Comparison Classifier of Condensed KNN and K-Nearest Neighborhood Error Rate Method


Abstract - Error rate of classification from condensed KNN, cross validation, K-nearest neighborhood error rate are applied to data of 228 pregnant women from 114 training set and the other 114 were the testing set used to estimate the true error rates of the two classification method. In the two procedure, k-nearest neighborhood k=3 and above has the lowest error rate of 0.439. Since, the lower the error rate, the better the procedure. Therefore the k-nearest neighborhood (KNN) gave the best performance classified; as the error rate was very small compared to others.

Keyword: Classifier, condensed, error rate, KNN and statistical pattern recognition.

I. INTRODUCTION

Many decision making fall into general category of classification [1], [2]. There are several ways of measuring classifier performance, the most common being error rate, although this has several limitations. Other measures, based on closeness of the estimates of the probabilities of class membership to the true probabilities may be more appropriate in many cases.

Estimating error rates techniques has been studied in pattern recognition [3], [4]. An error rate estimator based on the bootstrap, which was designed for the two-group problem but can easily be modified for more groups was proposed [5]. But experimenting with different designs and basing the choice on criteria that in addition to classification error can include other issues such as computational complexity and feasibility of efficient hardware implementation can find the best classifiers for a given task.

The k-nearest neighbor (KNN) classification rule is one of the most well-known and widely used nonparametric pattern classification methods.

II. PROCEDURE FOR PAPER SUBMISSION

The error rate is the probability of making a definite erroneous classification for a future randomly chosen sample, previously celled pmc. The easiest way to assess the error rate is to choose a test independent of the training set (and validation set if used), to classify its examples, count the errors and divide by the size M of the test set.

The simplest technique for honestly estimating error rate, the holdout or it method is a single train and test experiment [3]. The sample cases are broken into two groups of cases: a training group and a test group. The classifier is independently derived from the training cases, and the error estimates is the performance of classifier on the test cases. In order to assess the performance of the classification methods, the error rate over the test set is used. [6]

The test set needs to be large for the error rate to be estimated at all accurately. The task of comparing the error rate of two classifiers is rather easier. We can also calculate the error rates conditionally for each class, by counting within each class. If we know the prior probabilities πk, the estimator

\[ \hat{pmc}(k) = \sum_k \pi_k n_k \]  

(1)

A. k-NEAREST NEIGHBOUR

It is a non-parametric classification method in statistics. Let “k” be some fixed integer. Given a data record “r” to be classified, instead of finding the single nearest neighbor, k-nearest neighbors to datum “r” are found. The classification for “t” will now be the class that appears most frequently among the k-neighbors.

B. CONDENSED KNN

The training set is partitions into a few cluster of neighbor. Each cluster has numerical value for posterior probability of all possible classes given the input attribute for the numbers of the cluster. A new item is classified by finding its nearest cluster and using that cluster’s posterior probability estimates to estimate the class for new item.

The aim of this approach is to obtain a small template that is a subset of the training set without changing the nearest neighbor decision boundary substantially. The idea is that the patterns near the decision boundary are crucial to the KNN decision, but
III. METHODS

In the course of this work, we have tested the classification performance of the statistical method of the Condensed KNN and k-Nearest Neighborhood to compare the error rates of different classifiers on a diverse collection of data sets. These have been selected in such a way that the total number of samples per data set, the feature dimensionality and the number classes cover an appropriate range of which will provide a comprehensive comparison of the two classification methods[9]. The data set were obtained from antenatal cases of Usman Danfodiyo University Teaching Hospital (UDUTH) Sokoto. A statistical software package called R was used to analyze the data on delivery in order to classify the delivery mode. The results for Condensed KNN and k-Nearest Neighborhood techniques were obtained from the data. The data frame of 228 columns was divided into two. That is columns 1 to 114 represents our training set while columns 115 to 228 represent our testing set. We design a classifier to minimize the true error rate.

K-nearest Neighbor Classification Rule

The equation \( \hat{p}(X) = \frac{k}{NV} \) gives an expression for a density estimate. We can now use it to derive a decision rule for the k-nearest neighbor method. Let us suppose \( k_m \) out of the k-nearest neighbor belongs to class \( \pi_m \), and \( \sum_{m=1}^{c} k_m = k \). Let also the total number of items in class \( \pi_m \) be \( n_m \) with

\[
\sum_{m=1}^{c} n_m = N .
\]

Then the estimates, of class-conditional density

\[
P(x|\pi_m) = \frac{k_m}{n_mN} \quad (2)
\]

and the prior probability are respectively given as

\[
\hat{p}(\pi_m) = \frac{n_m}{N} \quad (3)
\]

Then the decision rule is: assign \( X \) to class \( \pi_m \) if

\[
\hat{P}(\pi_m|x) > \hat{P}(\pi_i|x) \quad \text{for} \ m, i \in c
\]

Using Bayes theorem

\[
P(\pi_m|x) = \frac{P(x|\pi_m)p(\pi_m)}{p(x)} \quad (5)
\]

With (5) in mind, using (3) and (4) we rewrite the decision rule as

\[
\frac{k_m n_m}{n_m \sqrt{N}} > \frac{k_i n_i}{n_i \sqrt{N}} \quad \text{for} \ m, i \in c
\]

That is assign \( X \) to class \( \pi_m \) if

\[
k_m > k_i , m, i \in c
\]

We can also write decision rule as:

Assign \( X \) to class \( \pi_m \) if \( k_m = \max \{ k_1, \ldots, k_c \} \rightarrow x \in \pi_m \)

Thus the decision is to assign \( X \) to the class that receives the largest number of votes from amongst the k-Nearest Neighbors. In case of ties, \( X \) may be assigned arbitrarily between tied classes or else it may be assigned to the class that is most compact, that is the class for which the distance to the \( k_j \)th member is the smallest. There are yet other methods of breaking the ties [3]. For \( K = 1 \), it is a straightforward case; assign \( X \) to the class of its nearest neighbor point.

Minimizing the Expected Error Rate

The usual way of formalizing a goodness criterion is by means of a loses function. Let \( L(K, L) \) be the loss incurred by making decision \( L \) if the true class is \( C = k \). Then \( L(k, k) = 0 \) and may be \( L(k, D) = d \) for all \( k \). And \( L(k, L) \)'s could in principle be any set of positive numbers. If every misclassification is equally serious, then

\[
L(K, L) = \{ 0 \text{ if } 1 = k \text{ (correct decision)} \}
\]

\[
\{ 1 \text{ if } l \neq k \text{ and } l \in \{ 1, \ldots, K \} \text{ (wrong decision)} \}
\]

\[
\{ d \text{ if } l = D \text{ (being in doubt)} \}
\]

For \( k = 1 \ldots k \), and \( l = 1 \ldots K \), \( D \) is a reasonable choice.

The risk function for classifier \( \hat{C} \) is the expected loss when using it, as a function of the unknown class \( k \):

\[
R(\hat{C}, k) = E \left[ L(K, \hat{C}(X))/C = K \right]
\]

\[
= \sum_{k=1}^{K} L(K, L) pr \{ \hat{C}(x) = l | C = k \} +
\]

\[
L(K, D) pr \{ \hat{C}(x) = D | C = K \}
\]

\[
= pmc(k) + d pd(k)
\]

The total risk is the total expected loss, viewing both the class \( C \) and the vector \( X \) as random

\[
R(\hat{C}) = E \left[ R(\hat{C}, C) \right]
\]

\[
= \sum_{k=1}^{K} \pi_k pmc(k) + d \sum_{k=1}^{K} \pi_k pd(k)
\]

The posterior probability of class \( K \) given \( X = x \)

\[
p(k|x) = pr \{ C = k | X = x \} = \frac{\pi_k p_k(x)}{\sum_{i=1}^{K} \pi_i p_i(x)}
\]

The error rate plays an important part in decision-making and classifier performance assessment. Therefore, estimation of
error rates is a problem of great interest in statistical pattern recognition. No matter how the decision regions are chosen the error rate may be regarded as a measure of a given decision rules performance.

IV. DISCUSSION OF RESULTS

Let us consider first the overall classification performance of both KNN and condensed KNN. The results are listed in Table 1 and are summarized as misclassification percentage rate. In the second and third column the error rate were obtained from it.

<table>
<thead>
<tr>
<th>TABLE I ERROR RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONDENSED KNN</td>
</tr>
<tr>
<td>K=1</td>
</tr>
<tr>
<td>K=2</td>
</tr>
<tr>
<td>K=3</td>
</tr>
<tr>
<td>K=4</td>
</tr>
<tr>
<td>K=5</td>
</tr>
</tbody>
</table>

Table I, gives the result of error rate. Since one important way of judging the performance of any classification-procedure is to measure its “error rate” or misclassification probabilities. The lower the error rate, the better the procedure. The error rates gotten were very low. Therefore, the procedure used that is Nearest Neighborhood approach is a better procedure and good in classifying the mode of delivery. Therefore, k-Nearest Neighborhood is a good classifier.

The result of the Graph which confirm that in the procedure, error rate was low, and k-Nearest Neighborhood at k=3 and above has the lowest error rate of 0.439. Since, the lower the error rate, the better the procedure. It can be said that k-Nearest Neighborhood is a better procedure than Condensed.

V. CONCLUSIONS

In this paper the error rate of two different classify of nearest neighborhood; condensed and K-nearest neighborhood were successfully applied and compared in the procedure of classifying mode of delivery data obtained from Usmanu Danfodiyo University Teaching Hospital (UDUTH). The aim of comparing the two classify is to found out the one that has the least error rate. The error rate for condensed was obtained and was found to be 0.202 at K = 1, for K = 2 the result is 0.152 and for K = 3 and above the result was found to be 0.0614. The error rate for k Nearest Neighborhood was obtained. It was found to be 0.0625 for K = 1, the result for K = 2, is 0.014 while the result for K = 3 and above was found to be 0.0439. It was also found out that k-Nearest Neighborhood at K = 3 have the least error rate. Since the lower, the error rates the better the procedure.

REFERENCES