DEPARTAMENTO DE MATEMÁTICA
DOCUMENTO DE TRABAJO

“Pattern recognition via projection – based k – NN rules”

Ricardo Fraiman, Ana Justel y Marcela Svarc

D.T.: Nº 53 Junio 2008
Pattern recognition via projection–based $k$–NN rules

Ricardo Fraiman*, Ana Justel ** and Marcela Svarc***

*Departamento de Matemática y Ciencias, Universidad de San Andrés, Argentina and CMAT, Universidad de la República, Uruguay

**Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain

***Departamento de Matemática y Ciencias, Universidad de San Andrés, Argentina

Abstract

We introduce a new procedure for pattern recognition, based on the concepts of random projections and nearest neighbors. It can be thought as an improvement of the classical nearest neighbors classification rules. Besides the concept of neighbors we introduce the notion of district, a larger set which will be projected. Then we apply one dimensional $k$-NN methods to the projected data on randomly selected directions. In this way we are able to provide a method with some robustness properties and more accurate to handle high dimensional data. The procedure is also universally consistent. We challenge the method with the Isolet data where we obtain a very high classification score.

Key words and phrases: High Dimensional Data; $k$NN–rules; Pattern Recognition; Robustness.


1 Introduction

The goal in pattern recognition problems (classification rules) is to classify individuals into groups. Information about these groups is provided by a training sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$, where each data $X_i$ has a label $Y_i$ attached that informs to which group it belongs to. One can consider $k$–nearest neighbor rules ($k$–NN), introduced by Fix and Hodges (1952), as the simplest non–parametric and widely spread rules. In them, each of the $k$–nearest neighbors in the training sample of a new observation $x$ votes in favor of the group to which it belongs to. The new observation is classified into the group that obtains more votes.

In spite of the simplicity of $k$–NN rules, performance of these methods on many real data examples is quite remarkable, being a universally consistent family of rules. Recently, with the aim of classifying tumors using data gene expressions, Dudoit et
al. (2002) compared performances of different discrimination methods. Their main conclusion is that in those cases, simple classifiers such as \( k \)-NN performed noteworthily better than more sophisticated methods.

In this paper, we propose a new classification family of rule based on the concepts of random projections and nearest neighbors: random projection-based \( k \)-NN rules (RP\( k \)NN). They can be thought as an improvement of the classical \( k \)-NN methods in order to attain robustness properties, and to handle high dimensional data.

The paper is organized as follows. In Section 2 we give a brief review on \( k \)-NN methods and introduce the new proposal. Section 3 contains a small simulation study where the robustness property is exhibited. Section 4 is devoted to the study of a real data set, the Isolet-data, which can be downloaded from the UCI Machine Learning Repository (Asuncion and Newman, 2007). Finally, in Section 5 we show universal consistency for the new classification rule.

2 Classification Rules

As mentioned in Devroye, Györfi and Lugosi (1996), “pattern recognition or discrimination is about guessing or predicting the unknown nature of an observation, a discrete quantity such as black or white, one or zero, sick or healthy, real or fake. An observation is a collection of numerical measurements such as an image (which is a sequence of bits, one per pixel), a vector of weather data, an electrocardiogram, or a signature on a check suitably digitized.”

Given a finite set \( \{1, \ldots, m\} \) and a vector space \( E \), an observation is a pair \((x, y) \in E \times \{1, \ldots, m\}\), where \( x \) is known and \( y \) is a class or label that denotes the unknown nature of the observation. A mapping \( g: E \rightarrow \{1, \ldots, m\} \) is called a classifier and represents our guess of the class \( y \) given its associated vector \( x \in E \). The classification is wrong if given an observation \((x, y)\), \( g(x) \neq y \).

Let \((X, Y) \in E \times \{1, \ldots, m\}\) be a random pair. Since an error occurs if \( g(X) \neq Y \), probability of misclassification for \( g \) is

\[
L(g) = P[g(X) \neq Y]. \tag{2.1}
\]

Then the best possible classifier is the function \( g^* \) that minimizes (2.1). The minimum error probability (the Bayes error) is denoted by \( L^* = L(g^*) \).

In order to obtain \( g^* \), distribution of \((X, Y)\) should be known, but this is not typically the case. One must build up a classifier based on a training sample of independent pairs \( \{(X_i, Y_i); 1 \leq i \leq n\} \), with the same distribution as the pair \((X, Y)\) and known \( Y_1, \ldots, Y_n \) values. Then a classifier is a function

\[
g_n(\cdot; X_1, Y_1, \ldots, X_n, Y_n): E \times (E \times \{1, \ldots, m\})^n \rightarrow \{1, \ldots, m\},
\]
with probability of missclassification given by the conditional error probability
\[ L_n(g_n) = P[g_n(X_1, X_1, Y_1, \ldots, X_n, Y_n) \neq Y | X_1, Y_1, \ldots, X_n, Y_n]. \]

A sequence of classifiers \( \{g_n; n \geq 1\} \) is called a rule. A rule is consistent when
\[ \lim_n L_n(g_n) = L^*. \]
Asymptotic results about consistency of discrimination rules can be found in Devroye, Györfi and Lugosi (1996).

2.1 \( k \)-NN rules

The \( k \)-nearest neighbors rule is a well studied method, which only depends on the distances between individuals. Formally, we define the \( k \)-NN rule by
\[ g_n(x) = \arg \max_{j=1,\ldots,m} \sum_{i=1}^{n} w_{ni} I\{Y_i=j\}, \]
where
\[ w_{ni} = \begin{cases} 1 & \text{if } X_i \text{ is among the } k \text{ nearest neighbors of } x \\ 0 & \text{otherwise.} \end{cases} \]

\( X_i \) is said to be the \( k \)-th nearest neighbor of \( x \) when distance \( ||X_i - x|| \) is the \( k \)-th smallest among \( ||X_1 - x||, \ldots, ||X_n - x|| \). Distance ties are broken by comparing indices, that is, if \( ||X_i - x|| = ||X_j - x|| \), \( X_i \) is considered closer to \( x \) when \( i < j \).

A proof of universal consistency for the case of non random nearest neighbor \( k \), if \( k \to \infty \) and \( k/n \to 0 \) as \( n \to \infty \), can be found in Devroye, Györfi and Lugosi (1996) (Theorem 6.4, p. 101). They also considered automatic nearest neighbor rules, where the parameter \( K \) is random and depends on the data sequence \((X_1, Y_1), \ldots, (X_n, Y_n)\). In this case, the same result is true if the sequence \( K_n \) satisfies that \( K_n \to \infty \) and \( K_n/n \to 0 \) with probability one, as \( n \to \infty \) (Theorem 26.1, p. 451).

2.2 Random projections and \( k \)-NN rules

We now introduce a new classification rule, which can be thought as an improvement of the classical \( k \)-NN rule, in order to attain robustness and to deal with high dimensional data.

In practice, many times we find that some observations in the training sample have wrong label. If there are a few of them together, the \( k \)-NN rule will be clearly affected and misclassify data close to them. Figure 1 exhibits an example of this type of data.

In high dimensional spaces, the problem is that distances among a set of points are very similar. This makes classification rules based on distances difficult to perform.

One dimensional projections, combined with \( k \)-NN rules, is a way to overcome these problems. The new RP\( k \)NN-method (Random Projections of \( k \) Nearest Neighbors) follows the next steps:
Figure 1: Contaminated data containing several clusters with small variance which are misclassified.

1. Given $x \in \mathbb{E}$ and the number of nearest neighbors $k$, we fix $r > 1$ and we find the $[rk]$–nearest neighbors of $x$ among $\{X_1, \ldots, X_n\}$. We denote this set by

$$J[x] = \{X_{i_1}, \ldots, X_{i_{[rk]}}\},$$

and call it a “district”, which has size $[rk]$. Distance ties are broken by comparing indices, that is, if $||X_i - x|| = ||X_j - x||$, $X_i$ is considered closer to $x$ when $i < j$.

2. Choose at random (according to a non–degenerate distribution on the unit sphere $\nu$) $N$ independent directions $f_1, \ldots, f_N$.

3. For each fixed $f_j$, project on that direction the district subset $J[x]$ and the new data $x$ that we want to classify. We obtain them $\{f_j(X_{i_1}), \ldots, f_j(X_{i_{[rk]}})\}$ and $f_j(x)$.

4. Define $n_1(f_j), \ldots, n_m(f_j)$, where $n_l(f_j)$ is the number of data with label $l$ on $J[x]$ that are among the $k$–nearest neighbors of $f_j(x)$ for the projected data on the direction $f_j$. If $J[f](x)$ denotes the set of the $k$–nearest neighbors of $f(x)$ among $\{f(X_{i_1}), \ldots, f(X_{i_{[rk]}})\}$,

$$n_l(f_j) = \sum_{i=1}^n I_{\{X_i \in J[x]\}} I_{\{f_j(X_i) \in J[f](x)\}} I_{\{Y_i = l\}}.$$

5. Define the average number of votes for label $l$ as

$$\bar{n}_l = \frac{1}{N} \sum_{j=1}^N n_l(f_j),$$

and classify $x$ into the group $l$ with a larger value of $\bar{n}_l$. 

4
In this way, we may have chosen some directions under which projections of the data that belong to the district are not adequate for a good classification. In order to avoid this problem, we reject those directions where the proportion of misclassified data among the projected district-data is above a given cut-off value, as for instance, 40% or 50%. This can be done by a leave one out procedure on the data in the district, since they belong to the training sample.

3 A small simulation study

In this section we present a small simulation study in order to exhibit the robustness property of the RPkNN method. For the sake of simplicity we considered a two-group structure in $\mathbb{R}^2$, represented in Figure 1. Group A was generated from a uniform distribution on the rectangle $[1.1, 3.1] \times [0, 2]$, while group B was generated from a uniform distribution on $([0, 1] \cup [3.2, 4.2]) \times [0, 2]$. The training sample has 720 observations: 600 of them generated according to the described distribution, one third of the points in each vertical rectangle; the remaining 120 points (contaminated misclassified data) were generated on eight clusters on group A and 16 clusters on group B with small variance, each one with five misclassified observations. The test sample has size 600 and was generated following the same distribution as the training sample without contamination.

We carried out 2000 replications, and each time both the training and the test samples were generated. Before running the simulation a preliminary study has been pursued so that the optimal parameters were found. Then the number of nearest neighbors and size of the district remained fixed through the simulation. In order to classify the testing sample 10 random directions were generated. Each of them was considered plausible if it classified correctly at least 40% of the points of the district data. The simulations results are presented in Figure 2.

Figure 2: Left: Density estimator for the classification rates of kNN and RPkNN. Right: Density estimator for the pairwise difference relative improvement rate of RPkNN respect to kNN.
The mean classification rate for \( k \)-NN was 87.75% when considering 3 neighbors. For the RP\( k \)NN method we obtained 93.40% of correct classifications, considering 5 neighbors and a district of size 20. This feature is exhibit in Figure 2 (left) where also it can be seen that the classification rate for RP\( k \)NN has less variance than the classification rate for kNN. In order to quantify the relative improvement achieved when considering the projection based rule we compute the relative improvement rate suggested by Croux et al. (2007), i.e.

\[
\text{Relative Improvement} = \frac{ER_{k-\text{NN}} - ER_{\text{RP}k\text{-NN}}}{ER_{k-\text{NN}}},
\]

where \( ER_{k-\text{NN}} \) and \( ER_{\text{RP}k\text{-NN}} \) are respectively the error rates for the \( k \)-NN and for the Projected \( k \)-NN classifications rules. In this case the relative improvement rate is 46.12%. Moreover, Figure 2 (right) exhibits the density estimation of the pairwise relative improvement, in every case RP\( k \)NN performed better than kNN, its the mean value is 46.04%.

4 A real data example: The Isolet data

Isolet is a data set of speech features from the UCI Machine Learning Repository (Asuncion and Newman, 2007), consisting on 617 attributes associated to English pronunciation of 26 letters from the alphabet. The data come from 150 people who spoke the name of each letter twice. There are three missing data, obviously not considered in the study. We try to classify each sound to the corresponding letter, so there are 26 groups. The 80% of the data were used in the training sample (6,238 data), and the remaining in the test sample (1,559 data). All the details about the experiment are discussed in Fanty and Cole (1991).

Features vectors include: spectral coefficients; contour features, sonorant features, pre-sonorant features, and post-sonorant features; and are described in Fanty and Cole (1991). Spectral coefficients account for 352 of the features. Exact order of appearance of the features is not known.

Our first approach was to carry out the procedure in one stage. We found the optimal parameters for each method considering a leave-one-out procedure on the training sample: 7 nearest neighbors for \( k \)-NN, and 5 nearest neighbors and district size 8 for RP\( k \)NN. Classification rates were 91.73% for \( k \)-NN and 92.37% for RP\( k \)NN. Relative improvement rate was 7.74%.

Automatic recognition of spoken letters is a difficult task due to the acoustical similarities among several letters, as mentioned in Fanty and Cole (1991). To improve our performance we consider a two–stage procedure. First, we classified the observations into three groups, following a classical partition considered in linguistics. The E-set conformed by the letters \{b, c, d, e, g, p, t, v, z\}; the MN-net, which is \{m, n\}; and the R-set with the remaining letters. Figure 3 shows the 617 attributes for three particular
Table 1: Classification and improvement rates on the second stage of the procedure for the three groups considered in the Isolet dataset.

<table>
<thead>
<tr>
<th></th>
<th>RP$k$NN</th>
<th>$k$-NN</th>
<th>Relative Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-set</td>
<td>87.38</td>
<td>86.36</td>
<td>7.48</td>
</tr>
<tr>
<td>MN-net</td>
<td>91.15</td>
<td>86.61</td>
<td>33.91</td>
</tr>
<tr>
<td>R-set</td>
<td>97.02</td>
<td>95.61</td>
<td>32.12</td>
</tr>
</tbody>
</table>

letters from each group: a, e and m. The solid line represents the deepest curve, as it is defined by Fraiman and Muniz (2001). The three curves are quite similar, making the identification of these three groups a challenging problem.

Figure 3: The 617 attributes for the letters a, e and m, for all the data in the training sample, the solid line represents the deepest observation.

On this first stage, we achieved a 99.17% of well classified data with RP$k$NN, while the classification rate with $k$-NN was 98.72%, with a relative improvement of 35.16%. Optimal number of neighbors in the first stage for $k$-NN was 7 and for RP$k$NN was 5 and a district of size 15.

On the second stage, we classified the observations inside the groups they were assigned to on the first stage. Table 1 shows the classification rate and the relative improvement for each group. We notice that projecting improves the error rate from the $k$-NN classification rules. Even in the case of the E-set, where the classification task is more difficult, our proposal outperforms the classical method. Table 2 shows the number of neighbors considered in each method.

Summing up, classification rate for the two-stages RP$k$NN method is 93.26% and for $k$-NN is 91.53%. These means that the overall improvement rate of our method is
Several attempts have been made to classify this data set by different methods, the results can be found at http://eden.dei.uc.pt/ bribeiro/UCI/isolet/isolet.info. It is important to notice that RP\(k\)NN performs better than all the other methods except for C4.5 107-bit ECOC soft pruned, Opt OPC, Opt 62-bit ECOC and Opt 30-bit ECOC, see Dietterich and Bakiri (1994) for details. They attained success rates from 96.73% (Opt 30-bit ECOC) to 93.39%. These methods use the conjugate-gradient implementation of a back propagation algorithm followed by an error-correcting output code. They are hard to apply and use extra information of the data set, as for instance the detection of the code words, being in some way "black-box methods". For the procedure proposed in this paper - which is universally consistent - we have just used the Euclidian distance between data points without any further information. Only with that we obtain a quite good classification rate.

5 Universal consistency of the RP\(k\)NN rule

The proposed rule based on random projections is weakly universally consistent if \(k\) is allowed to grow with \(n\), such that \(k = k(n) \rightarrow \infty\) and \(k/n \rightarrow 0\). To simplify notation, we will assume \(m = 2\).

We fix \(x \in \mathbb{R}^p\) and reorder the training data \(D_n =: ((X_1, Y_1), \ldots (X_n, Y_n))\), according to increasing values of \(||X_i - x||\). The reordered sequence is

\[
(X^{(1)}(x), Y^{(1)}(x)), \ldots, (X^{(n)}(x), Y^{(n)}(x)),
\]

where \(X^{(k)}(x)\) is the \(k\)-th nearest neighbor of \(x\). Distance ties are broken by comparing indices, that is, in the case of \(||X_i - x|| = ||X_j - x||\), \(X_i\) is considered closer to \(x\) if \(i < j\).

Let \((X, Y)\) be a random vector with the same distribution as \((X_1, Y_1)\) and independent of the training sample. Given a direction \(f\) on the unit sphere, let \(f(X) = \langle f, X \rangle\) be the projection of \(X\) on direction \(f\), and let

\[
D_{n,f} =: \left( (f(X_{i_1}), Y_{i_1}), \ldots (f(X_{i_{|k|}}), Y_{i_{|k|}}) \right)
\]

be the projected sub-sample of \(J[x]\) on \(f\). Reorder the projected sample according to

<table>
<thead>
<tr>
<th></th>
<th>(k)-NN neighbors</th>
<th>RP(k)NN neighbors</th>
<th>district</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-set</td>
<td>7</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>MN-net</td>
<td>5</td>
<td>7</td>
<td>21</td>
</tr>
<tr>
<td>R-set</td>
<td>7</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 2: Optimal parameters for \(k\)-NN and for RP\(k\)NN when classifying the Isolet dataset

18.50%
increasing values of $|f(X_h) - f(x)|$, for $h = 1, \ldots, [rk]$, and the new sequence is

$$\left( f^{(1)}(x), Y^{(1)}(x) \right), \ldots, \left( f^{[rk]}(x), Y^{[rk]}(x) \right),$$

where $f^{(l)}(x)$ stands for the $l$-th nearest neighbor of $f(x)$ among $f(X_{i1}), \ldots, f(X_{i[rk]})$.

Finally, as stated above, $J[f](x)$ denotes the set of the $k$-nearest neighbor of $f(x)$ among $f(X_{i1}), \ldots, f(X_{i[rk]})$.

A classification rule is weakly consistent for a certain distribution of $(X, Y)$ if

$$E(L_n) = P(g_n(X, D_n) \neq Y) \rightarrow L^*, \text{ as } n \to \infty.$$  

Since $L_n$ is a random variable bounded between $L^*$ and 1, this convergence is equivalent to the convergence of $L_n$ to $L^*$ in probability.

A rules is called universally consistent if it is consistent for any distribution of the pair $(X, Y)$. In order to prove universal consistency we use a theorem given by Stone (1977) for classification rules of the form

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} W_{ni}(x) I_{\{Y_i=1\}} > \sum_{i=1}^{n} W_{ni}(x) I_{\{Y_i=0\}} \\ 0 & \text{otherwise,} \end{cases}$$

or equivalently,

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} W_{ni}(x) Y_i > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

where weights $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ verify certain assumptions stated in the following theorem.

**Theorem 1. (Stone, 1977)** Assume that for any distribution of $X$, weights $W_{ni}(x)$ satisfy the following three conditions:

(i) There is a constant $c$ such that, for every nonnegative measurable function $f$ satisfying $E(f(X)) < \infty$,

$$E\left( \sum_{i=1}^{n} W_{ni}(X) f(X_i) \right) \leq cE(f(X)).$$

(ii) For all $a > 0$,

$$\lim_{n \to \infty} E\left( \sum_{i=1}^{n} W_{ni}(X) I_{|X_i - X| > a} \right) = 0.$$

(iii)

$$\lim_{n \to \infty} E\left( \max_{1 \leq i \leq n} W_{ni}(X) \right) = 0.$$

Then $g_n$ is universally consistent.
We will also need the following result.

**Lemma 1.** (Stone, 1977) For any integrable function $g$, any $n$, and any $k \leq n$,

$$\sum_{i=1}^{k} E \left( |g(X^{(i)}(X))| \right) \leq k \gamma_p E \left( |g(X)| \right),$$

where $\gamma_p \leq \left( 1 + 2/\sqrt{2} - \sqrt{3} \right)^p - 1$ only depends upon the dimension $p$.

The random projection based $k$-NN rule, $\tilde{g}_n(x)$, can be written in terms of weights. For each direction $f_j$, $1 \leq j \leq N$, let define the weights

$$W_{nij}(x) =: W_{nij}(x, f_j, D_n) = \begin{cases} 1/k & \text{if } f_j(X_i) \in J[f_j](x) \\ 0 & \text{otherwise.} \end{cases}$$

Rule $\tilde{g}_n(x)$ can be expressed in terms of the average weights along all random directions,

$$\tilde{g}_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} W_{ni}(x)Y_i > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

where

$$W_{ni}(x) = \frac{1}{N} \sum_{j=1}^{N} W_{nij}(x).$$

**Theorem 2.** (Universal consistency of $\tilde{g}_n$) If $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then for all distribution $E(L_n(\tilde{g}_n)) \to L^*$, i.e. $\tilde{g}_n$ is universally consistent.

**Proof of Theorem 2.**

The proof is mainly the same as for the universal consistency of $k$-NN rules, we proceed by checking the conditions of Theorem 1.

We start with condition (i). It is a consequence of Lemma 1 and the fact that $J[f_j](x) = J[f_j](x) \cap J[x]$, since

$$E \left( \sum_{i=1}^{n} W_{ni}(X)g(X_i) \right) \leq \frac{1}{k} E \left( \sum_{i=1}^{[rk]} g(X^{(i)}(X)) \right) \leq r \gamma_p E (g(X)).$$

Condition (ii) holds since

$$\lim_{n \to \infty} E \left( \sum_{i=1}^{n} W_{ni}(X)1_{\{|X_i-X|>a\}} \right)$$

holds whenever

$$P \left( \|X^{(rk)}(X) - X\| > \epsilon \right) \to 0,$$

which holds for all $\epsilon > 0$ if $rk/n \to 0$. Finally, condition (iii) follows since $\max_{1 \leq i \leq n} W_{ni}(x) \leq 1/k \to 0$. 

10
6 Conclusions

We introduce a new family of classification rules, which can be thought as an improvement of the classical $k$–NN rules, in order to attain robustness and to deal with high dimensional data. The new RP$k$NN rules are based on the concepts of random projections and nearest neighbors. The main idea is to expand the set of $k$–NN considering a new set call district that includes the $rk$–NN. This set is projected on random directions and then classical $k$-NN is applied to the projected data. Under mild conditions, the RP$k$NN method is universally consistent. Simulations and real data examples show an improvement in classification score when compared with the classical $k$-NN method.

Acknowledgements

Ricardo Fraiman and Ana Justel have been partially supported by the Spanish Ministerio de Educación y Ciencia, grant MTM2007-66632. Ana Justel is also partially supported by Comunidad Autónoma de Madrid, grant CCG07-UAM/ESP-1761. We thank Adolfo Hernández for his very interesting comments.

References


