PROCEEDINGS OF SCIENCE

Morphological boundary detection for cluster analysis and HEP data classification

Mostafa Mjahed [1] [2]
Ecole Royale de l'Air, Maths and Systems Dept, Marrakech, Morocco
E-mail: mmjahed@hotmail.com

This paper gives a new mode boundary approach for cluster analysis, based on mathematical concepts. It consists of a fast Parzen estimation of the underlying p.d.f. and a smoothing using numerical morphological transformation. The performance of this approach, with respect to a defined classification rule, is demonstrated on SM Higgs boson identification using generated LHC events.

XXIV International Symposium on Lattice Field Theory
Tucson, Arizona US
23-28 July, 2006

[1] Speaker
[2] Also at LPTN, Faculty of Sciences Semlalia, Marrakech, Morocco

PoS(LAT2006)033

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 this paper, we give a new mode boundary approach for cluster analysis, based on mathematical concepts.

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].

In this paper, 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 $p\bar{p} \rightarrow HX \rightarrow W^+W^-X \rightarrow l^+\nu l^-\nu X$. We seek improved classification performance using a coupling of c*lustering technique* and *morphological processing* of data.

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 [10, 11, 12]. The dominant mechanism is the gluon fusion process, $pp \rightarrow gg \rightarrow H$, which provides the largest production rate for the entire *Higgs* mass range of interest. For large *Higgs* masses, the fusion process $qq \rightarrow WW, ZZ \rightarrow H$ 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 $p\bar{p} \rightarrow HX \rightarrow W^+W^-X \rightarrow l^+\nu l^-\nu X$. The decay channel chosen is $H \rightarrow W^+W^- \rightarrow e^+\mu^-\nu\nu,\ e^-\mu^+\nu\nu,\ e^+e^-\nu\nu,\ \mu^+\mu^-\nu\nu$. 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) $t\bar{t}$ *production,* with $t \rightarrow Wb \rightarrow l\nu j$. 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,*

$$T_{nm} = \sum_{i \in event} \eta_i^{\,n} \cdot p_{iT}^{\,m} \qquad n,m=1,2,3,\ \ldots \tag{1}$$

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 modelled with PYTHIA6.1 [13], 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. We defined two classes: the *Higgs boson* process, (signal, denoted C_{Higgs} : $p\bar{p} \rightarrow HX \rightarrow W^+W^-X \rightarrow l^+\nu l^-\nu X$) and the *background* events (C_{Back}: $t\bar{t}$ 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 morphological boundary detection. Our approach consists of five basic steps.

First of all, a Parzen estimate [14] of the underlying p.d.f. is obtained. 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.

To estimate the performance of this approach, three parameters are used: the efficiency β_i , the purity η_i and the error ε_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:

$$\beta_i = \frac{A_{ii}}{\sum_l A_{il}} \ , \qquad \eta_i = \frac{A_{ii}}{\sum_l A_{li}} \ , \qquad \varepsilon_i = 1 - \beta_i \tag{2}$$

3. Analysis

3.1 Non Parametric Estimation of the Density Function

The underlying p.d.f. of a random variable is estimated in unsupervised context with a non parametric technique [16]. In this paper, a fast algorithm for determining the uniform kernel estimate of the p.d.f. from the set of Q available observations: $X_q = \left[x_{q,1}, x_{q,2}, ..., x_{q,n}, ...x_{q,N}\right]^T$ = $[\Delta\eta_{ll}, \Delta\phi_{ll}, \Delta\eta_{jj}, \Delta\phi_{jj}, M_{ll}, M_{jj}, T_{11}, T_{21}, T_{31}, T_{41}]^T_q$ $q= 1,2,...,Q$, ($N= 4000, Q=10$) is used in order to reduce the computational burden usually associated with others techniques as the Parzen kernel estimate [14].

The fast algorithm [16] consists at first to normalise the range of each component to the interval [*0, K*], with an integer $K \geq 2$ (resolution), by the transformation defined as:

$$y'_{q,n} = K \frac{x_{q,n} - \min\limits_{q} x_{q,n}}{\max\limits_{q} x_{q,n} - \min\limits_{q} x_{q,n}} \qquad (3)$$

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

Let $Y'_q = \left[y'_{q,1}, y'_{q,2}, \cdots, y'_{q,n}, \cdots, y'_{q,N}\right]^T$, $q=1,2,...,Q$ be the Q observations in the scaled data space. Each of them falls into a non-empty hypercube. If several observations fall into the same hypercube, this one appears many times in the list of non-empty hypercubes. It is easy to find the number of data points $q[H(X)]$ which fall into $H(X)$ by counting the number of times the hypercubes $H(X)$ appears in that list. Subsequently, the value of the local density estimate is $p(X) = q[H(X)]/Q$, since the volume of $H(X)$ is equal to unity.

3.2 Mode boundary extraction:

The efficiency of the proposed algorithm for mode boundary detection has been demonstrated using the above defined PYTHIA generated events. The raw data set, shown in Figure 1, consists of *4000* observations (*2000* C_{Higgs} events and *2000* C_{Back} events). The fast non-parametric estimation procedure yields the raw estimate of the underlying p.d.f. $p(X)$ shown in Figure 2 and obtained with $K=24$ which is the middle of the largest range where the number of detected mode boundary remains constant.

3.2.1 Smoothing of the p.d.f. :

The approach, exposed in this paper, is based on mathematical morphology [15]. This theory consists of analysing the relationships between an object (subset of *N*-dimensions) and its environment, using pre-defined geometrical sets, called structuring elements. Various interactions of the original set with the structuring element form the basis of all morphological operators.

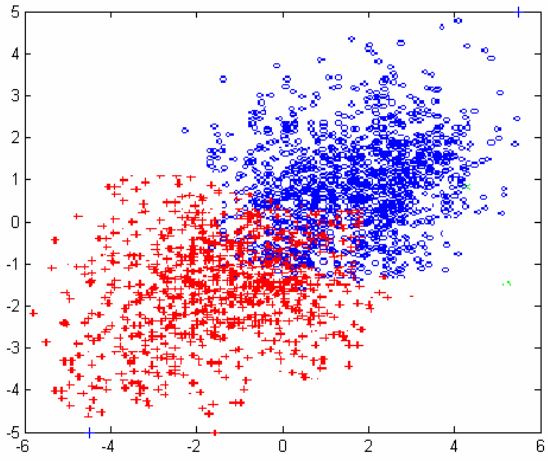

Figure 1: C_{Higgs} and C_{Back} raw data set

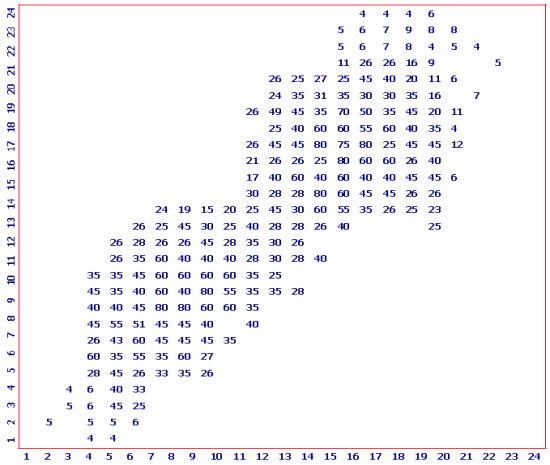

Figure 2: The p.d.f. estimator: p(X)

The raw estimate is smoothed by means of an original combination of the binary and numerical morphological transformations giving $h(X)$ [15]. This process removes all the minima of the p.d.f., localised 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.

3.2.2 Morphological gradient:

By using the characteristics of the smoothed function, the detection of the location of high amplitude changes in $h(X)$ is easily accomplished by operating the morphological gradient, $g(X)$ (Figure 3), obtained by the operation:

$$g(X) = [h(X) \oplus \underline{H}] - [h(X) \Theta \underline{H}] \tag{4}$$

where

$$h(X) \Theta H = \inf\{p(Y); Y \in \underline{H}_X\} \tag{5}$$

denotes the numerical erosion [15]of $p(X)$ by the structuring element $\underline{H}$.

3.2.3 Mode Boundaries Detection:

The analytical definition of the numerical thinning of multivalue density function by such structuring element is given by:

$$\begin{cases} (p \circ \underline{S})(X) = \sup\limits_{Y \in \underline{S_{0X}}} p(Y) & if \quad \sup\limits_{Y \in \underline{S_{0X}}} p(Y) < p(X) \le \inf\limits_{Y \in \underline{S_{1X}}} p(Y) \\ (p \circ \underline{S})(X) = p(X) & otherwise \end{cases} \tag{6}$$

where $S=(S_0, S_1)$ is a two-phases structuring element.

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 (Figure 4) 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 rule

Once the different mode boundaries of the p.d.f. are identified, 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 neighbour classification rule [17].

The classification results achieved by this procedure and their statistical parameters are consigned in Table 1 and Table 2 (respectively). The total efficiency and purity are equal to *69.55 %,* Compared to other multivariate analyses [7], we can see that the proposed approach yields good results for a non supervised statistical clustering.

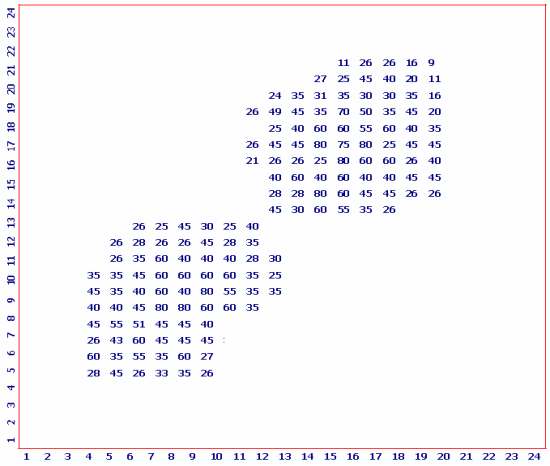

Figure 3: The gradient function g(X)

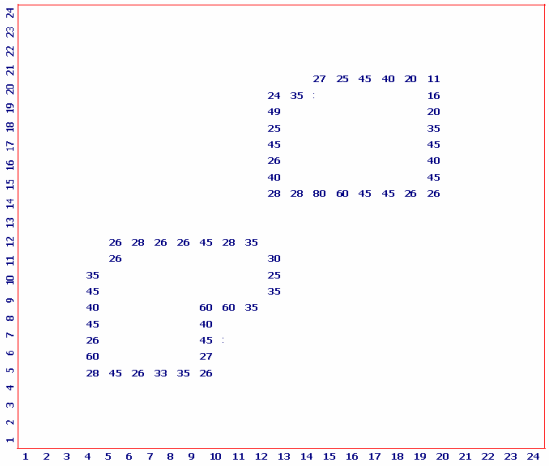

Figure 4: Mode boundaries of the p.d.f.

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

Table 1. Classification matrix obtained with the proposed approach

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

Table2. Efficiencies and purities of classifications

4. Conclusion

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

The final classification, which assigns the remaining data points to their respective clusters by means of the nearest neighbour classification rule, gives the good results compared with various classical classification schemes.

This analysis improves that the morphological mapping is much more efficient in difficult clustering situations such as non-spherical clusters and clusters with bridges between them, particularly when no a priori information is available. Thus it appears that multivalue morphology has fund a new development in the area of cluster analysis.

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. KTTLER *"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), (1976) 1109

[6] M. Rosenblatt. *Remarks on some Nonparametric Estimates of a Density Function.* Ann. Math. Stat., vol. **27**, (1956).832.

[7] M. Mjahed, Nucl. Instrum. and Methods **A559** (2006)172.

[8] M. Mjahed, Nucl. Instrum. and Methods. **A 481** (1-3) (2002) 601.
M. Mjahed, Nucl. Physics B Vol **106-107C**, (2002) 1094.

[9] J. Kittler et al, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. **20**, No. 3, March 1998.

[10] S.L. Glashow, Nucl. Physics. **B22** (1961) 579.

[11] S.L. Glashow, J. Iliopoulos and L. Maiani, Phys. Rev. **D2** (1970) 1285.

[12] P.W. Higgs, Phys. Letters. **12** (1964) 132.

[13] T. Sjostrand et al., *High-Energy-Physics Event Generation with PYTHIA 6.1*, Comp. Phys. Comm. **135** (2001) 238.

[14] E. Parzen. *On Estimation of a Probability Density Function and Mode.* Ann. Math. Stat., vol. **33**, (1962) 1065,.

[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**, n°6, (1982) 663.

[17] T.-M. Cover & P.-E. Hart. *Nearest Neighbour Pattern Classification*. I.E.E.E. Trans. Info. Theory, vol. **IT-13**, n°1, (1967) 21.