International Journal of Modern Physics A
© World Scientific Publishing Company

LHC DATA CLASSIFICATION USING A NEW MORPHOLOGICAL BOUNDARY DETECTION

MOSTAFA MJAHED
Ecole Royale de l'Air, Maths and Systems Dept
40000 Marrakech, Morocco
also at LPTN, Faculty of Sciences Semlalia, 40000 Marrakech, Morocco
mmjahed@hotmail.com

A new morphological boundary detection approach is used to separate the signal from the background in the Standard Model Higgs boson search at LHC. Based on mathematical concepts, this method consists of a fast computation of probabilistic density functions of events and a smoothing using a combination of dilatation and erosion operators. In a binary search approach, the performances are improved and the results compare favourably with other multivariate analysis.

Keywords: Morphological boundary; Classification; Higgs.

PACS numbers: 11.25.Hf, 123.1K

1. Introduction

Cluster analysis consists in partitioning a collection of data points into a number of groups, where the objects, inside a cluster, show a relatively high degree of closeness.

Statistical clustering algorithms [1-3] are usually classified according to the method they use to find clusters within the data set X. In a Hierarchical clustering, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. The end result is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained.

A Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters. A commonly used criterion function is the average squared distance of the data items from their nearest cluster centroids. This criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure.

Many statistical clustering approaches have been developed based on fundamental assumption that the patterns are drawn from a multidimensional probability density function p.d.f., each mode of this function corresponding to a cluster [4,5]. Another vision of the mode detection problem is stated as locating the boundary which separates a mode from its environment [6].

In previous works, multivariate analysis methods, as neural networks, discriminant analysis [7] and genetic algorithms [8] have been used to identify Higgs boson events at LHC. Several attempts have been made to combine different gender classifiers containing complementary information to improve the classification accuracy. Experiments show that the combined classifiers generally outperform individual classifiers [7, 8, 9].

After a first attempt, where we used a multidimensionnal morphological technique [10], we aim in this paper, to separate between Higgs signal and background by using a new approach for boundary detection. We shall not consider all the possible Higgs decay channels, but we shall limit this analysis to some specific case studies. We mainly focus on the detection of the Higgs boson in the channel $p\overline{p} \to HX \to W^+W^-X \to l^+\nu l^-\nu X$.

The paper is organized as follows: The next section (section 2) is devoted to an introduction on the data and variables used. In section 3, we give a short description of the morphological boundary detection method, followed by its application to Higgs signal identification problem. The overall results are also discussed. Section 4 concludes this paper.

2. Data and Variables

The search for the Higgs boson is one of the primary tasks of the experiments at the Large Hadron Collider (LHC). Indeed several mechanisms contribute to the production of SM Higgs bosons in proton collisions [11, 12, 13].

The dominant mechanism is the gluon fusion process, $p\overline{p} \to gg \to H$, which provides the largest production rate for the entire Higgs mass range of interest. For large Higgs masses, the fusion process $qq \to WW, ZZ \to H$ becomes competitive, while for Higgs particles in the intermediate mass range $M_Z < M_H < 2M_Z$ the Higgs-strahlung off top quarks and $W; Z$ gauge bosons are additional important production processes.

As introduced above, we will identify the SM Higgs boson in the channel $p\overline{p} \to HX \to W^+W^-X \to l^+\nu l^-\nu X$. The decay channel chosen is $H \to W^+W^- \to e^+\mu^-\nu\nu, e^-\mu^+\nu\nu, e^+e^-\nu\nu, \mu^+\mu^-\nu\nu$.

The basic signature of this process is:

- Two charged oppositely leptons with large transverse momentum P_T.
- Two energetic jets in the forward detectors.
- Large missing transverse momentum P_T'.

A number of backgrounds are relevant to the considered channel:

a) $t\overline{t}$ production, with $t \to Wb \to l\nu j$. In this process a pair of W and a pair of jets (j) are produced.

b) QCD $W^+W^- + jets$ production: This is due to QCD emissions to the production of W^+W^-.

The physical observables used for the separation between signal and backgrounds are:

- $\Delta\eta_{ll}$, $\Delta\phi_{ll}$: the pseudo-rapidity and the azimuthal angle differences between the two leptons,
- $\Delta\eta_{jj}$, $\Delta\phi_{jj}$: the pseudo-rapidity and the azimuthal angle differences between the two jets,
- M_{ll}, M_{jj}: the invariant mass of the two leptons and jets
- $T_{nm}, (n, m = 1, 2, 3...)$ some rapidity weighted transverse momentum,

$$T_{nm} = \sum_{i \in event} \eta_i^n p_{iT}^m \tag{1}$$

  where η_i is the rapidity of the leptons or jets, p_{iT} their transverse momentums.

The production of signal and background processes has been modeled with PYTHIA6.1 [14], in the Higgs mass range, $115 < M_H < 200\ GeV/c^2$. To achieve this analysis, some selection cuts are made to the generated Monte Carlo events, but no detector effects have been considered. For convenience, we defined two classes of events: the Higgs boson process, (signal, denoted C_{Higgs} : $p\overline{p} \to HX \to W^+W^-X \to l^+\nu l^-\nu X$) and some background events (C_{Back}: $t\overline{t}$ and $QCD\ W^+W^- + jets$ production).

After the selected cuts made to the PYTHIA6.1 generated events, the samples retained amounted to 4000 events (2000 samples for each class).

Our purpose is to classify signal and backgrounds by using a new morphological boundary detection.

Notice that the considered data provides a basis for measuring the achievable performance of the proposed approach. The method should be extended in the future by taking into account other physical aspects as the detector effects, the Higgs Mass dependence and a complete background consideration.

To estimate the performance of this approach, two parameters are used: the efficiency E_i and the purity P_i of classification. Based on the confusion matrix $A(A_{ij})$, (A_{ij} being the value of events of genuine class C_i classified as class C_j), and for each class C_i we have:

$$E_i = \frac{A_{ii}}{\sum_l A_{il}} \tag{2}$$

$$P_i = \frac{A_{ii}}{\sum_l A_{li}} \tag{3}$$

3. Morphological boundary detection

Our approach for classifying Higgs signal and backgrounds by using a new morphological boundary detection, consists of four basic steps.

First of all, a fast estimation of the probabilistic density function is computed. In a second step, the raw estimate of the underlying p.d.f. is smoothed by means of an original combination of the binary and numerical morphological transformation [15]. The two others steps of the proposed technique are mode boundary extraction as connected components and definition of the classification rule.

3.1. *Estimation of the Density Function*

In a first attempt to compute the p.d.f. [10] a multidimensional space is used. Indeed, from the set of $N \times r$ available observations (N is the number of events and r the considered attributes) : $X_i = [\Delta\eta_{ll}, \Delta\phi_{ll}, \Delta\eta_{jj}, \Delta\phi_{jj}, M_{ll}, M_{jj}, T_{11}, T_{21}, T_{31}, T_{41}]_i$, $i = 1, 2, ..., N$, ($N = 4000$, $r = 10$), the algorithm consists to normalize the range of each component to the interval $[0, K]$, according to a chosen resolution ($K > 2$).

Each axis of this normalized data space is then partitioned into K exclusive intervals of unity width. This discretization defines a set of K^r hypercubes lattices of side length unity. The centers of these hypercubes constitute a regular lattice of sampling points denoted X. Each hypercube is defined by its coordinates $x_1, x_2, ..., x_N$ which are the integer parts of the coordinates of its centre X.

The results of the multidimensional approach are summarized in Table 3.

In the new approach proposed in this paper, and instead of a multidimensional space, we consider a two dimensionnal image obtained from the projection of events with respect to the two principal components (λ_1, λ_2) (Fig. 1).

The simplest definition of the p.d.f. is then connected to the resolution of the image. For $K = 24$, the value of the p.d.f. $p(\lambda_1, \lambda_2)$ correspond to the number of events lying in each of 24×24 pixels.

The efficiency of the proposed algorithm for boundary detection has been demonstrated using the above defined PYTHIA generated events. The raw data set, shown

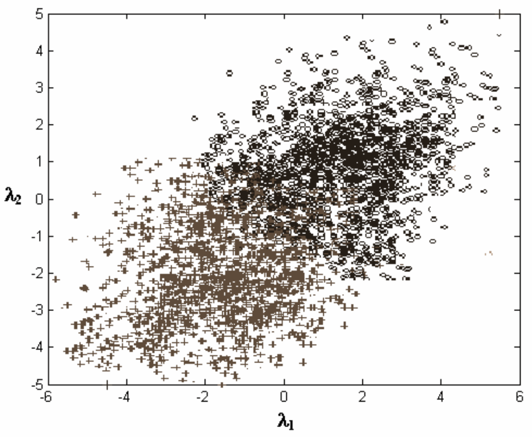

Fig. 1. C_{Higgs} (+) and C_{Back} (o) raw data set with respect to the first principal components

in Fig. 1, consists of 4000 observations (2000 C_{Higgs} events and 2000 C_{Back} events). The estimation yields the raw estimate of the underlying p.d.f. $p(\lambda_1, \lambda_2)$ shown in Fig. 2a and obtained for a 24 × 24 pixels image.

3.2. *Boundary detection*

The approach, used in this paper to detect boundary in images, is based on mathematical morphology [15].

Mathematical morphology is theoretically founded on set theory. It contributes a wide range of operators to image processing, based on a few simple mathematical concepts. The operators are particularly useful for the analysis of binary images, boundary detection, image enhancement, and image segmentation.

A morphological operation uses two sets of pixels, i.e., two images: the original data image to be analyzed and a structuring element H which is a set of pixels constituting a specific shape such as a line, a disk, or a square. Its shape can be regarded as a parameter to a morphological operation.

The basic morphological operations are dilation and erosion. They work conceptually by translating the structuring element H over the image points and examining the intersection between the translated kernel coordinates and the image coordinates.

Morphological operations are very effective for detection of boundaries in a binary image.

In mathematical morphology, neighborhoods are, therefore, defined by the structuring element, i.e., the shape of the structuring element determines the shape of the neighborhood in the image.

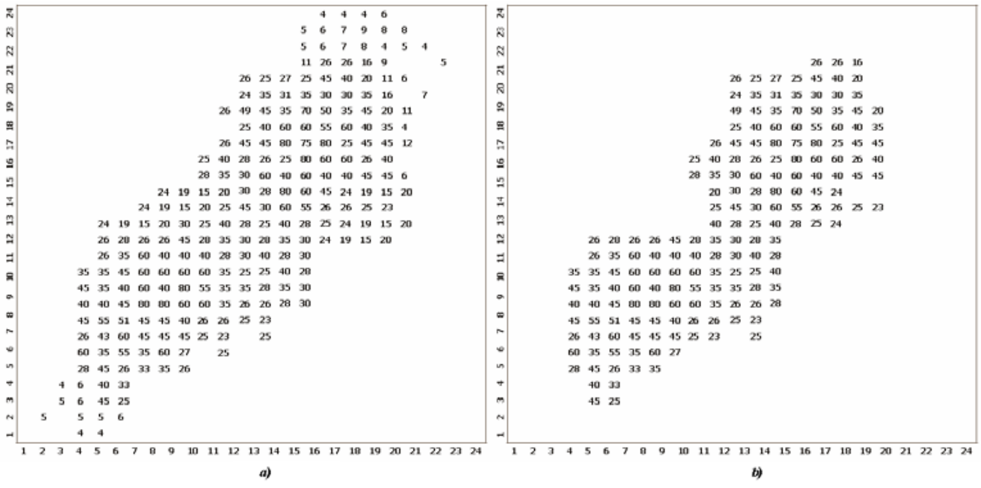

Fig. 2. a) The p.d.f. estimate: $p(\lambda_1, \lambda_2)$ and b) The gradient function $g(\lambda_1, \lambda_2)$

In our case, the raw estimate is smoothed by means of an original combination of the binary and numerical morphological transformations giving $h(\lambda)$ [15, 16]. This process removes all the minima of the p.d.f., localized in the valleys, which surround modal domains of this density function, and gives to the resulting function constant values within each core of modal region.

By using the characteristics of the smoothed function, the detection of the location of high amplitude changes in $h(\lambda)$ is easily accomplished by operating the morphological gradient, $g(\lambda)$ (Fig. 2b), obtained by the operation:

$$g(\lambda) = [h(\lambda) \oplus H] - [h(\lambda) \ominus H] \qquad (4)$$

where $[h(\lambda) \oplus H]$ and $[h(\lambda) \ominus H]$ denote respectively the numerical dilatation and erosion of $p(\lambda)$ by the structuring element H. Notice that, instead of an hypercube chosen for H in the multidimensionnal approach [10], the structuring element here is a 3×3 pixels image.

The mode boundaries of a function can be constructed through successive homotopic thinning of the gradient function.

With the characteristic of this resulting function, mode are easily extracted as connected components (Fig. 3) by a chaining approach where chains are constructed according to the configuration and to the function value's of all detected mode boundary lattices.

3.3. *Classification rules*

Once the different mode boundaries of the p.d.f. are identified, the events are classified according to two rules:

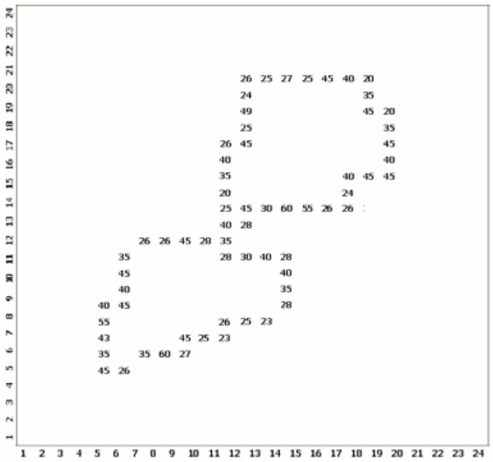

Fig. 3. Mode boundaries of the p.d.f.

Rule 1: The data points falling in interior of each of them are considered to be prototypes. The remaining observations assigned to their respective clusters by means of the nearest neighbor classification rule [17].

Rule 2: Only events inside mode boundaries are considered and assigned to the corresponding class.

The classification results achieved by this two procedures and their statistical parameters are consigned in Table 1 and Table 2 (respectively).

Table 1. Classification matrix and parameters obtained with the new approach, using Rule 1.

Processes	Classified C_{Higgs}	as C_{Back}	
C_{Higgs}: 2000	1401	599	
C_{Back}: 2000	591	1409	

Parameters	C_{Higgs}	C_{Back}	All
Efficiency (%)	70.05	70.45	70.25
Purity (%)	70.33	70.17	70.25

Table 2. Classification matrix and parameters obtained with the new approach, using Rule 2.

Processes	Classified C_{Higgs}	as C_{Back}	
C_{Higgs}: 1363	1105	258	
C_{Back}: 1384	292	1092	

Parameters	C_{Higgs}	C_{Back}	All
Efficiency (%)	81.07	78.90	79.98
Purity (%)	79.09	80.88	79.98

As introduced above, the results of the multidimensional image based method are summarized in Table 3. The total efficiency and purity of the two approaches are equal or higher of 69.55 %, Compared to other multivariate analyses [7], we can see that the proposed approach yields good results for a non supervised statistical clustering.

The advantage of the new approach with an efficiency of 70.25 % is due to a pre-processing of data by searching the two principal components. The corresponding image is more representative than the multidimensional image and its hyper-pixels.

In addition, the second rule, where only events inside boundaries are considered, allows to enhance the Higgs purity from 70.33 to 79.09 %.

This result means that events having their projection values, with respect to the two principal components (λ_1, λ_2), inside C_{Higgs} boundary are Higgs events with a probability of 79.09 %.

Obviously, in spite of the complexity of the Higgs classification problem and the used tool, the performances are strongly connected to some mathematical and physical parameters.

From mathematical point of view, the results are sensitive to several parameters as the principal components computation, the image resolution and the structuring element choice.

Further work should include an accurate selection of events according to the known cross section and its dependence on Higgs mass. In addition, we should include additional processes and variables and an accurate computation of detector effects.

Table 3. Classification matrix and parameters obtained with the multidimensional approach [10].

Processes	Classified C_{Higgs}	as C_{Back}	
C_{Higgs}: 2000	1387	613	
C_{Back}: 2000	605	1395	

Parameters	C_{Higgs}	C_{Back}	All
Efficiency (%)	69.35	69.75	69.55
Purity (%)	69.62	69.47	69.54

4. Conclusion

In this paper, a new approach for pattern classification has been proposed, which make concepts of mathematical morphology suitable for mode boundary detection in cluster analysis.

The new morphological approach shows significantly reduced computations and higher or equal performance compared to standard morphological operations and other linear multivariate methods.

Boundary detection is performed based on implicit morphological erosion and dilatation with a significantly reduced number of computations.

Two rules are used to classify the simulated LHC data. The final classification, which assigns the remaining data points to their respective clusters by means of the nearest neighbor classification rule, gives the good results compared with various classical classification schemes. When only events inside boundaries are considered, the classification purity of Higgs events is improved.

Further work should include an accurate computation of detector effects. The dependence of performance with regard to Higgs mass and other physical parameters will also be examined.

Some improvements are possible. First, others variables can be used with this multivariate approach, to increase the efficiencies of the different classifications. In addition, it would be interesting to see if these methods give the same results in others high energy physics classification tasks.

Acknowledgments

I am very grateful to the Organizing Committee of the CTP'2007 workshop for its support.

References

1. Van Ryzin, *Classification and clustering*, Academic Press, NewYork, 1977.
2. J. Hartigan, *Clustering Algorithms*, New York, Wiley, 1975.
3. A.K. Jain and al., *Algorithms for Clustering Data.* Prentice-Hall, NJ, 1988.
4. P-A. Devijver., J. Kittler *Pattern recognition: A statistical approach.* Englewood Cliff, NJ, Prentice -Hall international, 1982
5. R. Mizoguchi, S. Shimura. Nonparametric Learning without a Teacher based on Mode Estimation. *I.E.E.E. Trans. Comput.*, **C-25 (11)**,1109 (1976).
6. M. Rosenblatt. Remarks on some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.,* **Vol. 27**, 832 (1956).
7. M. Mjahed, *Nucl. Instrum. and Methods* **A559** 172 (2006).
8. M. Mjahed, *Nucl. Instrum. and Methods.* **A 481 (1-3)**601 (2002).
   M. Mjahed, *Nucl. Physics B* **Vol 106-107C**, 1094 (2002).
9. J. Kittler et al, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, 1998.
10. M. Mjahed, in *proceedings of LATTICE 2006*, POS (LAT20006)033.
11. S.L. Glashow, *Nucl. Physics.* **B22** 579 (1961).
12. S.L. Glashow, J. Iliopoulos and L. Maiani, *Phys. Rev.* **D2** 1285 (1970).
13. P.W. Higgs, *Phys. Letters.* **12** 132 (1964).
14. T. Sjostrand et al., High-Energy-Physics Event Generation with PYTHIA 6.1, *Comp. Phys. Comm.* **135** 238 (2001).
15. J. Serra. *Image Analysis and Mathematical Morphology.* Academic Press, New-York, 1982.
16. J.-G. Postaire, C.-P.-A. Vasseur. A Fast Algorithm for Nonparametric Probability Density. *I.E.E.E., Trans. Pattern Anal. Machine Intell.*, vol. **PAMI-4, n6,** 663 (1982).
17. T.-M. Cover, P.-E. Hart. Nearest Neighbour Pattern Classification. *I.E.E.E. Trans. Info. Theory,* **Vol. IT-13, n1**, 21 (1967).