SELECTION OF VARIABLES FOR NEURAL NETWORK ANALYSIS.
COMPARISONS OF SEVERAL METHODS WITH HIGH ENERGY PHYSICS DATA.

ALEPH 94-186
PHYSIC 94-158
J. Proriol
7 December 1994

J.PRORIOL
Laboratoire de Physique Corpusculaire
Clermont-Ferrand

ABSTRACT.

This paper compares five different methods for selecting the most important variables with a view to classifying high energy physics events with neural networks.The different methods are:the F-test,Principal Component Analysis (PCA),a decision tree method:CART,weight evaluation, and Optimal Cell Damage (OCD).

The neural networks use the variables selected with the different methods.We compare the percentages of events properly classified by each neural network.The learning set and the test set are the same for all the neural networks.

1) INTRODUCTION.

In high energy physics experiments,one can compute many different variables from experimental event data.The problem of picking out the most discriminating variables from a large set of candidate variables is important for a classification task.

Different methods of variable selection for neural networks have been suggested [1-6].

The F-test method [7,8] was first used [1] to get a discriminating power of the variables.The choice of variables for a real world problem [2] was done by selecting the variables given by the program of classification CART [9] applied to data.A first run of neural network was used [3] to get a variables discriminating classification while using the weights of the neural network.Recently [4] a new approach was proposed using a more sophisticated method.

We propose,in this paper,to compare the different methods with the same learning and test sets.The different learnings

will be conducted in the same way.We'll add some other methods.The comparison will be done with previous results on the same sets.[10]

We first recall the origin of the variables.The study of the different methods is performed.With the selected variables we train a neural network and we test the result with a test set.In conclusion,we compare the different results.In appendices,the formulas of some methods are given.

2)HIGH ENERGY EVENTS AND VARIABLES

The objects of the study are the high energy LEP events:

e^+e^- -> quark+antiquark -> hadrons.

We would like to identify the flavour of the event quark :b quark event,c quark event,light quark event.The information comes from the data connected to the tracks of the events recorded in the ALEPH detector.The quark masses are different,so the event shapes and the jet shapes are different.

For each fully simulated ALEPH event we can compute variables using the data connected to the event tracks .A large set of variables was computed for each event.The number of variables is 150.It is difficult to describe all the variables.Some variables are very classical ones such as sphericity,aplanarity,Fox-Wolfram coefficients and describe the event shape.Some variables connected to the vertex detector were also computed[12].Some other variables were designed specially to study ALEPH e^+e^- events [1,10].

We also compute some variables with the tracks of the event's 2 most energetic jets such as the 2 jets sphericity product [13].We also compute variables describing the jets shape [10],variables connected to one jet's vertex or one event's hemisphere [12],variables connected to the event's most energetic lepton ,some combinations of the tracks of the jets giving directed sphericities and invariant masses [14].

3)SELECTION METHODS

We present now the different methods of variable selection.

3-1) F-test METHOD

The first method adopted for variable selection is the F-test method [1,7,8].We recall the main formulas in appendix 1.

We have applied the F-test formula to the 150 variables,we have classified the variables according to the F-test value.As first variable we have chosen the variable with the highest value;as second variable,we have taken the next one if its correlation with the previous variable is lower than 0.55;we apply the same method to the following ones.We have kept the first 20 variables.

3-2) PCA METHOD

To avoid correlated variables,it is possible to use variables chosen after a principal component analysis (PCA) [11].

We replace the original variables x_i which are correlated with new variables:the principal components c_m which are x_i linear combinations ;they are not correlated and their variance is maximum.In appendix 2,we give the different formulas of this method.We have also kept the first 20 variables.

3-3) CART

The choice of the variables with CART [2,9],is possible.This method was used in a real world problem [2].

The CART program uses a binary decision tree method.Binary trees are constructed by repeated splits of a set S of events into two descendant subsets beginning with the set S itself.Each split is achieved according to the values of one variable chosen by the program.The partition is stopped when the subset event number is too small.The terminal subsets form a partition of S.Each terminal subset is identified with a

class label.There may be two or more terminals with the same class label [9].

To build the decision tree,we have first kept the first 70 variables given by the F-test classification without elimination of the correlated variables.We have limited the number of variables to 70 for memory size reasons,but the most interesting variables are among this variable subset .We have sent these 70 variables into the CART program without backward pruning.A decision tree was built using 16 variables.

3-4) WEIGHTS

The variables are classified according to the value of the weights between the input layer and the first layer [3].The weight W_{ij} is defined between the input layer neuron j and the first hidden layer neuron i ,we classify the variables of the input layer according to the S_j value:

$$S_j=\sum_i (W_{ij})^2$$

After a first training we have computed the S_j values using the first 70 variables given by F-test,we have chosen the first 20 variables given by the S_j values.

3-5) OCD

The OCD method is based on Optimal Cell Damage (OCD) considerations [4].It is also a method based on the weigths.We call E the mean square error (MSE) of the MLP neural network.The S value (called saliency [4]) is computed according to the relation:

$$S_j=\sum_i \frac{\partial^2 E}{\partial W_{ij}^2}(W_{ij})^2$$

We have computed the S_j values,using the method of the previous paragraph.We have also chosen 20 variables with the highest values.

4)LEARNING AND TEST

The method used for learning and test was proposed in [10];we recall the different steps.

After the selection of 20 input variables,the learning is done with a 20-20-8-3 four-layer neural network (a 16-20-8-3 for the CART method).The 3 classes are: b quark events,c quark events and light quark events.This choice of network gives good results.

The learning set is composed of 10000 events in each class.A validation set is composed of 10000 events in each class.We stop the training when the cost of the validation set begins to increase.

The test set is a set of 73376 events :16086 b quark events,12757 c quark events and 44533 light quark events.This set has the proportions of a set of LEP1 e^+e^- events giving hadrons .

The class label of a test set event is the class label of the highest value neuron output.We know the event original class ,we then get a classification matrix similar to a statistical method matrix.

5)RESULTS

The trace of the classification matrix gives the percentage of well classified events.This percentage gives an idea of its selection method power.We have computed the traces in 2 cases:the 3-class case (b,c,uds classes) and the 2-class case (b and udsc classes).

The numerical results of the well classified events percentage for the different methods are given in table 1.The errors were estimated using statistical methods [11].

The most discriminating variables chosen by OCD are:

-the vertex variable for all event tracks [12],

-the vertex variable for jet N° 2 tracks,

-the vertex variable for hemisphere N°1 tracks (the split is perpendiculat to the thrust axis),

-the vertex variable for hemisphere N°2 tracks,

-the variable $\beta(9)$ describing jet shapes [10],

-the variable built with the jets particles products of longitudinal momentum and of transverse momentum [1,10],

-the variable $\beta(14)$,

-the most energetic lepton transverse momentum,

-the variable $\beta(16)$,

-the jet N°1 most energetic particle longitudinal momentum ,

-the jet N°1 charged tracks number,

-the jet N°1 second most energetic particle longitudinal momentum,

-the variable $\beta(8)$,

-the jet N°2 most energetic lepton longitudinal momentum.

The first 10 variables were also chosen according to the weights method.

In each method,the first variable is the same,and some jet shape variables β were chosen.The other variables differ according to the method.We thus find the 2 jets sphericity product [13],and also directed sphericities and invariant masses [14].

The variable choices can explain the differences between the results of the different methods.The results are improved when we use MLP-RBF instead of MLP.

6) CONCLUSION

We can compare the results obtained with different variable selection methods.The variables were used to feed the same neural network.All the results are good but some are better than others.

6-1)Comparison

We have computed the percentage of good classification with the different methods.To improve the comparison, the 3-class purity obtained from the classification matrix was added.The event sample purity is defined by the ratio of a class true event number and of this class classified event number.

If we compare the 3-class classification percentages ,the OCD method seems best for variable selection.If we consider the 2-class classification ,we see that the PCA method gives good results.

But for physics reasons,we are also interested in the purity of b and c quark event samples.We see that OCD works well in these 2 cases.

The F-test is a linear method.This method is not meant for variable classification ,so the results are good for such a simple method.The F-test can be used for a first approach to a problem:this method is fast.

The PCA method which is a linear one gives good results but handling the huge files thus generated is difficult.

The other methods are non-linear ones.They seem slightly better than the other methods;we see that the best one is OCD.The weights method gives good results and is faster than OCD;the CART method is also faster than OCD.

6-2)Saliency

To explain the OCD choice we can write the saliency [15]:

$$S_j=\sum_i \delta E_{ij}.(W_{ij})^2,$$

where δE measures the saliency sensitivity to small perturbations in W_{ij}.In the method called "Optimal Brain Damage" [4],the sensitivity measure is approximated by the MSE second derivative;this choice improves the classification method .

6-3)Extension

We can extend the OCD method to a neural network pruning [15].A large neural network is trained,for each layer l neuron j we compute the saliency:

$$S_j^l = \sum_i \frac{\partial^2 E}{\partial(W_{ij}^{(l+1)})^2}.(W_{ij}^{(l+1)})^2 ,$$

where $W_{ij}^{(l+1)}$ is the weight between layer l neuron j and layer (l+1) neuron i ;for the input layer,l=1.

The small saliency neurons can be deleted.A work is in progress on this point.

7) AKNOWLEDGMENTS

I would like to thank Professor F. FOGELMAN-SOULIE for sending me the OCD paper before publication.This work is part of a work done within ALEPH collaboration;I would like to thank the physicists of the collaboration for their support.

e-mail:proriol@frcpn11.in2p3.fr

REFERENCES.

[1]J.PRORIOL,A.FALVARD,P.HENRARD,J.JOUSSET,B.BRANDL
Tagging b quark events in ALEPH with neural networks.
Proceedings of the workshop 'neural networks from biology to high energy physics' Isola d'Elba june 1991
O.BENHAR,C.BOSIO,P. del GUIDICE,E.TABET editors
ETS EDITRICE (PISA 1992) p 419

B.BRANDL,A.FALVARD,C.GUICHENEY,P.HENRARD,J.JOUSSET,J.PRORIOL
Multivariate analysis methods to tag b-quark events at LEP/SLC
NUCLEAR INSTR. AND METHODS A 324 (1993)307

[2]X.DING,T.DENOEUX,F.HELLECO
Tracking rain cells in radar images using multilayer neural network
Proceedings ICANN'93 Amsterdam sept 93
SPRINGER -VERLAG (Berlin 1993) p 962

[3]C.GUICHENEY
Selection non-linéaire de variables discriminantes au moyen de réseaux neuro-mimétiques.
Preprint Clermont PCCI RI 94-01

[4]T.CIBAS,F.FOGELMAN-SOULIE,P.GALLINARI,S.RAUDYS
Variable selection with optimal cell damage.
Proceedings ICANN'94 Sorrento may 94
SPRINGER VERLAG (Berlin 1994) p 1464

Y. Le CUN,J.S. DENKER,S.A. SOLLA
Optimal brain damage
Advances in neural information processing systems II
Morgan Kaufmann (San Mateo 1990)p 598

[5]G.STIMPFEL,P.YEPES
Higgs search and neural network analysis
Computer Physics Comm 78(1993)1

[6]H.PI,C.PETERSON
Finding the embedding dimensions and variables dependances in time series.
Neural computation 6(1994)509

[7]S.BRANDT
Statistical and computational methods in data analysis
North-Holland (Amsterdam 1983)

[8]M.S.SRIVASTAVA,E.M.CARTER
Applied multivariate statistics
North-Holland (Amsterdam 1983)

[9]L.BREIMAN,J.H.FRIEDMAN,R.A.OLSHEN,C.J.STONE
Classification and regression trees
Wadsworth (Monterey CA 1984)

[10]J.PRORIOL
Classification of hadronic events in e^+e^- collisions with neural networks
Nucl. Inst. and Meth. A335(1993)288

J.PRORIOL
Multi-modular neural networks for the classification of e^+e^- events.
Nucl. Inst and Meth. A337(1994)566

[11]G.SAPORTA
Probabilités,analyse des données et statistiques.
Editions TECHNIP (Paris 1990)

[12]D.BROWN,M.FRANCK
Tagging b hadrons using track impact parameters.
Aleph note 92-135

[13]SAU LAN WU
Nuclear Physics B3(proc suppl.)(1988)102

[14]L.BELLANTONI,J.S.CONWAY,J.E.JACOBSEN,Y.B.PAN,SAU LAN WU
Using neural networks with jet shapes to identify b jets in e^+e^- interactions.
Nuclear Inst and Methods A310(1991)618

[15] LiMin FU
Neural networks in computer intelligence.
Mc Graw Hill,Inc (N.Y. 1994)

APPENDICES

A1) F-test formulas [7,8]

A set of events is defined with n events and k classes;for each event, *l* variables are computed.All the data are contained in the matrix x_{ij} with i=1..n and j=1..*l*.

The class c event number is n_c with

$$n=\sum_{c=1}^{k} n_c .$$

We define the total center of gravity g_j for variable j:

$$g_j=\frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

The center of gravity of the class c for variable j is h_j

$$h_j=\frac{1}{n_c}\sum_{i\in class\ c} x_{ij}$$

We define the within W value for variable j

$$W_j=\sum_{c=1}^{k}\sum_{i\in class\ c}\frac{1}{n}(x_{ij}-h_j)^2$$

which describes the dispersion of the classes and the between B value

$$B_j=\frac{1}{n}\sum_{c=1}^{k} n_c(h_j-g_j)^2$$

which describes the distance of a class to the center of gravity .

If W is small and B is large the variable j is able to discriminate the classes and we define the F-test value [7,8]:

$$F_j=\frac{(n-k)}{(k-1)}\cdot\frac{B_j}{W_j} \qquad (j=1..l).$$

A2) PCA formulas [11]

The notation of A1 are used.

The coefficients of correlation s_{kj} are defined by

$$s_{kj}=\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-g_k).(x_{ij}-g_j) \qquad (k,j=1..l)$$

We also get the elements r_{kj} of the correlation matrix R:

$$r_{kj}=\frac{s_{kj}}{(s_{kk}\cdot s_{jj})^{1/2}} \qquad (k,j=1..l)$$

A new set of variables z_{ij} is defined by a linear combination of the original variables x_{ij}:

$$z_{ij}=\frac{(x_{ij}-g_j)}{(s_{jj})^{1/2}} \qquad (i=1..n,j=1..l)$$

In the PCA method,the eigenvectors $\overrightarrow{u_k}$ and the eigenvalues λ_k of the R matrix are given by the relation

$$R\ \overrightarrow{u_k} = \lambda_k\ \overrightarrow{u_k}$$

We classify the eigenvectors according to the decreasing λ_k values.

With the u_m^k ,components of the $\overrightarrow{u_k}$ eigenvectors (k,m=1..*l*) we build a new matrix V such as $V_{mj}= u_m^j$.

Using this matrix V,we define the new variables c_{im} by

$$c_{im}=\sum_{j=1}^{l} z_{ij}.V_{jm} \qquad (i=1..n,m=1..l).$$

METHODS	% 3 classes b/c/uds	% 2 classes b/udsc	pur b	pur c	pur uds
F-test(MLP)	71.2±0.3	91.4±0.2	80.8	33.6	86.6
F-test(MLP-RBF)	73.5±0.3	91.7±0.2	82.7	35.9	86.1
CART (MLP)	73.2±0.3	91.6±0.2	81.8	35.9	86.5
PCA (MLP)	73.3±0.3	91.9±0.2	83.3	35.8	86.7
WEIGHT(MLP)	72.9±0.3	91.5±0.2	81.9	35.3	86.2
OCD (MLP)	.73.8±0.3	91.5±0.2	82.1	35.9	85.7
OCD (MLP-RBF)	.74.4±0.3	91.7±0.2	82.6	36.7	85.8

TABLE 1. The two % columns give the percentage of well classified events for the 2 cases:3 classes(b/c/uds) and 2 classes (b/udsc).The other columns give the purity of the samples of b,c and uds events from the classification matrix.