ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Visualizing Dependencies of Spectral Features using Mutual Information

Andrej Gisbrecht[1], Yoan Miche[2],
Barbara Hammer[1], and Amaury Lendasse[2,3] *

1- University of Bielefeld - CITEC Centre of Excellence, Germany
2- Aalto University - School of Science,
Department of Information and Computer Science, Finland
3- IKERBASQUE, Basque Foundation for Science, Spain

Abstract. The curse of dimensionality leads to problems in machine learning when dealing with high dimensionality. This aspect is particularly pronounced if intrinsically infinite dimensionality is faced such as present for spectral or functional data. Feature selection constitutes one possibility to deal with this problem. Often, it relies on mutual information as an evaluation tool for the feature importance, however, it might be overlaid by intrinsic biases such as a high correlation of neighbored function values for functional data. In this paper we propose to assess feature correlations of spectral data by an overlay of prior dependencies due to the functional nature and its similarity as measured by mutual information, enabling a quick overall assessment of the relationships between features. By integrating the Nyström approximation technique, the usually time consuming step to compute all pairwise mutual informations can be reduced to only linear complexity in the number of features.

1 Introduction

Modern technology has made cheap and precise sensors available for a broad use. This development resulted not only in large amounts of data samples, but also in an increased complexity of each sample. Very prominent examples of this observation can be found when considering spectral data such as mass spectrometry, NMR, or spectra of light. Since the underlying data is given as a function of wavelength to intensity in the limit, improved sensor technology has led to samples with multiple thousands of features, in which neighboring features are usually highly correlated to each other. This observation leads to significant problems for classical machine learning tasks due to the curse of dimensionality. One possible way to deal with this problem is to perform feature selection.

Often, feature selection (FS) relies on a quantitative measure of the usefulness of the features. Mutual information (MI) constitutes an information theoretic criterion which has been successfully applied for this task [5]. In the case of supervised FS, i.e. for labeled data, one can compute the MI between each feature and the target. Features can then be ranked based on these values and the top ones can be chosen for further processing. One problem with this approach is that the top ranked features can be highly correlated with each other carrying redundant information. Hence, instead of an iterative greedy selection of single features, it might be advisable to look for the MI between sets

*This work has been supported by the DFG under grant number HA 2719/7-1 and by the Cluster of Excellence 277 Cognitive Interaction Technology.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

of features and the target. Unfortunately, there are exponentially many subsets and the computation becomes infeasible, already basic settings being NP hard [11]. This gives rise to greedy approaches such as forward search [6].

In the unsupervised case there is the possibility to compute the MI between all pairs of features and cluster the most similar features into groups [7]. Each group can then be represented by a single feature, thus reducing the amount of used features significantly. It was suggested in [8], that this approach might not be sound for spectral data, where features from different parts of the spectrum should not be grouped together. Instead, one should regard only consecutive features by the means of a hierarchical clustering algorithm. Obviously, due to the functional nature of spectral data, a natural bias towards a grouping of neighbored bands is given. The question now occurs how this functional grouping of features correlates to groupings based on MI, and how feature selection schemes can be based on this accumulated information.

In this contribution we present a first step in this direction by introducing an unsupervised visualization technique which allows to assess the relationship between the features based on their functional nature and pairwise mutual information quickly. This offers an interface based on which an expert can then derive a suitable approach for feature selection and feature grouping. The idea is to compute the MI between all pairs of features resulting in a quadratic similarity matrix and to visualize this matrix using e.g. classical multidimensional scaling (CMDS) [12]. This is overlaid by the spectral ordering, so that one can directly inspect whether consecutive features are correlated with each other under the MI criterion. Since the construction of the matrix is quadratic in the number of features, it might become problematic for high dimensional data. Thus, we propose to use the Nyström approximation to reduce the complexity.

Now we first give a short description of MI estimation, based in which the visualization framework is introduced. Then the Nyström approximation to reduce the computational effort is explained, and tested in two experiments.

2 Mutual information

Let X and Y be two random variables and denote $\mu_{X,Y}$ the joint probability density function of X and Y. The marginal density functions are given by $\mu_X(x) = \int \mu_{X,Y}(x,y)dy$ and $\mu_Y(y) = \int \mu_{X,Y}(x,y)dx$. The uncertainty on Y is then given by its entropy $H(Y)$ defined as $H(Y) = -\int \mu_Y(y) \log \mu_Y(y)dy$. If knowledge on Y is obtained indirectly by knowing X, the uncertainty on Y knowing X is given by the conditional entropy $H(Y|X)$ defined as

$$H(Y|X) = -\int \mu_X(x) \int \mu_Y(y|X=x) \log \mu_Y(y|X=x)dxdy.$$

The mutual information $I(X,Y)$ between X and Y can be considered as a measure of the amount of knowledge on Y provided by X [10]:

$$I(X,Y) = H(Y) - H(Y|X),$$

which is exactly the reduction of the uncertainty of Y when X is known.

Since estimating the marginal density functions μ_X and μ_Y is not trivial in practical cases, Kraskov *et al.* in [9] propose to use a k-nearest neighbor approach to estimate directly the mutual information.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Assuming metrics are given for each of the spaces spanned by X, Y and using the maximum norm on $Z = (X, Y)$ ($\|z - z'\|_{\max} = \max\{\|x - x'\|, \|y - y'\|\}$), then for every point $z_i = (x_i, y_i) \in Z, 1 \leq i \leq N$, its neighbors can be sorted according to their distance to the considered point. Denote then by $\varepsilon_i^Z/2$ the distance from the considered point z_i to its k-th neighbor and by $\varepsilon_i^X/2$ and $\varepsilon_i^Y/2$ the distances between z_i and its k-th neighbor, but projected in the X and Y subspaces, respectively denoted x_i and y_i.

The idea of the first Kraskov estimator $I^{(1)}$ of the mutual information, is then to count the number $n_i^x$of points whose distance from x_i is less than $\varepsilon_i^Z/2$, as well the number n_i^y of points whose distance from y_i is also less than $\varepsilon_i^Z/2$.

Given these, the first estimator $I^{(1)}$ of the mutual information between random variables X and Y proposed by Kraskov is given as

$$I^{(1)}(X, Y) = \psi(k) - \frac{1}{N}\sum_{i=1}^{N} E\left[\psi(n_i^x + 1) + \psi(n_i^y + 1)\right] + \psi(N),$$

where $\psi(x) = d\ln(\Gamma(x))/dx$ denotes the digamma function.

The choice of k remains non-trivial, and while Kraskov in [9] proposes the heuristic of $k = 6$, it is likely that the choice of the optimal k remains application-specific. For more details the reader is referred to the original article [9].

3 Visualization of spectral feature correlations

Spectral data have a functional form, which is characterized by a real-valued function $t \to x(t)$ mapping wavelengths (or similar such as time or mass) to intensity. Depending on the resolution of the spectrometric instruments, measurements result in samples $(x_1, \ldots, x_n)$ for n values of the index parameter t. Obviously, it is possible to estimate the pairwise MI of these features based on the Kraskov estimator. In addition, one can assume smoothness of the mapping $t \to x(t)$, such that neighbored features x_i and x_{i+1} are likely correlated.

We propose to overlay these two information sources in the following way: The possibly complex relation of features as measured by MI can be visualized by referring to the full MI matrix and displaying a low dimensional approximation by referring to classical (possibly nonlinear) embedding techniques for the matrix [12]. Here we will use classical multi-dimensional scaling (CMDS). Correlations as induced by functional smoothness follow a linear principle only, such that its display is easily possible by connecting neighbored features with lines. One potential problem of this approach consists in the complexity to compute the Kraskov estimator. Here, approximation techniques such as the Nyström technique which reduce the effort to linear time, might be beneficial.

4 Nyström approximation

The Nyström approximation is a powerful technique to approximate positive semi-definite (psd) matrices. It was proposed for kernel methods in machine learning by [1]. We give a short review. By the Mercer theorem a psd kernel

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

$k(\mathbf{x}, \mathbf{y})$ can be expanded by orthonormal eigenfunctions ϕ_i and eigenvalues λ_i

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y}). \tag{1}$$

The number of non zero eigenvalues can be infinite, but typically, since the kernels are given by a matrix $\mathbf{K}$, it is finite and given by the rank of this matrix. The eigenequation of a kernel $\int k(\mathbf{y}, \mathbf{x}) \phi_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda_i \phi_i(\mathbf{y})$ defines eigenfunctions and eigenvalues and can be approximated by sampling $\mathbf{x}_k$ from $p(\mathbf{x})$:

$$\frac{1}{m} \sum_{k=1}^{m} k(\mathbf{y}, \mathbf{x}_k) \phi_i(\mathbf{x}_k) \approx \lambda_i \phi_i(\mathbf{y}).$$

This equation together with the matrix eigenproblem $\mathbf{K}^{(m)} \mathbf{U}^{(m)} = \mathbf{U}^{(m)} \mathbf{\Lambda}^{(m)}$ of the $m \times m$ Gram matrix $\mathbf{K}^{(m)}$ can be used to derive the approximations:

$$\lambda_i \approx \frac{\lambda_i^{(m)}}{m}, \quad \phi_i(\mathbf{y}) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}} \mathbf{k}_y \mathbf{u}_i^{(m)},$$

where $\mathbf{u}_i^{(m)}$ is the ith column of $\mathbf{U}^{(m)}$. Now we can plug this approximation into the equation 1, which allows us to compute $k(\mathbf{y}_1, \mathbf{y}_2)$, if we know the vectors $\mathbf{k}_{y_1} = (k(\mathbf{x}_1, \mathbf{y}_1), ..., k(\mathbf{x}_m, \mathbf{y}_1))^T$ and $\mathbf{k}_{y_2}$ similarly. This results in the approximated kernel matrix $\tilde{\mathbf{K}} = \sum_{i=1}^{m} 1/\lambda_i^{(m)} \mathbf{K}_{n,m} \mathbf{u}_i^{(m)} (\mathbf{u}_i^{(m)})^T \mathbf{K}_{m,n}$, where we write $\mathbf{K}_{m,n}$ for $k(\mathbf{x}_i, \mathbf{y}_j)$ with $i = 1..m$ and $j = 1..n$. This corresponds to the part of the kernel matrix consisting of m rows and n columns, which evaluates the kernel between m sampling points, called landmarks, and n points. We can simplify this even further:

$$\tilde{\mathbf{K}} = \mathbf{K}_{n,m} \mathbf{K}_{m,m}^{-1} \mathbf{K}_{m,n},$$

where $\mathbf{K}_{m,m}^{-1}$ denotes the Moore-Penrose pseudoinverse of the matrix $\mathbf{K}^{(m)}$.

In practice, for n given points, to approximate the $n \times n$ matrix, we only need to sample m landmarks and compute the kernel between the landmarks and all n points. This reduces the complexity from $O(n^2)$ to $O(mn)$ for evaluation of $\mathbf{K}_{n,m}$ plus $O(m^3)$ for inversion of $\mathbf{K}_{m,m}$, leading to linear complexity in the number of data points. Note, that there exist many different sampling techniques to improve the quality of approximation ([3], [4]) but for simplicity we use random sampling. It should be mentioned, that the Nyström approximation was proposed for psd matrices, but for MI there is no guarantee that the resulting matrix will be psd. Fortunately, as stated in [2], the Nyström technique can be applied for an arbitrary symmetrical matrix.

5 Experiments

To demonstrate the usability of the proposed technique we report experimental results on two data sets. Both data sets were provided by Prof. Marc Meurens, Université catholique de Louvain, BNUT unit.

oj - depicts near-infrared spectra of orange juice together with the level of saccharose. There are 150 different spectra each sampled at 700 positions.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

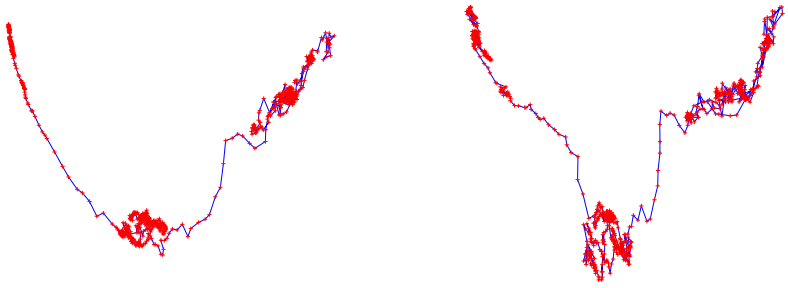

Fig. 1: Visualization of the features for **oj** data set. The mutual information matrix was fully computed (left) or approximated (right). Crosses denote features and the lines connect neighbouring features in the spectrum.

wine - depicts the mean infrared spectrum and the corresponding level of alcohol. The data set consists of 94 different wines and the spectrum is sampled at 256 positions.

The MI was computed for all features, resulting in a D-by-D square matrix, or, via the Nyström technique using 10 landmarks. Afterwards the matrices are projected into two dimensions using CMDS. Any different visualization technique could be used as well, but we used a simple linear method, since it already preserves 79% and 73% of variance in the data, for **oj** and **wine** respectively. The projected points are overlaid by connecting functional neighbors.

The features of **oj** show clear structure (Fig. 1). Features which are close in the spectrum are also close in the projection. Hence different parts of the spectrum provide different information about the data. Still, some features are grouped into clusters and each cluster represents a spectral band. The Nyström approximation is able to retain the overall structure of the data and reduces the run time from 1452 to 21.2 seconds.

The features of **wine** (Fig. 2) seem to be closely related to each other. They build a tight cluster in the center with a few outliers. The same image results using approximated MI with a run time of 21.2 instead of 131.6 seconds. The features in the cluster seem to be very strongly correlated to each other and thus they can be represented by only a few features. Depending on the goal one could now focus on the center cluster or outliers, observable in the display.

6 Conclusion

We presented a technique which can be used as a first tool to quickly analyze a new spectral data set, to assess the structure of the feature correlations and to discover possibilities for feature selection. Our method is based on the MI, which is widely used in the field to select features in supervised as well as in unsupervised scenarios. Although the presented approach is unsupervised, it is possible, as shown in [6], to compute the pairwise MI with regard to the auxiliary information, thus turning it into a supervised technique.

From the experiments we can see, that the Nyström approximation on MI

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

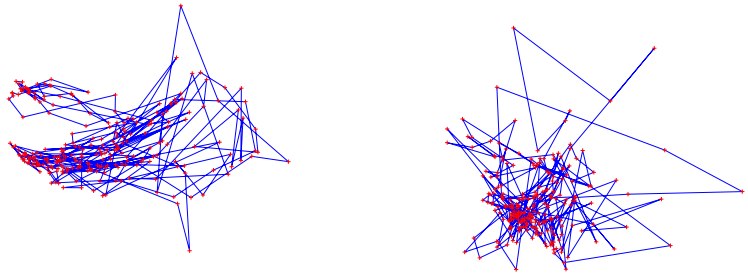

Fig. 2: Visualization of the features for **wine** data set. The mutual information matrix was fully computed (left) or approximated (right). Crosses denote features and the lines connect neighbouring features in the spectrum.

matrices is feasible, since it does not lose too much information, at the same time reducing the complexity from a quadratic to linear one (note that it can be directly integrated into CMDS such that the full setting is effectively linear.) This is especially crucial for MI, which is slow to compute such that approximations are mandatory if thousands of features should be judged this way.

References

[1] C. K. I. Williams, M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems 13*: 682-688, 2001.

[2] A. Gisbrecht, B. Mokbel, and B. Hammer. The Nystrom approximation for relational generative topographic mappings. In *NIPS workshop on challenges of Data Visualization*, 2010.

[3] S. Kumar, M. Mohri and A. Talwalkar. Sampling Methods for the Nyström Method. *J. Mach. Learn. Res.*: 981–1006, 2012.

[4] K. Zhang and J. T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *Trans. Neur. Netw.*, 21(10): 1576–1587, 2010.

[5] G. D. Tourassi, E. D. Frederick, M. K. Markey and C. E. Floyd, Jr. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28(12): 2394–2402, 2001.

[6] M. Verleysen, F. Rossi and D. François. Advances in Feature Selection with Mutual Information. *Similarity-Based Clustering*, Th. Villmann, M. Biehl, B. Hammer, M. Verleysen (Ed.): 52–69, 2009.

[7] G. Van Dijk and M.M. Van Hulle. Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis. *International Conference on Artificial Neural Networks*, 28(12): 31–40, 2006.

[8] C. Krier, D. François, F.Rossi and M. Verleysen. Feature clustering and mutual information for the selection of variables in spectral data. *European Symposium on Artificial Neural Networks*: 157-162, 2007.

[9] A. Kraskov, H. Stögbauer and P. Grassberger. Estimating mutual information. *Physical Review E*, 69: 066138, 2004.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, 1991.

[11] B. Hammer, T. Villmann, Input pruning for neural gas architectures, in: M.Verleysen (ed.), European Symposium on Artificial Neural Networks'2001, D-facto publications, 283-288, 2001.

[12] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.