ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 21-26

High-dimensional labeled data analysis with Gabriel graphs

Michaël Aupetit*
CEA - DAM

Département Analyse Surveillance Environnement
BP 12 - 91680 - Bruyères-Le-Châtel, France

Abstract. We propose the use of the Gabriel graph for the exploratory analysis of potentially high dimensional labeled data. Gabriel graph is a subgraph of the Delaunay triangulation, which connects two data points v_i and v_j for which there is no other point v_k inside the open ball with diameter $[v_i v_j]$. If all the Gabriel neighbors of a datum have a different class than its own, this datum is said to be ”isolated”. While if some of its Gabriel neighbors have the same class as its own and some others have not, then this datum is said to be ”border”. Isolated and border data together with Gabriel graph, allow to get informations about the topology of the different classes in the data space. It is complementary with “classical” and “neural” projection techniques.

1 Introduction

Exploratory data analysis is an important and active field of researches, which has to deal with the exponential growth of high-dimensional data bases in many domains. Here we consider the case of labeled data in potentially high dimensional space, and propose the use of a topological graph to extract some information about the topology of the different classes.

Duda et Hart [5] define a ”class-similarity” graph as a graph which connect pairs of centers of gravity of data of different classes, which are closer than a threshold. Another approach proposed by Melnik [10] creatse links between data if all the points sampled on a link are given the same label by a classifier. If the sampling is sufficiently high, the different connected components of the graph represent regions of different classes.

We propose the use of a topological graph called Gabriel graph to extract informations about connexity of the different classes from a labeled data set. Gabriel graph [8] is a subgraph of the Delaunay triangulation [7] of the data set. It has been studied in the field of classification, as a way to edit and

*aupetit@dase.bruyeres.cea.fr

condense large data sets [12], by keeping only relevant data which take part to the decision boundary of a nearest neighbor classifier. Recently, it has been proposed in [13], that these data which are preserved in the edited data set, and that we call ”border” data hereafter, are similar to support vectors in the context of Support Vector Machines [2], which determine the decision boundary.

We define ”normal”, ”border” and ”isolated” data and study their use with Gabriel graphs to reveal the topology of high dimensional labeled data, which is complementary with “classical” (Principal Component Analysis. . .) or “neural” (Curvilinear Component Analysis [3], Self-Organizing Maps [9]. . .) projection techniques.

2 Normal, border and isolated data

Considering a set of N data $\underline{v}$ and a graph $\mathrm{G}(\underline{v}, \underline{L})$ with $\underline{L}$ a set of edges or links between the data. Two data v_i and v_j are neighbors of each other if $l_{ij} = \{v_i, v_j\} \in \underline{L}$. We define different ”qualities” *w.r.t.* $\mathrm{G}(\underline{v}, \underline{L})$ (Figure 1a) :

Border: a datum $v \in \underline{v}$ for which there is at least one of its neighbors through G, which has not the same class as its own.

Isolated: a datum $v \in \underline{v}$ for which all the neighbors through the graph G, have a class different from its own. “Isolated” data are also “Border”.

Normal: a datum $v \in \underline{v}$ for which all the neighbors through the graph G, have the same class as its own.

In the following, we choose a topological graph G for which the previous definitions have a relevant meaning *w.r.t.* the topology of the classes in the data space.

3 Voronoï , Delaunay and Gabriel graphs

Considering a set $\underline{v}$ of N data in E, a bounded domain of $\mathcal{R}^D$, the Voronoï region $\mathcal{V}_i$ associated to a datum v_i is defined as the region of E which contains all the points for which v_i is the closest datum among all the data $\underline{v}$ [7]. The Delaunay triangulation $\mathrm{DT}(\underline{v})$ of $\underline{v}$ is the set of edges drawn between pairs of data whose Voronoï regions share a common boundary [7] (Figure 1b).

Although DT would be ideally suited for being the graph G we expect, it has a $O(N^{\lceil \frac{D}{2} \rceil})$ computing time, making DT construction intractable for high-dimensional data spaces. We propose to use the Gabriel graph [8] of $\underline{v}$ which approximates $\mathrm{DT}(\underline{v})$ and which takes $O(D.N^3)$ time for its computation.

The Gabriel graph $\mathrm{GG}(\underline{v})$ is the set of edges l_{ij} subset of $\mathrm{DT}(\underline{v})$, for which the open ball with diameter $[v_i v_j]$ contains no other data than v_i and v_j:

$\mathrm{GG}(\underline{v}) = \left\{ l_{ij} \subseteq \underline{v} \mid \forall\, v_k \in \underline{v}, (v_k - v_i)^2 + (v_k - v_j)^2 > (v_i - v_j)^2 \right\}$ (Figure 1c)

The brute force algorithm to compute $\mathrm{GG}(\underline{v})$ is straight forward from the above definition. A heuristic proposed in [1] may improve noticeably the computing time.

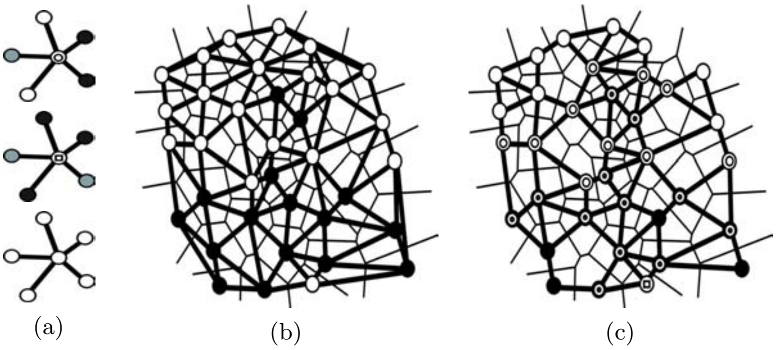


Figure 1: (a) from top to bottom: “border”, “isolated” and “normal” data (circle at the center) (classes in gray levels). (b) Voronoï diagram (thin lines) and Delaunay triangulation (bold lines) of the data (circles). (c) Gabriel graph (bold lines) of the data with border ◎ and isolated ◉ data of class ○, and border data ⦿ of class ●.

4 How to use Gabriel graph

Applying the definitions given in section 2 with the Gabriel graph, allows to analyze the topology of the labeled data set (Figure 2).

4.1 Basic insights

Border data are data of each class, closest to the class decision boundary than any other data. The identification of a border datum is a warning signal to the expert to pay more attention about its labeling.

The middle point of each edge joining two border data of different classes, is a sample of the decision boundary in the sense of the nearest neighbor classifier[1]. Hence, by projecting these middle points (*e.g.* using CCA [3]), we can visualize the decision boundary.

Isolated data are subset of border data. Isolated data may reveal either outliers (error in labeling) or overlapping of the classes. The identification of an isolated datum is a warning signal to the expert that its labeling of this datum is probably erroneous because all its neighbors have a different class than its own.

Normal data labeling need no particular attention not touching decision boundaries.

4.2 Advanced analysis

We propose to prune the original graph G in order to reveal the different connected components *w.r.t* classes and qualities. All the following constructions hold for any graph G but we focus on Gabriel graph here.

[1] A new datum not in $\underline{v}$ is labeled with the class of the closest datum to it among $\underline{v}$

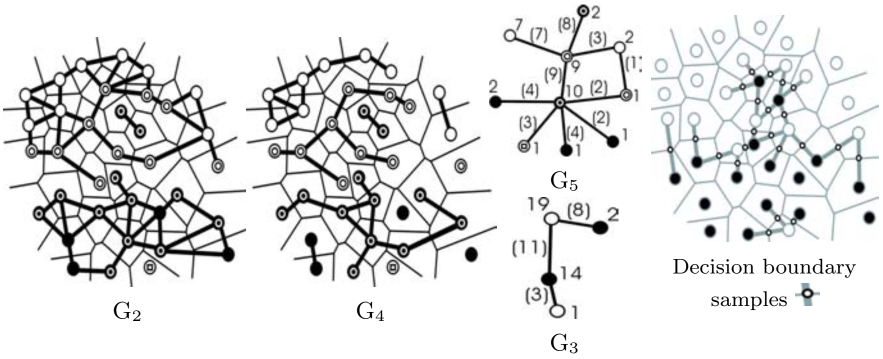


Figure 2: Number of nodes “x” (edges “(y)”) in (between) corresponding connected components of G$_2$ and G$_4$, are indicated beside nodes (edges) of G$_3$ and G$_5$.

i) G : Construct the Gabriel graph of $\underline{v}$.

ii) G$_2$: Clear edges in G connecting data pairs of different classes.

iii) G$_3$: Create the graph G$_3$ which associates a vertex to each connected components of G$_2$. If two of these connected components are linked in G, connect the corresponding vertices in G$_3$. G$_3$ is the ”class” graph.

iv) G$_4$: Clear edges in G$_2$ which connect data pairs of different qualities.

v) G$_5$: Create the graph G$_5$ which associates a vertex to each connected components of G$_4$. If two of these connected components are linked in G, connect the corresponding vertices in G$_5$. G$_5$ is the ”class&quality” graph.

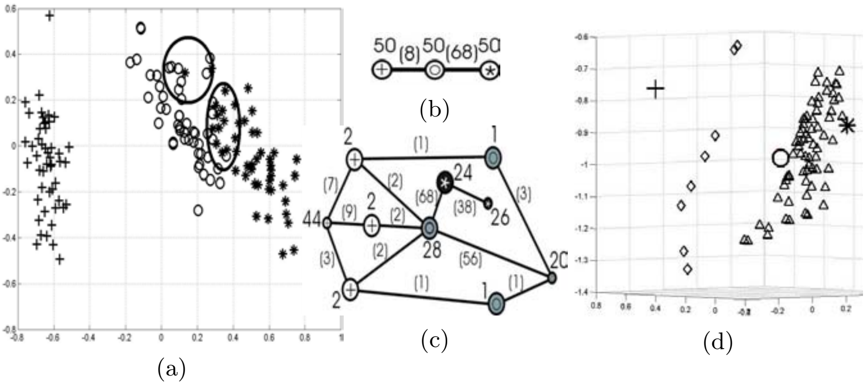
The drawing of the ”class” graph shows the topology of the classes, the way they are connected or not, and the density of the connections between the different components. Considering the ”class&quality” graph, the same analysis may be done and allows to visualize the number of border and isolated components.

5 Analysis of Iris database

The Iris benchmark database [6, 11] is a set of 150 data, with 4 attributes (sepal and petal length and width, of Iris plants) and 3 classes (Iris Setosa (*St*), Versicolor (*Vs*) and Virginica (*Vg*)). There are 50 data of each classes.

After a prior normalization of the data, the analysis technique we propose gives the following results (Figure 3):

- There is no isolated data in any of the classes which each consists of only one connected component (b). Hence, there is no overlapping in the data space despite what suggests 2-dimensional linear projection using PCA



(a)

(b)

(c)

(d)

Figure 3: Iris data base: (a) Projection onto the two first principal components (*St* "+", *Vs* "O", *Vg* "∗"). (b) "class" graph. (c) "class&quality" graph. (d) Nonlinear projection using CCA [3], of the center of gravity of each G_2's connected component, and of the decision boundary samples between *St* and *Vs* (diamonds), and between *Vs* and *Vg* (triangles).

(a) or the class-preserving projection presented in [4]. So the decision boundary considering each pair of classes, is homeomorph to a 3-flat.

- Classes *St* and *Vg* are not connected (b) so two different clasiffiers may be considered: one separating *St* from *Vs*, and another *Vs* from *Vg* (d).
- There are 68 links between *Vs* and *Vg*, while only 8 between *St* and *Vs* (c). And there are only 6 border data in *St* (12% of this class), while 30 in *Vs* (60%) and 24 in *Vg* (48%). The same border component (28 data) of *Vs* is connected to both *St* and *Vg*. Moreover, the component of 44 *St* normal data is connected to *St* border data with only 19 links. This is the opposite for *Vs* (20 normal data and 60 links) and *Vg* (26 normal data and 38 links). All this suggests that *St* is well clustered having a little part in contact with *Vs*, while *Vs* and *Vg* are more spread along their common boundary.

This approach helps us to grasp the topology of the data in the data space, and to detect eventual misleading projections of the data in lower dimensional spaces: it is complementary to "classical" and "neural" projection techniques.

6 Conclusion

We propose to define the quality of data as "isolated", "border" or "normal" according to the class of their neighbors on a graph. These qualities together with Gabriel graphs allow to discover the way the different classes are connected, the number of data of different classes which are in contact, to identify outliers allowing to notice data likely to have an erroneous label, to pay more attention to the labeling of data near decision boundaries, and to visualize a sampling of the decision boundary in the sense of the nearest neighbor. This approach is complementary with "classical" and "neural" projection techniques.

New visualizations tools remain to create to apprehend complex graphs obtained in case of strong overlapping of the classes. We experiment this approach on a large set of 5-dimensional data in the field of sismic events classification.

References

[1] Bhattacharya, B.K., Poulsen, R.S. & Toussaint, G.T.(1981) Application of Proximity Graphs to Editing Nearest Neighbor Decision Rule. *Int. Symp. on Information Theory*, Santa Monica.

[2] Burges, C.(2002) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery,***2**(2):121-167.

[3] Demartines, P. & Hérault, J. (1997) Curvilinear Component Analysis: a Self-Organising Neural Network for Non-Linear Mapping of Data Sets. *IEEE Trans. on Neural Networks,***8**(1):148-154.

[4] Dhillon, I.S., Modha, D.S. & Spangler, W.S. (2002) Class Visualization of high-dimensional data with applications *Computational Statistics & Data Analysis,***41**:59-90.

[5] Duda, R.O., & Hart, P.E. (1973) *Pattern Classification ans Scene Analysis.* J. Wiley & Sons eds.

[6] Fisher, R.A.(1936) The use of multiple measurements in taxonomic problems *Annual Eugenics,***7**(2):179-188.

[7] Fortune, S. (1992) Voronoï diagrams and Delaunay triangulations. *Computing in Euclidean geometry.* D.Z. Du, F. Hwang eds, World Scientific, 193-233.

[8] Gabriel, K.R. & Sokal, R.R.(1969) A new statistical approach to geographic variation analysis. *Syst. Zoology,***18**:259-278.

[9] Kohonen, T. (1997) Exploration of very large databases by self-organizing maps. *Proc. of IEEE Int. Conf. on Neural Networks, ICNN'97*, **1**:PL1-PL6.

[10] Melnik, O. (2002) Decision Region Connectivity Analysis: A method for analyzing high-dimensional classifiers . *Machine Learning.* **48**(1-3):321-351.

[11] Merz, C.J., Murphy, P.M. & Aha, D.W. (1997) UCI repository of machine learning databases. *Dept. of Inf. and Comp. Sc., Univ. of California at Irvine,* http://www.ics.uci.edu/pub/machine-learning-databases/

[12] Toussaint, G. (2002) Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress. *Proc. of INTERFACE'2002,* 34th *Symp. on Computing and Statistics,* Montreal, Canada.

[13] Zhang, W., & King, I. (2002) A Study of the Relationship Between Support Vector Machine and Gabriel Graph. *Proc. of IEEE Int. Joint Conf. on Neural Networks, IJCNN'2002.*