PROCEEDINGS OF SCIENCE

Separation of Higgs boson signal from Drell-Yan background with self-organizing maps

Aatos Heikkinen*
Helsinki Institute of Physics
E-mail: aatos.heikkinen@cern.ch

We demonstrate the use of the self-organizing feature maps, SOMs, in tagging b jets associated with heavy neutral MSSM Higgs bosons at the Large Hadron Collider, LHC. A Drell-Yan background discriminating power of SOM method using SOM_PAK software package is compared with our previous neural network studies.

XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research
April 23-27 2007
Amsterdam, the Netherlands

*Speaker.

PoS(ACAT)065

1. Introduction

1.1 Higgs boson production at LHC

At LHC, in the Minimal Supersymmetric Standard Model, MSSM, the heavy neutral Higgs boson production in association with two b quarks is the dominant Higgs boson production mechanism at large values of tanβ. These associated b jets can be used to extract the Higgs events from the Drell-Yan Z/γ^* background, for which the associated jets are mostly light quark and gluon jets. (More detailed description of physics relevant to this work can be found from [1] and references therein.)

1.2 b-tagging

In standard methodology a jet can be identified as a b jet using lifetime based tagging algorithm, which relies on displaced secondary vertices and track impact parameter, ip. Impact parameter is the closest approach of the track trajectory to the primary vertex. For a review of the main algorithms for inclusive b-tagging based on track ip and secondary vertex, see refs. [2] and [3]. Figure 1 demonstrates the case in the Compact Muon Solenoid experiment at LHC.

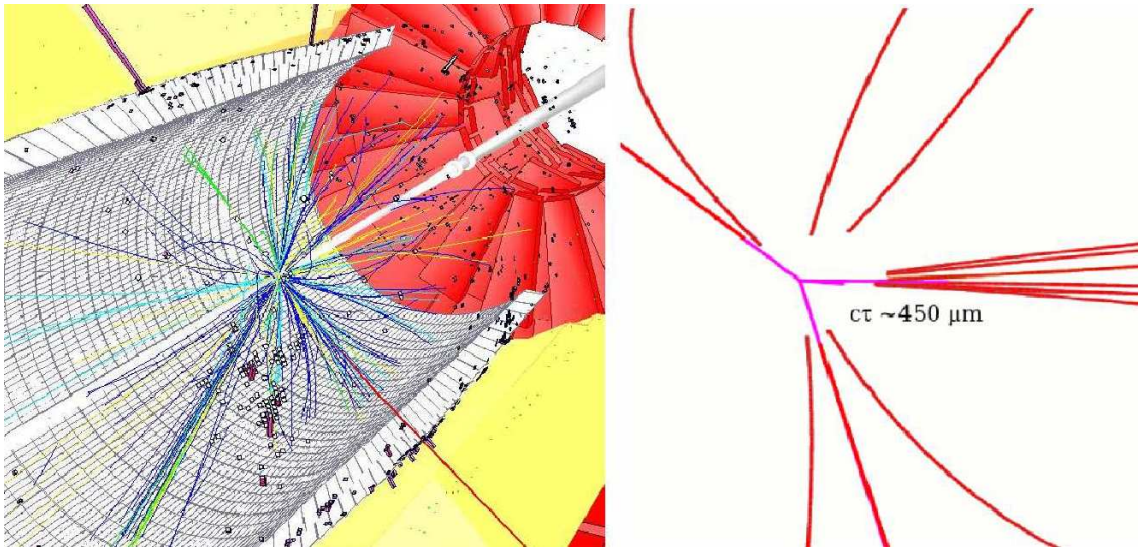


Figure 1: *Left:* Geant4 based simulation of a SUSY event in the CMS detector containing missing transverse energy, jets and several leptons in the barrel detector. (Picture: IguanaCMS.) *Right:* A displaced secondary vertex in a $b\bar{b}H$ event with $H \to \tau\tau$ in the CMS detector. The second b jet is not reconstructed due to a low jet energy and track multiplicity.

In high energy physics, contrary to popular neural network approach using Multi Layer Perceptron technique (see for example refs. [4, 5, 6, 7]), only few authors have reported on the use of the self-organized maps to separate a background from a signal. Though some promising results have been reported [8, 9, 10, 11]. This work is motivated by these few realizations of SOM based data analysis in HEP domain, and by the fact that SOM, also known as Kohonen network, can provide computationally more simple algorithm, with learning rate faster than what MLPs have.

2. Self-organizing maps

The most popular unsupervised neural network algorithm SOM [12, 13], provides a computationally simple algorithm, with fast learning rate. SOM defines a mapping from m -dimensional input data space onto a regular two-dimensional array of neurons:

- Every neuron of the map is associated with an m-dimensional reference vector.
- The neurons of the map are connected to adjacent neurons by a neighborhood relation, which dictates the topology of the map.
- During the unsupervised training phase, the SOM forms an elastic net that folds onto the cloud formed by input data and approximates the density of the data.

2.1 Competitive process

The SOM defines a mapping from the input data space $\mathbf{x} = [x_1, x_2, ..., x_m]^T$ onto a regular two-dimensional array of nodes. The synaptic weight vector $\mathbf{w}_j = [w_{j1}, w_{j2}, ..., w_{jm}]^T$, $j = 1, 2, ..., l$ of each neuron j has the same dimension as the input space; l is a total number of neurons.

Selecting the neuron with the largest inner product $\mathbf{w}_j^T\mathbf{x}$, is mathematically equivalent to minimizing the Euclidean distance between the input vectors $\mathbf{x}$ and $\mathbf{w}_j$. Thus, the winning neuron c is defined as:

$$c = arg\ min_j ||\mathbf{x} - \mathbf{w}_j||.$$

Essentially this sums up the competition process among the neurons, where the best-matching node locates the center of a topological neighborhood [12].

2.2 Adaptive process

During the learning, those nodes that are topographically close to a certain distance will activate each other to learn from the same input. Using discrete-time formalism, weight vector at time t is written as $\mathbf{w}_j(t)$, and updated weight vector is defined as:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \mathbf{h}_{jc}(t)[\mathbf{x} - \mathbf{w}_j(t)],$$

where $\mathbf{h}_{jc}(t)$ is neighborhood kernel. For details of SOM see ref. [14].

2.3 Advantages of using unsupervised learning

Unsupervised neural methods can be used in exploratory data-analysis, when we want to postpone the usual assumptions about what kind of model the data follow. In HEP applications we often have particularly challenging data mining problems where a priori information (for example the number of clusters in data) of the data sample is limited. Thus, unsupervised clustering has a potential to help us in searches of a signal of supersymmetry or another kind of new physics at LHC.

3. b-tagging with SOMs

3.1 Event data

In our SOM approach we feed SOM network with the same events and seven variables as used in the traditional track counting algorithm:

- Number of tracks in the jet cone (In the following we denote this with variable index 2 or $v2$)
- Impact parameters, ips, ($v4$, $v7$, $v10$, see fig. 2) and related ip significances, σ_{ip}s, ($v5$, $v8$, $v11$) for three leading tracks.

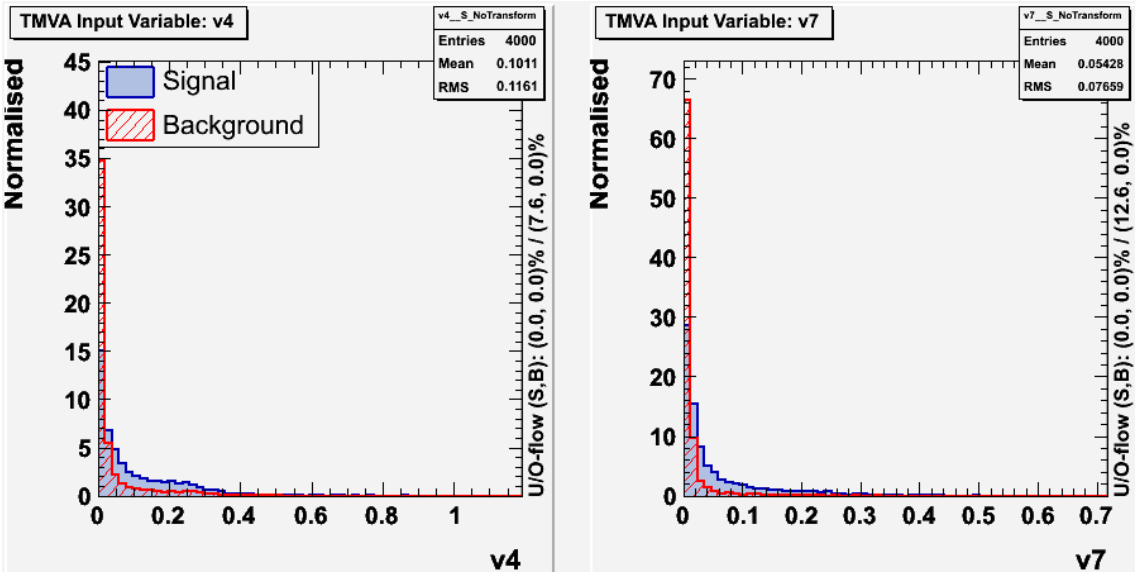


Figure 2: An example of variables used in the SOM teaching. *Left:* A leading track ip distribution for a signal and background events ($v4$). *Right:* ip for next leading track ($v7$).

3.2 Data preprocessing

We made a brief study using TMVA [15] to preprocess the training data, to assess basic properties of the discriminating variables used as input. The linear correlation coefficients of the input variables were calculated and displayed (fig. 3) and a preliminary ranking was derived.

Since the findings supported our previous understanding of the data (see fig. 5) we proceeded without additional preprocessing of data, and used the straightforward approach adopted in the previous study [16]. In future studies we plan to perform more detailed linear transformation of the variables into a non-correlated variable space, before entering to SOM learning phase.

3.3 SOM_PAK

We used a SOM_PAK [18] tool to analyze data created with a CMS ORCA [2] simulation package, using full simulation with track and jet reconstruction.

The signal and background event variables described above were fed to the SOM_PAK using a robust ASCII data format (fig.4). 40k events were used for teaching and 40k events were reserved for testing.

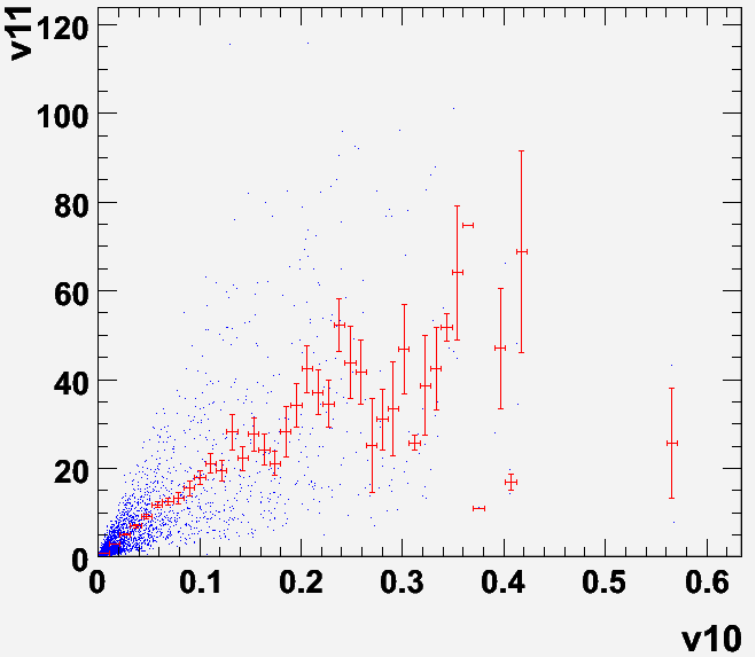


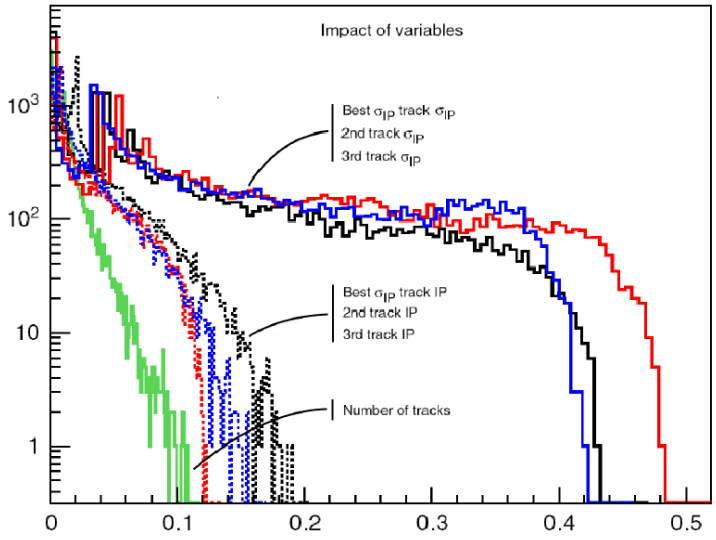
Figure 3: Correlation between 3rd best reconstructed track ip ($v10$) vs. σ_{ip} ($v11$).

```
#v2  n tracks
#v4  track 1  ip              <- Best sigmaIp track
#v5  track 1  sigmaIp
#v7  track 2  ip              <- 2nd best sigmaIp track
#v8  track 2  sigmaIp
#v10 track 3  ip              <- 3rd best sigmaIp track
#v11 track 3  sigmaIp
# v2 v4        v5      v7          v8       v10        v11
7
3  0.0492659   3.27892 0.0355801   1.01689  0.00540329 0.277054
20 0.0878318   9.07692 0.0787228   6.65229  0.0155514  3.83442
7  0.0228924   3.19142 0.0158648   1.44095  0.010939   1.24347
# only one track
1  0.00210019  1.02787 x           x        x          x
# two tracks
2  0.0120839   3.41975 0.00532341  0.952081 x          x
12 0.0921307  17.2172  0.0891074  13.958    0.108702   10.0725
```

Figure 4: Data format used by SOM_PAK. We see how SOM provides a natural way to treat incomplete event data.

4. Results

The b tagging efficiency with SOM was found to be 73 % with 11 % mistagging rate. We were able to filter 45 % of the background events with 0.2% misclassification probability for the signal. In figs. 6 and 7, which visualize SOM activation with test data, a clear separation to signal and background regions is seen. These results can be compared with typical counting algorithms performance 35 % efficiency with 1 % mistagging propability reported in [1, 16, 17].




```
Epoch: 3980 learn=0.267373 test=0.26963
Epoch: 3990 learn=0.267372 test=0.269627
Epoch: 3999 learn=0.267371 test=0.269629
Training done.
Network with structure:
v3,v5,v6,v8,v9,v11,v12:12:5:type
inputs with low values
in the differences plot may not be needed
v3 -> 0.2471 +/- 0.183578
v5 -> 0.104971 +/- 0.12343
v6 -> 0.133543 +/- 0.140757
v8 -> 0.104496 +/- 0.114607
v9 -> 0.110139 +/- 0.109817
v11 -> 0.142836 +/- 0.150484
v12 -> 0.0290961 +/- 0.0257619
tagger64111.h and tagger64111.cxx created
Mon Jan 15 12:47:17 EET 2007
========== Job ends here  ==============
```

Figure 5: The significance of variables found using supervised MLP networks. We notice that σ_{ip}s are particularly significant for correct classification result. (Figure from [17].)

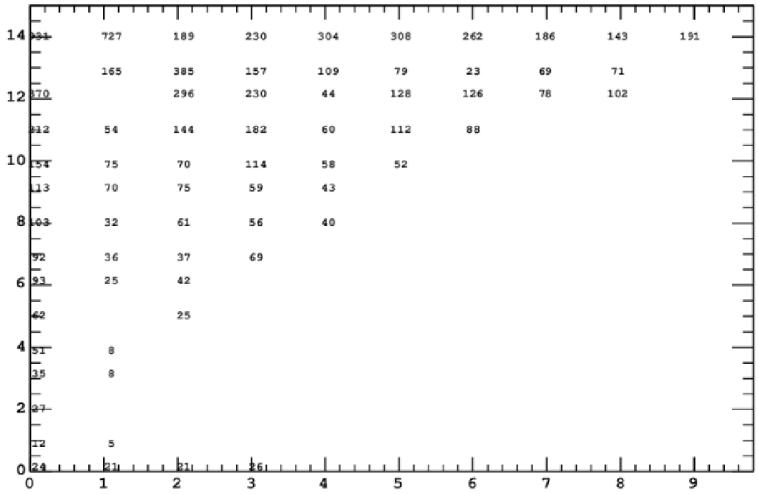


Figure 6: Number of signal events associated to winning node, while testing mapping of 15x15 node SOM performance.

5. Conclusion

We have shown how unsupervised classification can be utilized successfully in b-tagging problems. In our study the self-organizing maps were able to separate the Higgs signal from the background, based on CMS Monte Carlo data.

SOM methodology, being somewhat orthogonal to other data-analysis methods, such as supervised neural methods, shows a promise in HEP data mining, particularly for model free cases.

Recently, we have started an another promising approach in tagging b-jets with the use of a ROOT Toolkit for Multivariate Data Analysis, TMVA, which is an exiting new tool working in transparent factory mode guaranteeing an unbiased performance comparison, since all classifiers

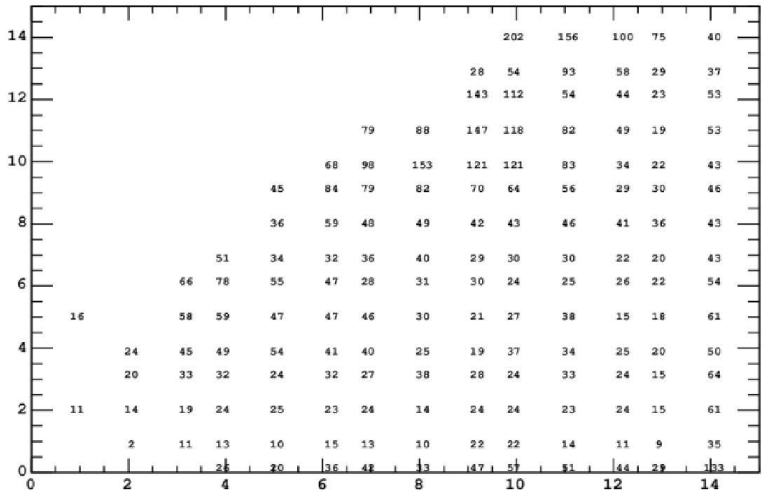


Figure 7: Number of background events associated to winning node, while testing mapping of 15x15 node SOM performance.

are evaluated with the same training and test data. TMVA, allows us to apply not only neural network classifiers but also projective and multi-dimensional likelihood estimators, linear discriminant analysis with H-Matrix/Fisher discriminants, and boosted/bagged decision trees. The first encouraging results applying TMVA transparent comparisons between various classifiers will be presented in [19].

References

[1] A. Heikkinen and S. Lehti, "Tagging b jets associated with heavy neutral MSSM Higgs bosons", Nuclear Instruments and Methods A 559 (2006) 195–198.

[2] G. Segneri and F. Palla, "Lifetime based b-tagging with CMS", CMS NOTE 2002/046, November 26, 2002.

[3] C. Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", CMS NOTE 2006/014, January 25, 2006.

[4] J. Zimmermann and C. Kiesling, "Statistical learning methods in high-energy and astrophysics analysis", Nuclear Instruments and Methods A 534 (2004) 204–210.

[5] K. Hultqvist *et al.*, "Using a neural network in the search for the Higgs boson", Nuclear Instruments and Methods A 364 (1995) 193–200.

[6] I. Iashvili and A. Kharchilava, CMS TN-1996/100.

[7] M. Mjahed, Nuclear Physics B 140 (2005) 799–801.

[8] A. De Angelis *et al.*, "Self-organizing networks for classification: developing applications to science analysis for astroparticle physics", arXiv:cs/0402014.

[9] K. Becks *et al.*, "Separation of hadronic W-decays from QCD-background with self-organizing maps", Nuclear Instruments and Methods A 426 (1999) 599–604.

[10] J. S. Lange, "Transputer self-organizing map algorithm for beam background rejection at the BELLE silicon vertex detector", Nuclear Instruments and Methods A 420 (1999) 288–309.

[11] M. J. Lang, "Application of a Kohonen network classifier in TeV γ-ray astronomy", J. Phys. G: Nucl. Part. Phys. 24 (1998) 2279–2287.

[12] S. Haykin, "Neural Networks – A Comprehensive Foundation", Prentice-Hall (1999).

[13] T. Kohonen, "Self-Organizing Maps", Springer-Verlag, Heidelberg (1995).

[14] J. Zurada, "Introduction to Artificial Neural Systems", Jaico Publishing House, Mumbai (2003).

[15] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis", arXiv::physics/0703039.

[16] A. Heikkinen and S. Lehti, "Self-organized maps for tagging b jets assosiated with heavy neutral MSSM Higgs bosons". (To be published in the proceedings of the CHEP 2006, Mumbai, India, February 13-17, 2006.)

[17] T. Linden, F. García, A. Heikkinen, and S. Lehti, "Optimizing Neural Network Classifiers with ROOT on a Rocks Linux Cluster". (To be published in the Lecture Notes in Computer Science.)

[18] T. Kohonen *et. al.*, "SOM_PAK: The Self-Organizing Map Program Package", Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Finland, 1996.

[19] A. Heikkinen *et al.*, "Testing TMVA software in search for MSSM Higgs bosons at the LHC". (To be published in the proceedings of the CHEP 2007.)