Comparison of Crisp and Fuzzy Classification Trees Using Chi-Squared ImpurityMeasure on Simulated Data

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

Doi:10.19044/esj.2018.v14n35p351 URL:http://dx.doi.org/10.19044/esj.2018.v14n35p351

Abstract

Classification trees are one of the most popular choices in classification and discriminant analysis. One chief reason is that they are distribution free methods. Recently, with the introduction of fuzzy theory,fuzzy classification trees are gaining popularity. In this paper we use Pearson's chi-squared impurity measure to compare the performance of crisp and fuzzy classification trees. This is done using simulated data. 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. Using the Pearson's chi-squared statistic for testing homogeneity in contingency tables showed that fuzzy classification trees algorithm makes two subnodes more heterogeneous than the crisp classification algorithm. Therefore fuzzy classification trees allocated observations to the correct population with fewer errors than did crisp classification tree.

Keywords: Pearson's chi-squared impurity measure, Crisp classification tree, Fuzzy classification tree, Fuzzy decision points, Crisp decision points,splitting variable.

1. Introduction

Classification trees have been used for prediction and decision making. The key factor in the performance of a classification tree is the choice of the splitting variable. Various criteria have been proposed for selecting the variable used for splitting data. 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 theory.Two approaches are used, that is either consider all the data as fuzzy or use fuzzy decision points only. 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. Muchai and Odongo(2014) compared the crisp and fuzzy classification trees using Gini impurity measure on simulated data.

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

2. Methodology

When generating a classification tree, in each recursive step in an algorithm, one must select a variable (an attribute) to test a condition. The more heterogeneous a split algorithm makes the two subnodes, the better the algorithm.In a binary tree, the composition of the subnodes can be treated as a J x 2 contingency table. Pearson's chi-squared statistic is used for testing homogeneity in contingency tables. Therefore this statistic can be used as a splitting measure. The variable and the value which gives the highest computed Pearson's chi-squared statistic gives the subnodes that are most heterogeneous and is therefore used as the splitting variable.

The Pearson's chi-squared impurity measure is based on the chi-square distribution given by the following formula

$$P(X_0)_D = \int_0^{X_0} p(x)_D \, dx$$

Where $p(x)_D$ is the chi-square distribution with D degrees of freedom and X_0 is the value of the statistic for a given variable. This may be approximated by

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \quad where \qquad e_{ij} = \frac{n_{.j} n_{i.}}{n}$$ Where

$n_{.j}$ is the number of individuals from population j
$n_i$is the number of individuals from population i
andn is the total number of individuals

Selecting the splitting variable and the splitting value

Calculate thePearson's' chi-squared statisticamong the child branches over all possible decision points for each variable $X_j$at each node. These decision points are either crisp or fuzzy, hence generating crisp or fuzzy classification trees.

- select the variable and the value of that variablewith the maximum chi squared statistic value, denoted by $X_{j0}$and use it for splitting .
- repeat this process at each node until splittingis completely done.

Performance of fuzzyclassification trees is compared with crisp classification.This is done on 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 Chi-squared 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

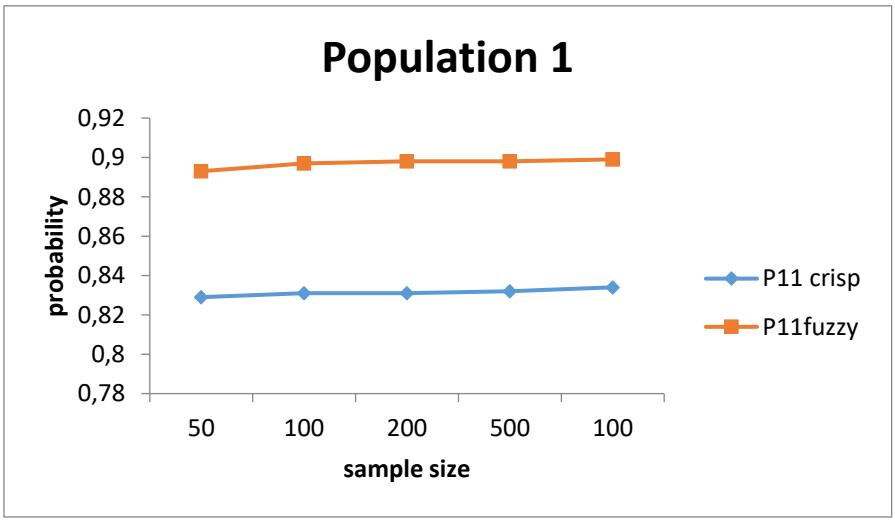
3. Results, Discussion and Conclusion

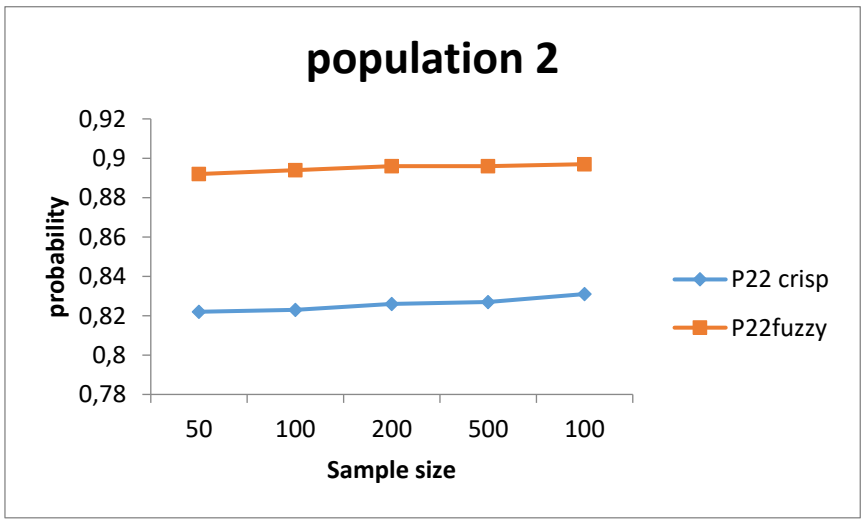
Two populations with three variables

Table1 gives the average probabilities of correct allocation from the 4000 samples, at different sample sizes, using crisp and fuzzy decision points.

Table 1: Probabilities of Correct Allocation

Sample size	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}
50	0.829	0.893	0.822	0. 892
100	0.831	0. 897	0.823	0. 894
200	0.831	0. 898	0.826	0. 896
500	0.832	0. 898	0.827	0. 896
1000	0.834	0. 899	0.831	0. 897





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 allthe sample sizes considered in the study. We also note that, as the sample size increases the average probabilities of correct allocation increases.

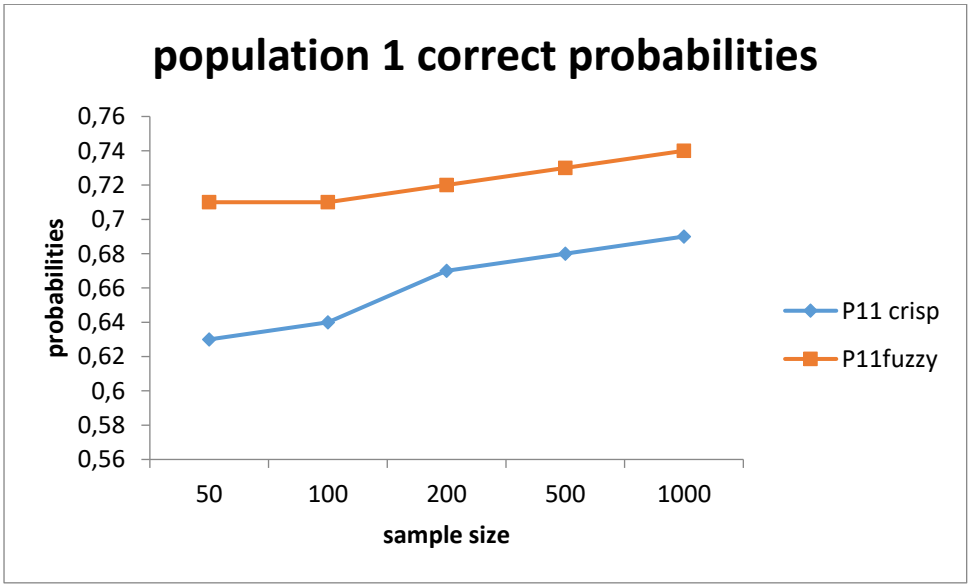
It is therefore reasonable to conclude that, for two populations with three variables, fuzzy Pearson's chi-squared classification tree performed better than the crisp Pearson's chi-squared classification tree.

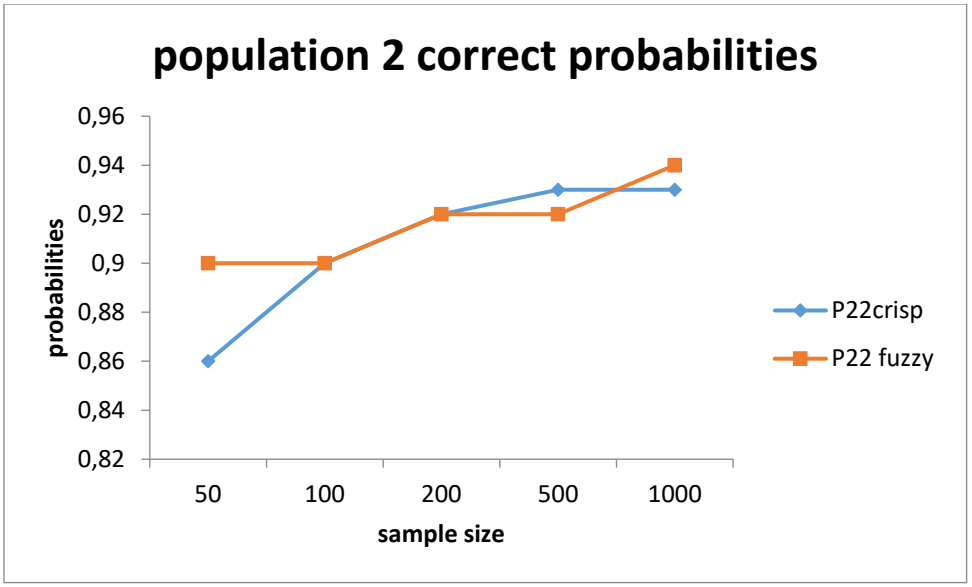
Three populations with four variables

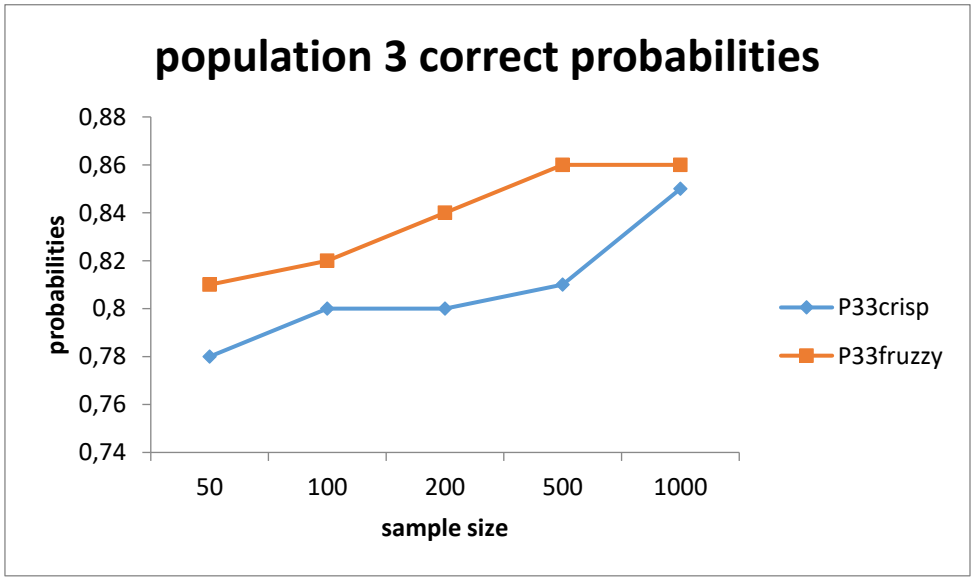
Table 2 gives the average probabilities of correct allocation from the 4000 samples of different sizes using crisp and fuzzy decision points.

Table2: Probabilities of Correct Allocation

Sample size	$P_{11\ crisp}$	$P_{11fuzzy}$	$P_{22crisp}$	$P_{22\ fuzzy}$	$P_{33crisp}$	$P_{33fruzzy}$
50	0.63	0.71	0.86	0.90	0.78	0.81
100	0.64	0.71	0.90	0.90	0.80	0.82
200	0.67	0.72	0.92	0.92	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







Comparing the columns of P_{11}, P_{22} and P_{33} in Table 2 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.

As in the case of two populations, fuzzy classification trees perform better when there are three populations. Therefore, observing the results in Tables 1-2 above, it can be concluded that Pearson's chi-squared fuzzy

classification tree perform better than Pearson's chi-squared crisp classification tree.

References:

1. Breiman, L., Friedman, J. H., Olshen, R. A.and Stone, C. J. *Classification and Regression Trees,* New York: Chapman & Hall, 1984.
2. Janikow.C.Z (1998). Fuzzy decision trees: Issues and methods.*IEEE Transactions on Systems, Man, and Cybernetics*, **28**: 1-14
3. Kass.G.V.(1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**: 119-127.
4. Kwakernaak, H. (1978). Fuzzy random variables: definitions and theorems. *Information Science,***15**:1-29.
5. Loh, W.Y. andVanichsetakul,N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion).*Journal of the American Statistical Association*, **83**: 715-728.
6. Loh, W.Y. and Shih, Y.S. (1997). Split selection methods for classification trees. *StatisticaSinica***7**: 150-156.
7. Muchai, E. and Odongo, L. (2014).Comparison of Crisp and Fuzzy Classification Trees Using GiniIndex Impurity Measure on Simulated Data. European Scientific Journal: **10**No.18, 130-134
8. 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*