ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The role of Fisher information in primary data space for neighbourhood mapping

H. Ruiz[1], I. H. Jarman[2], J. D. Martín[3], P. J. Lisboa[1]

1 - School of Computing and Mathematical Sciences - Department of Mathematics and Statistics - LJMU, Liverpool L3 3AF - UK
2 - Centre for Public Health - LJMU, Liverpool L3 2ET - UK
3 - Escuela Técnica Superior de Ingeniería - Departamento de Ingeniería Electrónica - Universidad de Valencia, Burjassot (Valencia) - Spain

Abstract. Clustering methods and nearest neighbour classifiers typically compute distances between data points as a measure of similarity, with nearby pairs of points considered more like each other than remote pairs. The distance measure of choice is often Euclidean, implicitly treating all directions in space as equally relevant. This paper reviews the application of Fisher information to derive a metric in primary data space. The aim is to provide a natural coordinate space to represent pairwise distances with respect to a probability distribution $p(c|x)$, defined by an external label c, and use it to compute more informative distances.

1 Introduction

The primary purpose of this work is to define framework to calculate the similarity between data points, in primary data space, using an auxiliary variable which is a class label. This will enable networks of data points to be arranged in a way that is informed about this variable. For the sake of illustration, we measure classification rate using k-NN to evaluate the near neighbour homogeneity of the data with respect to the auxiliary variable.

In the standard formulation, each observation consists of a set of N variables, and therefore represents a point in the N-dimensional space. A very intuitive and widely used way to compute distances between data points is to use the Euclidean metric. This distance assigns equal relevance to all directions and, by extension, to all variables, but in reality each attribute will have a different degree of influence over the auxiliary label c. In this work, similarity between data points is defined with respect to some auxiliary data comprising observations of a dichotomous variable c which divides the dataset into two classes. Data points are considered close to each other if they have similar class membership probabilities, and this definition also applies to groups of points and areas of the dataspace. The goal of this definition of similarity is to form clusters or divide the data into classes that are homogeneous with respect to the label c. That is precisely what a learning metric does. Effectively, such a metric resizes each dimension in space expanding those corresponding to relevant features and compressing those related with less important ones.

While the bulk of statistical work on Fisher information is focused on the space of model parameters, there also is some previous work on learning metrics defined in primary data space [1,2,3] with successful applications to self-organizing maps and standard clustering algorithms such as k-means. In common with the literature, the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

work presented in this paper shares the objective of developing an intelligent metric that improves the performance of the algorithms, but differs in three key aspects.

Firstly, the way the Fisher metric is obtained. By definition, the metric is derived from the probability density *p(c|x)*, which must therefore be estimated [1,2,3]. The estimations used in this work are drawn directly from the posterior distributions of class membership, with either generalised linear models or generic (semi-parametric) non-linear inference models, namely a linear logistic regressor and a multilayer perceptron (MLP).

The second main difference is the approach used to compute distances. In the non-Euclidean space resulting from the application of the Fisher metric, the shortest distance requires the explicit optimisation of the distance measured across a geodesic path. This is discussed in [3,4,5], but the two most related solutions [3] make strong simplifying assumptions, one using a single straight line between distant points and the other depending on the particular layout of the data points. We propose a new method which is efficient in iteratively adjusting the path towards the shortest distance, as explained in section 2.2.

Finally, the motivation for the construction of the Fisher metric in this work is different than that of the existing literature. In previous work, the concept of learning metrics is included into existing clustering and classification methods to improve their performance. That is not the case here. The methodology that this paper presents has been developed with the intention of applying it to the construction of graphs from datasets with auxiliary variables.

2 Methodology

This section describes the different concepts involved in the metric building process. First the Fisher information is introduced and derived for linear and non-linear estimators assuming a logistic regression transfer function of the output. Second, the problem of finding geodesics is addressed and introduces the proposed generic approach for distance estimation in primary data space with non-linear metrics.

2.1 Fisher information in the primary data space

The FI value [6] at a particular data point *x* in the space is the difference between the information that the probability distributions *p(c|x)* and *p(c|x+dx)* carry, where *dx* is infinitesimally small. In other words, a large FI value at a certain point means that a slight change in the position of that point strongly influences the posterior density function and thus that area of the space is very relevant with respect to the auxiliary data *c*.

The metric is defined by the matrix *G(x)* in the well-known quadratic differential form [7]:

$$ds^2 = dx^T G(x) dx = \sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}(x)\, dx_i dx_j \qquad (1)$$

The Fisher information matrix in primary data space is defined equivalently by:

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

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The calculation involves the conditional expectation over the values of the external label *c* with respect to the probability function *p(c|x)*. Limiting *c* to a discrete variable simplifies the calculation because the expectation, which would be computed as an integral in the continuous case, becomes a summation. We further assume a specific structure for the form of the posterior probability *p(c|x)*, namely

$$p_c = p(c|x) = \frac{c + (1-c)e^{-a(x)}}{1 + e^{-a(x)}} \quad , \quad c = \{0,1\} \tag{2}$$

The dependence on *x* is contained in the activation variable *a*, which defines the complexity of the estimator. In the logistic regressor, *a* is just a linear combination of the input vector *x* and the coefficient vector β, while in the case of the MLP it is given by a non-linear, but differentiable function of the inputs. Once *p(c|x)* is defined, the FI can be expressed in matrix form, assuming column vectors, as follows:

$$FI(x) = \big(\nabla a(x)\big)\big(\nabla a(x)\big)^T p_{c=1}(1 - p_{c=1})$$

Returning to (1) yields the distance between infinitely close points. A general formula for the distance between two points is obtained by solving the path integral:

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

The next section provides a solution of integral in (3) in closed form for dichotomous classifiers of the assumed form, whether linear or not.

2.1.1 Fisher distance with a linear estimator

We start with the logistic regression, with $a=\beta^T x+\beta_0$. Since *a* is linearly dependent on *x*, its first derivative is constant, resulting in the following expression for the integral:

$$d(x_A, x_B) = \left| \int_0^1 \sqrt{\dot{x}(t)^T \beta\beta^T \dot{x}(t) \cdot p_{c=1}(1 - p_{c=1})}\, dt \right| \tag{4}$$

which is readily solved by substituting *a* as the integration variable giving:

$$d(x_A, x_B) = \left| 2\left[arctg\left(e^{-\frac{a(t)}{2}} \right) \right]_{a(t=0)}^{a(t=1)} \right| \tag{5}$$

It is important to note that distance is independent of the particular path from x_A to x_B., in effect collapsing the data space onto the projections along the vector defined by the weights β.

2.1.2 Extension to a non-linear estimator

The natural next step is to develop an expression for the distance when using a non-linear estimator of the posterior density distribution by solving equation (3) as in the previous section, but with *a* as a non-linear function of *x.* Using the first two terms of the Taylor expansion of *a* produces a linear approximation for which the distance expression (5) applies. Equation (5) is thus globally applicable for the linear case, but only locally valid for a non-linear estimator. In this work, the so-called free points approach is used to overcome this limitation.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2.2 Geodesic distances. The free points approach

The algorithm described in this section performs two important functions: it ensures that the Taylor approximation used previously holds and it finds the geodesic path between x_A and x_B. Figure 1 shows an illustrative sketch of the approach.

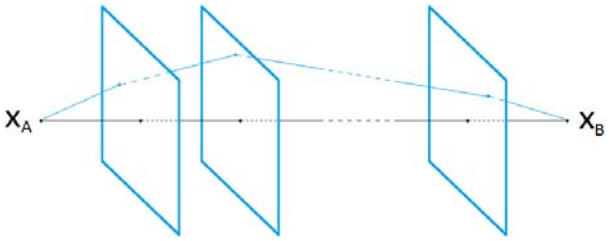

Fig. 1: The free points approach for 3-dimensional data.

The method starts by dividing the straight line joining x_A and x_B into segments. A hyperplane is then defined between each pair of consecutive segments. All these hyperplanes are parallel between each other and orthogonal with respect to the straight line path. Then, a point is defined in each hyperplane, with the idea of forming a path from x_A to x_B by joining all the points as shown in fig. 1. Since the points can move freely within their respective hyperplanes, any path can be formed by appropriately choosing the number of hyperplanes and the position of the points.

The points move as a result of the minimization of the objective function, defined as the overall length of the path computed as the sum of each segment's length. Each of these individual distances is calculated using (5). Since this applies locally for the non-linear case, every free point must be close to its two neighbours, and that is guaranteed by choosing a large enough number of hyperplanes.

3 Experimental results

In this section, the Fisher metric is put into practice in a classification problem using synthetic data. Two versions of the standard k-nearest neighbours (kNN) classifier are compared: one computes pairwise distances using the Euclidean metric (E-kNN) and the other uses the Fisher distance (F-kNN) derived from a MLP.

The method is benchmarked using a kNN classifier to assess the homogeneity of the resulting network with respect to the external label, not because the classifier itself brings any originality. Fisher metric based classifiers can be found in the literature, the most important being the SVM-Fisher kernel methods [8].

The dataset consists of two classes generated by two Gaussian distributions with same means but different standard deviations (0.9 and 2). One distribution contains the other, creating a non-linear border. This is a large dataset (10^4 samples/class) that provides the MLP with enough training episodes to accurately estimate $p(c|x)$, which is critical for the estimation of the Fisher information. A validation dataset is generated using the original generating functions of the data. This smaller dataset (250 samples/class) contains the points to be classified, calculating distances using either the Euclidean or Fisher metric and taking a majority vote in the usual way.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Table 1 shows the results of the simulations. The first row in each cell shows the percentage of correctly classified points using E-kNN and F-kNN in that order. The relative increase of accuracy when using the Fisher metric appears in the second row.

k \ N	2	5	10	15	25	40
3	69.4/72 +3.75%	87.4/73.6 -15.79%	88.8/92.2 +3.83%	81.6/93.2 +14.21%	66.4/94.6 +42.47%	51.6/94.2 +82.56%
5	72.4/74.8 +3.31%	88.2/72.8 -17.46%	89/92.2 +3.6%	80/93.4 +16.75%	64.4/95 +47.52%	50.6/94 +85.77%
7	74/74 +0%	88.2/74.4 -15.65%	88.8/92.2 +3.83%	77.6/93 +19.85%	61.8/95.2 54.05%	50.4/94 +86.51%
11	73.6/75.6 +2.72%	88.8/76.6 -13.74%	87.2/93.2 +6.88%	74.8/93.2 +24.6%	57/95.2 +67.02%	50.2/94 87.25%
15	75.4/76.8 +1.86%	88.2/79 -10.43%	86/93.4 +8.6%	73/93.8 +28.49%	56.2/95.2 +69.4%	50.2/94.2 +87.65%
21	75/76.6 +2.13%	88.6/79.4 -10.38%	85.8/93.8 +9.32%	71.4/93.2 +30.53%	54.4/95.2 +75%	50/94.4 +88.8%

Table 1: Simulation results for input dimensionality N and k neighbours.

In low dimensions, the two methods perform similarly. However, the accuracy of the Euclidean classifier increases until N=10 and decreases from then on. To understand this behaviour, a histogram of the pairwise distances is plotted in fig. 2 for different values of N, comparing interclass and intraclass distances. The performance of kNN is best when intraclass distances are small compared to interclass distances.

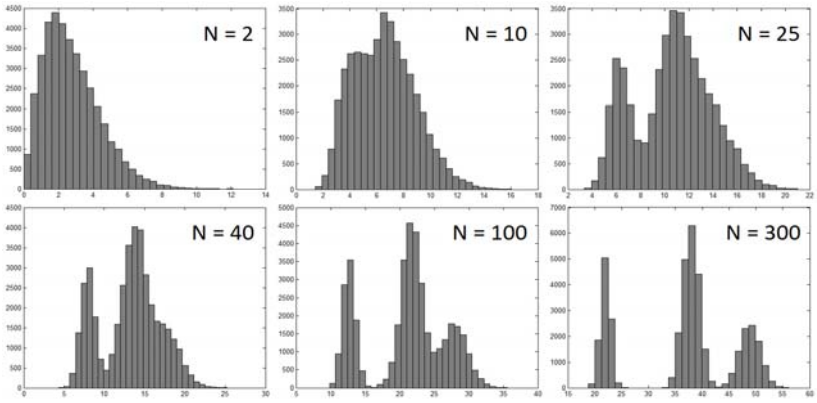

Fig. 2: Histograms of the pairwise distances for different values of N.

The 2-dimensional case in Figure 2 shows all three distributions overlapping. Individual histograms show that intraclass distances have their peak slightly more to the left than interclass distances. In the next plot, N=10, the shape shifts right and starts splitting up into two humps, the left one corresponding to class 1 intraclass distances and the other formed by class 2 intraclass distances and interclass distances. The increase of the distances is related to the increase of the diagonal of a hypercube when N grows and is caused by the nature of the high dimensional space.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The reason for class 1 distances to shift more slowly is their smaller standard deviation. At this point, the classification of class 1 members becomes easier because their intraclass distances remain small with respect to interclass ones. For class 2 the situation is similar as when N=2, so the overall result is an increase of the accuracy.

In the last four cases the distances keep growing as mentioned. Very important is the fact that class 2 distances increase faster than interclass distances. This results in a clear division of the three groups with increasing width and mean when going from class 1 to class 2, causing the classification of all points as class 1 members. In the case of an actual class 1 member, intraclass distances are much smaller than interclass ones, so the k chosen neighbours always belong to class 1. For class 2, interclass distances are smaller, resulting in a wrong choice of neighbours and a bad prediction.

The Fisher metric compensates this effect by resizing the space dimensions. On top of that, the MLP estimates $p(c|x)$ much better in high dimensions, so the result obtained is a very accurate classification. Also notice the stability of the percentages achieved with F-kNN when the parameter k varies for large values of N.

4 Conclusions

This paper outlines the construction process of the Fisher metric from the choice of the probability estimator to the development of a distance expression. Unique from any previous work, the Fisher information is derived from sigmoidal output estimators, and from this an analytical expression is obtained for the Fisher matrix.

In addition, a closed form expression for the geodesic distance is obtained by solving the path integral for a linear estimator. This opens the door to a distance expression for the general non-linear case by local linearization of the response surface of the MLP. Then the free points approach is introduced to find the geodesic between points with the new metric and also to ensure that the approximations hold.

References

[1] S. Kaski, J. Sinkkonen and J. Peltonen, Bankruptcy analysis with self-organizing maps in learning metrics, *IEEE Transactions on Neural Networks*, 12(4):936-947, 2001.

[2] J. Salojärvi, S. Kaski and J. Sinkkonen, Discriminative clustering in Fisher metrics. In *Artificial Neural Networks and Neural Information Processing – Supplementary proceedings* (ICANN/ICONIP 2003), June 26-29, Istanbul (Turkey), 2003.

[3] J. Peltonen, A. Klami and S. Kaski, Improved learning of Riemannian metrics for exploratory analysis, *Neural Networks*, 17:1087-1100, 2004.

[4] R. Kimmel and J. A. Sethian, Computing geodesic paths on manifolds, *Proceedings of the National Academy of Sciences of the USA*, 95(15):8431-8435, 1998.

[5] J. Baek, A. Deopurkar and K. Redfield, Finding geodesics on surfaces, unpublished, 2007.

[6] S. Kullback. *Information theory and statistics,* John Willey and Sons, New York, 1959.

[7] M. H. J. Gruber, Some applications of the Rao distance to shrinkage estimators, *Communications in Statistics – Theory and Methods*, 37:180-193, 2008.

[8] T. Jaakkola, M. Diekhans and D. Haussler, Using the Fisher kernel method to detect remote protein homologies, *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* (ISMB-99), 149-158, August 6-10, Heidelberg (Germany), 1999.