COMPARISON OF CRISP AND FUZZY CLASSIFICATION TREES USING GINI INDEX IMPURITY MEASURE ON SIMULATED DATA

Eunice Muchai
Leo Odongo
Department of Statistics and Actuarial Science,
Kenyatta University, Nairobi, Kenya

Abstract

Crisp classification trees have been used to model many situations such as disease classification. With the introduction of fuzzy theory, fuzzy classification trees are gaining popularity especially in data mining. Very little work has been done in comparing crisp and fuzzy classification trees. This paper compares crisp classification trees and fuzzy classification trees using Gini index as the impurity measure. The objective is to determine which of the two classification trees gives fewer errors of classification. The data used consisted of two sets of observations from multivariate normal distributions. The first set of data were from two 3-variate normal populations with different mean vectors and common dispersion matrix. From each of the two populations 5000 samples were generated. 1000 samples out of the 5000 were used to create the trees. The remaining 4000 samples from each population were used to test the trees. The second set of data were from three 4-variate normal populations with different mean vectors and common dispersion matrix. A similar sampling and testing procedure as for the case of first set of data was employed. Computations were implemented using R statistical package. The results from the test showed that fuzzy classification trees allocated observations to the correct population with fewer errors than did crisp classification tree.

Keywords: Crisp classification tree, Fuzzy classification tree, Gini index, Fuzzy decision points, Crisp decision points

Introduction

Various criteria have been proposed for selecting the variable used for splitting tha data in creating classification trees. Kass(1980) used a testing procedure based on Pearson's chi-squared statistic to choose the best multiway split. Breiman, et.al. (1984) introduced CART which provided the

Gini index and towing criterion. Loh and Vanichsetakul (1988) and Loh and Shih (1997) employed statistical test to select splits. Singh, et.al. (2010) applied Gini index to feature selection for text classification.

The concept of fuzzy random variable was introduced at the end of 1970's Kwakernaak(1978). This was to deal with situations where the outcomes cannot be observed with exactness.

Fuzzy decision trees differ from traditional trees by using splitting criteria based on fuzzy restrictions. Fuzzy sets defining fuzzy terms are imposed on the splitting algorithm. Janikow(1998) presented fuzzy trees using information gain as impurity measure and studied the performance of the tree when some data are missing. Wang, et.al. (2007) gave a survey of the different impurity measures that are currently in use.

In this paper the performance of crisp and fuzzy classification trees is compared. We only fuzzify the decision boundary using triangular membership function.

The organisation of the paper is as follows: Section 2 explains methodology and section 3 contains the results, discussions and conclusions.

Methodology

The Gini Index

When using decision/classification trees, the variable used at every node to split the tree affects the performance of the decision tree. The problem of selecting the splitting variable is therefore not trivial. After the variable has been selected the value of the variable that gives the best split is then selected. The objective of classification is to allocate individuals to the correct population with the minimum classification error. Using classification trees this is done so as to arrive at the different classes with the least number of splits with the least misclassification errors.

At every node the best splitting variable and the variables' best value are selected. The best variable and value are the ones that reduce the node non-uniformity commonly known as node impurity. Different impurity measures have been proposed and are in use, examples include Gini index, information gain, gain ratio, χ^2 statistic and the G statistic among others. In this paper the Gini impurity measure normally referred to as the Gini index is discussed and applied to simulated data.

A data set T containing individuals from n classes has the Gini impurity measure (Gini index), denoted by G (T), and given by

$$G(T) = 1 - \sum_{j=1}^{n} p_j^2$$ (see Breiman pg 104)

Where p_j is the relative frequency of class j in T.

Suppose the data set T is split into two subsets T_1 and T_2 with sizes N_1 and N_2 respectively, the Gini index of the split data is given by

$$\text{Gini}_{\text{split}}(T) = \frac{N_1}{N} G(T_1) + \frac{N_2}{N} G(T_2)$$

Using the Gini index the variable that gives the best split is applied. The best split is the one that has the least value of $\text{Gini}_{\text{split}}(T)$.

Splitting Procedure

The following procedure is used to select the splitting variable and the splitting value.

- calculate the $\text{Gini}_{\text{split}}(T)$ among the child branches over all possible decision points for each variable X_j at each node.
- select the variable and the value of that variable with the least $\text{Gini}_{\text{split}}(T)$, denoted by X_{j0} and use it for splitting .
- repeat this process at each node until splitting is completely done.

Comparison of the performance of Gini index having fuzzy decision points with Gini index having crisp decision points is carried out using simulated data. The first set of observations was generated from two 3-variate normal populations with different mean vectors and common dispersion matrix. The second set of observations was generated from three 4-variate normal populations with different mean vectors and common dispersion matrix.

Two populations with three variables

5000 Samples of different sizes from each population were generated. The populations were assumed to be normally distributed with different mean vectors but a common dispersion matrix. 1000 samples from each of the populations were used to create the classification tree. This was done using the splitting criteria discussed above. The splitting variable and value were obtained using Gini split. After the tree was created, the remaining 4000 samples from each population were used to test the performance of the tree. This was done by calculating the probabilities of correct allocation, that is P_{11} and P_{22} for both crisp and fuzzy decision points. .

Three populations with four variables

Simulation similar to the above scenario was done except in this case there were three populations with three variables. The probabilities of correct allocations P_{11} ,P_{22} and P_{33}, were calculated and are given below. Simulation and coding was done using the statistical package R and implemented on Pentium IV using windows 7 environment

Results, Discussion and Conclusion

Two populations with three variables

Table1 gives the average probabilities of correct allocation from the 4000 samples, at different sample sizes, using crisp and fuzzy decision points. The proportion of times probabilities of correct allocation was higher when using crisp cut points than triangular fuzzy decision points is given in Table2.

Table 1 : Probabilities of Correct Allocation

Sample size	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}
50	0.834	0.893	0.822	0. 895
100	0.831	0. 897	0.823	0. 894
200	0.825	0. 898	0.830	0. 896
500	0.832	0. 896	0.826	0. 892
1000	0.829	0. 895	0.831	0. 896

Table 2: Proportion of times crisp probabilities are higher than fuzzy probabilities

Sample size	P_{11}fuzzy<P_{11}crisp	P_{22}fuzzy<P_{22}crisp
50	0.074	0.079
100	0.013	0.027
200	0.001	0.002
500	0	0
1000	0	0

From Table 1, we observe that the average probabilities of correct classification using fuzzy decision points are higher than when using crisp decision points for all the sample sizes considered in the study. We also note that, as the sample size increases the average probabilities of correct allocation increases.

Also the proportion of times the probabilities of correct allocation were higher when using crisp decision points is quite low as seen in Table 2. As the sample size increases, the proportion of times crisp classification tree outperforms fuzzy classification tree tends to zero

It is therefore reasonable to conclude that, for two populations with three variables, fuzzy Gini classification tree performed better than the crisp Gini classification tree.

Three populations with four variables

Table 3 gives the average probabilities of correct allocation from the 4000 samples of different sizes using crisp and fuzzy decision points. The proportion of times probabilities of correct allocation was higher when using crisp decision points than triangular fuzzy decision points is given in Table 4.

Table3: Probabilities of Correct Allocation

Sample size	$P_{11\ crisp}$	$P_{11fuzzy}$	$P_{22crisp}$	$P_{22\ fuzzy}$	$P_{33crisp}$	$P_{33fruzzy}$
50	0.66	0.71	0.86	0.90	0.78	0.80
100	0.64	0.71	0.90	0.90	0.81	0.82
200	0.67	0.72	0.92	0.91	0.80	0.84

500	0.68	0.73	0.93	0.92	0.81	0.86
1000	0.69	0.74	0.93	0.94	0.85	0.86

Table 4 : : Proportion of times crisp probabilities are higher than fuzzy probabilities

Sample size	P_{11}fuzzy<P_{11}crisp	P_{22}fuzzy<P_{22}crisp	P_{33}fuzzy<P_{33}crisp
50	0.096	0.100	0.102
100	0.050	0.060	0.068
200	0.010	0.015	0.020
500	0.001	0.003	0.008
100	0	0	0

Comparing the columns of P_{11}, P_{22} and P_{33} in Table 3 above, we observe that the average probabilities of correct classification using fuzzy decision points are higher than when using crisp decision points. As observed in the case of two populations, as the sample size increases the average probabilities of correct allocation increases.

Also the proportion of times the probabilities of correct allocation when using crisp decision points are higher than using fuzzy decision points is quite low as is observed in Table 4. This proportion gets lower as the sample size increases.

As in the case of two populations, fuzzy classification trees perform better when there are three populations. Therefore, observing the results in Tables 1-4 above, it can be concluded that Gini fuzzy classification tree perform better than Gini crisp classification tree.

References:

Breiman, L., Friedman, J. H., Olshen, R. A.and Stone, C. J. *Classification and Regression Trees*, New York: Chapman & Hall, 1984.

Janikow. C.Z (1998). Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man, and Cybernetics*, **28**: 1-14

Kass.G.V.(1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**: 119-127.

Kwakernaak, H. (1978). Fuzzy random variables: definitions and theorems. *Information Science,* **15**:1-29.

Loh, W.Y. and Vanichsetakul,N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion).*Journal of the American Statistical Association*, **83**: 715-728.

Loh, W.Y. and Shih, Y.S (1997). Split selection methods for classification trees. *Statistica Sinica* **7**: 150-156.

Singh, S.R., Murthy, A.H. and Gonsalves, A. T.(2010). Feature selection for text classification based on Gini coefficient of inequality. *Proc. 4th workshop on feature*
*selection in data mining,***10***: 76-85*