ALEPH 90-95
PHYSIC 90-58
J. Proriol et al.
23.6.1990

EVENT RECOGNITION: MULTIVARIATE ANALYSIS METHODS TO TAG B QUARKS EVENTS IN ALEPH

J. PRORIOL, J. JOUSSET, C. GUICHENEY,
A. FALVARD, P. HENRARD, D. PALLIN,
P. PERRET, F. PRULHIERE

Laboratoire de Physique Corpusculaire
de Clermont-Ferrand
IN2P3/CNRS Université Blaise Pascal
F-63177 AUBIERE CEDEX FRANCE

ABSTRACT :

Two multivariate analysis methods : a linear discriminant method and a classification tree have been performed to get classifiers.

These classifiers have been applied to tag b quark events in ALEPH.

Two decays of the b quark have been considered : semi-leptonic decays involving an electron and purely hadronic decays.

The ratios of number of classified events $n(b \to e)/n(e)$ and $n(b \to hadrons)/n(events)$ are computed with Monte-Carlo events and ALEPH events.

Presented at the International Workshop on Software Engineering, Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics.

LYON March 1990

CONTENTS

1 – Introduction

An important application of artificial intelligence is the pattern recognition [1],[2].

Among the different methods of pattern recognition the statistical multivariate analysis methods had lead to a lot of applications [3]-[6].

1.1 Previous works

The methods of pattern recognition have already been used in high energy physics [7], in particular the multivariate analysis methods have been extensively used in recent studies with Monte Carlo simulations such as the identification of top quark events in UA2 and LEP experiments [8], the determination of the number of jets of an event and the tagging of the quark jet flavour [9],[11]; however these results have not been yet confronted to the data.

The method of classification tree have been used to identify charged clusters in ALEPH electromagnetic calorimeter[10].

The common purpose of these works is to classify an event by building a set of rules called classifier, the discrimination methods described later on are a way to get such rules.

1.2 The discrimination Baye's rule

The basic purpose of a classification is to get an accurate classifier, that is to characterize the conditions allowing to determine whether an object is in one class or an another.

Considering two classes C_1 and C_2 with the probabilities $M_1 = P(C_1)$ and $M_2 = P(C_2)$ for an object to belong to C_1(resp C_2), the goal is to get a decision rule sharing the space of the variables in two regions R_1 and R_2, R_1 being filled with the objects of C_1 and R_2 with those of C_2.

If $C(1,2)$ is the cost in classifying an object of C_1 in R_2 and $C(2,1)$ the cost in classifying an object of C_2 in R_1, $C(1,2)$ and $C(2,1)$ are the misclassification costs.

Let $P(1,2)$ the probability for an object of class 1 to be classified in class 2 and $P(2,1)$ the probability for an object of class 2 to be classified in class 1. The mean misclassification cost is :

$$P(2,1)\,C(2,1) + P(1,2)\,C(1,2) \tag{1}$$

This cost can be computed by the knowledge of prior probability M_i of $C_i\,(i = 1, 2)$ and of the density function f_i of $C_i\,(i = 1, 2)$ inside the space of the variables.

The probability of good classification for an object of C_1 is $M_1 = \int_{R_1} f_1(x)\,dx$ and the probability of misclassification for the same object is $\int_{R_2} f_1(x)\,dx$. Thus the mean misclassification cost can be expressed as

$$M_1 \int_{R_2} f_1(x)\,dx \,.\, C(1,2) + M_2 \int_{R_1} f_2(x)\,dx \,.\, C(2,1) \tag{2}$$

A classification process which minimizes this cost is called a bayesian classification process.

1.3 Present work

In the following we present two discriminant methods fulfilling the Baye's rule:

-the linear discriminant analysis [3]–[6]

-the method of classification trees [6],[17]

these two methods have been applied here to the tagging of b quark events in ALEPH.

Performing a pattern recognition method needs a first step called the learning phase in which one uses some events for which the class they belong to is known. In the present case these events are Monte Carlo generated events. From this first step one gets the rules of classification, called classifier, an unknown event can then be classified according to such rules.

Once that the learning phase has been performed, a validation of the rules must be done via a large sample of known events, called hereafter the test sample, to minimize the statistical error. In this work, the test sample is done with a number of Monte Carlo events much larger than the learning sample.

After applying the classifier to these events several ratios have been computed. In the case of leptonic b quark decay, the ratio given by the number of classified b events divided by the number of events with an electron : $n(b \rightarrow e)/n(e)$; in the case of hadronic decay, the ratio $n(b \rightarrow had)/n(events)$.

We seek events classified as b events by processing the ALEPH data with the same classifier as for the Monte Carlo test sample. The same ratios $n(b \rightarrow e)/n(e)$ $n(b \rightarrow hadron)/n(events)$ are computed and compared to the Monte Carlo results.

1.4 Outline of the paper

In the chapter 2 we describe the methods used to select the events of the Monte Carlo samples and the variables used throughout our study.

In the chapter 3 after a short recall of the discriminant analysis method we describe the procedure used to get the classification rule from the learning sample. We then compare Monte Carlo test sample results with ALEPH data results in the two cases of b quark decay.

In the chapter 4 we give a short approach of the classification tree methods applying two programs to the b quark semi-leptonic and purely hadronic decays.

2 – Events and variables

This section is devoted to the origin of the learning sample events, the method used to identify the electrons in an event, and the variables used throughout the study.

2.1 Learning sample events and test sample events

The Monte Carlo events were generated in the ALEPH collaboration by Annecy, Clermont and Marseille. These events have been reconstructed through JULIA, and can thus be fully compared to the data.

A set of b quark events, a set of c quark events and a set of mixed u, d, s quark events have been used, assuming for the test sample the following proportions :

$$b : 21,9\% \; ; \; c : 17,1\% \; ; \; uds : 60,9\% \text{ versus } q\bar{q}.$$

2.2 Identification of the events with an electron

The identification of a leptonic event is made with the standard subroutines already used in ALEPH [12] to tag the $b \rightarrow e$ events.

2.3 Variables associated to an event

We have used a set of variables which are computed with the components (momentum, energy) of the tracks of the event.

-Variables connected with the shape of the event : sphericity, aplanarity, oblatness, thrust [13]. The Fox-Wolfram variables H_2, H_3, H_4 and H_5 [13] being also included.

-Three variables already used to separate the b quark events : the momentum of the electron, the transverse momentum of the electron [12] in the leptonic b decay case, and the double sphericity [14].

-Variables connected to the most energetic jet requiring at first that the angle of the thrust axis of the event must be greater than 25^0 to get a clean jet. The charged particles are clustered by the QJMMCL algorithm [13], and the jet with the greatest energy is considered, extracting then the energy, mass, charge, and sum of the absolute value of the transverse momenta of the tracks.

The visible energy of an event, the number of charged particles, the missing p_T, and the transverse mass [15] are included too.

The comparison of the distribution of all these variables with a Monte Carlo test sample and the with ALEPH real events leads to a good agreement, the Monte-Carlo events can thus be used to get an accurate classifier.

2.4 The input learning sample

Generally speaking the learning sample is done with n events (statiscal objects) belonging to k known classes $\mathrm{C_i}$ $(i = 1...k)$ and are associated to p variables; the population of the class Ci is n_i with $n = \sum_i n_i$

In this work the p variables have been choosen according to the results given by previous experiments dealing with jets and heavy flavour studies as explained above.

3 – Linear discriminant analysis

Writing the events as p-components vectors $\vec{x_1}, \vec{x_2} \ldots \vec{x_n}$ with

$$\vec{x_i} = (x_{i1}, x_{i2} \ldots x_{ip})$$

the input is the learning matrix of elements x_{ij}.

3.1 The FISHER linear discriminant function [3]

Assuming a normal distribution for $\vec{x}$, we write : $\vec{x} \sim N_p(\vec{\mu}, \Sigma)$,

$\vec{\mu}$ being the p-components vector of the the mean values and Σ the symmetric variance matrix.

Considering the simple case of 2 classes C_1 et C_2, $\vec{\mu_1}$ and $\vec{\mu_2}$ are the $\vec{\mu}$ vectors of these classes, the Σ matrix is assumed to be the same for C_1 and C_2.

The Fisher linear discriminant function [3] maximizes the norm of the vector $\vec{\delta} = \vec{\mu_1} - \vec{\mu_2}$ which gives the distance of the two classes.

In a matrix notation the linear discriminant function L for $\vec{x_0}$ can be written [3]

$$L(\vec{x_0}) = \vec{\delta}[\Sigma^{-1}](\vec{x_0})^T \tag{1}$$

$\vec{x_0}$ will be classified inside the class C_1 if:

$$L(\vec{x_0}) > \frac{1}{2}(\vec{\mu_1} - \vec{\mu_2})[\Sigma^{-1}](\vec{\mu_1} + \vec{\mu_2})^T \tag{2}$$

and inside the class C_2 if:

$$L(\vec{x_0}) \leq \frac{1}{2}(\vec{\mu_1} - \vec{\mu_2})[\Sigma^{-1}](\vec{\mu_1} + \vec{\mu_2})^T \tag{3}$$

It can be shown that such a rule of classification is a rule minimizing the Baye's risk of misclassification.

This rule of classification can otherwise be differently stated.

Taking the MAHALANOBIS distance D_{oi} defined by the relation

$$D_{oi}^2 = (\vec{x_0} - \vec{\mu_i})[\Sigma^{-1}](\vec{x_0} - \vec{\mu_i})^T \tag{4}$$

D_{oi} can be understood as a generalized distance of $\vec{x_0}$ to the center of mass of the class C_i.

The rule of classification of $\vec{x_0}$ into C_1 is no longer given by the value of $L(\vec{x_0})$ but by the value of D_{oi}^2, which means, for two classes, that if:

$$D_{02}^2 < D_{01}^2 \tag{5}$$

the event $\vec{x_0}$ is thus closer to C_1 than to C_2 and will be classified into C_1.

The generalization of this rule to k classes is straight forward. Computing $\vec{\mu_i}$ and Σ for all the classes $i = 1, ..., k$, $\vec{x_0}$ will be in class C_i if

$$D_{oi}^2 = \min_{\substack{1 \leq j \leq k \\ j \neq i}} D_{oj}^2 \tag{6}$$

3.2 Methodology

The purpose of this work is to tag the b quark events which means that we would like to distinguish b quarks from c quarks and light quarks events.

The classification rule has been obtained from a learning sample of 500 b events, 500 c events and 500 uds events, giving an equal weight to the three classes.

3.2.1 Selection of the variables

Usually, all the variables are not needed for a well doing discrimination, a good practice is to check which variables are meaningfull. Several criteria are used to select the significant ones [5].

In this study the selection of variables has been done with the program SELDISC from the library MODULAD [16]. The program uses the statistical parameter F [3] to select the significant variables. A limited number of variables (3 to 8) among the 19 original ones has been retained.

3.2.2 Rules of discrimination

In a second step, the selected variables are handled in a program of discrimination. The programm DISC[16] of MODULAD allowing a discrimination between three classes has been used.

The classification rule used in DISC is given in terms of the probability of an event $\vec{x_0}$ to belong to the class Ci:

$$P_{oi} = N\, exp\,(-\frac{1}{2}D^2_{oi}), \tag{7}$$

where N is the normalization factor such that

$$\sum_{i=1}^{h} P_{oi} = 1. \tag{8}$$

The rule (6) to classify the event $\vec{x_0}$ into class Ci can be rewritten with P_{oi} :

$$P_{oi} = \underset{\substack{j=1,\,k \\ i \neq j}}{Max}\, P_{oj} \tag{9}$$

When performing a classification one gets from DISC the value n_{ij} of events of genuine class i classified as class j. The purity of the sample classified as class 1 is thus:

$$Pur_1 = \frac{n_{11}}{n_{11} + n_{21} + n_{31}} \tag{10}$$

and the efficiency

$$E_{ff_1} = \frac{n_{11}}{n_{11} + n_{12} + n_{13}} \tag{11}$$

The program DISC generates the Σ^{-1} matrix and the $\vec{\mu_i}$ vectors with the selected variables. The learning sample events are classified into the 3 C_i classes.

In order to classify an unknown event, the outputs of DISC running on the learning sample are gathered in a file which contains Σ^{-1} and $\vec{\mu_i}$.

This output file can be read from any analysis program in a standard way or can be put in a Bos bank through the ALPHA CARDS allowing the classification to be processed inside an ALPHA program.

3.2.3 New rule of discrimination

When the overlap of the distributions functions for the different classes is important, the purity inside a class is low; this purity can however be improved as explained below.

Selecting a value P_c, considered as a lower probability limit in the range

$$0.3 < P_c < 1 \tag{12}$$

an event will be classified inside the class C_i if the two following conditions are fullfilled :

$$P_{oi} = \underset{\substack{j=1,\ k \\ i \neq j}}{Max}\ P_{oj} \tag{13}$$

and

$$P_{oi} \geq P_c \tag{14}$$

The application of (14) can reject some events of class C_i but the whole purity of this class will be increased.

3.2.4 Monte Carlo test sample and ALEPH data sample

The conditions (13) and (14) give the rule of classification which have been looking for, a check of consistency is then performed by classifying events of both the decays $n(b \to e)$ and $b \to hadrons$ for 60 000 events coming from Monte-Carlo test sample (20 000 for each b, c, and u d s quark flavours) and from ALEPH data (10 000 events).

3.3 Results with $b \to e$

This decay has been studied otherwise by using cuts on the momentum of the electron and on the p_T of the electron [12]. Such a method is very well established and can thus be compared with the present results of the discriminant analysis which has been done in figure 2 and in figure 8.

The variables selected by SELDISC for the discriminant analysis DISC are the double sphericity, electron momentum, p_T of the electron, sum of the transverse momenta of the jet's tracks and H_2, the second Fox-Wolfram momentum [13].

We can get an insight into the discrimination by comparing the ratio $n(b \to e)/n(e)$; the results are presented in table 1.

The ratio $n(b \to e)/n(e)$ versus the probability limit P_c for the case of the leptonic decay is given on figure 1 and the purity of a selected sample of b events versus the efficiency $n(b \to e)/n(events)$ is given in figure 2; the errors are statistical.

We note, for $n(b \to e)/n(e)$, a good agreement between the Monte Carlo and the data showing that the purity versus efficiency evolution computed on the test sample can be applied to the data.

3.4 Results with $b \to hadron$

In such purely hadronic events the light quark background is important due to the fact that 76% of the light quarks have no electronic decay comparing with only 60% for the b quarks.

Up to now any method is known allowing to separate, in an inclusive way, the purely hadronic decays of b quark.

The selection has been made in two way. The discriminant variables are taken at first with no extra cut, then to improve the purity, a double-sphericity lower than 0.13 has been required.

The variables selected by the program SELDISC, in the first case are : double sphericity, number of charged particles and the transverse momentum of the tracks inside the jet. The selected variables, in the second case, are : double sphericity, H_5 and the energy of the jet.

Table 2 shows the ratio $n(b \to had)/n(events)$ computed from the Monte Carlo test sample and from the datas when no cut is applied to the double-sphericity.

Figure 3 give the values $n(b \rightarrow had)/n(events)$ versus P_c computed with Monte Carlo events and with ALEPH data; the figure 4 give the purity of the test sample versus the efficiency.

The purity obtained with no double sphericity cut is rather low : 50% compared to the value of 60% with cut. This last value is yet lower than in the case of the semi-leptonic b decay due to the background of light quarks.

A better separation will be provided by the use of the mini vertex detector in ALEPH, removing most of the light quark background.

4 – Classification tree [18]

A different method of classification is the construction of binary trees. Such trees provide a hierarchical type of representation of the data space that can be readily used as a basis for the classification by following the appropriate branches of the tree.

4.1 Method of binary tree

Let X a set of objects to be classified, the so-called binary tree structured classifier is constructed by repeated split of X into two descendant subsets beginning with X itself. Such a process is pictured in figure 5.

The sets X_2 and X_3 are disjoint, with $X = X_2 U X_3$, similarly X_4 and X_5 are disjoint with $X_2 = X_4 U X_5$, and $X_6 U X_7 = X_3$. Those subsets which are not split, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{14}, X_{15} are called terminal subsets (rectangular boxes).

These terminal subsets provide a partition of X, a class label is associated to each terminal subset. There may be two or more terminal subsets with the same class label.

To explain how this split is made at each node let us consider at first the one dimensional case. Let $f_1(x)$ and $f_2(x)$ the two continuous density function associated with two classes of objects.

$$F_i(x) = \int_0^x f_i(x)dx \quad (i = 1, 2) \tag{15}$$

is the associated distribution.

It can be shown[18] that the value x^* of x which minimize the Kolmogorov-Smirnoff distance

$$D(x^*) = |F_1(x) - F_2(x)| \tag{16}$$

minimize also the mean cost of misclassification, according to the Baye's rule.

Splitting the space in two subsets according to the comparison of x to x^* ($x \leq x^*$; $x > x^*$) gives two subsets purer than the parent.

In the more complex case of p variables associated with one object, p Kolmogorov-Smirnoff distances are computed at each node

$$D(x_j^*) = Max_{xj}|F_1(x_j) - F_2(x_j)| \tag{17}$$

and a cut is made on the greater of x_j^*.

4.2 Programs

Two programs can be used to build such trees : DNP [16] (Discrimination Non Paramétrique) and CART [19] (Classification and Regression Tree).

The difference between DNP and CART is that DNP minimizes the cost at each node while CART minimizes the global cost and give a pruned tree for which this global cost is minimum.

The final result is that the DNP tree is larger than the CART one but the tree provided by DNP can be handled more easily in the sense of purity versus efficacity evolution.

The tree is used as a classifier. If an event belonging to an unknown class is dropped into a tree and ends up in a terminal node labelled as class j, it is classified as a class j event.

4.3 Results with DNP

Binary trees have been built for the two cases $b \to e$ and $b \to hadrons$.

4.3.1 $b \to e$

The learning sample contains events belonging to the 3 classes, the proportions beeing those of real events which decay with an electron.

The contents of the learning sample is

2 706 quark b events
1 224 quark c events
2 072 quark uds events

This sample is the input of the DNP program, the output is a tree given in figure 5. The two branches with the best purity for the b quark sample have been kept only, these two branches are pictured in figure 6. Each final node is reached by a set of successive cuts giving the classification rule.

These rules applied to the Monte Carlo test sample and to the ALEPH data sample are checked by the comparison of the ratio $n(b \to e)/n(e)$ in both cases as for the linear discriminant method. The results are given in table 3.

With the branch 1, the agreement between the Monte Carlo and the data is good : we can get a high purity sample of events classified as b events.

With the branch 2, the agreement between the different results is poor : the number of events classified in every segment is small and the statistical error more important.

4.3.2 $b \to$ hadrons

The learning sample is made of purely hadronic events.

The proportion between the different classes

626 b quark events
546 c quark events
2 157 *uds* quark events

corresponding to the natural proportion.

As for the $b \to e$ case two branches of the output tree have been kept (Fig. 7).

Branch 1 allows the classification of b quark events with a maximum purity of 65%, in branch 2 one can extract the light quarks events with a purity of 84%.

4.4 Results with CART

Due to the automatic pruning of branches in the CART program to get a minimal global cost, the tree is very small.

The tree with the classification rule is given in table 4.

On the same table we give the results of a computation of $n(b \to e)/n(e)$ with a Monte Carlo test sample and an ALEPH data sample.

The agreement is good.

5 – Conclusion

In this work we were aiming to the construction of a method to classify precise types of events in ALEPH.

Two methods have been used : linear discrimination and segmentation both giving classification rules for b quark events in ALEPH.

The results are compared in figure 8 where the plot purity versus efficiency is given for different methods. It is worth noting that a high purity classified sample can be found with discriminant analysis methods.

When considering the parameters connected only to the discrimination a good agreement has been found between the test Monte Carlo sample and the ALEPH data sample for both the ratios $n(b \rightarrow e)/n(e)$ and $n(b \rightarrow had)/n(events)$.

Stating a first conclusion it must be pointed out that a good discrimination of such ALEPH event needs :

– a good Monte Carlo for the learning sample. Actually our variables are very well described by the Monte Carlo apart from a slight disagreement in electron identification.

– an accurate set of variables to describe the type of events. A large number of variables must be reduced with appropriate algorithms before performing the classification rules.

For the present study the performances are comparable with those of the electron p_T cut. On the other hand, for the purely hadronic decay for which there is up to now any known method, the pattern recognition provide a good way to get b quark sample.

Some ameliorations are possible:

– a better electron identification can be done by reconstructed data and a new version of the Monte Carlo generated events.

– the new variable connected with the number of vertices will play the main role for lowering the light quarks background.

The same method can be applied to other problems like the discrimination of a quark jet from a gluon jet.

Acknowledgements

We would like to thank Dr G. CELEUX and Y. LECHEVALLIER from INRIA (PARIS) for helpfull comments and discussions.

REFERENCES

[1] J.T. TOU, R.C. GONZALEZ
Pattern recognition principles
Addison Wesley N.Y. (1974)

[2] J.C. SIMON
La reconnaissance des formes par algorithmes
Masson Paris (1985)

[3] M.S. SRIVASTAVA, E.M. CARTER
An introduction to multivariate statistics
North Holland Amsterdam (1983)

[4] L. LEBART, A. MORINEAU, K.M. WARWICK
Multivariate descriptive statistical analysis
Wiley N.Y. (1984)

[5] E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU
Eléments d'analyse de données
Dunod Paris (1982)

[6] G. CELEUX, E. DIDAY, G. GOVAERT, Y. LECHEVALLIER, H. RALAMBONDRAINY
Classification automatique des données
Dunod Paris (1989)

[7] H. GROTE
Rep. Prog. Phys. 50 (1987) 473

[8] R. ODORICO
Phys. Lett. B120 (1983) 219
G. BALLOCHI, R. ODORICO
Nucl. Phys. B229 (1983) 1
P. LUTZ
DELPHI-NOTE 85-3 PHYS 3
D. DELIKARIS, P. LUTZ
DELPHI-NOTE 87-3 PHYS 15
D. DELIKARIS
Thèse, C de F Paris (1987)
P. LUTZ
Cours de GIF IN2P3 Paris (1988)

[9] M. MJAED
Thèse L.P.C. Clermont (1987)

M. MJAED, J. PRORIOL
Phys. Lett. B127 (1989) 560
Ch. de la VAISSIERE Annals of N.Y. Academy of Sciences (1988)

[10] A. EALET
Thèse Marseille (1987)

[11] C. GUICHENEY
Rapport DEA Clermont (1989)

[12] ALEPH Collaboration Preprint CERN-EP 90-54

[13] H. ALBRECHT, E. BLUCHER
ALPHA USER'S GUIDE - ALEPH 89-151

[14] SAU LAN WU
Nucl. Phys. B3 (Proc. Suppl.) (1988) 102

[15] P. HENRARD
Preprint Clermont PCCF RI 88-08

[16] Library MODULAD
INRIA Paris

[17] G. CELEUX
Private communication

[18] J.H. FRIEDMAN
IEEE Trans. Comp. C26 (1977) 404

G. CELEUX, Y. LECHEVALLIER
in COMPSTAT 82
PHYSICA-VERLAG (1982)

G. CELEUX, Y. LECHEVALLIER
Revue de Stat. Appl. 30 (1982) 39

L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN, C.J. STONE
Classification and regression trees
Wordworth California (1984)

[19] CART Program
California Statistical Software Inc (USA)

TABLES

Table 1 Monte Carlo results and experimental results for the case $b \rightarrow e$.

Table 2 Monte Carlo results and experimental results for the case $b \rightarrow hadrons$ (without cut).

Table 3 Monte Carlo results, experimental results, DNP results for two branches of a DNP tree, in the $b \rightarrow e$ case.

Table 4 Monte Carlo results, experimental results, CART results, in the $b \rightarrow e$ case.

FIGURES

$b \to e$

Discriminant analysis

Pc	Monte Carlo		Data
	Purity %	$\frac{n(b \to e)}{n(e)}$	$\frac{n(b \to e)}{n(e)}$
0.44	75.1 ± 3.7	0.305 ± 0.012	0.278 ± 0.012
0.50	77.9 ± 4.1	0.250 ± 0.013	0.217 ± 0.013
0.72	83.9 ± 7.2	0.088 ± 0.004	0.084 ± 0.004

Table 1

$b \rightarrow hadrons$

Discriminant analysis

EVENTS	Monte Carlo b	c	uds	mean value %	Data
generated	17 194	17 937	36 521		9 846
without e	59.8 %	68.9 %	76.3 %	71.44 ± 0.51 %	73.68 ± 0.87 %

Pc	Monte Carlo $\frac{n(b \rightarrow had)}{n(events)}$ %	Purity %	Data $\frac{n(b \rightarrow had)}{n(events)}$ %
0.50	10.05 ± 0.20	41.5 ± 1.5	9.85 ± 0.32
0.60	6.4 ± 0.16	44.3 ± 2.0	6.11 ± 0.25
0.70	3.8 ± 0.12	46.9 ± 2.7	3.68 ± 0.19
0.80	1.84 ± 0.08	50.0 ± 4.1	1.81 ± 0.14

Table 2

$b \to e$

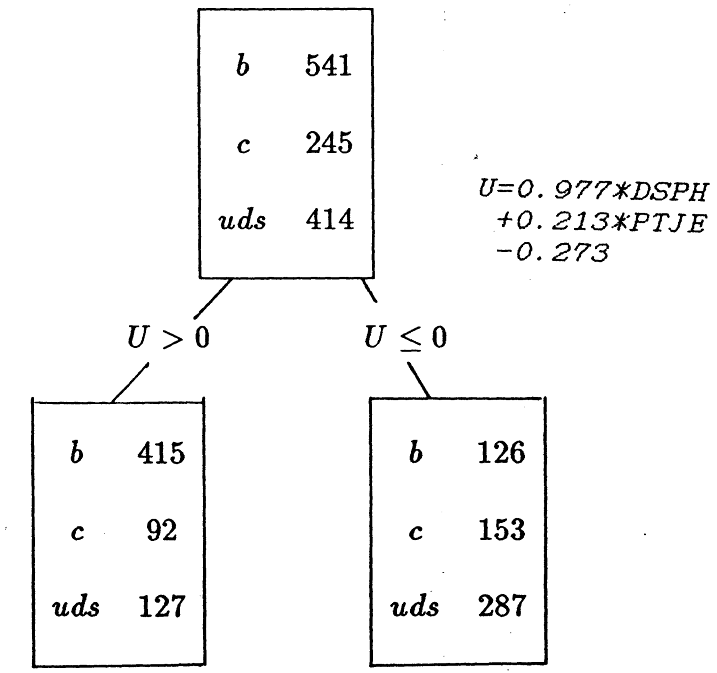
D.N.P. with 2 branches

SEGMENTS	Monte Carlo Purity %	Monte Carlo $\frac{n(b \to e)}{n(e)}$	Data[12] $\frac{n(b \to e)}{n(e)}$
Branch 1			
Node 3	65.7 ± 2.9	0.4079 ± 0.0181	0.4325 ± 0.0353
Node 7	78.5 ± 4.4	0.2253 ± 0.0115	0.2066 ± 0.0266
Node 15	85.8 ± 5.6	0.1449 ± 0.0084	0.1510 ± 0.0177
Node 31	89.8 ± 6.9	0.0988 ± 0.0054	0.0931 ± 0.0126
Branch 2			
Node 2	31.7 ± 1.4	0.5930 ± 0.0238	0.5685 ± 0.0429
Node 5	47.3 ± 2.8	0.2797 ± 0.0141	0.2516 ± 0.0241
Node 11	53.4 ± 3.7	0.1985 ± 0.0110	0.1788 ± 0.0197
Node 23	58.4 ± 4.8	0.1298 ± 0.0081	0.1039 ± 0.0140
Node 46	64.9 ± 6.9	0.0745 ± 0.0055	0.0632 ± 0.0106
Node 92	66.3 ± 9.1	0.0436 ± 0.0041	0.0300 ± 0.0063

Table 3

$b \rightarrow e$

CART



Purity : 65.4%

Monte Carlo		Data	CART	
Purity %	$\frac{n(b \rightarrow e)}{n(e)}$	$\frac{n(b \rightarrow e)}{n(e)}$	Purity %	$\frac{n(b \rightarrow e)}{n(e)}$
$64.6 \pm 6.3\,\%$	0.5336 ± 0.0207	0.5364 ± 0.0410	$65.4\,\%$	0.5283

Table 4

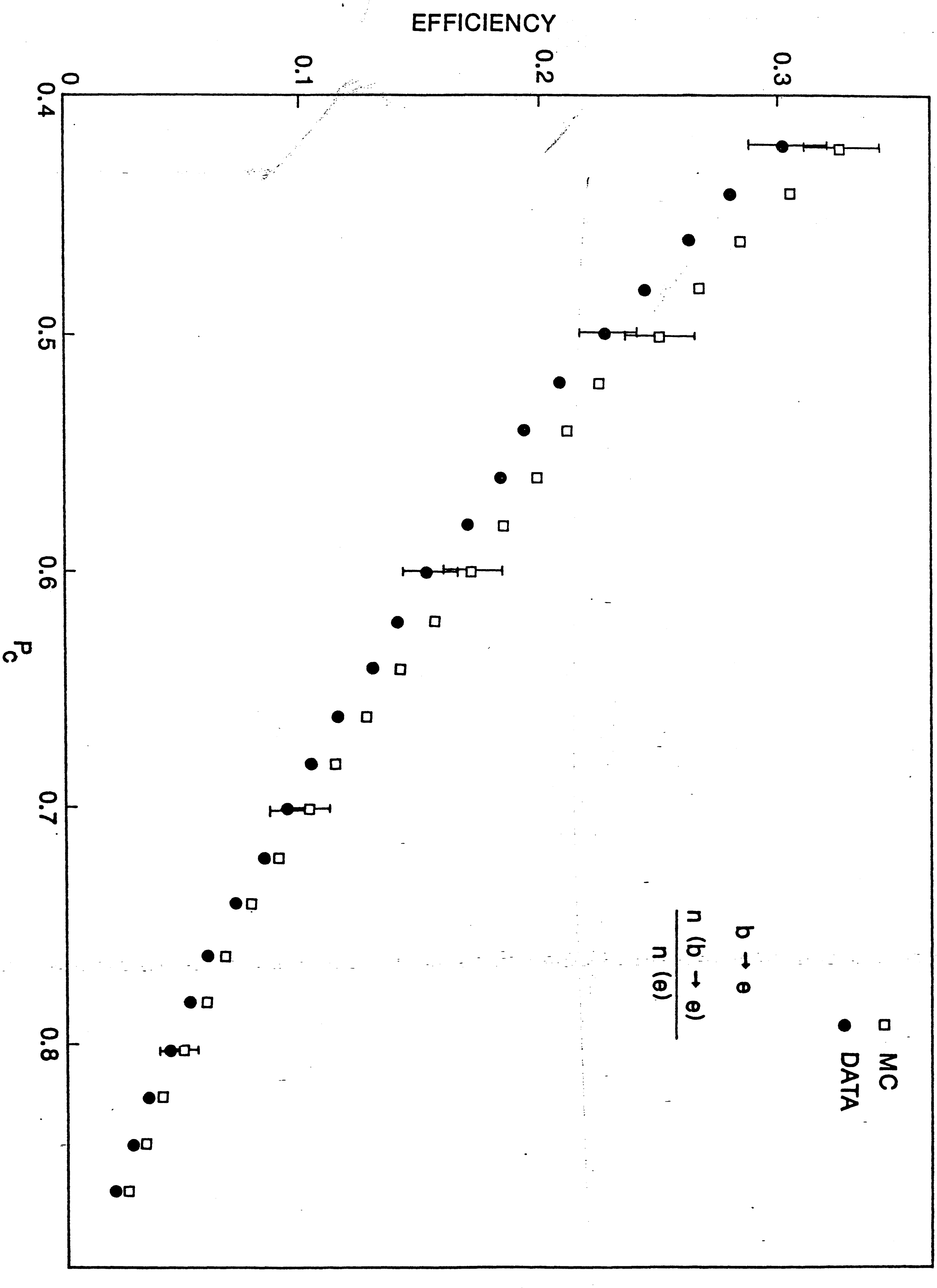


Fig.1

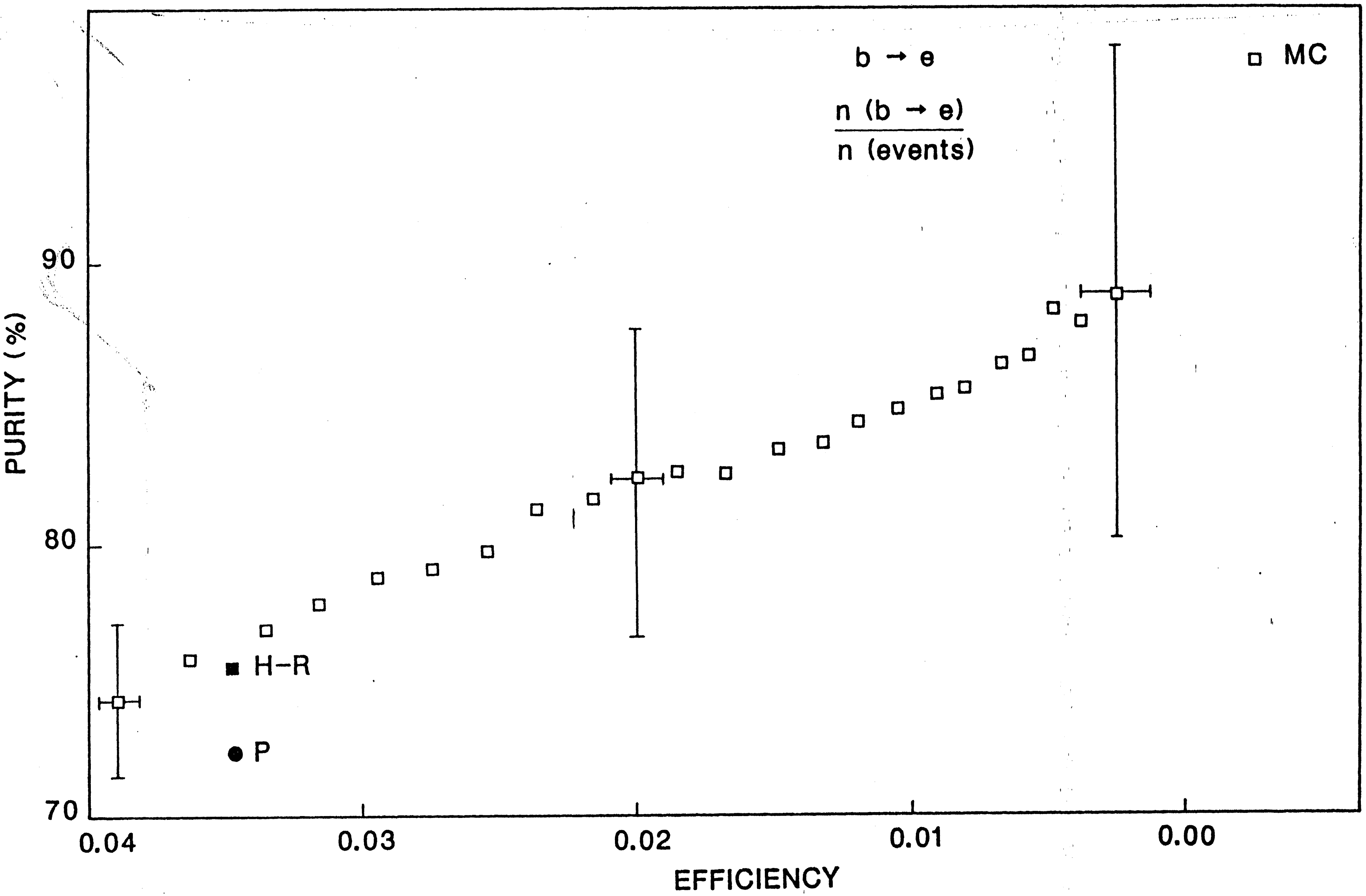
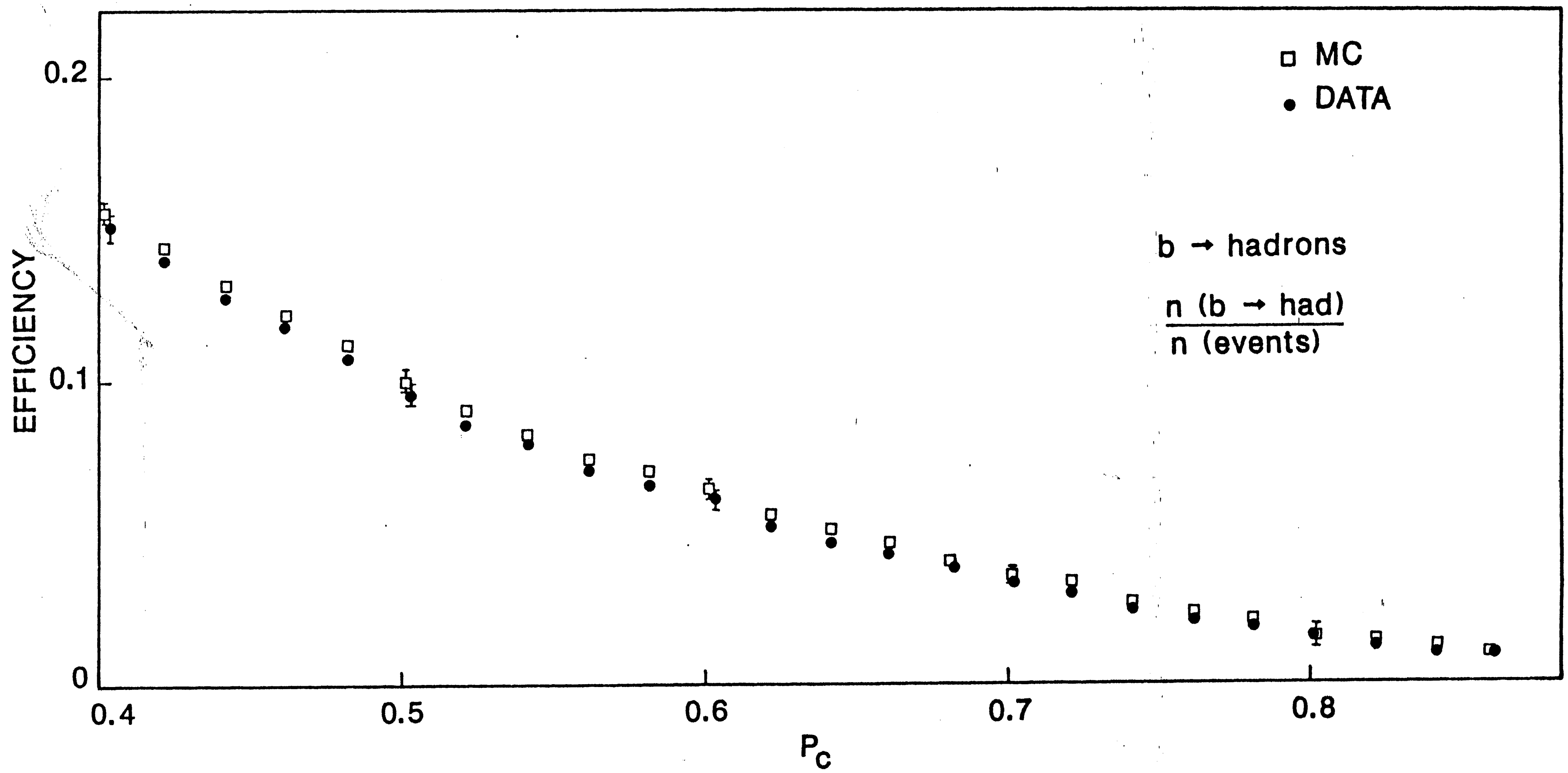
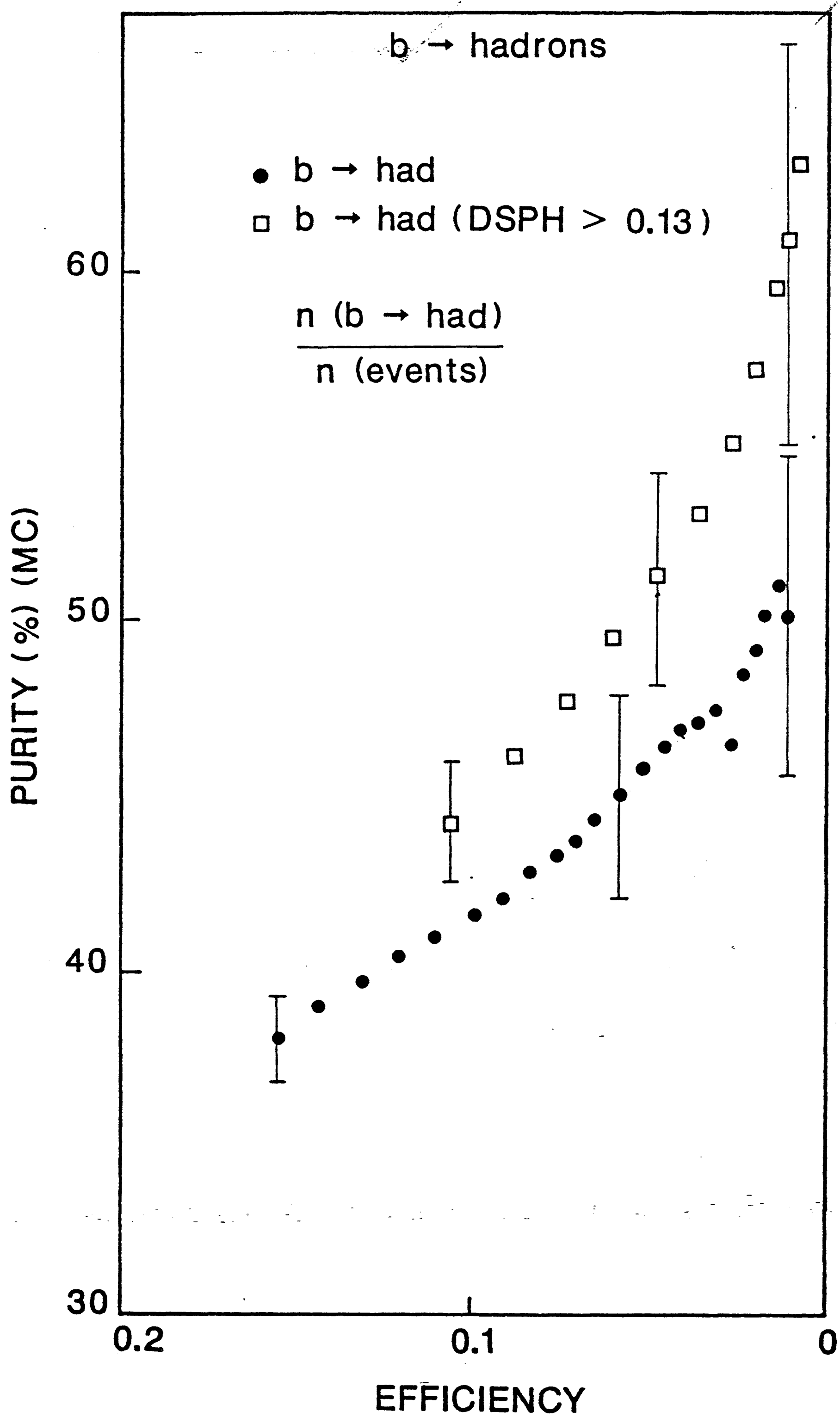

PURITY (%)
90
80
70
0.04
0.03
0.02
0.01
0.00
EFFICIENCY
b → e
n (b → e)
n (events)
MC
H–R
P

Fig.2



MC
DATA
b → hadrons
n (b → had) / n (events)
EFFICIENCY
0.2
0.1
0
0.4
0.5
0.6
0.7
0.8
P_c

Fig.3



b → hadrons
b → had
b → had (DSPH > 0.13)
n (b → had) / n (events)
PURITY (%) (MC)
60
50
40
30
0.2
0.1
0
EFFICIENCY

Fig.4

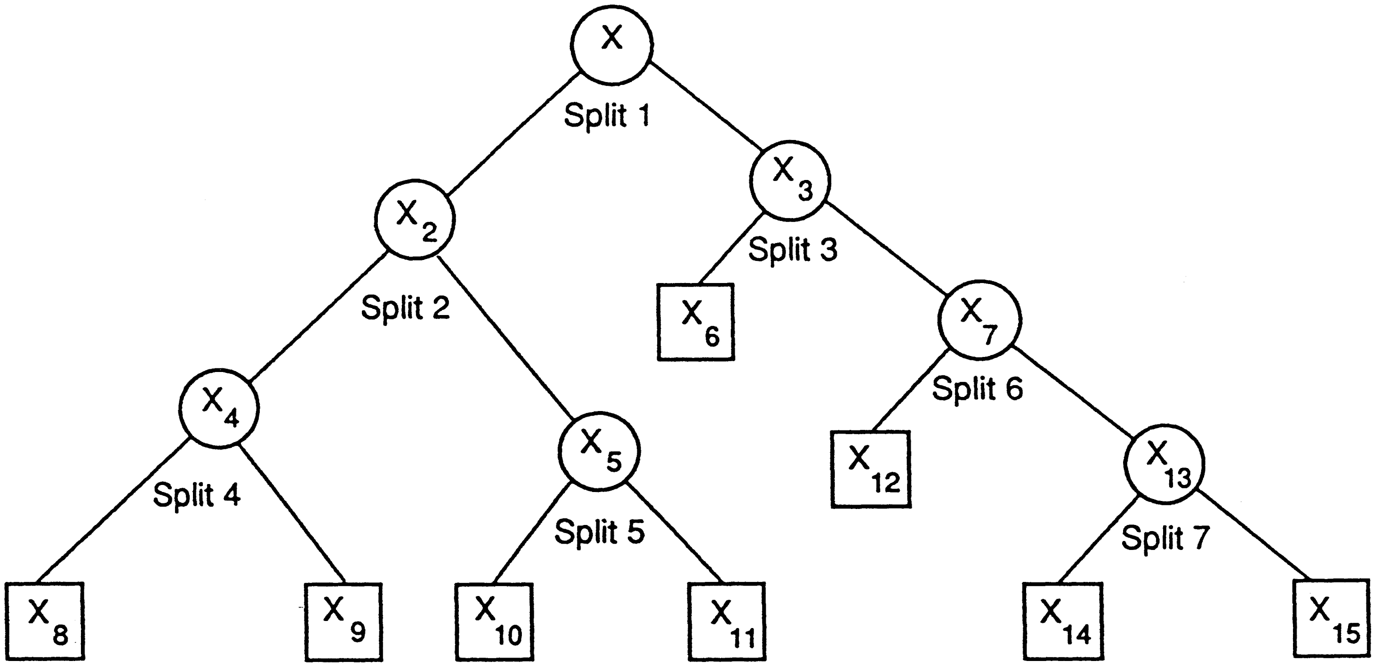

X
Split 1
X2
X3
Split 2
Split 3
X6
X7
X4
Split 6
X5
X12
X13
Split 4
Split 5
Split 7
X8
X9
X10
X11
X14
X15

Fig. 5

2 BRANCHES OF A DNP TREE (b --> e)

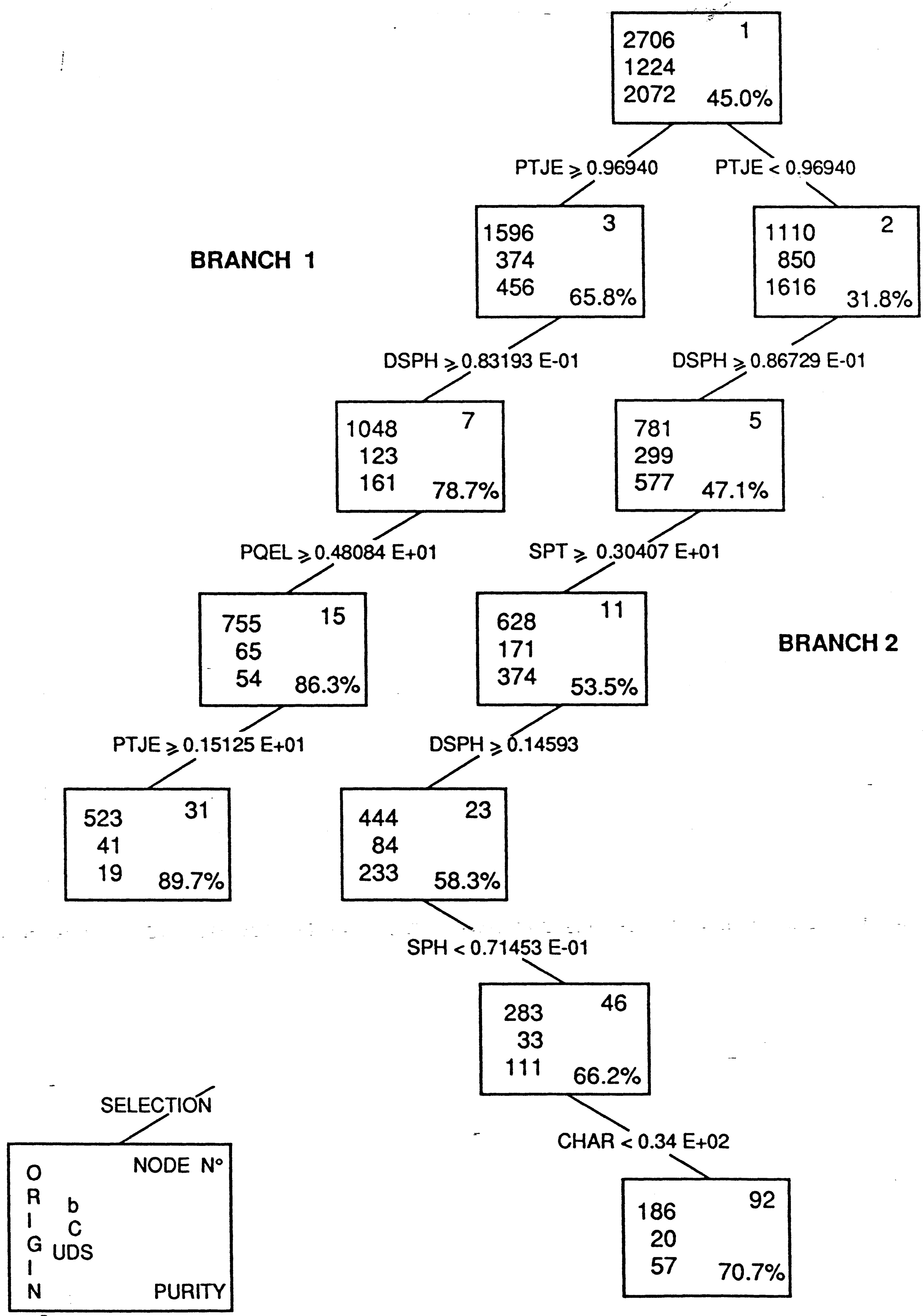


Fig.6

2 BRANCHES OF A DNP TREE (b --> hadrons)

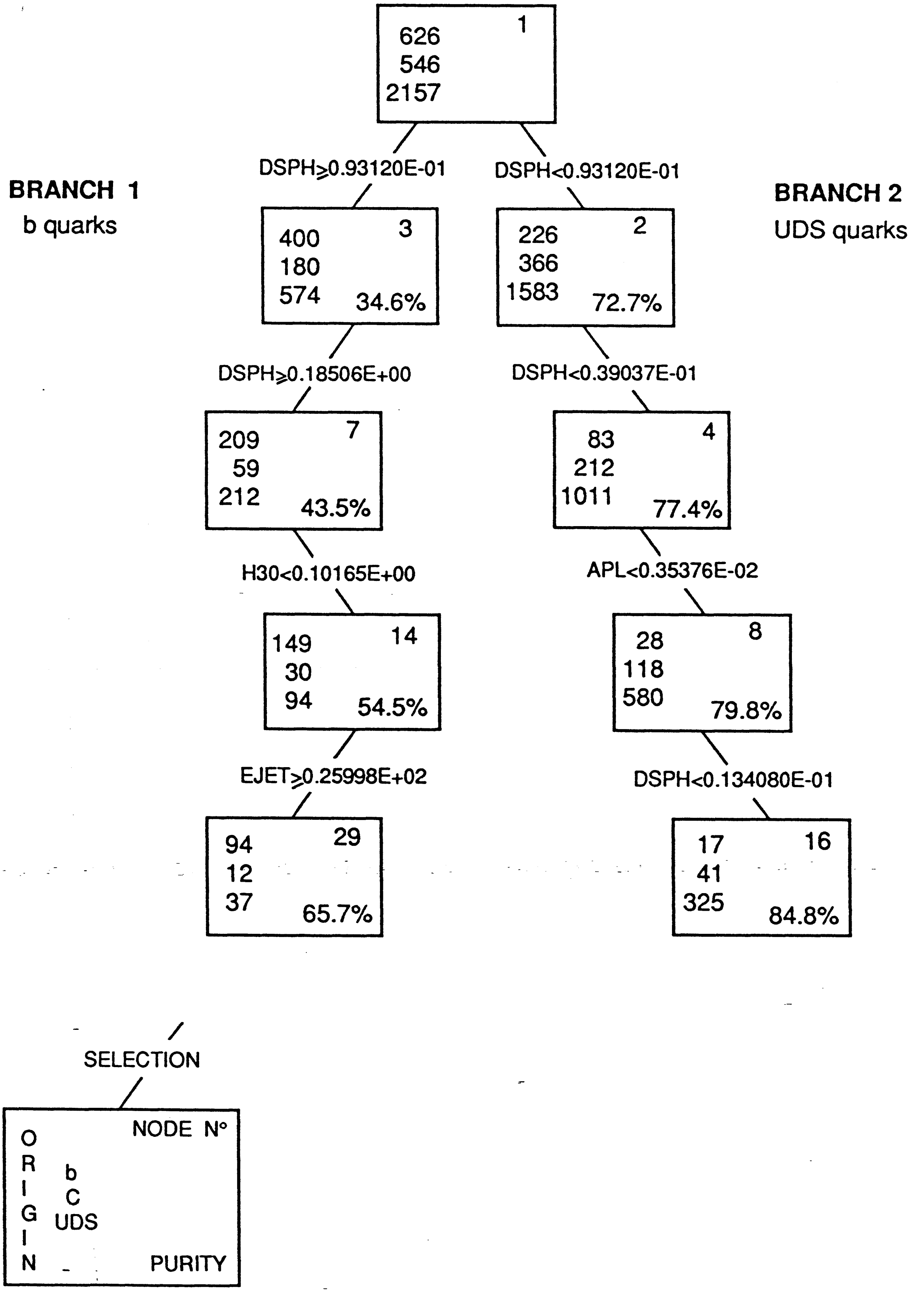


Fig.7

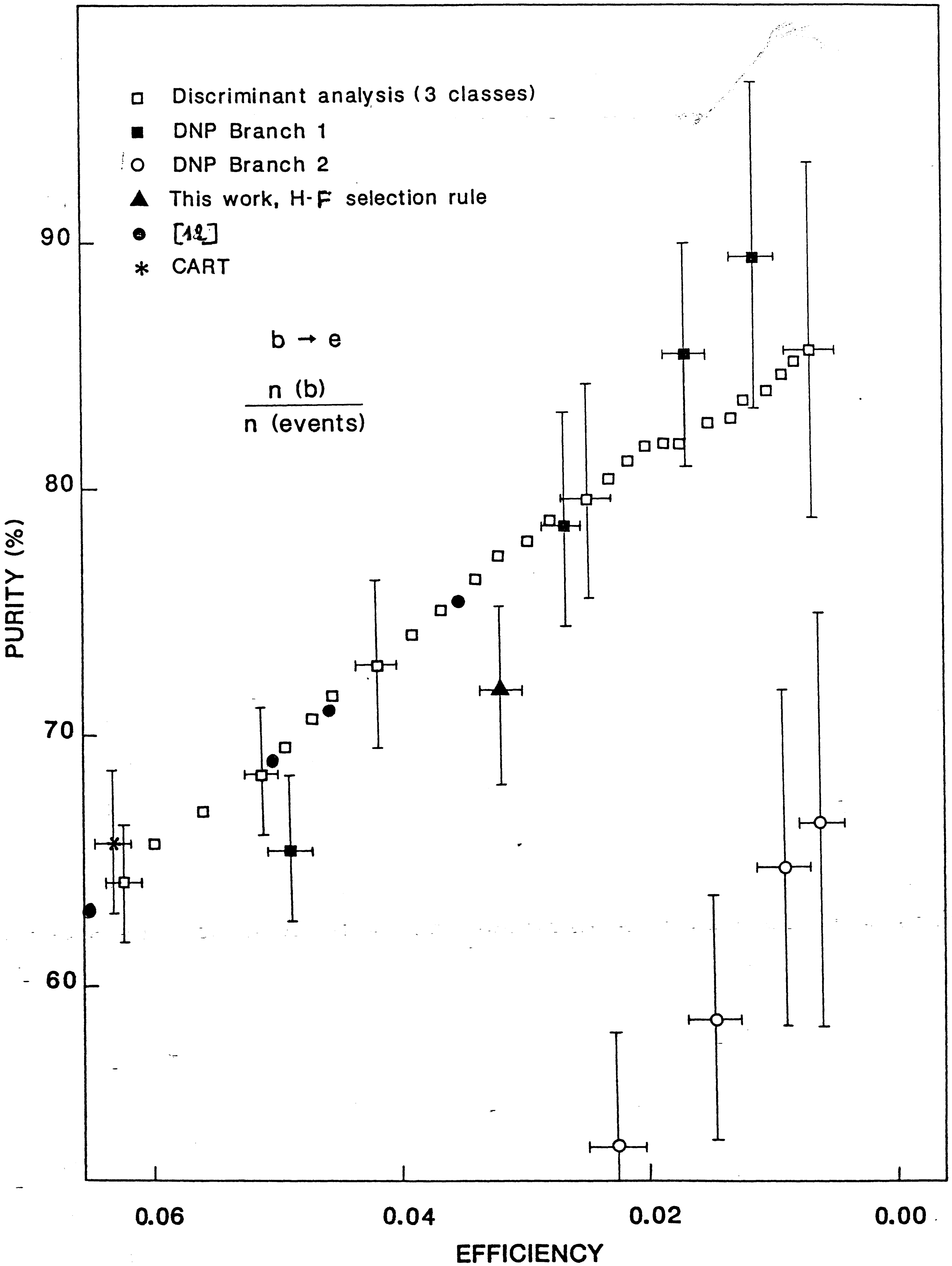

Discriminant analysis (3 classes)
DNP Branch 1
DNP Branch 2
This work, H-F selection rule
[12]
CART
b → e
n (b) / n (events)
PURITY (%)
90
80
70
60
EFFICIENCY
0.06
0.04
0.02
0.00

Fig.8

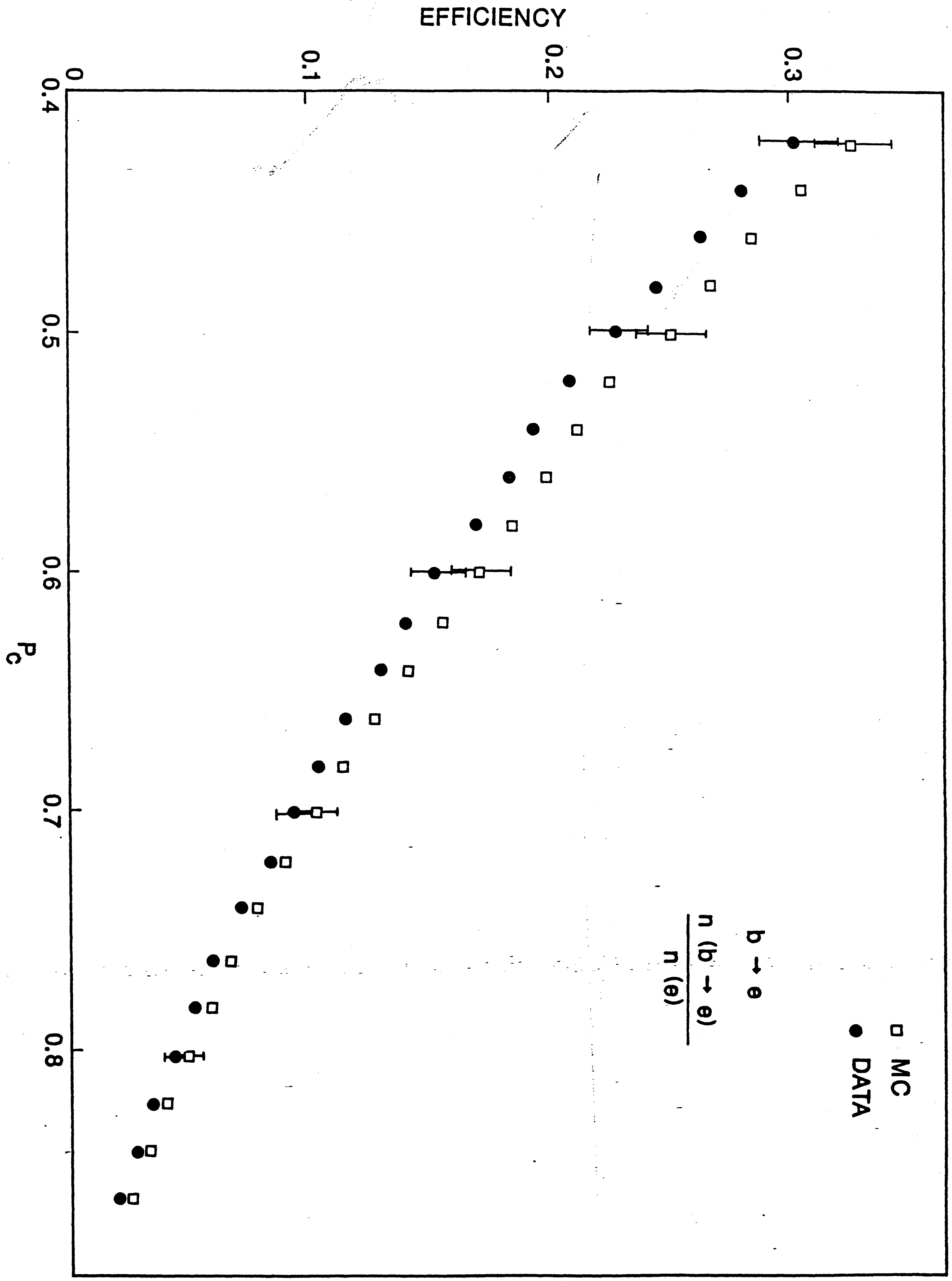


Fig.1

$b \rightarrow e$

D.N.P. with 2 branches

SEGMENTS	Monte Carlo Purity %	Monte Carlo $\frac{n(b \rightarrow e)}{n(e)}$	Data[12] $\frac{n(b \rightarrow e)}{n(e)}$
Branch 1			
Node 3	65.7 ± 2.9	0.4079 ± 0.0181	0.4325 ± 0.0353
Node 7	78.5 ± 4.4	0.2253 ± 0.0115	0.2066 ± 0.0266
Node 15	85.8 ± 5.6	0.1449 ± 0.0084	0.1510 ± 0.0177
Node 31	89.8 ± 6.9	0.0988 ± 0.0054	0.0931 ± 0.0126
Branch 2			
Node 2	31.7 ± 1.4	0.5930 ± 0.0238	0.5685 ± 0.0429
Node 5	47.3 ± 2.8	0.2797 ± 0.0141	0.2516 ± 0.0241
Node 11	53.4 ± 3.7	0.1985 ± 0.0110	0.1788 ± 0.0197
Node 23	58.4 ± 4.8	0.1298 ± 0.0081	0.1039 ± 0.0140
Node 46	64.9 ± 6.9	0.0745 ± 0.0055	0.0632 ± 0.0106
Node 92	66.3 ± 9.1	0.0436 ± 0.0041	0.0300 ± 0.0063

Table 3

$b \to e$

Discriminant analysis

Pc	Monte Carlo		Data
	Purity %	$\frac{n(b \to e)}{n(e)}$	$\frac{n(b \to e)}{n(e)}$
0.44	75.1 ± 3.7	0.305 ± 0.012	0.278 ± 0.012
0.50	77.9 ± 4.1	0.250 ± 0.013	0.217 ± 0.013
0.72	83.9 ± 7.2	0.088 ± 0.004	0.084 ± 0.004

Table 1

TABLES

Table 1 Monte Carlo results and experimental results for the case $b \rightarrow e$.

Table 2 Monte Carlo results and experimental results for the case $b \rightarrow hadrons$ (without cut).

Table 3 Monte Carlo results, experimental results, DNP results for two branches of a DNP tree, in the $b \rightarrow e$ case.

Table 4 Monte Carlo results, experimental results, CART results, in the $b \rightarrow e$ case.

REFERENCES

[1] J.T. TOU, R.C. GONZALEZ
Pattern recognition principles
Addison Wesley N.Y. (1974)

[2] J.C. SIMON
La reconnaissance des formes par algorithmes
Masson Paris (1985)

[3] M.S. SRIVASTAVA, E.M. CARTER
An introduction to multivariate statistics
North Holland Amsterdam (1983)

[4] L. LEBART, A. MORINEAU, K.M. WARWICK
Multivariate descriptive statistical analysis
Wiley N.Y. (1984)

[5] E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU
Eléments d'analyse de données
Dunod Paris (1982)

[6] G. CELEUX, E. DIDAY, G. GOVAERT, Y. LECHEVALLIER, H. RALAMBONDRAINY
Classification automatique des données
Dunod Paris (1989)

[7] H. GROTE
Rep. Prog. Phys. 50 (1987) 473

[8] R. ODORICO
Phys. Lett. B120 (1983) 219
G. BALLOCHI, R. ODORICO
Nucl. Phys. B229 (1983) 1
P. LUTZ
DELPHI-NOTE 85-3 PHYS 3
D. DELIKARIS, P. LUTZ
DELPHI-NOTE 87-3 PHYS 15
D. DELIKARIS
Thèse, C de F Paris (1987)
P. LUTZ
Cours de GIF IN2P3 Paris (1988)

These rules applied to the Monte Carlo test sample and to the ALEPH data sample are checked by the comparison of the ratio $n(b \to e)/n(e)$ in both cases as for the linear discriminant method. The results are given in table 3.

With the branch 1, the agreement between the Monte Carlo and the data is good : we can get a high purity sample of events classified as b events.

With the branch 2, the agreement between the different results is poor : the number of events classified in every segment is small and the statistical error more important.

4.3.2 $b \to$ hadrons

The learning sample is made of purely hadronic events.

The proportion between the different classes

626 b quark events
546 c quark events
2 157 uds quark events

corresponding to the natural proportion.

As for the $b \to e$ case two branches of the output tree have been kept (Fig. 7).

Branch 1 allows the classification of b quark events with a maximum purity of 65%, in branch 2 one can extract the light quarks events with a purity of 84%.

4.4 Results with CART

Due to the automatic pruning of branches in the CART program to get a minimal global cost, the tree is very small.

The tree with the classification rule is given in table 4.

On the same table we give the results of a computation of $n(b \to e)/n(e)$ with a Monte Carlo test sample and an ALEPH data sample.

The agreement is good.

5 – Conclusion

In this work we were aiming to the construction of a method to classify precise types of events in ALEPH.

Two methods have been used : linear discrimination and segmentation both giving classification rules for b quark events in ALEPH.

Figure 3 give the values $n(b \to had)/n(events)$ versus P_c computed with Monte Carlo events and with ALEPH data; the figure 4 give the purity of the test sample versus the efficiency.

The purity obtained with no double sphericity cut is rather low : 50% compared to the value of 60% with cut. This last value is yet lower than in the case of the semi-leptonic b decay due to the background of light quarks.

A better separation will be provided by the use of the mini vertex detector in ALEPH, removing most of the light quark background.

4 – Classification tree [18]

A different method of classification is the construction of binary trees. Such trees provide a hierarchical type of representation of the data space that can be readily used as a basis for the classification by following the appropriate branches of the tree.

4.1 Method of binary tree

Let X a set of objects to be classified, the so-called binary tree structured classifier is constructed by repeated split of X into two descendant subsets beginning with X itself. Such a process is pictured in figure 5.

The sets X_2 and X_3 are disjoint, with $X = X_2 U X_3$, similarly X_4 and X_5 are disjoint with $X_2 = X_4 U X_5$, and $X_6 U X_7 = X_3$. Those subsets which are not split, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{14}, X_{15} are called terminal subsets (rectangular boxes).

These terminal subsets provide a partition of X, a class label is associated to each terminal subset. There may be two or more terminal subsets with the same class label.

To explain how this split is made at each node let us consider at first the one dimensional case. Let $f_1(x)$ and $f_2(x)$ the two continuous density function associated with two classes of objects.

$$F_i(x) = \int_0^x f_i(x)dx \quad (i = 1, 2) \tag{15}$$

is the associated distribution.

It can be shown[18] that the value x^* of x which minimize the Kolmogorov-Smirnoff distance

$$D(x^*) = |F_1(x) - F_2(x)| \tag{16}$$

minimize also the mean cost of misclassification, according to the Baye's rule.

$$Pur_1 = \frac{n_{11}}{n_{11} + n_{21} + n_{31}} \tag{10}$$

and the efficiency

$$E_{ff_1} = \frac{n_{11}}{n_{11} + n_{12} + n_{13}} \tag{11}$$

The program DISC generates the Σ^{-1} matrix and the $\vec{\mu_i}$ vectors with the selected variables. The learning sample events are classified into the 3 C_i classes.

In order to classify an unknown event, the outputs of DISC running on the learning sample are gathered in a file which contains Σ^{-1} and $\vec{\mu_i}$.

This output file can be read from any analysis program in a standard way or can be put in a Bos bank through the ALPHA CARDS allowing the classification to be processed inside an ALPHA program.

3.2.3 New rule of discrimination

When the overlap of the distributions functions for the different classes is important, the purity inside a class is low; this purity can however be improved as explained below.

Selecting a value P_c, considered as a lower probability limit in the range

$$0.3 < P_c < 1 \tag{12}$$

an event will be classified inside the class C_i if the two following conditions are fullfilled :

$$P_{oi} = \underset{\substack{j=1,\ k \\ i \neq j}}{Max}\ P_{oj} \tag{13}$$

and

$$P_{oi} \geq P_c \tag{14}$$

The application of (14) can reject some events of class C_i but the whole purity of this class will be increased.

3.2.4 Monte Carlo test sample and ALEPH data sample

The conditions (13) and (14) give the rule of classification which have been looking for, a check of consistency is then performed by classifying events of both

$\vec{\mu}$ being the p-components vector of the the mean values and Σ the symmetric variance matrix.

Considering the simple case of 2 classes C_1 et C_2, $\vec{\mu_1}$ and $\vec{\mu_2}$ are the $\vec{\mu}$ vectors of these classes, the Σ matrix is assumed to be the same for C_1 and C_2.

The Fisher linear discriminant function [3] maximizes the norm of the vector $\vec{\delta} = \vec{\mu_1} - \vec{\mu_2}$ which gives the distance of the two classes.

In a matrix notation the linear discriminant function L for $\vec{x_0}$ can be written [3]

$$L(\vec{x_0}) = \vec{\delta}[\Sigma^{-1}](\vec{x_0})^T \tag{1}$$

$\vec{x_0}$ will be classified inside the class C_1 if:

$$L(\vec{x_0}) > \frac{1}{2}(\vec{\mu_1} - \vec{\mu_2})[\Sigma^{-1}](\vec{\mu_1} + \vec{\mu_2})^T \tag{2}$$

and inside the class C_2 if:

$$L(\vec{x_0}) \leq \frac{1}{2}(\vec{\mu_1} - \vec{\mu_2})[\Sigma^{-1}](\vec{\mu_1} + \vec{\mu_2})^T \tag{3}$$

It can be shown that such a rule of classification is a rule minimizing the Baye's risk of misclassification.

This rule of classification can otherwise be differently stated.

Taking the MAHALANOBIS distance D_{oi} defined by the relation

$$D^2_{oi} = (\vec{x_0} - \vec{\mu_i})[\Sigma^{-1}](\vec{x_0} - \vec{\mu_i})^T \tag{4}$$

D_{oi} can be understood as a generalized distance of $\vec{x_0}$ to the center of mass of the class C_i.

The rule of classification of $\vec{x_0}$ into C_1 is no longer given by the value of $L(\vec{x_0})$ but by the value of D^2_{oi}, which means, for two classes, that if:

$$D^2_{02} < D^2_{01} \tag{5}$$

the event $\vec{x_0}$ is thus closer to C_1 than to C_2 and will be classified into C_1.

The generalization of this rule to k classes is straight forward. Computing $\vec{\mu_i}$ and Σ for all the classes $i = 1, ..., k$, $\vec{x_0}$ will be in class C_i if

$$D^2_{oi} = \min_{\substack{1 \leq j \leq k \\ j \neq i}} D^2_{oj} \tag{6}$$

1.4 Outline of the paper

In the chapter 2 we describe the methods used to select the events of the Monte Carlo samples and the variables used throughout our study.

In the chapter 3 after a short recall of the discriminant analysis method we describe the procedure used to get the classification rule from the learning sample. We then compare Monte Carlo test sample results with ALEPH data results in the two cases of b quark decay.

In the chapter 4 we give a short approach of the classification tree methods applying two programs to the b quark semi-leptonic and purely hadronic decays.

2 – Events and variables

This section is devoted to the origin of the learning sample events, the method used to identify the electrons in an event, and the variables used throughout the study.

2.1 Learning sample events and test sample events

The Monte Carlo events were generated in the ALEPH collaboration by Annecy, Clermont and Marseille. These events have been reconstructed through JULIA, and can thus be fully compared to the data.

A set of b quark events, a set of c quark events and a set of mixed u, d, s quark events have been used, assuming for the test sample the following proportions :

$$b : 21,9\% \; ; \; c : 17,1\% \; ; \; uds : 60,9\% \text{ versus } q\bar{q}.$$

2.2 Identification of the events with an electron

The identification of a leptonic event is made with the standard subroutines already used in ALEPH [12] to tag the $b \rightarrow e$ events.

2.3 Variables associated to an event

We have used a set of variables which are computed with the components (momentum, energy) of the tracks of the event.

1 – Introduction

An important application of artificial intelligence is the pattern recognition [1],[2].

Among the different methods of pattern recognition the statistical multivariate analysis methods had lead to a lot of applications [3]-[6].

1.1 Previous works

The methods of pattern recognition have already been used in high energy physics [7], in particular the multivariate analysis methods have been extensively used in recent studies with Monte Carlo simulations such as the identification of top quark events in UA2 and LEP experiments [8], the determination of the number of jets of an event and the tagging of the quark jet flavour [9],[11]; however these results have not been yet confronted to the data.

The method of classification tree have been used to identify charged clusters in ALEPH electromagnetic calorimeter[10].

The common purpose of these works is to classify an event by building a set of rules called classifier, the discrimination methods described later on are a way to get such rules.

1.2 The discrimination Baye's rule

The basic purpose of a classification is to get an accurate classifier, that is to characterize the conditions allowing to determine whether an object is in one class or an another.

Considering two classes C_1 and C_2 with the probabilities $M_1 = P(C_1)$ and $M_2 = P(C_2)$ for an object to belong to C_1(resp C_2), the goal is to get a decision rule sharing the space of the variables in two regions R_1 and R_2, R_1 being filled with the objects of C_1 and R_2 with those of C_2.

If $C(1,2)$ is the cost in classifying an object of C_1 in R_2 and $C(2,1)$ the cost in classifying an object of C_2 in R_1, $C(1,2)$ and $C(2,1)$ are the misclassification costs.

Let $P(1,2)$ the probability for an object of class 1 to be classified in class 2 and $P(2,1)$ the probability for an object of class 2 to be classified in class 1. The mean misclassification cost is :

CONTENTS

ALEPH 90-95
PHYSIC 90-58
J. Proriol et al.
23.6.1990

EVENT RECOGNITION: MULTIVARIATE ANALYSIS METHODS TO TAG B QUARKS EVENTS IN ALEPH

J. PRORIOL, J. JOUSSET, C. GUICHENEY,
A. FALVARD, P. HENRARD, D. PALLIN,
P. PERRET, F. PRULHIERE

Laboratoire de Physique Corpusculaire
de Clermont-Ferrand
IN2P3/CNRS Université Blaise Pascal
F-63177 AUBIERE CEDEX FRANCE

ABSTRACT :

Two multivariate analysis methods : a linear discriminant method and a classification tree have been performed to get classifiers.

These classifiers have been applied to tag b quark events in ALEPH.

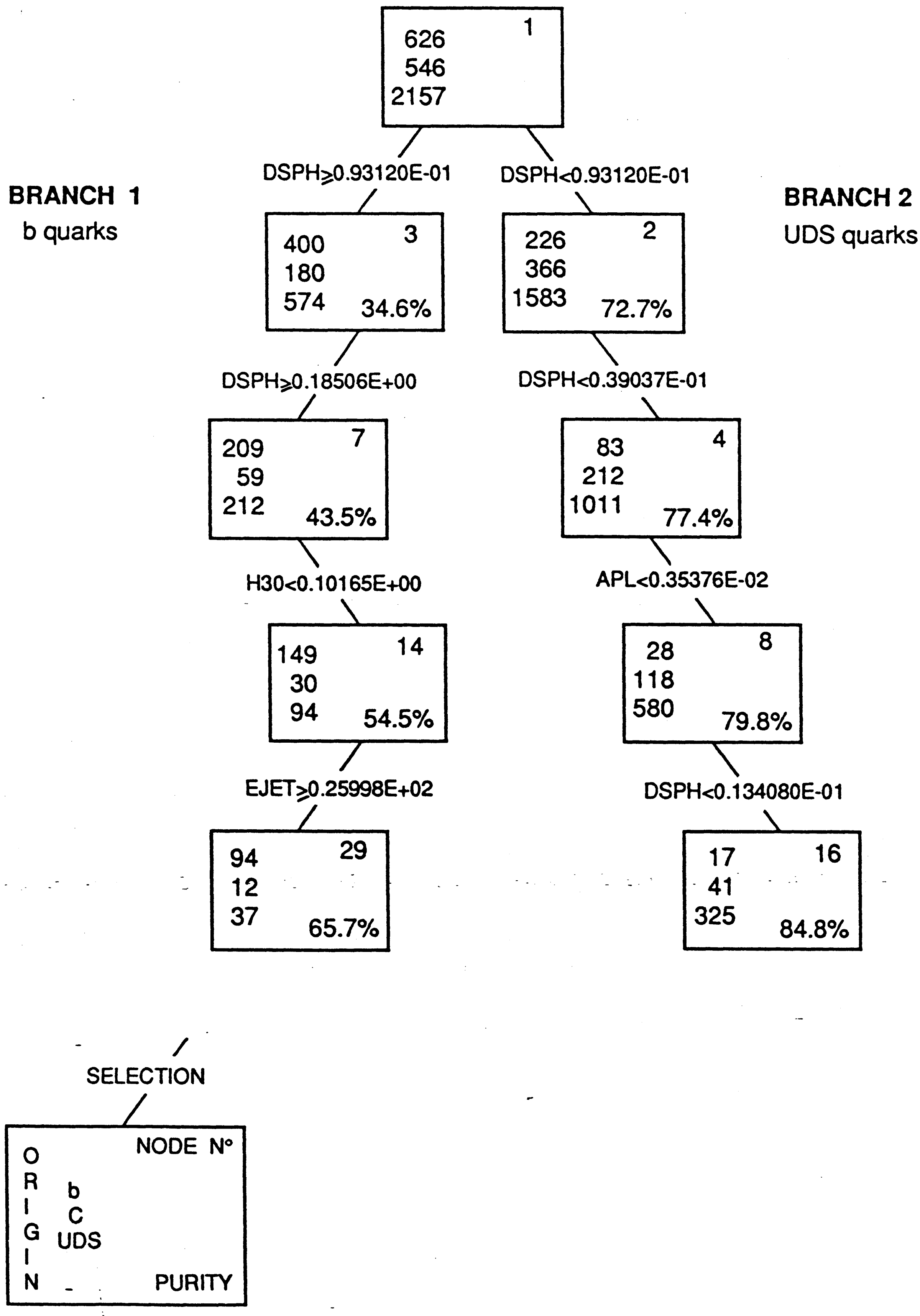
Two decays of the b quark have been considered : semi-leptonic decays involving an electron and purely hadronic decays.

The ratios of number of classified events $n(b \to e)/n(e)$ and $n(b \to hadrons)/n(events)$ are computed with Monte-Carlo events and ALEPH events.

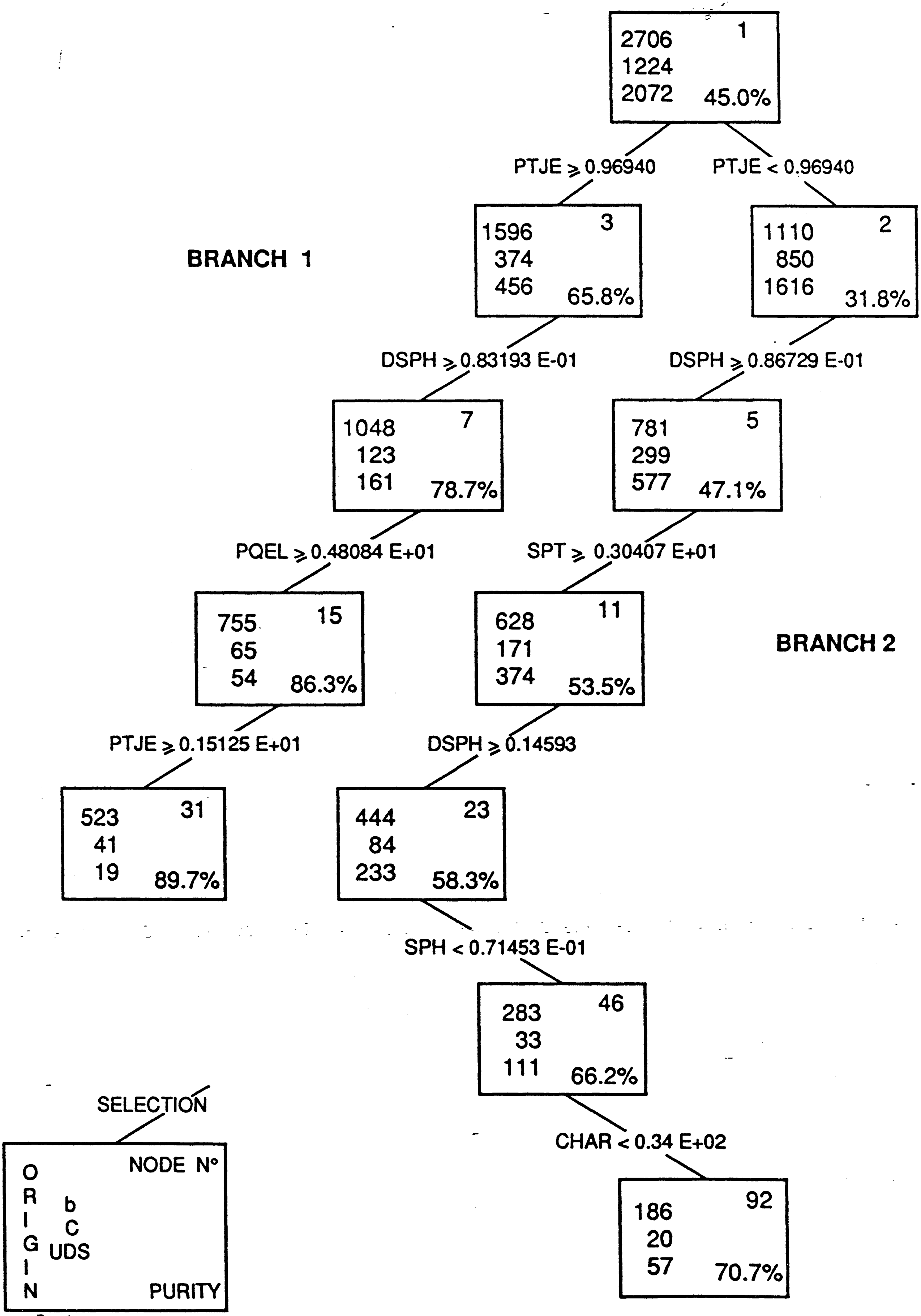
Presented at the International Workshop on Software Engineering, Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics.

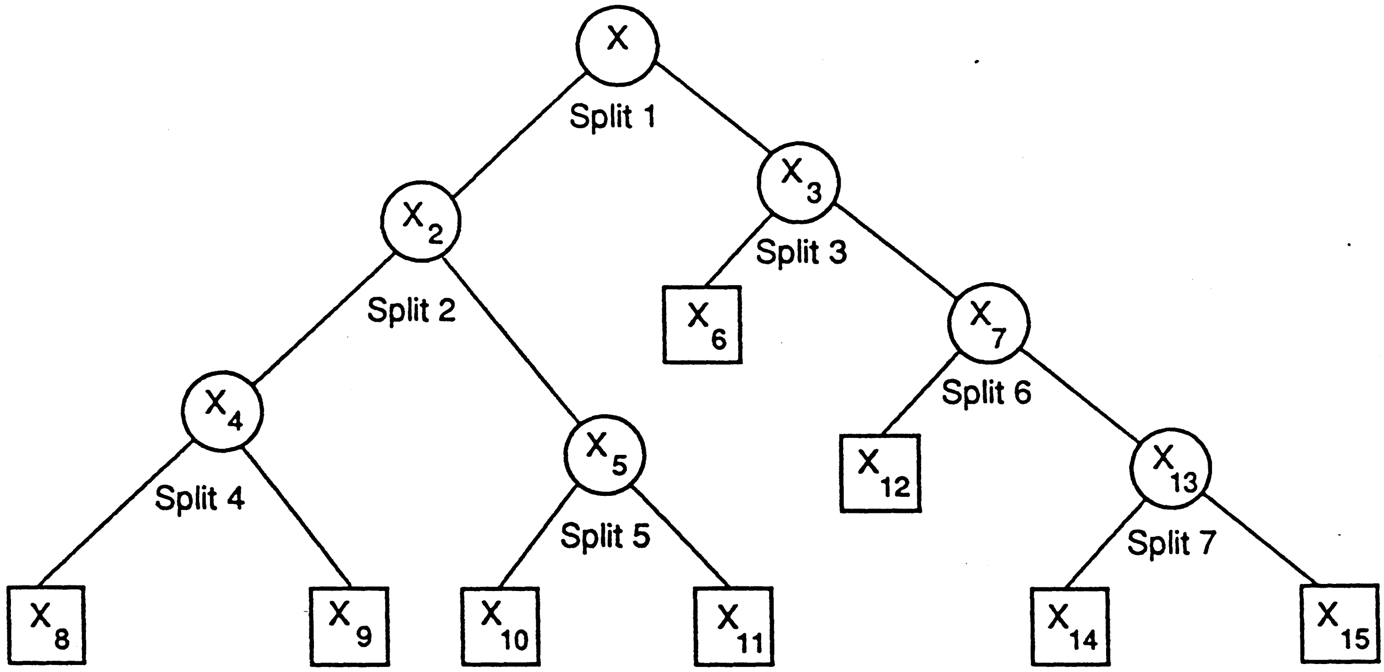
LYON March 1990

2 BRANCHES OF A DNP TREE (b --> hadrons)



2 BRANCHES OF A DNP TREE (b --> e)





X
Split 1
X_2
X_3
Split 2
Split 3
X_6
X_7
X_4
X_5
Split 6
X_{12}
X_{13}
Split 4
Split 5
Split 7
X_8
X_9
X_{10}
X_{11}
X_{14}
X_{15}

Fig. 5

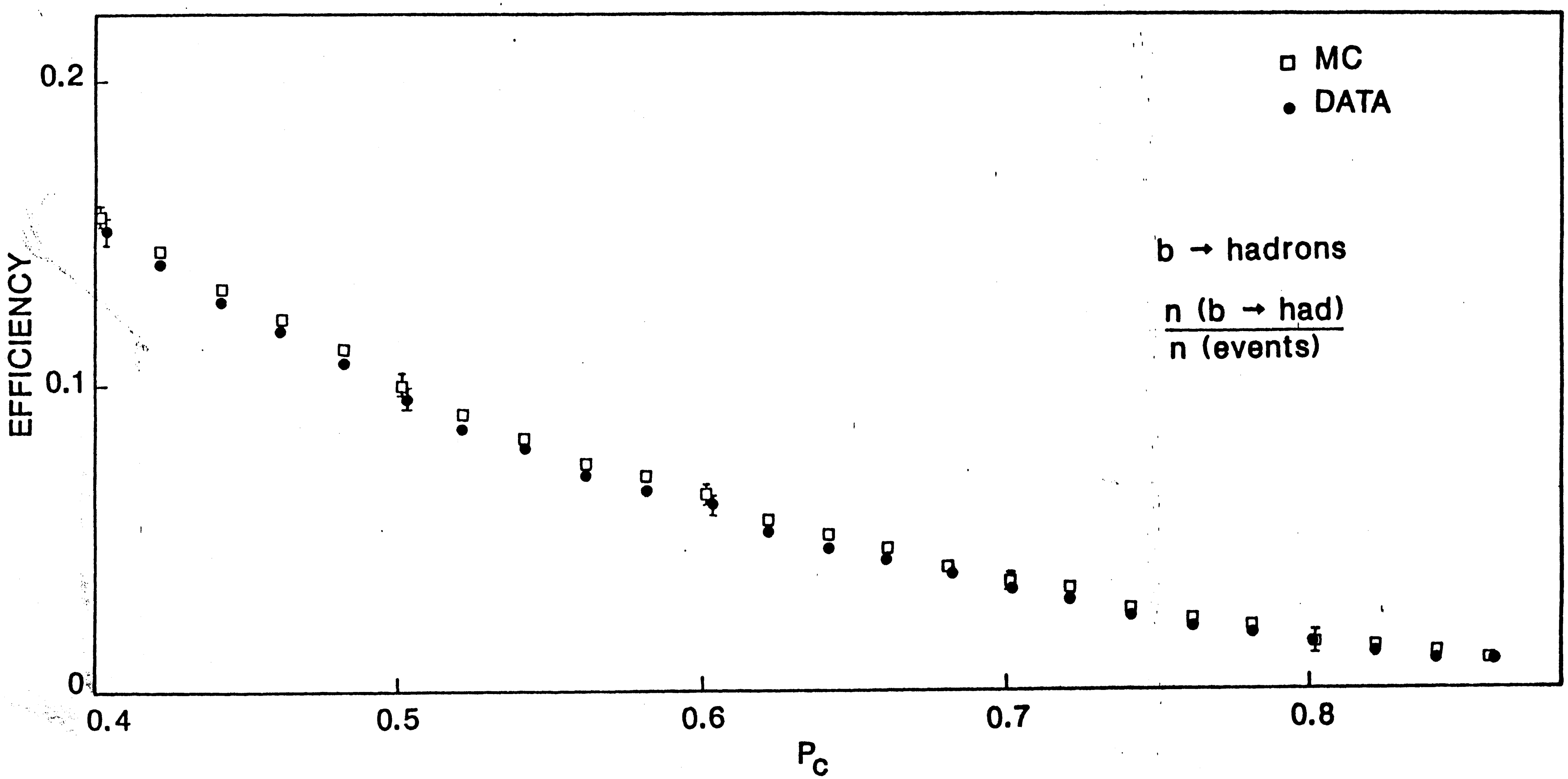

MC
DATA
b → hadrons
n (b → had) / n (events)
EFFICIENCY
0.2
0.1
0
0.4
0.5
0.6
0.7
0.8
Pc

Fig.3