Advanced learning in massive fusion databases: nonlinear regression, clustering, dimensionality reduction and information retrieval

G. Verdoolaege, G. Van Oost

Department of Applied Physics, Ghent University, Ghent, Belgium

Introduction

Present-day fusion devices generate large amounts of data that are stored in massive databases. Pattern recognition techniques are very useful for learning data structures of interest directly from the data, either off-line or in real time. In the present paper we propose a unified framework for discovering or retrieving data structures based on two pillars: probability and geometry.

Pattern recognition for fusion data

The term pattern recognition encompasses several concepts. *Regression* refers to learning a—possibly nonlinear—relation between variables. *Clustering* and *classification* are used to group data points according to similarity of their characteristics. Searching a database for a pattern in a given query is called *information retrieval*. All of these tasks basically require a *similarity measure* between data points.

Pattern recognition for fusion data is hampered by several data characteristics. First, the databases are vast, therefore learning algorithms need to work sufficiently fast. Second, the dimensionality of the data space is often large, causing learning algorithms to perform poorly. Data dimensionality reduction is essential and can be used for visualization purposes as well. Third, there is a considerable redundancy between measured quantities due to complex, *nonlinear* interactions. Finally, the measurements are often subject to substantial uncertainty, both statistical (e.g. measurement noise) and systematic. In this work, we describe an integrated framework that tackles these various challenges.

Manifold learning

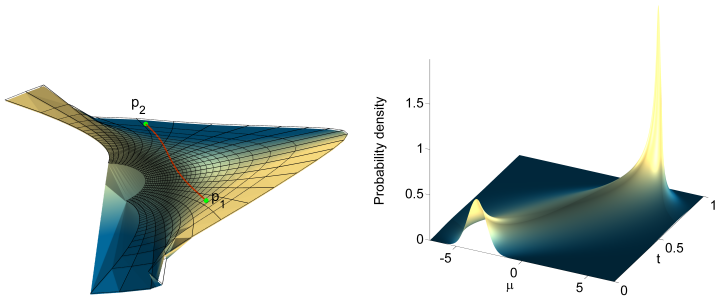
Data manifolds

Measurements may be represented as data points in a multidimensional Euclidean space. The next step in the learning process is to recognize that the data often are not merely randomly distributed throughout this space, but lie scattered (due to the statistical uncertainty) around one or more *manifolds*, in general nonlinear, of reduced dimensionality embedded in the Euclidean space. This is referred to as the concept of (data) *manifold learning*. A very simple example is linear regression. The intrinsic geometry of the manifold can be learned for instance by calculating geodesic distances between the data points [1]. Often a coordinate system can be found

on the manifold that is related to the underlying physical degrees of freedom that independently govern the dynamical behavior of the system. Hence, manifold learning can contribute substantially to the physical understanding of the system. In addition, usually pattern recognition tasks are considerably more effective on the manifold. Thus, data manifold learning simultaneously addresses two of the difficulties with fusion data mentioned above: reducing data dimensionality and redundancy in a natural way.

Probabilistic manifolds

The fundamental object in the measurement process is a probability distribution for the measured quantity (mostly a voltage over a sensor). In the field of *information geometry*, probability density families are interpreted as differentiable manifolds [2]. A point on the manifold corresponds to a specific PDF within the family and the family parameters provide a coordinate system on the manifold. The Fisher information provides a metric tensor allowing the calculation of geodesic distances on the manifold. We will show that modeling the data uncertainty in this way provides a distinct advantage for pattern recognition tasks. Learning of probabilistic manifolds may be combined with regular (data) manifold learning by studying the submanifold spanned by the data on the probabilistic manifold.



(b)

Figure 1: (a) Embedding of the univariate Gaussian manifold and geodesic between two arbitrary Gaussians p_1 ($\mu_1 = -4$, $\sigma_1 = 0.7$) and p_2 ($\mu_2 = 3$, $\sigma_2 = 0.2$). (b) Visualization of the distributions along the geodesic, showing the change in mean and standard deviation as a function of the parameter t along the geodesic.

Confinement regime identification

We now apply the concepts of probabilistic and data manifolds to the identification of plasma confinement regimes. Particularly real-time regime classification will be important for ITER. We demonstrate the performance of our classification method using data from the ITPA Global H Mode Confinement Database (DB3) [3]. The only plasma parameters that were examined in this work in order to differentiate between L and H mode discharges were the central line-averaged electron density n_e and the total power loss P_{loss} from the plasma.

The key element in our analysis is the modeling of the in the database mentioned error bars for n_e and P_{loss}, which are approximate estimates. In line with the principle of maximum entropy we identify the measurement value itself with the mean of a Gaussian distribution and the error bar with the standard deviation. If the noise on n_e and P_{loss} is considered independent, then the total data distribution is given by the product of two univariate Gaussians. By way of illustration, the two-dimensional univariate Gaussian manifold is shown embedded in Euclidean space in Figure 1a. A geodesic between two arbitrary Gaussians is visualized and the distributions along this geodesic are drawn in Figure 1b.

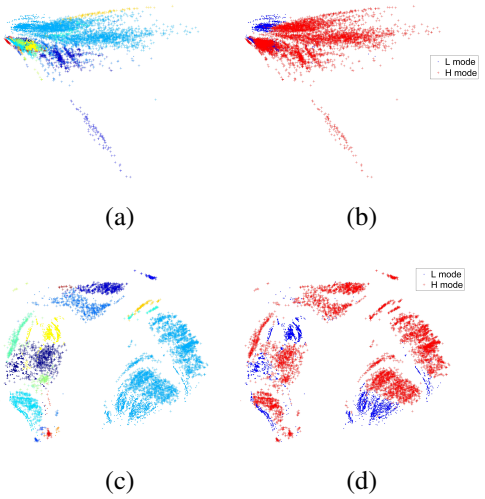


Figure 2: Two-dimensional embeddings of the DB3 data using the Euclidean distance without measurement error ((a), per machine and (b), per confinement mode) and using the geodesic distance with measurement error ((c) and (d)).

Figure 2 shows a series of two-dimensional projections of the DB3 data. The projections obtained with the geodesic distance, which take into account the measurement error, clearly exhibit much more structure and more clear clustering of machines and confinement modes compared to the ones calculated via the Euclidean distance without consideration of the error bars. A similar three-dimensional projection of the DB3 data is displayed in Figure 3, showing a complicated data geometry. Geodesic distances on this manifold would thus have to be computed numerically. Therefore in the sequel we only employed the probabilistic geometrical information.

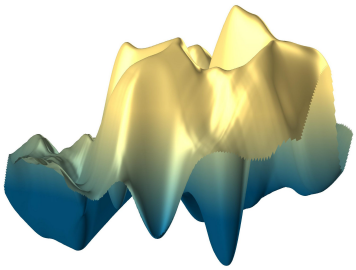


Figure 3: Three-dimensional embedding of the DB3 data.

We next performed a series of classification experiments with two classes (L and H mode) using 5% of the data for training. We first carried out k-nearest neighbor (kNN) classification ($k = 1$), the results of which are shown in Table 1. Next a support vector machine (SVM) algorithm was used with a Gaussian kernel (optimized standard deviation), see Table 1. Both experiments were performed once without and once with consideration of the measurement error. The results are clearly better if the measurement error is considered, even using the Euclidean distance. The best results are obtained with the geodesic distance, since it properly takes into account the geometry of the probabilistic manifold. It is remarkable that even this approximate and limited

	kNN			SVM	
Mode	Eucl. w/o err.	Eucl. with err.	GD with err.	W/o err.	With err.
L	85.1	87.7	91.0	85.6	93.1
H	88.6	89.4	93.0	89.1	96.6

Table 1: Correct classification rates (%) using a kNN and an SVM classifier.

information on the underlying probability distribution is beneficial to the classification task. Additionally learning the spatial distribution of the data points (regular manifold learning) on the probabilistic manifold may still improve the classification results.

Conclusion and outlook

We have discussed some of the difficulties related to pattern recognition from fusion data and we have proposed the technique of probabilistic and data manifold learning to address these issues. The identification of confinement regimes via classification has been shown to clearly benefit from information on the measurement uncertainty. The next step in this research program is to include statistical information in the wavelet domain of plasma time series. Geodesic distances will be calculated on wavelet distribution manifolds, allowing fast calculation of geodesic distances for the purpose of dimensionality reduction and pattern recognition for fusion data [4]. Real-time applications are envisaged to confinement mode identification employing the D_α time series as well as to disruption prediction.

References

[1] J.B. Tenenbaum, V. de Silva, J.C. Langford, Science **290**, 2319 (2000)

[2] R.E. Kass, P.W. Vos, 'Geometrical Foundations of Asymptotic Inference', Wiley (1997)

[3] http://efdasql.ipp.mpg.de/HmodePublic

[4] G. Verdoolaege, P. Scheunders, 'Geodesics on the Manifold of Multivariate Generalized Gaussian Distributions With an Application to Multicomponent Texture Discrimination', International Journal of Computer Vision, accepted (2011)