ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Constructing similarity networks using the Fisher information metric

H. Ruiz[1], S. Ortega-Martorell[2], I. H. Jarman[1], J. D. Martín[3], P. J.G. Lisboa[1]

1 - School of Computing and Mathematical Sciences - Department of
Mathematics and Statistics - LJMU, Liverpool L3 3AF - UK
2 - Departament de Bioquímica i Biología Molecular – Universitat
Autònoma de Barcelona, Cerdanyola del Vallés (Barcelona) - Spain
3 - Escuela Técnica Superior de Ingeniería - Departamento de Ingeniería
Electrónica - Universidad de Valencia, Burjassot (Valencia) - Spain

Abstract. The Fisher information metric defines a Riemannian space where distances reflect similarity with respect to a given probability distribution. This metric can be used during the process of building a relational network, resulting in a structure that is informed about the similarity criterion. Furthermore, the relational nature of this network allows for an intuitive interpretation of the data through their location within the network and the way it relates to the most representative cases or prototypes.

1 Introduction

Measures of similarity between data points are central to pattern recognition and data mining methodologies, although they are not always explicitly calculated. Nevertheless, using a distance function to measure similarity between pairs of elements of a space is an intuitive way to understand their relationship in the context of a particular problem domain. The Euclidean distance is a common choice because of its simplicity and little computational cost, even though the equal weighting of each dimension, for instance in clustering, leads to results that can be heavily dependent on the choice of data representation.

The Fisher information (FI) is a natural choice of metric in the space of probabilistic density functions [1]. In the case of the space of the covariates, a natural similarity measure between points is provided by the symmetric divergence between the posterior distributions $p(c|\boldsymbol{x})$ of classifiers fitted to the class labels, which are categorized by a discrete random variable C. The concept of a metric defined by differentiating a posterior distribution $p(c|\boldsymbol{x})$ with respect to the coordinates is reported in [2] as a natural extension of the metric defined in parameter space, e.g. in [3]. In a recent publication [4], we explained in detail the process of deriving a metric from the Fisher information using linear and non-linear models and presented a novel approach to the problem of finding geodesic distances in non-Euclidean metrics. The idea in the present paper is to follow the same process and use the FI metric to go from the dataset in the original high-dimensional space to a network where data points are nodes connected to each other by edges based on their similarity.

The resulting networks are analysed in terms of prediction accuracy and structure and the closing section discusses the interpretability of the classifier by identifying relevant reference cases.

2 Methodology

This section provides a brief description of the FI metric derivation and discusses the choice of the method used to build the networks.

2.1 Derivation of the Fisher information metric

The FI is a local measure of the variation that an infinitesimal displacement of a point produces on the value of a probability distribution when evaluated at that point. Traditionally, the space where this displacement takes place is that of some parameter vector **θ** upon which the probability function depends. We are, however, more interested in the approach introduced in [2], where the space of interest is the primary data space, i.e. the space where the dataset under study lies. The data is assumed to be divided into classes, with $p(c|\mathbf{x})$ representing the posterior probability of the class variable given a point in the data space.

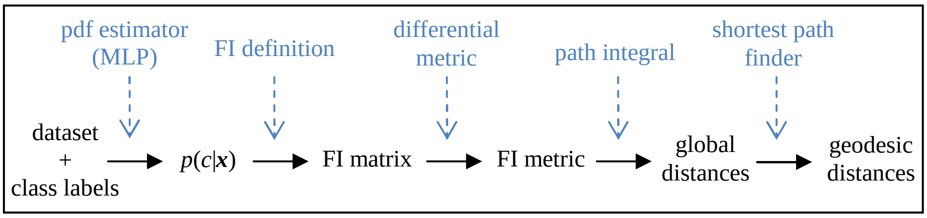

Fig. 1: Derivation of the FI metric

Figure 1 outlines the process of deriving the FI metric. Initially we have a dataset along with the class membership of each of the points. Using a density estimator, we obtain the posterior distribution $p(c|\mathbf{x})$. This estimate completely determines the metric in the sense that only a good model will produce a metric that reflects similarity accurately with respect to the true probability density. We use a multilayer perceptron (MLP) for this purpose because of its versatile architecture, which makes it ideal for non-linear data distributions.

The FI takes the form of a square matrix of the same dimensionality as the data. It is obtained from the estimated density according to either of the two equivalent definitions in (1).

$$\boldsymbol{FI}(\boldsymbol{x})_{p(c|\boldsymbol{x})} = \begin{cases} E_{p(c|\boldsymbol{x})}\{(\nabla \ln p(c|\boldsymbol{x}))^2\} \\ -E_{p(c|\boldsymbol{x})}\{(\nabla^2 \ln p(c|\boldsymbol{x}))\} \end{cases} \qquad (1)$$

where $E_{p(c|\mathbf{x})}$ is the conditional expectation over the values of the class label c with respect to $p(c|\mathbf{x})$. The matrix defines a differential metric for the calculation of infinitesimal distances:

$$d(\boldsymbol{x}, \boldsymbol{x} + \Delta\boldsymbol{x})^2 = \Delta\boldsymbol{x}^T \boldsymbol{FI}(\boldsymbol{x}) \Delta\boldsymbol{x} \qquad (2)$$

This can be integrated to calculate the distance between any pair of points by using the path integral

$$d(\boldsymbol{x_A}, \boldsymbol{x_B}) = \left| \int_0^1 \sqrt{\dot{\boldsymbol{x}}(t)^T \boldsymbol{FI}\big(\boldsymbol{x}(t)\big) \dot{\boldsymbol{x}}(t)}\, dt \right| \qquad (3)$$

where $\boldsymbol{x}(t)$ is a path that goes from $\boldsymbol{x_A} = \boldsymbol{x}(t = 0)$ to $\boldsymbol{x_B} = \boldsymbol{x}(t = 1)$. At this point we can compute global distances in the space along a given path. The last part of the process is to find geodesic paths between points and to calculate their length. To do so, we use the free points approach. The reader is referred to [4] for an explanatory section on this algorithm.

2.2 Construction of the networks

Usual methods to build networks are k-nearest neighbours (kNN), where each data point is connected to the k nearest points, and ϵ-neighbourhood, where a connection is present when points are closer than a constant distance ϵ. kNN is preferred over ϵ-neighbourhood because it is adaptive to scale and density, while the use of the latter can result in disconnected graphs.

During the experiments carried out for this work, we applied kNN and b-matching. The b-matching method [5] is more rigorous than kNN in that it ensures that the final number of neighbours of each node is always the same. However, it only guarantees to converge if the linear programming relaxation of the formulation of the b-matching problem is tight [6]. In practice, when applying the algorithm to our data we found that it did not converge most of the time. For this reason, we only use kNN in this work.

3 Experimental results

In this section, we study the implications of the use of the FI metric in the construction of networks. Three aspects are discussed: the visualization power of networks, classification accuracies using kNN and the presence of network substructure.

The synthetic data analysed in this study are modelled from samples extracted from a data base used in a previous publication [7]. Class (tumour type) labelling was used to generate posterior distributions of the data density, using single multivariate normal models fitted to the mean and variance/covariance matrices of class specific cohorts of single-voxel proton Magnetic Resonance Spectroscopy (SV ^{1}H-MRS) from brain tumour patients.

This synthetic set included samples of the generated data for 78 glioblastoma-like (GL), and 31 metastasis-like (ME) cases. The data dimensionality is 195 reflecting the clinically-relevant frequency intensity values measured in parts per million (ppm) that are typically sampled from each spectrum in the [4.24,0.50] ppm interval. A second dataset was generated for the validation of the methods. In this dataset, each class has 50 samples generated using the same means and covariance matrices used for the training set. The discrimination between GL and ME, on the basis of SV ^{1}H MRS information, is a very challenging problem due to their radiological similarities. The appearance of both pathologies is often dominated by large peak intensities corresponding to neutral lipids, a byproduct of necrosis [8].

3.1 Visualizing data using networks

Using networks to represent data is a powerful visualization tool, especially when the data space is high-dimensional and direct examination is not possible. Similarity between points in the original space is captured by the connections in the network, enabling the viewer to see how the data looks like in terms of structure and clustering. Figures 2 and 3 show the kNN networks built from the dataset using Euclidean and Fisher distances respectively, with black nodes representing GL cases and white corresponding to ME. It is immediate to see a much clearer structure in the Fisher network (Fig.3), with edges connected in a very local manner. The Euclidean network (Fig.2), on the other side, presents a very fuzzy arrangement of the edges, and the grouping of the nodes is quite weak in terms of connectivity within/between groups.

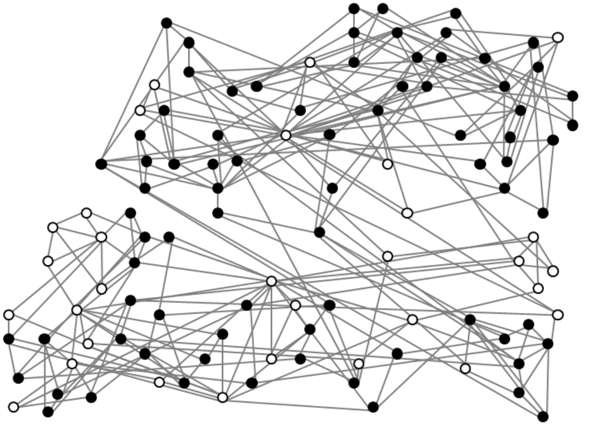

Fig. 2: kNN network (k=3) using Euclidean distances. Black = GL, white = ME.

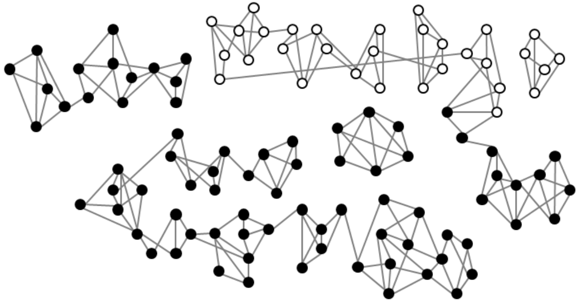

Fig. 3: kNN network (k=3) using Fisher distances. Black = GL, white = ME.

The information contained in the class labels is put in the form of a distance measure by the FI metric and is captured in the network, producing an informative and intuitive visualization of the data that otherwise would be difficult to interpret.

3.2 Classification rates

Tables 1 and 2 contain the classification accuracies using Euclidean and Fisher kNN classification (E-kNN and F-kNN, respectively) for different values of k. Table 2 corresponds to the results with the validation dataset. Fisher kNN obtains very good accuracies as reported in the first table because these are the training samples of the MLP. The second table provides a more realistic impression.

	1	3	5	7	9	11	13	15
E-kNN	0.83	0.81	0.81	0.83	0.76	0.74	0.77	0.74
F-kNN	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98

Table 1: Classification rates on the training dataset

	1	3	5	7	9	11	13	15
E-kNN	0.72	0.79	0.77	0.78	0.74	0.69	0.69	0.68
F-kNN	0.77	0.80	0.80	0.81	0.82	0.80	0.79	0.80

Table 2: Classification rates on the validation dataset

During the validation stage, the use of the FI metric brings little improvement on the Euclidean kNN results for small values of k, but the difference becomes more significant when the size of the neighbourhood increases. When this happens, the performance of E-kNN deteriorates, so points correctly classified for a small number of neighbours are misclassified when more neighbours are taken into account, reflecting heterogeneity in the local structure of the network. The stability of F-kNN under this variation is caused by the FI metric moving the areas of the space with a high density of points from the same class away from the border regions between classes. This means that points away from the border form very compact and homogeneous groups that are far away from the areas of mixed membership, therefore having more stable neighbourhoods with respect to k.

3.3 Class substructure

Going back to Fig.3, we stated that a clear structure in the network is easy to see, not only because the classes are well separated, but also because within each class, nodes are arranged forming small groups or clusters. In this section, we briefly look into some of these clusters to find out the differences between them.

The plots in Fig.4 are the mean spectra of the points in each cluster, the first plot corresponding to the real GL spectrum. The four clusters are part of the GL class, and are circled in red in the miniature view of the network.

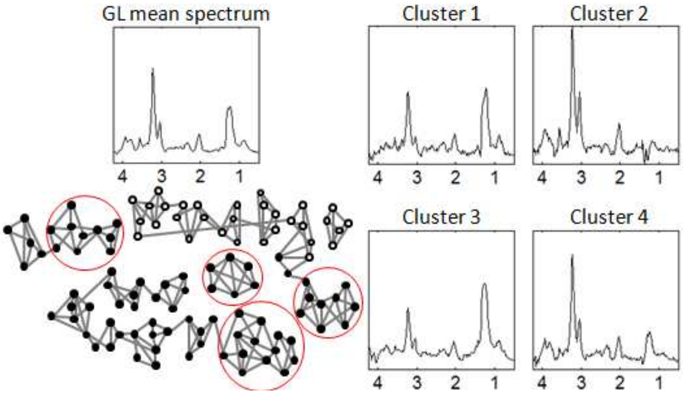

Figure 4: Mean spectra of the different clusters.

The plots show how clusters from the same class are different in terms of the height of the two main peaks of the spectrum, which vary from group to group giving rise to a different “prototype” in each of them.

The shape of each spectrum has a specific meaning and corresponds to a different medical condition, so within the same GL class we can find different subtypes of brain tissue. In other words, a sample classified as GL could be further subclassified depending on where it lies within the network.

4 Conclusions

The FI can work as a measure of how similar points in the space are with respect to some class membership probability distribution. To do so, we derive a metric from the FI matrix, and therefore a distance measure. This can be used to build a relational network that captures similarity in the original data space and translates it into node to node connections, resulting in a more interpretable representation of the data, especially when the original data space is of high dimensionality.

The structure of the network contains useful information on how the data is distributed in the space. Section 3.3 presented a very simple analysis of some of the substructures found in the dataset. Our motivation for the use of networks is to develop a way of interpreting new data by mapping it into the base network and relating it to the reference cases in it. By doing so, we go from just a scalar that represents the probability of a point belonging to a certain class to a much more informative tool that not only predicts a category for the data, but also puts it into context by telling how it relates to the most representative cases.

It is important to bear in mind the small sample size of the data (109 points in a 195-dimensional space). We chose to keep the original size in the synthetic dataset because it is not crucial for our aim of showing the interpretability of the methodology. However, if the estimation of the probability surfaces was required to be very precise, a larger dataset would be necessary.

References

[1] S. Amari (2001). Information geometry on hierarchy of probability distributions. IEEE Information Theory, 47, vol. 5: 1701-1711.

[2] S. Kaski, & J. Sinkkonen (2000). Metrics that learn relevance. IJCNN 2000 proceedings, vol 5:547-552.

[3] S. Kullback (1959). Information theory and statistics. Wiley. New York, 1959.

[4] H. Ruiz, I. H. Jarman, J. D. Martín, & P. J. Lisboa (2011). The role of Fisher information in primary data space for neighbourhood mapping. ESANN 2011 proceedings.

[5] T. Jebara, J. Wang, & S. F. Chang (2009). Graph construction and b-matching for semi-supervised learning. ICML 2009 proceedings.

[6] B. Huang, & T. Jebara (2011). Fast b-matching via sufficient selection Belief Propagation. AISTATS 2011 proceedings.

[7] S. Ortega-Martorell, A. Vellido, P.J.G. Lisboa, M. Julia-Sape, & C. Arus (2011). Spectral decomposition methods for the analysis of MRS information from human brain tumours. IJCNN 2011 proceedings.

[8] A. Vellido, E. Romero, M. Julia-Sape, C. Majos, A. Moreno-Torres, J. Pujol, & C. Arus (2011). Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel ^{1}H MRS. NMR in Biomedicine. In press. DOI: 10.1002/nbm.1797