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

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.