Electron identification with the Nearest Neighbor Rule

M. Pallavicini, C. Patrignani, M. Pontil, A. Verri

October 7, 1996

Abstract

In E-760 electron identification was based on the Electron Weight (EW), described elsewhere [1].

In this note a different method, based on the *nearest neighbor rule*, is described together with its performances, tested on E-760 data at the χ_2, ψ', ${}^1P_1$, η'_c.

We show that this approach is better for background rejection while keeping comparable efficiencies on the signal.

This note is strucured as follows: section 1 is a brief introduction to this method; section 2 gives all results obtained using E-760 data; section 3 is a more theoretical and complete description of this technique, together with some references to the classical literature on this subject.

1 Introduction

Electron identification in our experiment can be very schematically stated as follows: each track, either "signal" or "background", is represented by a vector whose elements carry all relevant information we get from the detector; tipically this vector includes H1 and H2 pulse heights, RPC dE/dx, Cerenkov photoelectrons, CCAL cluster mass, cluster fraction 3/5, second moments and so on.

Electrons and backgrounds populate this sample space according to some unknown distributions, say $f_e(\mathbf{x})$ and $f_b(\mathbf{x})$. Particle identification means obviously to compute what is the probability that a measured track $\mathbf{y}$ comes from "signal" or "background".

In E-760 the technique used was the Electron Weight, that is the ratio of the two hypothesis' likelyhoods; despite its good performances, we should point out that this method is very crude for it completely ignores all correlations between different vector components or, stated differently, it assumes that the functions $f_e(\mathbf{x})$ and $f_b(\mathbf{x})$ are factorizable. This assumption is probably quite good for $f_e(\mathbf{x})$ but it's definitely wrong for background[1], so we may expect a different approach to give even better results.

[1] Our background is two-folded: we have hadrons and e^+e^- couples; this fact

The Nearest Neighbor Rule (NNR), in its simplest form, is the following: we start from two clean samples of "electrons" and "backgrounds" that we call E and B; given a suitable metric in sample space, for each new track $\mathbf{y}$ we compute the distances:

$$d_e = min\{||\mathbf{y} - \mathbf{x}|| : \quad \forall\, \mathbf{x} \in E\} \tag{1}$$

$$d_b = min\{||\mathbf{y} - \mathbf{x}|| : \quad \forall\, \mathbf{x} \in B\} \tag{2}$$

Track $\mathbf{y}$ is identified according to the smallest between d_e and d_b, i.e. it is identified as an electron if $d_e < d_b$.

This method can be slightly modified to reduce the effect of noise or inefficiencies in real life measurement[2]; instead of finding just the nearest point, k nearest neighbor are searched for and a fiducial cut is used to identify the track. For example, you may call a track an "electron" if four Nearest Neighbor out of five belong to set E, allowing one wrong point at most; the best trade off between efficiency and background rejection has to be found choosing the appropriate k and the best fiducial cut.

2 Results on E-760 data

We have tested the NNR on E-760 J/ψ inclusive data, using 1991 runs[3] only.

Each track is described as a point in an 8 dimensional space whose components are[4]:

1. $H1$ pulse height
2. $H2$ pulse height
3. RPC dE/dx
4. Cerenkov photoelectrons
5. CCAL cluster fraction 3/5[5]
6. Second moment CCAL cluster in phi

---

implies the existence of a strong correlations between dE/dx measurements and Cerenkov response because hadrons have low signals on all these detectors, while e^+e^- couples have larger ones. Shower shape in CCAL is different as well for hadrons and e^+e^- couples.

[2]For example, if one of the measurements is missing, the effect on the distance can be large enough to switch from one sample to the other one.

[3]Because of the lack of the RPC measurements, 1990 data require an independent sample both for background and for the electron signal.

[4]In order to make a meaningfull comparison with EW method, we decided to use the very same set of variables; there is no special reason to use this set instead of a different one.

[5]This is the ratio of the energy associated to the cluster using a 3×3 grid or a 5×5 one.

7. Second moment CCAL cluster in theta
8. CCAL cluster mass

As "good" electron sample we used about 800 χ_2 events selected requiring good exclusive kinematic fit (prob. > 0.2).

The background sample has been selected using ${}^1P_1$ data (stack number 1 mini DSTs) keeping all events with invariant mass above 2.4 GeV[6].

We ended up with 1830 background tracks and 1666 "good" electron tracks.

Then we applied the NNR to the whole ${}^1P_1$ data, to the whole η_c' sample and to all χ_2 data not already used as sample electrons. Figures 2, 2 show a comparison between the NNR performance and the EW at the ${}^1P_1$ energy and at the η_c' energy. Both pictures show that the NNR works better for low invariant mass events while has comparable rejection power in the "signal" region.

Figure (2) shows a comparison between the two methods at the χ_2 peak; the efficiency of the NNR is 90 ± 2 %, to be compared to EW efficiency that is 92 ± 2 %.

More interestingi, figure 2 shows that this two methods reject different background events, i.e. you can get a powerful selection criteria imposing both cuts (Electron weight and NNR).

In this case efficiency on the signal decreases to $86 \pm 2\%$, but the background is clearly reduced even in the high invariant mass region. This statement is also supported by figure 2 which shows the weak correlation between EW and NNR at the χ_2 peak.

3 The Nearest Neighbor Rule

In this section the Nearest-Neighbor Rule (NNR) is reviewed. Strengths and weaknesses of NNR are also discussed. A more detailed account of NNR can be found in [2].

3.1 How it works

The NNR is a non-parametric classification technique for the solution of supervised learning problems. In what follows NNR is illustrated as an alternative to parametric techniques. Let us first establish some basic notation and recall the essence of parametric techniques.

Let $(\mathbf{x}_i, y_i)$ with $i = 1, \dots, n$ be a set of pairs in which $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (the *training set*) is a subset of a certain sample set and the y_i are labels. The $\mathbf{x}_i$ can be regarded as N-tuples of real numbers for some fixed N and the y_i as integers in the set $[1, m]$ with $m \in N$. The supervisor assigns $\mathbf{x}_i$ the label y_i for $i = 1, \dots, n$, and the problem of supervised learning reduces to which label assign to a sample $\bar{\mathbf{x}}$ not in the training set S.

[6]The goal is to identify background events that simulate almost perfectly the signal, so all events that do not fulfill minimum requirements have not been included in the background sample beacuse can be easily rejected without using electron identifications.

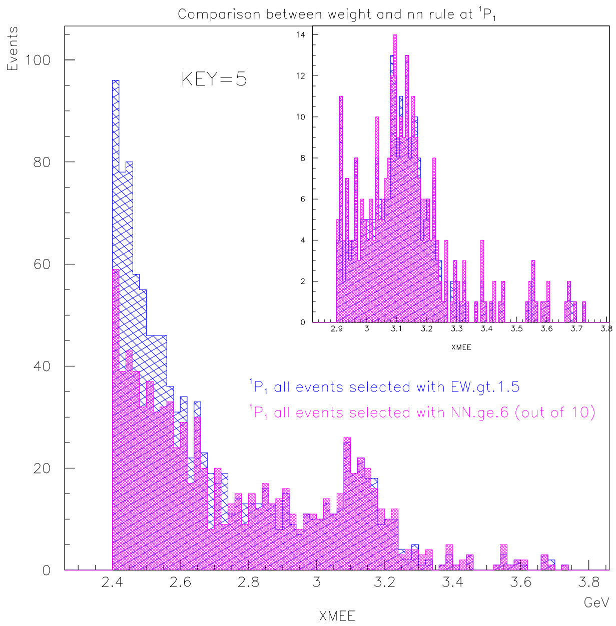

Figure 1: This picture shows the improved background rejection obtained with the NNR in the ${}^1P_1$ region; the histogram is the e^+e^- invariant mass using all 16 stacks at the ${}^1P_1$ formation energy. Crossed area are events selected with the electron weight only (cut 1.5); the filled one are events selected with the NNR with $(n_1 + n_2 >= 6)$. The background is much reduced in the low invariant mass region while keeping comparable efficiency around 3.1 GeV. The small box shows a magnified view of the "signal" region. KEY is the number of nearest neighbors computed for each track.

A parametric classification procedure for the solution of this problem usually consists of four steps. First, it is assumed that the probability distributions of the samples given each possible label, or $P(\mathbf{x}|y_i)$ with $i = 1, \ldots, m$, are known up to some parameter to be determined. Second, the pairs $(\mathbf{x}_i, y_i)$ with $i = 1, \ldots, n$ are used for estimating the unknown parameters (typically through maximum likelihood procedures). Third, Bayes theorem is applied to obtain the *a posteriori* probabilities $P(y_i|\mathbf{x})$ in terms of the estimated probability distributions $P(\mathbf{x}|y_i)$ and the *a priori* probabilities $P(y_i)$. Fourth and finally, the sample $\bar{\mathbf{x}}$ is assigned the label y_j that maximizes the *a posteriori* probabilities, or

$$P(y_j|\bar{\mathbf{x}}) = \max_i P(y_i|\bar{\mathbf{x}}).$$

It is straightforward to show that this last step, sometimes called Bayes decision rule, leads to the smallest possible error rate in the hypothesis that all the probability distributions are perfectly known.

The NNR, which bypasses probability estimation, tackles directly the decision problem in terms of the available information and can be stated as follows.

> Let $\mathbf{x}_j \in S$ be the sample nearest to $\bar{\mathbf{x}}$, and y_j the label of $\mathbf{x}_j$. The *nearest neighbor rule* for classifying $\bar{\mathbf{x}}$ is to assign $\bar{\mathbf{x}}$ the label y_j.

Implicit in the definition of NNR is a suitable definition of distance between points of the sample space. Since in our analysis the N=8 components are neither independent nor measured in comparable units, we have found convenient to subtract to each sample $\mathbf{x}_j$

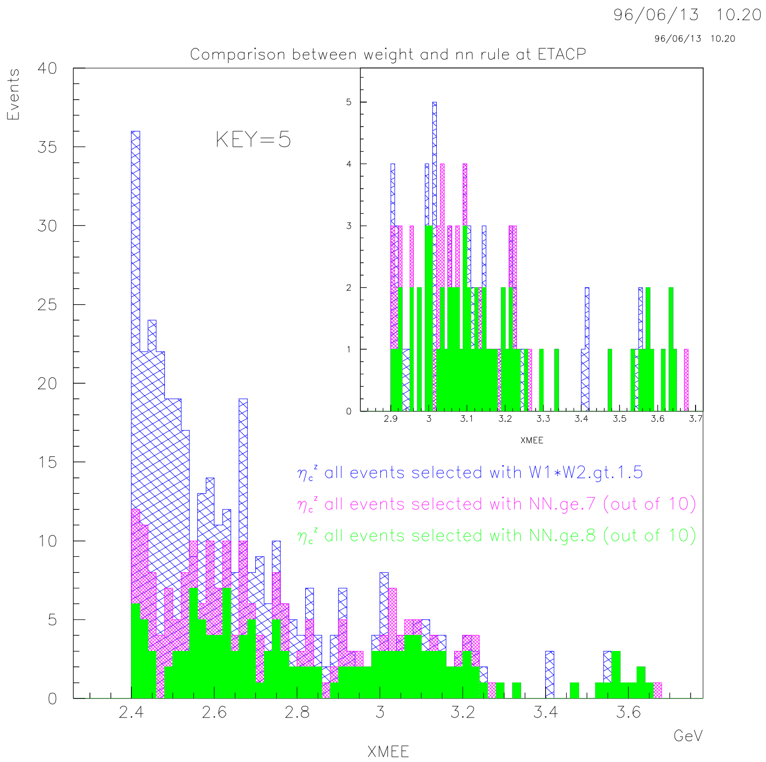

Figure 2: The same as in figure 2 in the η_c' region.

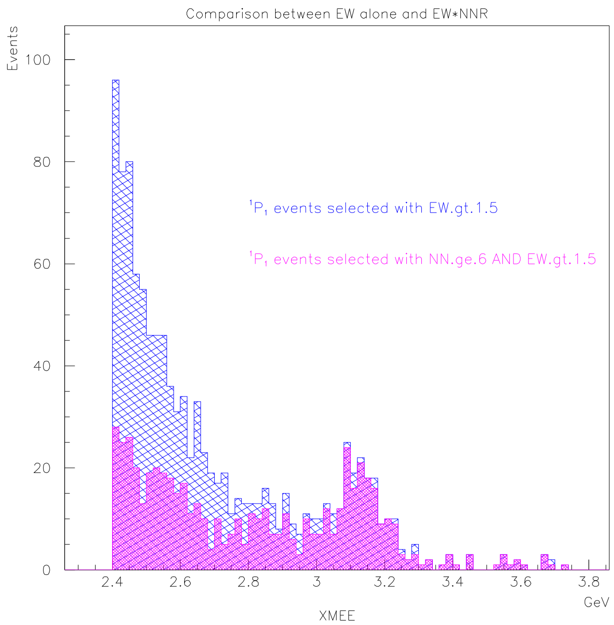

Figure 3: This picture shows how the background can be reduced using both the EW and the NNR in the $^1P_1$ region. The efficiency of the two combined cuts is around 86% while NNR alone is 90%

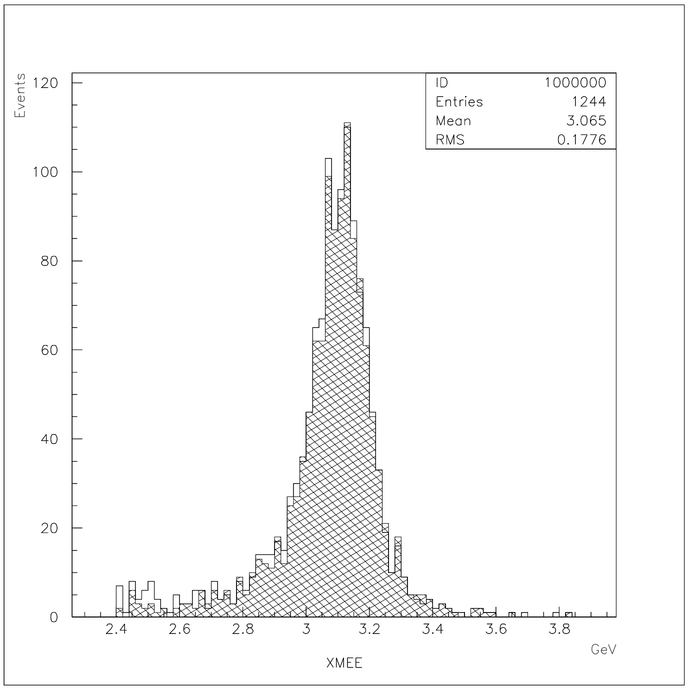

Figure 4: Events selected at the χ_2 formation energy with the EW (white histogram) and with NNR (filled); the background in the low energy part of the histogram is reduced while keeping comparable efficiency in the signal region. EW efficiency is 92%, NNR efficiency is 90%

.

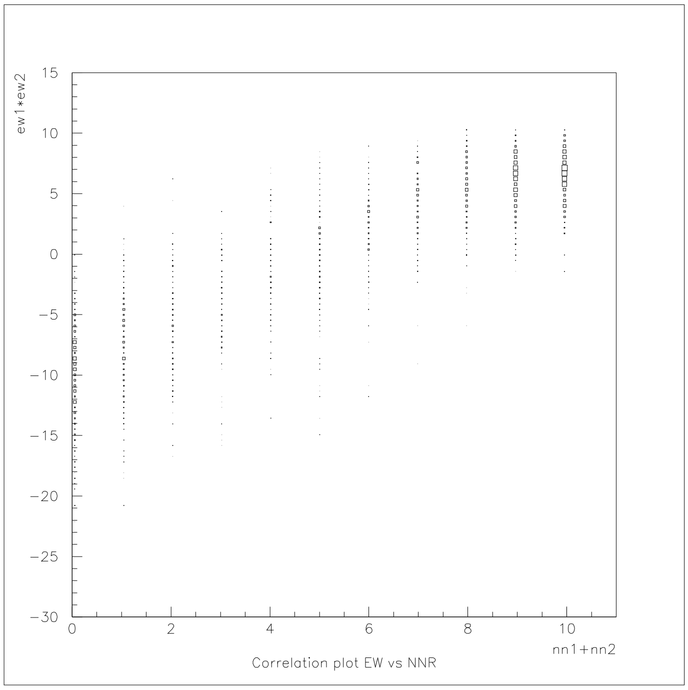

Figure 5: Correlation plot between EW and NNR on χ_2 events.

in the training set S the centroid $< \mathbf{x} >$ and compute the covariance matrix Σ of all the samples $\mathbf{x}'_j$ with:

$$\mathbf{x}'_j = \mathbf{x}_j - < \mathbf{x} >$$

and $\mathbf{x}_j \in S$. The distance d between the sample point $\mathbf{x}$ and the point $\mathbf{x}_j \in S$ is computed as:

$$d^2 = (\mathbf{x}'_j - \mathbf{x}')\Sigma^{-1}(\mathbf{x}'_j - \mathbf{x}') \tag{3}$$

with $\mathbf{x}' = \mathbf{x} - < \mathbf{x} >$. The distance between $\mathbf{x}$ and $\mathbf{x}_j$ given by equation 3 (known in the pattern recognition literature as Mahalanolis distance [2]) can be thought of as the Euclidean distance that can be obtained after a linear trasformation of the original sample space in which the new axis are mutually orthogonal (and given by the eigenvectors of Σ) and the scaling factors along each axis proportional to the square root of the inverse of the eigenvalues of Σ. In practice, the metric induced by the inverse of the covariance matrix compesates for the linear correlation between the original components.

In the attempt to effectively deal with close decisions, it is customary to assign $\bar{\mathbf{x}}$ the label associated with the k samples ($k > 1$) nearest to $\bar{\mathbf{x}}$. A number of different NNRs, named k-nearest-neighbor techniques (k-NN), can then be obtained depending on the rule with which a particular label is selected from the labels of the k nearest neighbors. A commonly adopted rule is to select the label mostly represented within the k nearest neighbors (with k odd to avoid ties).

3.2 Why it works

As with most nonparametric techniques, error rates of NNR can be easily computed in the hypothesis of unlimited number of samples. Let P_B be the error rate obtained by applying the Bayes decision rule (that is, the smallest possible error rate) and P_{NN} the asymptotic nearest-neighbor error rate. Clearly,

$$P_B \leq P_{NN}.$$

Under the assumption that with $n \to \infty$ the distance between $\mathbf{x}$ and its nearest neighbor $\mathbf{x}_j$ becomes arbitrarily small, it can be shown [2] that, for any possible metric,

$$P_{NN} \leq P_B(2 - \frac{m}{m-1}P_B). \tag{4}$$

where m, as before, is the number of different labels. Hence, the asymptotic nearest-neighbor error rate is always less than twice the smallest possible error rate. The upper bound of (4) is tight since it is possible to produce examples in which $P_{NN} \to P_B(2 - P_B m/(m-1))$ as $n \to \infty$ [2]. As a consequence, NNR is only a suboptimal procedure.

Intuitively, NNR can be cast into the familiar probabilistic framework by observing that the probability that the sample $\bar{\mathbf{x}}$ has label y is the a posteriori probability $P(y|\bar{\mathbf{x}})$. In the event of a large number of labeled samples, it appears reasonable to assume that since $\mathbf{x}_j$, the sample nearest to $\bar{\mathbf{x}}$, lies close to $\bar{\mathbf{x}}$, $P(y|\mathbf{x}_j)$ is close to $P(y|\bar{\mathbf{x}})$. Therefore, NNR can be thought of as a randomized decision rule which assigns $\bar{\mathbf{x}}$ the label y with probability $P(y|\mathbf{x}_j)$ (since $P(y|\bar{\mathbf{x}}) \simeq P(y|\mathbf{x}_j)$). From this, it follows easily that if the *a posteriori* probability are peaked around a specific label, then NNR almost always agrees with the Bayes (optimal) decision rule. Instead, if all the labels are equally likely, NNR often disagrees with the Bayes selection but the probability of error is approximately $1 - 1/m$ for both.

3.3 When to use it

Let us finally discuss the main advantages and drawbacks that can be expected when applying NNR to a generic classification problem.

The NNR has at least two appealing properties. First, NNR can be applied to virtually any problem of supervised classification independently of the complexity of the shape of the underlying probability distributions. This point might be crucial in the presence of probability distributions that can only be approximately described by means of parametric distributions, not to mention the case in which the probability distributions are either completely unknown or hard to describe in parametric terms.

Second, NNR can be easily implemented and applied with no appreciable change to very different problems of supervised classification. This should be compared with the substantial effort required for the implementation of parametric techniques in the analysis of different experiments.

On the other hand, NNR has two main shortcomings. First, the classification time grows linearly with the size of the training set. Of the many methods proposed in the literature to deal with this problem, let us mention a variation of NNR, named *condensed nearest-neighbor technique* (CNN), in which the training set is effectively reduced to a smaller set before the classification stage [3]. In essence, CNN retains in the condensed training set only the samples that are necessary for correctly classifying the entire training set. Not surprisingly, most of the samples of the condensed training set happen to lie not too far away from decision boundaries.

A second weakness of NNR is that there is no known upper bound on the error rate in the case of finite number of samples. Therefore, the decision of whether or not the training set is representative of the entire sample set is mostly based on empirical considerations. This lack of theoretical insight on the behavior of NNR in the finite case is probably the reason for which this technique, despite its undisputable simplicity, is not as popular as parametric techniques within the particle physics community.

4 Conclusions

We have used the *nearest neighbor rule* for electron identification on E-760 data; the method shows similar performance compared to the Electron Weight for events with invariant mass above 2.9 GeV but gives better result at lower energy. Furthermore, NNR seems to be only weakly correlated to the EW, therefore, by using both cuts, background rejection improves significantly while keeping comparable efficiency on the signal.

It is very simple, does not require any assumption on unknown probability distributions and is easly applicable to other classification problems as, for instance, the p^0 - γ discrimination. The very same code can be used on any classification problem, provided that an independent way to select a clean training sample is available or that a reliable montecarlo program exists. The code is written in standard C and is already available in *offline fashion*; just two calls at routine

nnfilter(iopt,vect,dist,nn)

are needed, where *iopt* is the standard E-760 / E-835 call parameter, *vect* is a real vector of dimension equal to the number of variables used to classify, *dist* is the distance of the nearest neighbor sample vector, *nn* is the number of nearest neighbor electrons. With *iopt*=-1 the algorythm is initialized while for *iopt*=0 you get the actual filter.

References

[1] Electron identification criteria : $EW_1 * EW_2$
E760 note - C. Patrignani

[2] Pattern recognition and scene analysis, Duda R.O. and Hart P.E., Wiley, 1973

[3] The condensed NN Rule, Hart P.E., IEEE trans. Information Theory, May 1968 pp. 515-516