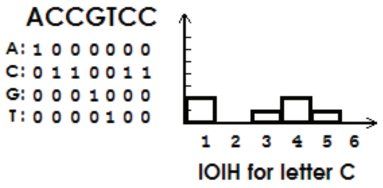
based on proximity graphs could be used, which has the advantage of being relatively simple and of providing a clear visualization of the data, from which one can directly determine whether or not the data support the existence of clusters.

BACKGROUND

A probe is a sequence of a fixed number of nucleotides, say 25 (Fig. 1), and it could simply be regarded as a string of 25 symbols from the alphabet {A, C, G, T}.

Probes are generated by sliding over, by one, a window of size 25 on the same nucleotide sequence. This procedure is not typical but particular to the Mei et. al dataset (Mei, Hubbell, Bekiranov, Mittmann, Christians, Shen, Lu, Fang, Liu, Ryder, Kaplan, Kulp, & Webster, 2003). In order to cluster the probes, one can either resort to classical clustering techniques, or to proximity-graph-based ones, or both. Moreover, for each of these sets of techniques, two approaches can be considered: the feature-based approach and the sequence-based approach. The first one builds the dataset from characteristics extracted from the probes, while the second one deals with the probe sequence directly. In the feature-based approach, classical probe features can be computed such as the frequency of different nucleotide types in the whole probe, and in either of its halves (Johnson, & Wichern, 2007). But current interdisciplinary work on the research theme reveals novel and creative probe-feature-extraction methods. One may, for instance, combine tools from computational music theory and bioinformatics, by considering features inspired from the rhythmic analysis proposed by Gouyon et. al. (Gouyon, Dixon, Pampalk, & Widmer, 2004) and extracting them from microarray probes. In this approach, for each probe, an inter-onset interval histogram (*IOIH*) (Toussaint, 2004; Toussaint, 2005) is built for each letter of the alphabet, and 8 features are computed for each histogram, namely the mean, the geometric mean, the total energy, the centroid, the flatness, the kurtosis, the high-frequency content, and the skewness (Gouyon, Dixon, Pampalk, & Widmer, 2004). The *IOIH* is a histogram that summarizes how many intervals there exist in a binary string of a fixed length, where 'interval' denotes the distance between two not necessarily successive set bits (or '1's). In the left-hand side of Fig. 2, four binary strings are first generated from the DNA sequence, one for each letter, by writing a '1' where that letter occurs and a '0' where it does not. Only the *IOIH* for letter C is shown, in the right-hand side of Fig. 2.

Figure 2. Example of inter-onset interval histogram. We see that there are two inter-onset intervals of length 1, one inter-onset interval of length 3, two inter-onset intervals of length 4, one inter-onset interval of length 5 and no inter-onset intervals of lengths 2, 6.



Then, 6 inter-histogram distances are computed using the Kolmogorov distance. Hence, in total, 32+6 = 38 features are extracted for each probe. Six classical distances are used for defining the dissimilarity of the points in the feature space. They are the standardized and unstandardized Manhattan distance, the standardized and unstandardized Euclidean distance, the Mahalanobis distance and the Mahalanobis-Manhattan distance (Johnson, & Wichern, 2007). The Manhattan distance is the distance between two points measured along axes at right angles. The Mahalanobis distance is effectively a weighted Euclidean distance where the weighting is determined by the sample correlation matrix of the point set.

The second approach is based entirely on the sequence of the symbols in the probe, and aims at producing a distance matrix that summarizes the distances between all pairs of probes, and which serves as input for a clustering algorithm. Sequence-based distances with normalization variations can be derived. These include the nearest neighbour distance, the edit distance (Levenshtein, 1966) and the directed swap distance (Colannino, & Toussaint, 2005). The nearest neighbor distance measures the dissimilarity of two binary strings via the concept of nearest set bit neighbor mapping. A first pass is done over the first string, say from left to right, and set bits are mapped to their closest set-bit neighbors in the second string, in terms of character count. The same kind of mapping is done in a second pass over the second string. The nearest neighbor distance is the accumulated distances of each mapping link without counting double links twice. The edit distance is a metric that measures the dissimilarity of two strings by counting the minimum number of editing operations needed to transform one string into another. The editing operations typically considered are 'replace', 'insert' and 'delete'. A cost can be associated with each operation, hence penalizing the use of one operation over another. Finally, the directed swap distance measures the dissimilarity of two binary strings via the concept of an assignment. It is equal to the cost of the minimum-cost assignment between the set bits of the first string and the set bits of the second string, where cost is taken to be the minimum number of swaps between adjacent bits needed to displace a set bit in the first string to the position of a set bit in the second string, thereby making one assignment.

Classical clustering techniques include the k-medoids clustering using partitioning around medoids (*PAM*) (Handl, & Knowles, 2005), the hierarchical clustering using single linkage, and classical multidimensional scaling (*CMDS*). *PAM* is an algorithm that clusters around medoids. A medoid is the data point which is the most centrally located in a point set. The sum of the distances of this point to all other points in a point set is less than or equal to the sum of the distances of any other point to all other points in the point set. *PAM* finds the optimal solution. It tends to find 'round' or 'spherical' clusters, and hence is not very efficient when the clusters are in reality elongated, or in line patterns. Hierarchical clustering is a traditional clustering technique that clusters in an agglomerative approach. First, each point is assigned to its own cluster and then, iteratively, the two most similar clusters are joined, until there is only one cluster left. If the 'single linkage' version is used, it becomes equivalent to clustering with minimum spanning trees. Other options include 'complete linkage' and 'average linkage'. Finally, *CMDS* is an algorithm that takes as input a distance matrix and returns a set of points such that the Euclidean distances between them approximate the corresponding values in the distance matrix. This method, in the context of clustering, allows one to try different dimensions in which clustering can be performed. On the other hand, it is a technique to reduce the dimensionality of the data to one that can be easily visualized.

As to proximity-graph-based clustering techniques, they are: *ISOMAP* (Tenenbaum, de Silva, Langford, 2000), and clustering using Gabriel graphs with the Zahn edge-cutting criterion. *ISOMAP* is a proximity-graph based algorithm similar in its goal to *CMDS*, but with the flexibility of being able to learn a broad class of nonlinear manifolds. A manifold is an abstract space in which every point has a neighborhood which resembles the Euclidean space, but in which the global structure may be more complicated. The idea of dimension is important in manifolds. For instance, lines are one-dimensional, and planes two-dimensional. *ISOMAP* is computationally efficient and ensures global optimality and asymptotic convergence. It tries to conserve the geodesic distances between the points, and for that it constructs the *k*-nearest neighbor graph. This set of techniques can help to find and visualize the presence or absence of any real clustering in the data.

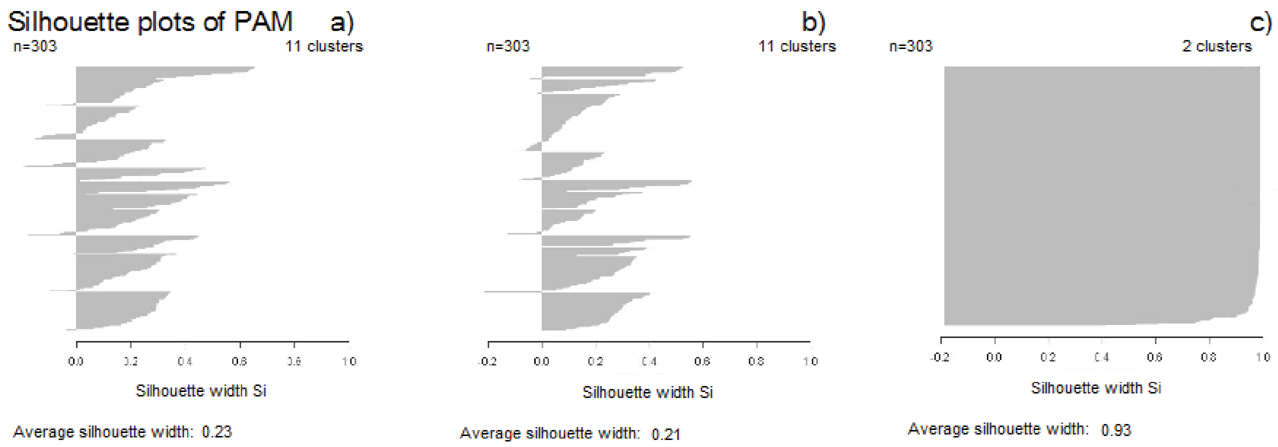
MAIN FOCUS

The rhythmic-analysis-based feature extraction applied to microarray probes, and the nearest neighbor distance for measuring probe dissimilarity are, in themselves, a novelty in the current research on microarray data mining. What is further emphasized here is the comparison between traditional clustering techniques and novel proximity-graph-based clustering techniques, the latter which give a simple and clear way of visualizing the clustering and, for our DNA microarray clustering example, it shows that the data lacks any real clustering.

Dataset

The dataset consists of a collection of genes, with sliding probes (Fig. 1) for each gene. The HTC (Human Test Chip) data set is used. This is a proprietary data set that Affymetrix used in its probe selection algorithm study (Mei, Hubbell, Bekiranov, Mittmann, Christians, Shen, Lu, Fang, Liu, Ryder, Kaplan, Kulp, & Webster, 2003) and was accessed through Simon Cawley and Teresa Webster (Affymetrix). A set of 84 transcripts (77 human and 7 bacterial) were used for probe selection analysis. Each human gene has approximately 500 probes made of 25 nucleotides each. Let us consider one representative gene from which 303 probes have been generated. The dataset is therefore the collection of the 303 probes.

Figure 3. PAM silhouettes using: a) Standardized Manhattan distance, b) Standardized Eucidean distance, c) Mahalanobis distance. A silhouette plot is a plot that displays a measure of how close each point in one cluster is to points in the neighboring clusters. No good clustering of the probes is found, as the average silhouette widths are either small (a and b), or they are high but with only 2 disproportionate clusters found (c).



Methods

PAM and hierarchical clustering using single linkage are used as a baseline to which proximity-graph-based clustering methods are compared.

Clustering of Microarray Probes using Classical Techniques

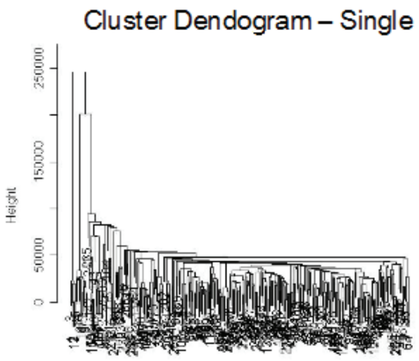
First, *PAM* is applied using the feature-based approach. No good clustering is obtained. Fig. 3 shows a silhouette plot for each of the three distance measures used. A silhouette plot is a plot that displays a measure of how close each point in one cluster is to points in the neighboring clusters. It takes into account both cohesion and separation. A detailed definition is given by (Choo, Jiamthapthaksin, Chen, Celepcikay, Giusti, & Eick, 2007). It is therefore an indicator of the 'goodness' of the clustering. The plots show that clusters do not look well defined; they hence give a first indication that there is no cluster structure in the data.

Next, hierarchical clustering with single linkage, using the feature-based approach, is applied. Fig. 4 shows one representative dendrogram. A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by the hierarchical clustering algorithm. The leafs represent the data points, children of the same parent are points in the same cluster, and the edge lengths correspond to the distances between clusters. The dendrogram in Fig. 4 shows, again, that no clustering structure seems to be present in the data, this time with long clusters, just as with spherical ones.

Now, let us apply *PAM* again, but with the sequence-based approach. The total average silhouette width was chosen as an indicator of the 'goodness' of the clustering using *PAM*. With the non-normalized edit distance, a clustering with 61 clusters yielded the best average width. With the nearest neighbor distance, a clustering with 5 clusters yielded the best average width. In both cases, however, the corresponding silhouette plots showed no real clustering. The dendrograms output by the hierarchical clustering with single linkage algorithm also showed no clustering for both the edit distance and the nearest neighbor distance.

The last method we can apply in classical clustering is *CMDS*. As previously defined, *CMDS* can reduce the dimensionality of the data to one that can be visualized. However, clustering would have to be visually performed, which is often difficult and inaccurate. Fig. 5 shows the probe set in a three dimensional space after *CMDS* is applied.

Figure 4. Representative dendrogram for single-linkage clustering, using the unstandardized Manhattan distance. No underlying good clustering is apparent.



Clustering of Microarray Probes using Proximity Graphs

Now, proximity-graph-based clustering techniques are applied on the same dataset and using the same two approaches. In the feature-based approach, clustering using the Gabriel graph – Zahn's criterion gives rise to the plots in Fig. 6. The result, for each distance measure, consists of one very large cluster, containing most of the probes and one or more smaller clusters containing too few probes to be considered. Therefore, we can consider this as more evidence to the theory that no natural clustering structure is present in the data. This time, edges allow an easy visualization of the clustering.

For the sequence-based approach, the *ISOMAP* algorithm is first used to embed the distance matrix in the best Euclidean space, taking into account the possibility that the data set is in reality manifold-shaped. *ISOMAP* found the 3D space to be the best space in which to embed our probes. This makes it possible to plot the points in 3D and to apply clustering using the Gabriel graph – Zahn's criterion algorithm as shown next in Fig. 7. If a higher dimensional space was found by *ISOMAP*, clustering using Gabriel graphs would still be possible, and visualizing it would require additional projection methods. Again, no good clustering is found, as one cluster turns to have most of the points, and the others only few. This time, the absence of real clustering in the data is confirmed.

FUTURE TRENDS

Proximity graphs will continue to play a crucial role in applications such as the calibration problem in microarray analysis, as well as in other clustering applications, especially in those applications where clustering is a preprocessing step for a problem where the need to discover the presence or absence of any real clustering in the data, and where the visualization of the data points and clusters plays a determining role to this end. The current most efficient manifold-learning algorithms are based on nearest neighbor types of graphs. Further research on incorporating Gabriel graphs in manifold learning algorithms, such as *ISOMAP,* should be considered, since Gabriel graphs have consistently been proven to be powerful and helpful tools in designing unsupervised learning and supervised instance-based learning algorithms. Moreover, answers to open questions that arise when

Figure 5. Classical mutidimensional scaling with: a) Edit distance (non-normalized), b) Nearest neighbor distance. No clustering is noticeable.

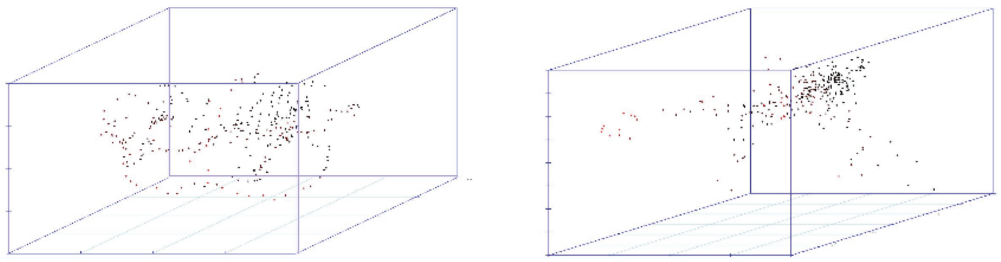


Figure 6.Gabriel graph with Zahn's edge cutting criterion: a) Standardized Manhattan distance, b) Standardized Euclidean distance, c) Mahalanobis distance. No real clustering of the probes is noticeable.

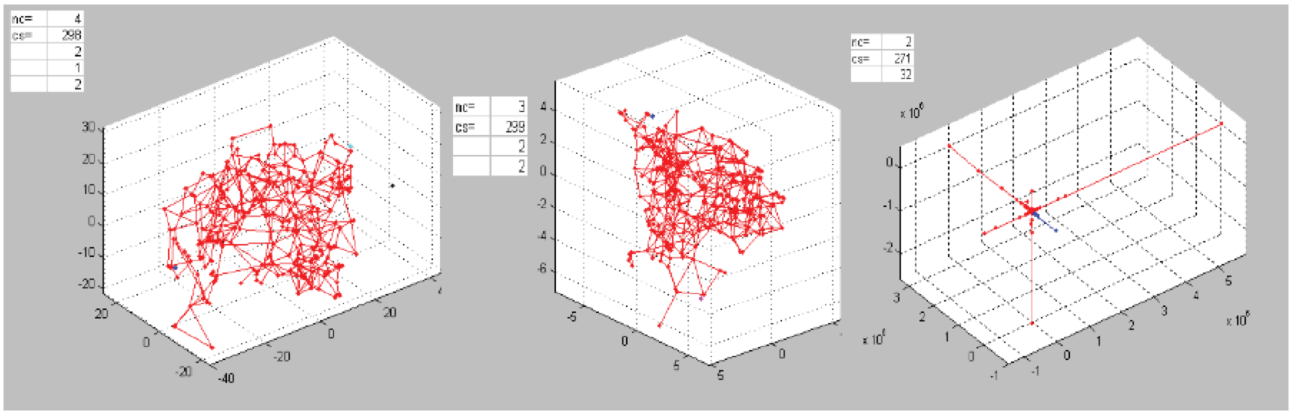
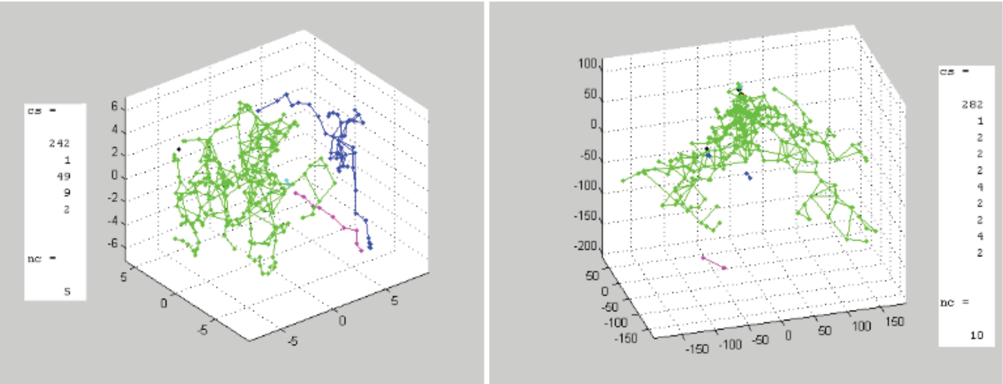


Figure 7. Gabriel Graph with Zahn edge cutting criterion: a) Edit distance, b) Nearest neighbor distance. No real clustering of the probes is noticeable.



applying Gabriel-graph-based clustering on a fixed-size set of points as the dimension gets higher, will have to be investigated. In fact, in a situation like this, Gabriel graphs can have edges between all pairs of points (Chazelle, Edelsbrunner, Guibas, Hershberger, Seidel, & Sharir, 1990; Jaromczyk, & Toussaint, 1992); a situation that may also be viewed from the standpoint of supervised learning, when features of a dataset, which correspond to dimensions, are so numerous, that a comparatively smaller number of data points is not enough to learn them, essentially leading to what is termed the curse of dimensionality. Which subgraph of the Gabriel graph would give the best result in this case, or how sparse the graph should be, will then be interesting future research problems to be solved in this field.

CONCLUSION

Proximity-graph-based clustering can be a very helpful preprocessing tool for the calibration problem in microarray analysis. Both classical and proximity-graph-based clustering methods can be used to cluster microarray probes. However, classical methods do not provide a simple, elegant, and clear way of visualizing the clustering, if it exists. Furthermore, unlike some proximity-graph-based algorithms, they almost always fail to detect any clusters of structurally complex higher level shapes, such as a manifold. Proximity-graph-based clustering methods can hence be efficient and powerful alternate or complementary tools for traditional unsupervised learning. These methods can also play a useful role in visualizing the absence (or presence) of any real clustering in the data that may have been found using classical clustering methods. Moreover, in this context, novel interdisciplinary probe-feature-extraction methods are being considered, and a sequence-based approach that defines novel distance measures between probes is currently under investigation.

REFERENCES

Bhattacharya, B., Mukherjee, K., & Toussaint, G. T. (2005). Geometric decision rules for high dimensions. *In Proceedings of the 55th Session of the International Statistical Institute.* Sydney, Australia.

Chazelle, B., Edelsbrunner, H., Guibas, L. J., Hershberger, J. E., Seidel, R., & Sharir, M. (1990).

Slimming down by adding; selecting heavily covered points. *Proceedings of the sixth annual symposium on Computational Geometry* (pp. 116-127). Berkley, California, United States.

Choo, J., Jiamthapthaksin, R., Chen, C., Celepcikay, O. U., Giusti, C., & Eick, C. F. (2007). MOSAIC: A Proximity Graph Approach for Agglomerative Clustering. In *Data Warehousing and Knowledge Discovery* of *Lecture Notes in Computer Science* (pp. 231-240). Regensburg, Germany: Springer Berlin / Heidelberg.

Colannino, J., & Toussaint, G. T. (2005). An algorithm for computing the restriction scaffold assignment problem in computational biology. *Information Processing Letters*, *95*(4), 466-471.

Dasarathy, B. V., Sanchez, J. S., & Townsend, S. (2000). Nearest neighbor editing and condensing tools - synergy exploitation. *Pattern Analysis and Applications, 3,* 19-30.

Dong, Y., & Zhuang Y. (2004). Fuzzy hierarchical clustering algorithm facing large databases. *Fifth World Congress on Intelligent Control and Automation: Vol. 5* (pp. 4282 - 4286).

Johnson, R. A., & Wichern, D. W. (Ed.) (2007). *Applied Multivariate Statistical Analysis.* New York, NY: Prentice Hall.

Gan, G, Ma, C., & Wu, J. (2007). Clustering Algorithms. In ASA-SIAM Series on Statistics and Applied Probability . Philadelphia, PA.: SIAM.

Gouyon, F., Dixon, S., Pampalk, E., & Widmer, G. (2004). Evaluating rhythmic descriptors for musical genre classification. *ES 25th International Conference* (pp. 17-19). London, UK.

Handl, J., & Knowles, J. (2005). Multiobjective clustering around medoids. *Proceedings of the Congress on Evolutionary Computation: Vol. 1* (pp. 632-639).

Jaromczyk, J. W., & Toussaint, G. T. (1992). Relative neighborhood graphs and their relatives. *Proceedings of the Institute of Electrical and Electronics Engineers : Vol. 80. No. 9* (pp. 1502-1517).

Koren, Y., North, S. C., & Volinsky, C. (2006). Measuring and extracting proximity in networks. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining :* (pp. 245-255).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, *10*(8), 707-710.

Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F. C., Shen, M. M., Lu, G., Fang, J., Liu, W. M., Ryder, T., Kaplan, P., Kulp, D., & Webster, T. A. (2003). Probe selection for high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America: Vol.* 100. (pp. 11237-11242).

Sanchez, J. S., Pla, F., & Ferri, F. J. (1997). On the use of neighborhood-based non-parametric classifiers. *Pattern Recognition Letters*, *18*, 1179-1186.

Sanchez, J. S., Pla, F., & Ferri, F. J. (1997). Prototype selection for the nearest neighbor rule through proximity graphs. *Pattern Recognition Letters*, *18*, 507-513.

Sanchez, J. S., Pla, F., & Ferri, F. J. (1998). Improving the k-NCN classification rule through heuristic modifications. *Pattern Recognition Letters*, *19*, 1165-1170.

Schena, M. (Ed.). (2003). *Microarray Analysis*. New York, NY: John Wiley & Sons.

Tenenbaum, J. B., de Silva, V., Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319-2323.

Toussaint, G. T. (2004). Computational geometric aspects of musical rhythm. *Abstracts of the 14th*

Annual Fall Workshop on Computational Geometry (pp. 47-48). Massachussetts Institute of

Technology.

Toussaint, G. T. (2005). The geometry of musical rhythm. In J. Akiyama et al. (Eds.), *Proceedings of the Japan Conference on Discrete and Computational Geometry: Vol. 3742. Lecture Notes in Computer Science* (pp. 198-212). Berlin, Heidelberg: Springer-Verlag.

Zahn, C. T. (1971). Graph theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers, C-20*(1), 68-86.

KEY TERMS

Clustering: Data mining technique falling under the unsupervised learning category. It is the process of organizing data elements into groups, called clusters, whose members are similar in some way

Gabriel Graph (GG): A proximity graph formed by joining by an edge all Gabriel neighbors.

Gabriel Neighbors: Two points are Gabriel neighbors if the hyper-sphere that has them as diameter is empty, i.e., if it does not contain any other point.

Inter-Onset Interval Histogram (IOIH): A histogram that summarizes how many intervals there exist in a binary string of a fixed length, where ‘interval’ denotes the distance between two (not necessarily successive) set bits.

Microarray: An array of ordered sets of DNA fragments fixed to solid surfaces. Their analysis, using other complementary fragments called probes, allows the study of gene expression.

Nearest Neighbor: The point, in a point set, that has the minimum distance to a given point, with respect to a certain distance measure.

Nearest Neighbor Distance: A distance measure that measures the dissimilarity of two binary strings via the concept of nearest set bit neighbor mapping. A first pass is done over the first string, say from left to right, and set bits are mapped to their closest set-bit neighbors in the second string, in terms of character count. The same kind of mapping is done in a second pass over the second string. The nearest neighbor distance is the accumulated distances of each mapping link without counting twice double links.

Nucleotide: A subunit of DNA or RNA. Thousands of nucleotides are joined in a long chain to form a DNA or an RNA molecule,. One of the molecules that make up a nucleotide is a nitrogenous base (A, G, C, or T in DNA; A, G, C, or U in RNA); hence a nucleotide sequence is written as a string of characters from these alphabets.

Probe: A sequence of a fixed number of nucleotides used for the analysis of microarrays. It is designed to bind to specific DNA fragments in a microarray, and emit fluorescent light as an indicator of the binding strength.

Proximity Graph: A graph constructed from a set of geometric points by joining by an edge those points that satisfy a particular neighbor relationship with each other. The most well-known proximity graphs are the nearest neighbor graph (*NNG*), the minimum spanning tree (*MST*), the relative neighborhood graph (*RNG*), the Urquhart graph (*UG*), the Gabriel graph (*GG*), and the Delaunay triangulation (*DT*).