The Nearest-Neighbor Technique for Particle Identification

Marco Pallavicini Claudia Patrignani
INFN and Università di Genova

Massimiliano Pontil Alessandro Verri
INFM - Dipartimento di Fisica dell'Università di Genova,
Via Dodecaneso 33, 16146 Genova (I).

January 29, 1997

Abstract

In this paper the *nearest-neighbor technique*, a classical nonparametric method of pattern recognition relatively unknown within the Particle Physics community, is compared in a problem of particle identification against a standard likelihood method. The analysis performed on real data indicates that the proposed technique has the same efficiency of the likelihood method on the *signal*, but a better rejection of the *background*. It is concluded that the nearest-neighbor technique, which is conceptually clear and easy to implement and use, is very useful for particle identification, possibly in conjunction with other classification methods.

1 Introduction

Typically, particle identification is tackled with statistical tools, like Fischer discriminant analysis and likelihood techniques, or neural networks [1]. In this paper we address the particle identification problem through the *Nearest-Neighbor Technique* (NNT) [2], a classical method of pattern recognition. Given a training set of events correctly identified and a distance measure between events, the NNT assigns a new event the *same* class of the *nearest* event in the training set. The NNT presents a number of attractive features that range from ease of implementation (since it does not depend at all on free parameters) to conceptual clarity. The NNT can be applied to correlated data with virtually no preprocessing and does not require assumptions on the shape of the various probability distributions. Last, but not least, the same code can be used to solve different classification problems.

Despite these good properties, NNT has received little attention from the Particle Physics community. This may be due to *(a)* the theoretical difficulty of estimating error bounds of NNT given a finite training set, and *(b)* the NNT algorithmic complexity which grows linearly with the size of the training set. For different reasons, however, it appears that both limitations *(a)* and *(b)* can be overcome. First, the NNT has tight asymptotic error bounds (the NNR asymptotic error bound can never be worse than twice the minimum Bayes error bound [2]). Therefore, in the presence of large amounts of data and small Bayes error, the NNT is expected to produce very good classification rates. Second, the computational power now available is sufficient to effectively deal with methods of linear complexity in the size of the training set.

In this paper we compare the performances of NNT against a standard likelihood technique (more precisely the Neimann-Pearson test) using real data. The obtained results indicate that NNT always produces better results and, apparently, at a much lower cost. The gap between the two methods is particularly evident in the rejection rates. Interestingly, we found that a hybrid method which combines NNT and likelihood techniques produce even better rejection rates with no significant loss in efficiency.

The paper is organized as follows. In Section 2 we review NNTs from the theoretical viewpoint and describe its implementation. Section 3 shows the results obtained in the comparison between NNT and likelihood techniques on a study case. Finally, Section 4 draws the conclusions.

2 The Nearest-Neighbor Technique

In this section the NNT is reviewed. A more detailed account of NNT can be found in [2]. The NNT is a *non-parametric* classification technique for the solution of supervised learning problems, *i.e.*, problems in which a sample of already classified events is available for each class. We start by establishing basic notation and recalling the essence of likelihood techniques, the most used parametric techniques. This will help to explain the difference between NNT and standard parametric techniques.

Let $(\mathbf{x}_i, y_i)$ with $i = 1, \ldots, n$ be a set of pairs in which $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ (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 $m \in N$. In what follows, for simplicity, we restrict the analysis to the case $m = 2$. Thus the supervisor assigns $\mathbf{x}_i$ the label $y_i = 1$ or $y_i = 2$ for $i = 1, \ldots, n$, and the problem of supervised learning reduces to which label assign to a sample $\bar{\mathbf{x}}$ not in the training set S.

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 $i = 1, 2$, are known up to some parameter to be determined. Second, the pairs $(\mathbf{x}_i, y_i)$ with $i = 1, \ldots, n$ and $\mathbf{x}_i \in S$ 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.

Contrary to parametric techniques, NNT bypasses probability estimation and tackles directly the decision problem in terms of the available information through the *nearest-neighbor rule*:

Let $\mathbf{x}_j \in S$ be the sample nearest to $\bar{\mathbf{x}}$, and y_j the label of $\mathbf{x}_j$. Assign $\bar{\mathbf{x}}$ the label y_j.

Implicit in the definition of NNT is a suitable definition of *distance* between points of the sample space. In particle identification, the components of $\mathbf{x}$ collect different physical information coming from a wide range of detectors (like energy loss measurement, Cerenkov photoelectrons, momentum measurement, β measurement from Rich detectors, and cluster shape in calorimeters, for example). Therefore, since these components are neither independent nor measured in comparable units, it is usually convenient to subtract to each sample $\mathbf{x}_j$ the centroid $< \mathbf{x} >$ of the training set S and compute the covariance matrix Σ of all the samples $\mathbf{x}'_j$ with:

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

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

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

with $\mathbf{x}' = \mathbf{x} - < \mathbf{x} >$. The distance between $\mathbf{x}$ and $\mathbf{x}_j$ given by equation 1 (known in the Pattern Recognition literature as Mahalanobis distance [2]) can be thought of as the Euclidean distance that can be obtained after a rotation of the original sample space in which the new axes are the eigenvectors of Σ and a scaling along each axis proportional to the square root of the inverse of the corresponding eigenvalues. In practice, the metric induced by the inverse of the covariance matrix compensates for the linear correlation between the original components and normalize the new variable variances.

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 NNTs, named k-nearest-neighbor techniques (k-NNT), 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).

As with most nonparametric techniques, error rates of NNT 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,

$$P_{NN} \leq P_B(2 - 2P_B). \tag{2}$$

Hence, the asymptotic nearest-neighbor error rate is always less than twice the smallest possible error rate. The upper bound of (2) is tight since it is possible to produce examples in which $P_{NN} \to P_B(2 - 2P_B)$ as $n \to \infty$ [2]. As a consequence, NNT is a suboptimal procedure but its error is never worse than twice the Bayes error. As we show in Section 3, in our test the NNT, with sufficiently high statistics, always gives better result than parametric techniques.

Intuitively, NNT 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, since $\mathbf{x}_j$, the sample nearest to $\bar{\mathbf{x}}$, lies close to $\bar{\mathbf{x}}$, $P(y|\mathbf{x}_j)$ is likely to be close to $P(y|\bar{\mathbf{x}})$. Therefore, NNT 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 that if the *a posteriori* probability are peaked around a specific label, then NNT almost always agrees with the Bayes (optimal) decision rule. Instead, if all the labels are equally likely, NNT often disagrees with the Bayes selection but the probability of error is approximately 1/2 for both.

We now turn to the comparison between NNT and a likelihood technique.

3 Results on an electron identification problem

To test the NNT we used data taken from experiment E-760 at Fermilab. This experiment was searching for events with inclusive J/ψ decaying to e^+e^- produced in $p-\bar{p}$ interactions. Since the cross section for these processes range from 10 pb to a few nb (to be compared to $p-\bar{p}$ total cross section of 70 mb), electron identification is crucial in order to achieve a good signal-to-noise ratio.

Each track is described as a point in an 8 dimensional space whose components are 3 dE/dx measurements, 1 threshold Cerenkov signal, and 4 variables describing the cluster shape in lead glass calorimeter. As "good" electrons for the training set we used about 800 $p-\bar{p} \longrightarrow \chi_2 \longrightarrow J/\psi\gamma \longrightarrow e^+e^-\gamma$ events selected requiring good exclusive kinematic fit (prob. > 0.2)[1]. The background sample for the training set was selected using 2 pb^{-1} data taken at energy far away from charmonium resonances ($E_{cm} = 3510 MeV$), keeping all events with invariant mass above 2.4 GeV[2]. We ended up with a training set of 1830 background tracks and 1666 "good" electron tracks.

In our implementation of NNT we measure the distances between tracks through equation (1) and with $k = 5$. To identify "signal" events in the background, we first compute the distance between each pair of tracks and the tracks in the training set. Then we count the number of electron nearest neighbors for either track (n_1 and n_2) and cut on the sum $n_1 + n_2$, namely $n_1 + n_2 > 5$.

We applied the NNT to the whole $^1P_1$ charmonium data (center of mass energy around 3526 Mev/c^2), to η_c' search data (around 3600 MeV/c^2) and to all χ_2 data ($E_{cm} = 3556 MeV/c^2$) not included in the training set. The $^1P_1$ data contain a small (about 80 events in 15 pb^{-1}) J/ψ inclusive sample coming from $J/\psi\ \pi^0$ events in a large background, so it is a significative benchmark for any particle identification method. On the contrary, the χ_2 sample provides a clear $J/\psi\ \gamma$ signal, useful to compute the method real efficiency.

[1] These selection criteria provide a very clean electron sample without introducing any bias in the sample variables.

[2] The goal is to identify background events that simulate almost perfectly the signal. Therefore all events not fulfilling minimum requirements were not included in the background sample, as they can be easily rejected without using electron identifications.

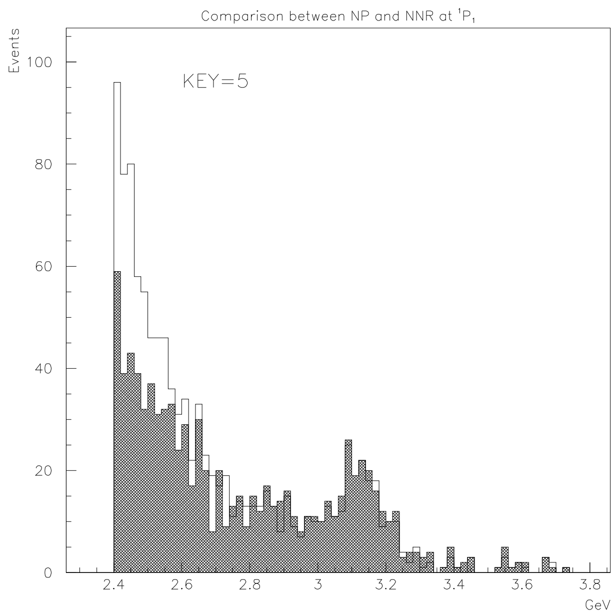

Figure 1: Improved background rejection obtained with the k-NNT in the $^1P_1$ region. The histogram is the e^+e^- invariant mass using about $15\,pb^{-1}$ taken at the $^1P_1$ formation energy (E_{cm} around 3526 MeV). White areas are events selected with the NPT, filled areas with the NNT with $n_1 + n_2 \geq 6$. The background is considerably reduced in the low invariant mass region while keeping comparable efficiency around 3.1 GeV, where a clear J/ψ inclusive "signal" is shown. KEY denotes the number of nearest neighbors computed for each track.

Figures 1 and 2 show a comparison between the NNT performance and the Neimann-Pearson test (NPT) at the $^1P_1$ and η_c' energy respectively. Both pictures show that the NNT works better, especially for low invariant mass events. Figure 3 illustrates a comparison between the two methods at the χ_2 peak. The efficiency of the NNT is 90 ± 2 %, to be compared to an EW efficiency of 92 ± 2 %.

Interestingly, Figure 4 shows that the two methods reject different background events. Consequently, a powerful selection criterion can be obtained by imposing both the Neimann-Pearson test and NNT cuts. By doing so, 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 5 which shows the weak correlation between EW and NNT at the χ_2 peak.

4 Conclusion

In this paper we have addressed the particle identification problem by means of Nearest-Neighbor Techniques (NNTs). Despite their undisputable simplicity and conceptual clarity, NNTs do not seem to be well-known within the Particle Physics community. We compared NNTs against a likelihood technique in a specific particle identification problem with real data and found that NNT yields the same efficiency on the signal but always better rejection

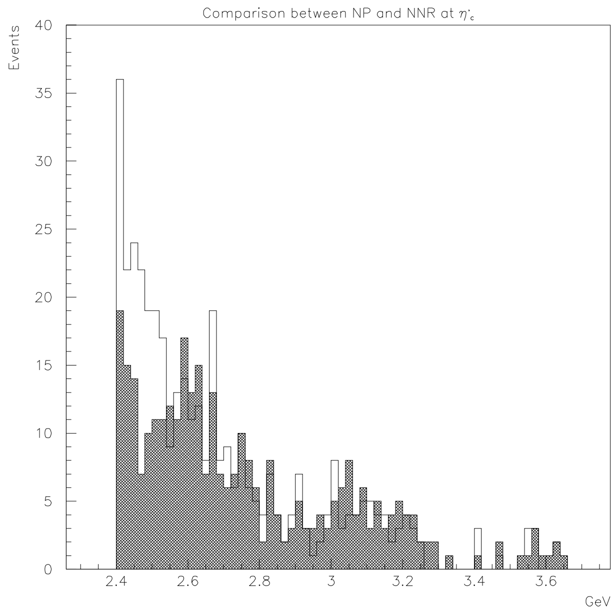

Figure 2: Improved background rejection obtained with the k-NNT in the η'_c region. Legend as in Figure 1.

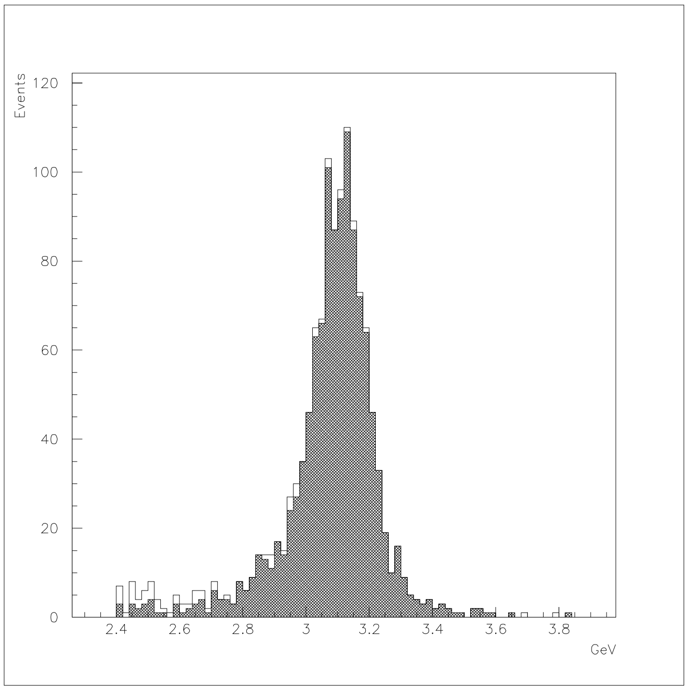

Figure 3: e^+e^- invariant mass for events selected at the χ_2 formation energy with the NPT (white) and with the NNT (filled histogram). The background in the low energy part is reduced while keeping comparable efficiency in the signal region. The efficiency is $92 \pm 2\%$ for the NPT and $90 \pm 2\%$ for the NNT.

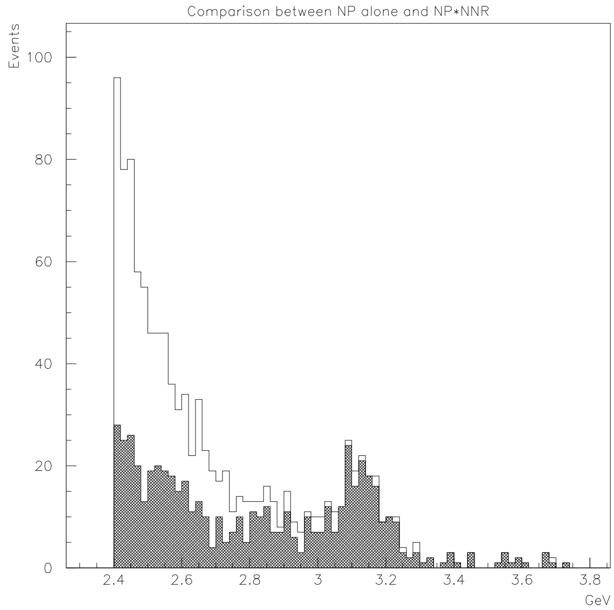

Figure 4: Background reduction using both the NPT and NNT in the $^1P_1$ region. The efficiency of both cuts combined is around 86% (while for NNT alone it is 90%).

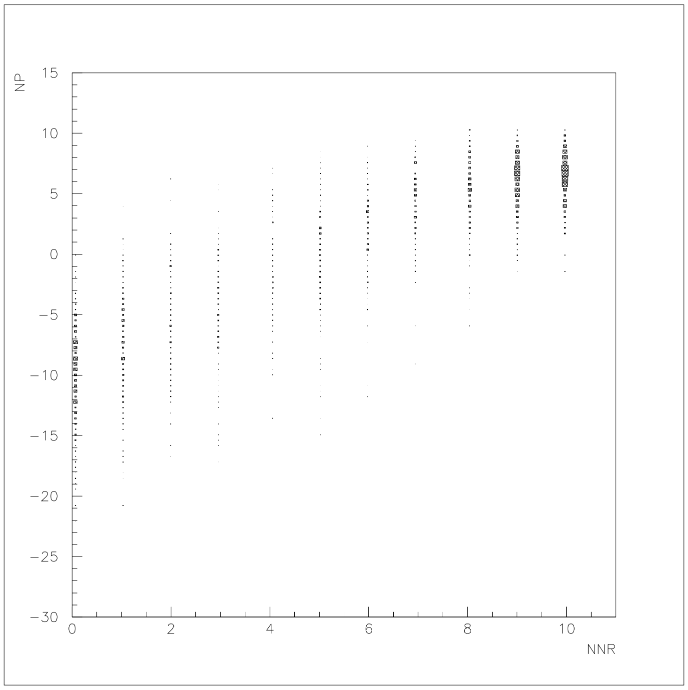

Figure 5: Correlation between NPT and NNT on χ_2 events.

of the background.

The NNT has a number of appealing properties. First, NNT can be applied to virtually any problem of supervised classification independently of the complexity of the shape of the underlying probability distributions. This point is crucial because typically the probability distributions are either unknown or, at best, approximately described by means of parametric distributions. Second, NNT can be easily implemented and applied with no appreciable change to very different supervised classification problems. While recognizing the sub-optimality of NNT, fully explained in Section 2, the simplicity and clarity of this approach should be compared with the substantial effort required for the implementation of parametric techniques or artificial Neural Networks in the analysis of different experiments. Finally, the very same code can be used on different classification problems, provided that an independent way to select a clean training sample is available or that a reliable montecarlo program exists.

It is thus concluded that NNTs can be usefully adopted for solving particle identification problems, especially when combined with other classification methods.

5 Acknowledgements

Two of us (Pallavicini and Patrignani) wish to thank the rest of the E-760 collaboration for having allowed us to use 1991 data as a real-life test of this technique.

References

[1] Neural Networks An Introduction
Springer-Verlag, Berlin 1991
Pattern Recognition in High Energy Physics with ANN
Lönnblad et. al.
Comp. Phys. Commun., 70, (1992), 167

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

[3] Electron identification criteria : $EW_1 * EW_2$
E835 internal note - C. Patrignani

[4] Observation of the $^1P_1$ state of charmonium
T. Armstrong et. al. Phys. Rev. Lett. 69,2337 (1992)
and references therein.