Signal detection in high energy physics via a semisupervised nonparametric approach*

Individuazione di un segnale fisico mediante un approccio non parametrico semi-supervisionato

Alessandro Casa and Giovanna Menardi

Abstract In particle physics, the task of identifying a new signal of interest, to be discriminated from the background process, shall be in principle formulated as a clustering problem. However, while the the signal is unknown, usually even missing, the background process is known and always present. Thus, available data have two different sources: an unlabelled sample which might include observations from both the processes, and an additional labelled, sample from the background only. In this context, semisupervised techniques are particularly suitable to discriminate the two class labels; they lies between unsupervised and supervised ones, sharing some characteristics of both the approaches. In this work we propose a procedure where additional information, available on the background, is integrated within a nonparametric clustering framework to detect deviations from known physics. Also, we propose a variable selection procedure that allows to work on a reduced subspace.

Abstract Nell'ambito della fisica delle particelle la ricerca di un segnale di interesse, che si manifesta come una deviazione dal processo di background, può essere formulata in termini di problema di raggruppamento. Tuttavia, mentre la presenza del segnale non è certa, lo è quella del background, che rappresenta un processo noto. Nelle analisi empiriche, si dispone non solo di dati non etichettati, che potrebbero contenere segnale, ma anche di un campione di dati etichettati, provenienti dal solo processo di background. Ha senso allora adottare un approccio semisupervisionato, che si colloca a metà strada tra i metodi supervisionati e non. In questo lavoro si propone una procedura che integra l'informazione aggiuntiva a disposizione a tecniche di clustering non parametrico per individuare deviazioni dalle teorie fisiche esistenti. Viene inoltre proposta una procedura di selezione delle variabili che permette di operare su un sotto-spazio ridotto.

Alessandro Casa, Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova
via C. Battisti 241, 35121, Padova; e-mail: casa@stat.unipd.it, menardi@stat.unipd.it

*This report is part of a project that has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement 675440. The authors wish to thank Dr. T. Dorigo, from the National Institue of Nuclear Physics (INFN) for providing the data.

Key words: high energy physics, nonparametric clustering, semisupervised classification

1 Introduction

Since the early Sixties, the *Standard Model* has represented the state of the art in High Energy Physics. It describes how the fundamental particles interact with each others and with the forces between them, giving rise to the matter in the universe. Despite its empirical confirmations, there are indications that the Standard Model does itself not complete our understanding of the universe. Model independent searches aim to explain the shortcomings of this theory by empirically looking for any possible *signal* which behaves as a deviation from the *background* process, representing, in turn, the known physics.

The considered problem can be recasted to a classification framework, although of a very peculiar nature. While the background process is known and a sample of virtually infinite size can be drawn from it, the signal process is unknown, possibly even missing. Available data have, consequently, two different sources: a first, labelled, sample from the background class only, and a second, unlabelled sample which might include observations from the signal. A semisupervised perspective [2] shall be then adopted, either by relaxing assumptions of supervised methods, or by strenghtening unsupervised clustering structures through the inclusion of additional information available from the labelled data.

In [5], the problem has been faced by building on a suitable adaptation of parametric density-based clustering to the semisupervised framework, according to the same logic of anomaly detection tasks. In this work we follow a similar route, yet in a nonparametric guise. Such formulation appears consistent with the physical notion of signal, *i.e.* a new particle would manifest itself as a peak emerging from the background process. Nonparametric *-modal-* clustering, in turn, draws a correspondence between groups and the modal peaks of the density underlying the observed data. Thus, the one-to-one relationship between clusters and modes of the distribution would provide an immediate physical meaning to the detected clusters.

The main idea underlying this work is to semisupervise nonparametric clustering by exploiting information available from the background process. Specifically, we tune a nonparametric estimate of the unlabelled data by selecting the smoothing amount so that the induced modal partition will classify the labelled background data as accurately as possible. As a side contribution we propose a variable selection procedure, specifically conceived for this framework, linked to the concept of stability of the distribution underlying the data.

We adopt the following notation: $\mathscr{X}_b = \{\mathbf{x}_i\}_{i=1,\dots,n_b}$ denotes the set of labelled data, supposed to be a sample of *iid* multidimensional observations from the background distribution f_b. Since the background is known and well explained by the existing physical theories, we may assume n_b to be as large as needed to estimate f_b arbitrarily well. $\mathscr{X}_{bs} = \{\mathbf{x}_i\}_{i=1,\dots,n_{bs}}$ has the same structure as $\mathscr{X}_b$ and denotes

the unlabelled set of data, assumed to be drawn from the distribution f_{bs} underlying the whole process. We assume that f_{bs} and f_b could be different just because of the presence of a signal which features as a new mode of f_{bs}, not arising from f_b.

2 The statistical framework

According to the nonparametric formulation of density-based clustering, the observed data $\mathscr{X} = \{\mathbf{x}_i\}_{i=1,\dots,n}$, $\mathbf{x}_i = (x_{i1},\dots,x_{ij},\dots,x_{id})' \in \mathbb{R}^d$ are supposed to be a sample from a random vector with unknown probability density function f, whose modes are regarded as the archetypes of the clusters, in turn represented by the surrounding regions. After building a nonparametric estimate $\hat{f}$ of f, the identification of the modal regions may occur according to different directions. One strand of methods looks for an explicit representation of the modes of f and associates each cluster to the set of points along the steepest ascent path towards a mode, e.g. via the mean-shift algorithm. A second class of methods does not attempt explicitly the task of mode detection but associates the clusters to disconnected density level sets of the sample space, as the modes correspond to the innermost points of these sets. See [4] for a review of these approaches.

Whatever direction is followed, any estimate of f leaves defined the modal structure and hence the clustering. However, nonparametric density estimation is a critical task, at least with respect to two aspects. First, the shape and the number of modes of the density estimate depend on the regulation of some smoothing parameter, whatever estimator is chosen. While not binding, in the rest of the paper, we focus on the specific case of product kernel estimator:

$$\hat{f}(\mathbf{x};\mathscr{X},h) = \frac{1}{n\cdot h^d}\sum_{i=1}^{n}\prod_{j=1}^{d}K\left(\frac{x_j - x_{ij}}{h}\right), \tag{1}$$

where K is the kernel, usually a symmetric probability density function, and $h > 0$ is the bandwidth. A large bandwidth tends to oversmooth the density, possibly pulling out its modal structure, while a small bandwidth favours the appearance of spurious modes. How to set the amount of smoothing is then an issue to be tailored.

A second critical aspect related to density estimation, and worth to be accounted for, is the dimensionality of the problem at hand. The curse of dimensionality is known to have a strong impact on nonparametric density estimators and this explains a worsened behaviour of modal clustering for increasing d. In high dimensions, much of the probability mass flows to the tails of the density, possibly giving rise to the birth of spurious clusters and averaging away features in the highest density regions. Since a typical aspect of high dimensional data is the tendency to fall into manifolds of lower dimension, dimension reduction methods are often advisable.

3 A nonparametric semisupervised approach

Our contribution, to include a source of supervision in nonparametric clustering, builds on the idea of exploiting available information on the known labelled process to aid the most critical aspects of the nonparametric framework, i.e. density estimation in high dimensions and selection of the smoothing amount.

To address the former issue, here we propose a variable selection approach specifically formulated for this context. The procedure is based on the idea that a possible different behavior between f_b and f_{bs} shall be only due to the presence of a signal of interest in f_{bs}; hence, a variable will be considered to be relevant if it contains any trace of signal. The approach here adopted consists in comparing repeatedly the estimates of multivariate marginal distributions of f_b and f_{bs}, at each step on a different, randomly selected subset of variables. In this way we operate in lower dimensional spaces, with a gain in density estimation accuracy, while accounting for relations among variables. The comparison is based on the use of the non parametric statistic [3] to test equality of two distributions. If a different behavior is detected, the procedure updates a counter for the variables selected at that step; at the end of the procedure the counter will indicate the relevance of each single variable. The procedure allows for selecting a smaller subset of variables to work with, leading both to interpretative and computational advantages.

To address the second critical aspect discussed above, we propose a procedure whose rationale is the following. We identify the modal partition of the unlabelled data associated with the nonparametric estimate $\hat{f}_{bs}$ which guarantees the most accurate classification of the labelled background observations. Given an estimate $\hat{f}_b$ of f_b, supposed to be arbitrarily accurate due to our knowledge of the background process, a partition $\mathscr{P}_b(\mathscr{X}_b)$ of the background data remains associated. Then, we get multiple estimates $\hat{f}_{bs}(\cdot;\mathscr{X}_{bs},h_{bs})$ of f_{bs} for h_{bs} varying in a range of plausible values. Each of these estimates identifies a partition $\mathscr{P}_{bs}(\mathscr{X}_{bs})$ and, eventually, also a partition $\mathscr{P}_{bs}(\mathscr{X}_b)$ of the background data, both defined by the modal regions of $\hat{f}_{bs}$. The latter classification is obtained by assigning a background observation to the cluster of $\hat{f}_{bs}$ for which its density is the highest. $\mathscr{P}_{bs}(\mathscr{X}_b)$ is then compared with $\mathscr{P}_b(\mathscr{X}_b)$ via the computation of some agreement index I. The bandwidth h_{bs} that maximizes I is then selected to estimate f_{bs} and identify the ultimate partition $\mathscr{P}_{bs}(\mathscr{X}_{bs})$. The main steps of the procedure are listed in the Pseudo-algorithm 1.

From an operational point of view we use, to obtain partitions, the clustering method [1] and, as agreement index, the *Adjusted Rand Index*. Furthermore, $\mathscr{P}_{bs}(\mathscr{X}_b)$ and $\mathscr{P}_b(\mathscr{X}_b)$ are not, in fact, compared on the whole background sample $\mathscr{X}_b$ but on a number of different bootstrap samples from $\mathscr{X}_b$; this allows us to get the empirical distribution of the agreement index I and obtain more reliable results.

Eventually we note that, besides the background process is known and $\mathscr{X}_b$ is arbitrarily large, the procedure presented above requires an estimate of the background density f_b, i.e. the relative choice of the bandwidth. To this aim we rely on the concept of stability of the density estimate and select the bandwidth that minimizes the *integrated squared distance* among density estimates computed from different samples drawn from the background process.

Pseudo-algorithm 1 *Semisupervised procedure for bandwidth selection*
Denote with: $\mathscr{X}_b$ the *background* sample, $\mathscr{X}_{bs}$ the *unlabelled* sample from the whole process; it is assumed that the dimensionality of both samples has been already reduced via variable selection. Let h_b be the *background* bandwidth; h_{bs} the whole process bandwidth (to be determined); h_{grid}: a grid of plausible values for h_{bs}. Finally let $\mathscr{P}_k(\mathscr{X})$ be a partition of data $\mathscr{X}$ identified by the modal structure of density f_k and $I(\mathscr{A},\mathscr{B})$ an agreement index between partitions $\mathscr{A}$ and $\mathscr{B}$

Input $\mathscr{X}_b$, $\mathscr{X}_{bs}$, h_b, h_{grid}.
compute $\hat{f}_b(\cdot;\mathscr{X}_b,h_b)$;
obtain $\mathscr{P}_b(\mathscr{X}_b)$;
for h in h_{grid} **do**
    compute $\hat{f}_{bs}(\cdot;\mathscr{X}_{bs},h)$;
    obtain $\mathscr{P}_{bs}(\mathscr{X}_b)$;
    compute $I(\mathscr{P}_b(\mathscr{X}_b),\mathscr{P}_{bs}(\mathscr{X}_b))$.
end for
$h_{bs} = argmax_{h\in hgrid} I(\mathscr{P}_b(\mathscr{X}_b),\mathscr{P}_{bs}(\mathscr{X}_b))$
compute $\hat{f}_{bs}(\cdot;\mathscr{X}_{bs},h_{bs})$;
obtain $\mathscr{P}_{bs}(\mathscr{X}_{bs})$;

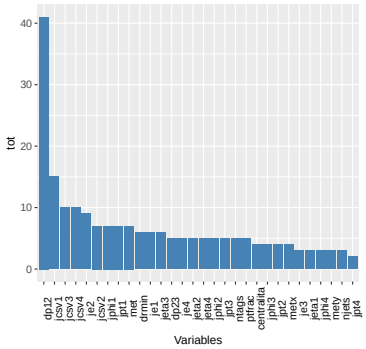
Output: $\mathscr{P}_{bs}(\mathscr{X}_{bs})$.

4 Empirical results

In this section, we show the results of the application of the proposed procedure on a *Monte-Carlo* physical process simulated within the CMS experiment; the experiment refers to high-energy proton-proton collisions where each observation corresponds to a single collision event and may produce particles from two different physical processes: the *QCD multijet background*, and a signal known as *top pair production*. $\mathscr{X}_b$ includes $n_b = 20000$ background observations, while $\mathscr{X}_{bs}$ include $n_{bs} = 10000$ observations, whose the 16% comes from the signal process. For each dataset we observe $d = 30$ variables related to the kinematic characteristics of the particles produced by the proton collisions. While both $\mathscr{X}_b$ and $\mathscr{X}_{bs}$ are labelled, labels of $\mathscr{X}_{bs}$ have been employed only for evaluating the quality of the results.

In Figure (1) the results of the variable selection procedure are displayed. Two features (*dp12* and *jcsv1*) show a remarkably different behavior between the background and whole process densities. In the subsequent analyses we have worked with these two variables only.

After estimating f_b based on a bandwidth selected to guarantee the density stability as explained in the previous section, we applied the procedure reported in Pseudo-algorithm 1. The obtained bandwidth was used to estimate the density f_{bs} of the whole process and thus obtaining a partition of $\mathscr{X}_{bs}$ via the subsequent application of nonparametric clustering. Results, reported in the right table of Figure 1, compare the obtained partition with the known actual labels. The procedure identifies four different groups: two of them clearly refer to the background process, while the other two mostly contain observations coming from the *top pair production* sig-



	Clusters			
Label	1	2	3	4
Bkg	0.410	0.421	0.003	0.004
Sgn	0.008	0.011	0.088	0.055

	Clusters									
Label	1	2	3	4	5	6	7	8	9	10
Bkg	0.056	0.180	0.066	0.073	0.097	0.099	0.092	0.033	0.139	0.004
Sgn	0.003	0.002	0.004	0.016	0.002	0.002	0.006	0.003	0.003	0.120

Fig. 1: Top left: results of variable selection procedure; variables are ordered decreasingly by importance (higher bar implies higher importance). Top right: true process labels vs clusters detected by the proposed semisupervised procedure. Bottom: true process labels vs clusters detected by the benchmark method [5].

nal. The overall misclassification error is equal to 2.6% with a true positive rate larger than 88%. For comparison purposes we also present results of the application of the competitive methodology proposed in [5]. Data dimensionality have been previously reduced by keeping four principal components, as proposed by the authors. Working in a parametric framework there is one-to-one relationship between mixture components and clusters; hence the method find 9 background clusters and an additional one capturing the signal. The overall error is equal to 4.5% with a true positive rate amounting to the 74.5%.

References

1. Azzalini, A., Torelli N.: Clustering via nonparametric density estimation. Stat Comput, 17(1): 71-80 (2007).
2. Chandola, V., Banerjee A., Kumar V.: Anomaly detection: A survey. ACM Comput Surv 41(3): 1-58 (2009).
3. Duong, T., Goud, B., Schauer, K.: Closed-form density-based framework for automatic detection of cellular morphology changes. P Natl Acad Sci Usa, 109(22): 8382-8387 (2012).
4. Menardi, G.: A review on modal clustering. Int Stat Rev, 84(3): 413-433 (2016).
5. Vatanen, T., Kuusela, M., Malmi, E., Raiko T., Aaltonen T., Nagai, Y.: Semi-supervised detection of collective anomalies with an application in high energy particle physics. Int Jt Conf Neural Netw: 1-8 (2012).