COMPARISON OF FUZZY AND CRISP CLASSIFICATION TREES USING GINI INDEX, CHI-SQUARE STATISTIC AND THE GAIN RATIO

MUCHAI, EUNICE WAMBUI

REG.NO. I84/13318/2009

A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF DOCTOR OF PHILOSOPHY (STATISTICS) IN THE SCHOOL OF PURE AND APPLIED SCIENCES OF KENYATTA UNIVERSITY

JULY 2017

DECLARATION

This thesis is my original work and has not been presented for any other degree or award in any university

Muchai, Eunice Wambui

Signature ……. Date……

This thesis has been submitted with our approval as university supervisors.

Prof. L. O. Odongo

Signature ……. Date……

Department of Statistics and Actuarial Science

Kenyatta University

Dr. J. K. Kahiri

Signature ……. Date……

Department of Statistics and Actuarial Science

Kenyatta University

DEDICATION

To my family, Peter, Timothy & Angeline, James & Mercy, David and Talia.

ACKNOWLEDGEMENT

I am most grateful to God for His Mercy and Grace extended to me through the years and in particular, during the period of this study.

To Prof. L.O., Odongo thank you for your unwavering patience, availability and invaluable guidance throughout the study. To Dr. J.K., Kahiri your timely comments and guidance made the completion of this work possible.

I really appreciate my husband and children who encouraged me when I almost gave up.

I wish to thank the entire staff of Statistics and Actuarial Science, Department of Kenyatta University for encouraging me.

To all other persons not mentioned here who may have played any part in my work, thank you and may God bless you all.

TABLE OF CONTENT

LIST OF TABLES

ABSTRACT

Discriminant (classification) analysis is a classification problem where a new individual is allocated into one of known populations or classes based on the measured characteristics of the individual. Different models are used in allocating the new individual into one of the populations (classes). Some of the models depend on the underlying distribution of the populations, hese are known as parametric models. If the model does not depend on any underlying distribution it is known as a distribution free or non parametric model. In this work a distribution free model known as classification tree is used. A classification tree is a presentation of edges and nodes. It is a model that is used to assign an individual to one of many classes or populations. At each node a test is applied on a value of one of the attributes (variables) of the individual. The individual moves to the next node (child node) along an edge depending on the result of the test. The attribute, on which the test is applied, is known as the splitting attribute and the value the splitting value. Tests are carried out at each node until it is not possible to carry out more tests. The final nodes are known as terminal or leaf nodes. Classification is done at the terminal nodes by assigning all the individuals on that node to a class. If the splitting value is a fuzzy value, then the tree is known as a fuzzy classification tree otherwise the tree is known as a crisp classification tree. When there are only two possible answers to the test at each node, the resulting tree is known as a binary tree. Classification trees have been used to model many situations. These include speech recognition, data mining and market surveys among others. In this study the performance of crisp and fuzzy classification trees was compared. The performance was based on probabilities of correct allocation and probabilities of misclassification. Simulated data and real data were used. Data was simulated using R and the real data was obtained from machine learning repository. Gini Index, Chi-Square Statistic and Gain Ratio impurity measures were applied to both the simulated data and real data. The performance of Gini Index, Chi-Square Statistic and Gain Ratio impurity measures was also compared. Finally the performance of the trees using varied sample sizes was compared. It was found that for the simulated data, fuzzy classification tree performed better than the crisp classification tree when all the three impurity measures were applied. It was found that the Gini Index and Chi-Square Statistic impurity measures were appropriate as impurity measures for the data used in the study and gave similar results. However the Gain Ratio impurity measure did not perform as well as the other two impurity measures. It was also found that there was no significant difference in the probabilities of misclassification irrespective of different sample sizes in the populations.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Classification is a branch of Statistics that identifies which of a set of categories or classes, a new individual belongs. This is done on the basis of measurements on one or more of the attributes (variables) of an individual. A criterion or classification rule is applied on the individual's measurements to determine which category (class) the individual observation belongs. Since this applies probability theory, it is possible that an individual maybe allocated to the wrong class. The probability of allocating an individual to the wrong class is known as probability of misclassification. A good classification rule minimizes the probabilities of misclassification.

An interest in classification permeates many scientific studies, and also arises in the context of many applications. Classification has been used in applications in speech and speaker recognition and problems in acoustics. In biological sciences applications to problems of taxonomy, problems of classifying diseases by symptoms in health sciences use classification techniques. Classifying artifacts in archaeology or identifying market segments in market research also apply classification techniques. The central interest is in classifying objects of some kind or assigning an object to a class. Discriminant and classification analysis have been applied in diverse fields in various spheres of life. Examples include: In the medical, field for instance, doctors have applied classification in deciding whether a patient has cancer or not (Banas *et al.*, 2007). Cases have also been cited in psychology where psychologists apply classification techniques to identify intellectually gifted children (Pyryt, 2004). In market survey researchers

have used classification techniques to predict whether a company will go bankrupt or not (Altman, 1968). Discriminant analysis has also been applied to a war data set, where a discriminant model was used to predict who would win a battle given some characteristics from the Second World War (Chalikias *et al.*, 2009). Classification has also been applied in artificial intelligence, especially in pattern recognition (Hastie *et al.*, 1995).

Many different models of classification are in use today. In all the models, the guiding principle is to minimize the probabilities of misclassification. Classification should also be done within acceptable time and with reasonable effort. Below is a brief description of some classification models in use.

1.2 Classical classification

Consider g mutually exclusive and exhaustive populations denoted by $\Pi_1, \Pi_2, ..., \Pi_g$ where g $\geq$2. Assume that each individual in population Π_i can be described by a p-dimensional random vector $\underset{\sim}{X} = (X_1, X_2, \ldots, X_p)'$ with corresponding density function given by $f_i(\underline{x})$ i=1, 2,. . . , g, defined on p-dimensional real space $\Re^p$. Let the corresponding mean vector and dispersion matrix be given by $\underline{\mu_i}$ and Σ_i respectively.

Then classification (or discrimination) procedures involve partitioning the sample space $\Re^p$ into g disjoint subsets R_1, R_2, . . ., R_g such that $\Re^p = \bigcup_{i=1}^{g} R_i$, and then allocating an individual observation $\underline{X}$ to Π_i if $\underset{\sim}{X} \in R_i$.

1.2.1 Linear Discriminant Rule

A rule d which allocates $\underline{X}$ to $\prod_i$ if $\underline{X} \in R_i$ is called a discriminant rule. The classical linear discrimnant rule allocates $\underline{X}$ to $\prod_j$ if it maximizes $\left(\underline{x}-\underline{\mu_i}\right)' \Sigma^{-1}\left(\underline{x}-\underline{\mu_i}\right),\ i=1, 2, \ldots, g$

That is, allocate an individual with characteristic $\underline{X}$ to $\prod_j$ if;

$$\left(\underline{x}-\underline{\mu_j}\right)' \Sigma^{-1}\left(\underline{x}-\underline{\mu_j}\right) = \max_i \left(\underline{x}-\underline{\mu_i}\right)' \Sigma^{-1}\left(\underline{x}-\underline{\mu_i}\right) \text{ i, j =1, 2,..., g} \qquad (1.1)$$

Under multivariate normal assumption this is equivalent to maximizing the likelihood function. That is allocate $\underline{X}$ to

$$\Pi_j \ if\ L_j(\underline{x}) = \max_i L_i(\underline{x}) \text{ i, j = 1, 2, . . . , g} \qquad (1.2)$$

Where L_i denotes the likelihood function of $\underline{X}$.

1.2.2 Bayes' Discriminant Rule

The Bayes' discriminant rule allocates a new observation $\underline{X}$ to $\prod_j$ if;

$$\pi_j L_j(\underline{x}) = \max_i \pi_i L_i(\underline{x}) \qquad i,\ j=1, 2, \ldots, g \qquad (1.3)$$

Where π_i is the prior probability of an observation belonging to population $\prod_i$.

Suppose that the g populations are normally distributed. That is if $\underline{X}$ comes from the population $\prod_i$, then $\underline{X}$ is normally distributed with mean vector $\underline{\mu_i}$ and covariance matrix $\sum_i$. Bayes'

allocation rule will allocate a new observation $\underline{X}$ to $\prod_j$ if ;

$$\pi_j L_j(\underline{x}) = \max_i \pi_i L_i(\underline{x}) \qquad i,\ j = 1, 2, \ldots, g$$

That is, allocate $\underline{x}$ to $\prod_j$ if;

$$\pi_j (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_j\right)' \Sigma_j^{-1}\left(\underline{x}-\underline{\mu}_j\right)\right\}$$

$$= \max_i \pi_i (2\pi)^{-\frac{p}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_i\right)' \Sigma_j^{-1}\left(\underline{x}-\underline{\mu}_i\right)\right\} \text{ i, j=1, 2, …,g} \qquad (1.4)$$

This implies that the Bayes' allocation rule will allocate $\underline{x}$ to $\prod_j$ if;

$$\pi_j |\Sigma_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_j\right)' \Sigma_j^{-1}\left(\underline{x}-\underline{\mu}_j\right)\right\} \geq \pi_i |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_i\right)' \Sigma_i^{-1}\left(\underline{x}-\underline{\mu}_i\right)\right\} \text{ i, j = 1,2, . ..g,}$$

(1.5)

That is, allocate $\underline{x}$ to $\prod_j$ if;

$$\frac{\pi_j}{\pi_i} \geq \frac{|\Sigma_j|^{-\frac{1}{2}}}{|\Sigma_i|^{-\frac{1}{2}}} \exp\ -\frac{1}{2}\left\{\left(\underline{x}-\underline{\mu}_i\right)' \Sigma_i^{-1}\left(\underline{x}-\underline{\mu}_i\right) - \left(\underline{x}-\underline{\mu}_j\right) \Sigma_j^{-1}\left(\underline{x}-\underline{\mu}_j\right)\right\} \text{ i, j= 1, 2, . . . , g} \qquad (1.6)$$

If g =2 then, in this case there are two populations $\prod_1$ and $\prod_2$ and the Bayes' rule will allocate $\underline{x}$ to $\prod_1$ if;

$$\frac{\pi_1}{\pi_2} \geq \frac{|\Sigma_1|^{-\frac{1}{2}}}{|\Sigma_2|^{-\frac{1}{2}}} \exp -\frac{1}{2}\left\{\left(\underline{x}-\underline{\mu}_1\right)' \Sigma_1^{-1}\left(\underline{x}-\underline{\mu}_1\right) - \left(\underline{x}-\underline{\mu}_2\right)' \Sigma_2^{-1}\left(\underline{x}-\underline{\mu}_2\right)\right\} \text{i,j = 1, 2, . . . , g} \qquad (1.7)$$

Otherwise it will allocate $\underline{x}$ to $\prod_2$.

1.3 Classification trees

A classification tree is a representation of edges and nodes. Each node is connected to a set of possible subnodes, which are either Intermediate nodes or terminal nodes. Classification is done at the terminal nodes. Classification trees were first introduced by Morgan and Sonquist (1963) in the application of data survey. Since then classification trees have been used a lot as nonparametric classification models. A classification tree is a nonparametric method of discriminant analysis. Figure 1.1 is an illustration of a classification tree that is used to classify customers into low risk, moderate risk and high risk.

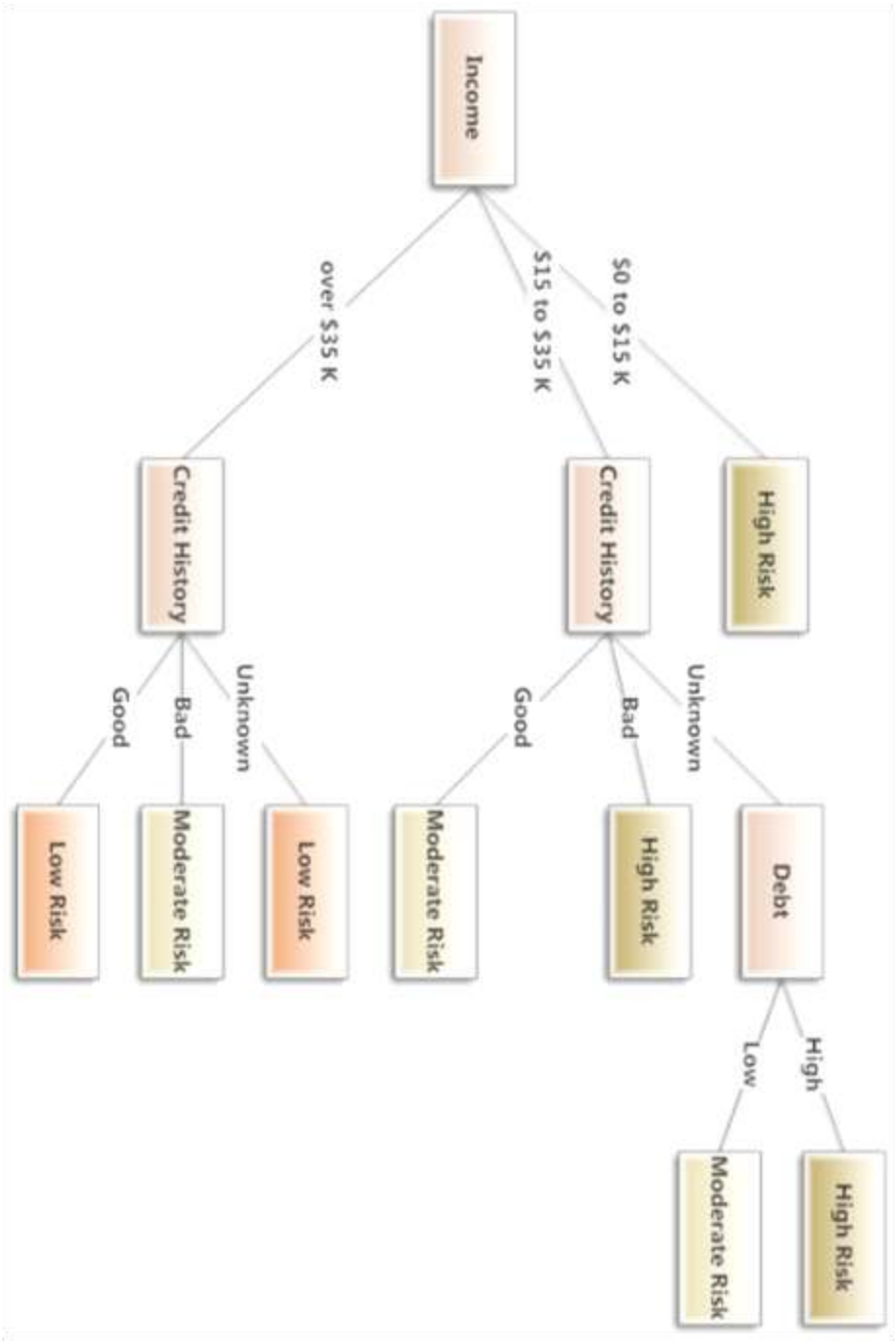

Figure 1.1: Classification tree for credit customers

Altman, (1968)

The original data, known as the root node, was partitioned into two or more non-overlapping sub-samples (or children nodes). The partitioning is done based on one of the independent variables known as the splitting attribute(variable). Each instance in the root node is sent down one of the branches (depending on its value of the splitting attribute) into one of the nodes. The splitting attribute is the attribute that will partition the original sample into sub-samples that are as homogenous (or pure) as possible. If a variable has more than one value, the value that gives the most homogenous sub-samples is used.

The process is repeated for each of the subsequent nodes. By considering the cases in a particular node the node is partitioned and new nodes are created. This process is repeated for each node until some rule is violated. When this happens, the node is not partitioned further and such a node is referred to as leaf or terminal node. The whole process is terminated when there are only leaf nodes left. Each leaf node is assigned one class representing the most appropriate target value. (Maimon and Rokach, 2010).

Three issues are to be considered in classification trees. These are; how to select the splitting attribute, when to stop splitting and how nodes are assigned to classes.

The splitting attribute is selected by using an impurity measure and classes are assigned at the terminal nodes. These will be further discussed in chapter three.

When the transition between attribute values is abrupt we have a crisp classification tree, otherwise when the transition between attribute values is gradual we have a fuzzy classification tree.

1.3.1 Fuzzy classification trees

Fuzzy classification trees are a fusion of fuzzy sets and decision trees. The fundamental difference between fuzzy and crisp trees is that with fuzzy decision trees, gradual transitions exist between attribute values. For the crisp classification the tree allows for abrupt transition. Fuzzy classification trees are the more natural approach. The early approaches to using fuzzy models were to fuzzify the whole data set. This is quite cumbersome and for big data sets is time consuming. A more recent approach is to fuzzify the decision point only (Janikow1996). Below is an introduction of fuzzy theory.

1.3.1.1 Fuzzy sets and approximate reasoning

Fuzzy sets were introduced by Zadeh (1965) to represent and manipulate data and process information when there are uncertainties which are non statistical. It was specifically designed to mathematically represent vagueness and to provide formalized tools for dealing with the imprecision intrinsic to many problems. Fuzzy logic provides an inference morphology that enables approximate human reasoning capabilities to be applied to knowledge-based systems. The theory of fuzzy logic provides a mathematical strength to capture the uncertainties associated with human cognitive processes, such as thinking and reasoning (Zadeh, 1965).

In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning and everything is a matter of degree, where knowledge is interpreted as a collection of elastic or, equivalently, fuzzy constraint on a collection of variables. In fuzzy reasoning, inference is viewed as a process of propagation of elastic constraints hence any logical system can be fuzzified.

Fuzzy systems are suitable for uncertain or approximate reasoning, especially for the system with a mathematical model that is difficult to derive. Also fuzzy logic allows decision making with

estimated values under incomplete or uncertain information. This gives fuzzy systems better performance for some applications.

In set theory, a subset A of a set X can be defined by a function χ_A as a mapping from the elements of X to the elements of the set $\{0, 1\}$, $\chi_A: X \rightarrow \{0, 1\}$.

This mapping is represented as a set of ordered pairs, with one ordered pair present for each element of X. The first element of the ordered pair is an element of the set X, and the second element is an element of the set $\{0, 1\}$. The value zero is used to represent complete non-membership, and the value one is used to represent complete membership. The truth or falsity of the statement, "x is in A", is determined by the ordered pair $(x, \chi_A(x))$. The statement is true if the second element of the ordered pair is 1, and the statement is false if it is 0.

A fuzzy subset A of a set X can be defined as a set of ordered pairs, each with the first element from X, and the second element from the interval $[0, 1]$, with exactly one ordered pair present for each element of X. This defines a mapping, μ_A, between elements of the set X and values in the interval $[0, 1]$. The value zero is used to represent complete non-membership, the value one is used to represent complete membership, and values in between are used to represent intermediate degrees of membership.

The set X is referred to as the universe of discourse for the fuzzy subset A. Frequently, the mapping μ_A is described as a function, the membership function of A. The degree to which the statement, "x is in A", is true is determined by finding the ordered pair $(x, \mu_A(x))$. The degree of truth of the statement is the second element of the ordered pair. The terms membership function and fuzzy subset get used interchangeably.

Definition 1.1 (Zadeh, 1965): Let X be a nonempty set. A fuzzy set A in X is characterized by its membership function μ_A, $X \rightarrow [0, 1]$ and $\mu_A(x)$ *is* interpreted as the degree of membership of element x in fuzzy set A for each $x \in X$

It is clear that *A* is completely determined by the set of tuples $A = \{(x, \mu_A(x)) | x \in X \}$

Frequently one writes simply $A(x)$ instead of $\mu_A(x)$. The family of all fuzzy subsets in *X* is denoted by $F(X)$. Fuzzy subsets of the real line are called fuzzy quantities.

Let *B* be a fuzzy subset of *X*, the support of *B*, denoted supp(*B*), is the crisp subset of *X* whose elements all have nonzero membership values in *B*.

That is,

$$Supp(B) = \{x \in X / B(x) > 0\}$$

Let *X* be a classical set, then a fuzzy subset *B* of X is called normal if there exists an $x \in X$ such that $B(x) = 1$. Otherwise *B* is said to be subnormal.

Definition 1.2(Zadeh, 1965): An α-level set of a fuzzy set *A* of *X* is a non-fuzzy set denoted by $[A]^{\alpha}$ and is defined by,

$$[A]^{\alpha} = \begin{cases} \{t \in X / A(t) \geq \alpha\}, & \alpha > 0 \\ cl(Supp\, A), & \alpha = 0 \end{cases}$$

where cl (supp*A*) denotes the closure of the support of *A*.

In many situations people are only able to characterize numeric information imprecisely. For example, people use terms such as, about 50, close to zero or greater than 100. These are

examples of *fuzzy numbers*. Using the theory of fuzzy subsets we can represent these fuzzy numbers as fuzzy subsets of the set of real numbers. More exactly, a fuzzy number *A* is a fuzzy set of the real line with a normal, (fuzzy) convex and continuous membership function of bounded support. The family of fuzzy numbers is denoted by *F*.

Definition 1.3 A fuzzy set A is called triangular fuzzy number with peak (or center) a, left width $\alpha > 0$ and right width $\beta > 0$ if its membership function has the following form

$$A(t) = \begin{cases} 1 - \dfrac{a-t}{\alpha}, & a - \alpha \le t \le a \\ 1 - \dfrac{t-a}{\beta}, & a \le t \le a + \beta \\ 0 & otherwise \end{cases} \qquad (1.8)$$

and is denoted by A = (a, α, β). (Zadeh, 1965)

Definition 1.4 A fuzzy set A is called trapezoidal fuzzy number with tolerance interval [a, b], left width α and right width β if its membership function has the following form

$$A(t) = \begin{cases} 1 - \dfrac{a-t}{\alpha}, & a - \alpha \le t \le a \\ 1, & a \le t \le b \\ 1 - \dfrac{t-b}{\beta}, & b \le t \le b + \beta \\ 0 & otherwise \end{cases} \qquad (1.9)$$

This is denoted by A = (a, b, α, β). The support of A is (a − α, b + β).

A trapezoidal fuzzy number may be seen as a fuzzy quantity "x is approximately in the interval $[a, b]$".(Francesco, 2010)

1.3.2 Crisp classification trees

This is a classification tree where the splitting value, x, is a crisp value. Suppose then a subset A of a set X can be defined by a function χ_A as a mapping from the elements of X to the elements of the set $\{0, 1\}$, $\chi_A: X \rightarrow \{0, 1\}$.

This mapping is represented as a set of ordered pairs, with one ordered pair present for each element of X. The first element of the ordered pair is an element of the set X, and the second element is an element of the set $\{0, 1\}$. The value zero is used to represent complete non-membership, and the value one is used to represent complete membership. The truth or falsity of the statement, "x is in A", is determined by the ordered pair $(x, \chi_A(x))$. The statement is true if the second element of the ordered pair is 1, and the statement is false if it is 0.

1.4 Statement of the problem

Work on crisp classification trees has been introduced as a nonparametric discriminant method. (Yu-Shin, 2004; Dobra, 2002). With the introduction of fuzzy logic, researchers started using fuzzy classification trees, which have the ability to deal with numeric, missing or inaccurate data (Dorokhov and Chernov,2011) . For example (Hashemi *et al*., 2008) applied fuzzy decision trees to classify data streams in the presence of noise. Zeinalkhani and Eftekhari (2011) presented a criteria for stopping fuzzy trees. According to Dorokhov and Chernov (201) fuzzy decision tree allows one to consider the uncertainty of estimations of decisions. Fuzzy trees may also supposes the use of qualitative presentations. Since the fuzzy classification trees are applied in cases where data is missing or inaccurate, it is important to compare whether their performance is

comprised. This may be done by comparing their performance to that of crisp classification trees. In this work performance of fuzzy classification trees and crisp classification trees are compared. This is done using Gini index, Pearson's Chi-squared statistic and gain ratio impurity measures. Probabilities of correct allocation were used to compare the performance.

1.5 Justification

The theory of fuzzy random variables has led to applications in fuzzy theory, especially in fuzzy classification trees. It is necessary to study under what conditions fuzzy classification trees perform better than crisp classification trees in terms of probabilities of misclassification. This will help researchers in deciding when to use fuzzy classification trees and when to use crisp classification trees.

1.6 Objectives

1.6.1 Main Objective

The main objective of this study is to compare the performance of fuzzy and crisp classification trees, using Gini Index, Chi-Square Statistics and Gain information. The study also compares Gini Index, Chi-Square Statistics and Gain Ratio.

1.6.2 Specific Objectives

i. To compare the performance of crisp classification trees with that of fuzzy classification trees, based on Gini Index, Chi-Squared Statistic, and Gain Ratio impurity measures, using probabilities of correct allocation and probabilities of misclassification.

ii. To compare the performance of Gini index, the chi-squared statistic and the gain ratio as impurity measures for fuzzy and crisp classification trees in terms of probabilities of misclassification.

1.7 Significance of the study

In this work fuzzy and crisp classification trees using different impurity measures was studied and compared. Classification is an area with varied applications, especially in data mining. The results will be useful for researchers in deciding when to use crisp or fuzzy classification trees. This work will also be useful in deciding which impurity measure to apply in different classification problems.

1.8 Outline of the thesis

The foregoing Chapter One has given the introduction to this work. Chapter Two is the literature review whereas, Chapter Three gives the methodology. Chapter Four applies Gini index, the Pearson's chi-squared statistic and gain ratio impurity measures to crisp and fuzzy classification trees. McNemar's test procedure is used in deciding if there is significant difference between the various trees. Chapter Five contains discussion, conclusion and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter an outline of previous work on classification is given. Classical classification and classification trees, both crisp and fuzzy, are presented. In section 2.2, related work on classical classification is given, in section 2.3, previous work on classification trees is given while fuzzy theory and fuzzy classification trees are outlined in section 2.4.

2.2 Classical classification

The most commonly used discriminant rule is the linear discriminant rule (LDA). This was proposed by R.A. Fisher in the 1950's. It assumes that the i^{th} population is normally distributed with mean $\underline{\mu}_i$ and variance- covariance matrix Σ, i=1, 2, …, g. This rule does well under the assumptions that:

i. The variances of the g populations are all equal
ii. The sample size is far much greater than the parameters describing the populations.
iii. The populations are normally distributed.

When the above assumptions are not met, then the LDA is not applicable.

When variances of the populations are not the same, quadratic discriminant procedures are applied (Wakaki,1992 and Ducinkas and Saltyte, 2001).

With the growth of interest in classification in artificial intelligence, problems are encountered where the dimension of the vector of parameters (feature vector) is greater than the sample size. As noted above, LDA is not applicable in such situations since the dispersion matrix is singular. Models that have been developed in such situations try to regularize the variance matrix, so that

the regularized matrix is not singular. An example of the regularizing of variance matrix Σ is done by replacing it with $\tilde{\tilde{\Sigma}}$ where $\tilde{\tilde{\Sigma}} = \lambda\Sigma + I_p$ (Hastie *et al.*,1995; Guo *et al.*, 2006).

Applications exist on discrimination when the sample size is close to the number of parameters describing the populations. For example (Tibshirani *et al.*, 2003) introduced a modified version of linear discriminant analysis, called the "nearest shrunken centroids" (NSC) and Guo *et al.*, (2006) generalized this idea to "shrunken centroids regularized discriminant analysis"(SCRDA) . Bayesian quadratic discriminant procedures have been applied in such situations where some prior information of the populations is available (Srivastava *et al.*, 2007).

Another model that has been used in discriminant analysis is the stepwise discriminant model. In this model, one begins by choosing the single best discriminating variable which is then paired with other variables one at a time until no more improvement on classification is possible (Qiu and Wu, 2005). An example where this model has been applied is in food science (Abdullah *et al*; 2001).

The major shortcoming of the above procedures is that they assume that the underlying populations are Gaussian. When it is not possible to assume that the populations are normally distributed, distribution free procedures are used. The most common ones are classification (decision) trees and the k- nearest neighbor.

2.3 Crisp Classification trees

Classification trees were studied extensively by Morgan and Sonqiust (1963). A classification tree is a nonparametric method of discriminant analysis. It portrays the discriminant problem in terms of a tree, usually a binary tree. The classification tree is a rule for assigning an object to a class based on the values of its variables. When the variable value has an exact decision point,

this is referred to as a crisp classification tree. The tree is constructed recursively by partitioning a learning sample of data. Each partition is represented by a node in the tree. The most popular approach of tree construction is to examine all possible binary splits of the data along each variable, and select the split that minimizes some node impurity (Morgan and Sonquist, 1963; and Breiman *et al*., 1984). There are many applications using crisp classification trees. For example, Venkatesan and Velmurugan (2015), used classification trees in breast cancer study.

2.4 Fuzzy classification trees

The classification trees described in section 2.3 are referred to as crisp classification trees. This is because they have sharp decision boundaries. Fuzzy sets were first studied by Zadeh (1965) to represent/manipulate data and information possessing uncertainties or vagueness. In general, vagueness is associated with the difficulty of making sharp or precise distinctions between alternatives Higashi and Klir (1983).The theory of fuzzy logic provides a mathematical tool to capture the uncertainties associated with human cognitive processes, such as thinking and reasoning Yager *et al.(*1992). Due to the growing popularity of fuzzy theory, researchers have proposed to apply fuzzy theory in decision trees. Classification tress using fuzzy decision points are referred to as fuzzy (classification) decision trees.

Fuzzy trees have been applied in a number of cases. Hashemi, *et al*. (2008) applied fuzzy decision trees for mining high speed data streams in computer science. Lien *et al*. (2011) applied fuzzy trees to study abnormal accessing of customer data for insurance cover in the customer database. Elyassami, *et al.* (2012) investigated fuzzy decision tree as a method for software effort estimation.

From the literature review, no such comparison has been done. In this study fuzzy decision trees are compared to crisp decision trees. The Gini Index, Chi-Square Statistic and Gain Ratio are used as the impurity measures and probabilities of correct allocation computed for comparison.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

In this chapter the theory of classification trees is presented. Different impurity measures used in this work are discussed.

3.2 Classification Trees

Denote by $P_{ij}d(x)$ the misclassification probability, that is, the probability of assigning $\underline{x}$ into class j when it actually belongs to class i using rule d(x). Given a classifier, that is a function d(x) defined on $\underline{X}$, the desire is to minimize $P_{ij}d(x)$. Since $P_{ij}d(x)$ is unknown, it is estimated using sample values, the rule that minimizes $P_{ij}d(x)$ is also the rule that minimizes the impurity of a node. A pure node contains only individuals from one class. Suppose there are g classes, let p(j) j=1, 2, . . . , g denote the probability of an individual belonging to the jth class. Let also p(j|t) denote the probability of an individual belonging to class j at node t. The rule that minimizes $P_{ij}d(x)$ is presented in terms of impurity measure. Different impurity measures are discussed in section 3.3.

3.2.1 Splitting a Node

The goal of splitting up a sample is to get sub-samples that are more homogenous than the original sample. We consider how homogenous the sub-samples are in relation to the original data. The perfect situation is when each sub-sample consists of instances that have the same value of the splitting attribute i.e. completely pure nodes. Splits can be done on nominal attributes as follows;

For nominal attributes the splits can be binary or n-ary in nature and are of the form {is X_m in $(c_1, c_{2,\ldots})$}.

A commonly used technique is to choose a split that will create the largest and purest child nodes by only looking at the instances in that node. This technique is referred to as the 'local optimization' approach. Its advantage is that it is computationally efficient regardless of the size of the problem.

3.2.2 Assigning Classes to Tree Nodes

Classification is done at the leaf nodes. Every leaf node in a tree is assigned to a particular class. The class for a leaf node is usually determined by a simple majority. In a given node, the class attached to it will be the class that is most well represented by the instances in the node.

Each leaf node will have an error rate, say e_i which is the proportion of misclassified instances in it. The probability that a particular classification will be correct is then simply $1\text{-}e_i$. The probability of a correct prediction from the model is then the weighted average of these probabilities from each leaf.

3.3 Impurity Measures

Definition 3.1

Let p(i) be the probability that an individual belongs to class Π_i. An impurity function is a function $\phi(p(i))$ defined on p(1), p(2), . . . , p(g) satisfying $p(j) \geq 0$, i= 1, 2, . . . ,g, and $\Sigma_j p(i) = 1$ with the properties:

i. $\phi(p(j))$ is a maximum only at the point $\left(\frac{1}{j}, \frac{1}{j}, \ldots, \frac{1}{j}\right)$

ii. $\phi(p(j))$ achieves its minimum only at the points(1, 0, 0, . . . 0), (0, 1, 0, . . . , 0), . . . ,(0, 0, . . . , 0, 1)

iii. $\phi(p(j))$ is a symmetric function of p(1), p(2), . . . , p(g). (Breiman *et al*., 1984)

When dealing with binary splits, it follows that $\phi(0) = \phi(1) = 0$ and $\phi(0.5)$ = maximum. However $\phi((p_1)) = \min (p(_1), 1\text{-}p(_1))$ does not sufficiently reward purer nodes. The class of node impurity functions which reward purer nodes is defined as the class ***F*** of functions, $\phi(p(_1))$, having continuous second derivatives on $0 \le p_1 \le 1$ and satisfying

$\phi(0) = \phi(1) = 0$

$\phi(p_1) = \phi(1\text{-}p(_1))$

$\phi''p_1)) < 0,\ 0 < p(_1) < 1$. (Yu-Shan, 1999)

The class ***F*** of functions is strictly a concave function.

Given an impurity function ϕ, of class ***F*** of functions, define the impurity measure i(t) at node t as

$$i(t) = \phi(p(1|t),\ldots,p(j|t)) \qquad (3.1)$$

where p(i|t) is the probability of an individual belonging to class i at node t.

In the case of binary splits, let a split S at node t send a proportion p_R of the data cases in t to t_R and proportion p_L to t_L, where t_R is the right subtree and t_L is the left subtree. The increase in impurity with respect to measure of impurity i(t) is defined as;

$\Delta_i (s, t) = i(t) - p_R i(t_R) - p_L i(t_L)$ where s is the split function. (3.2)

(Breiman *et al*., 1984)

The goodness of a split S at node t and impurity function ϕ is taken to be Δ_i (s, t).

3.3.1 The Gini Index

The Gini coefficient is derived from the Lorenz Curve (Ceriani and Verme, 2012). The Lorenz curve has been used extensively in studying income distribution. To plot a Lorenz curve, rank the observations from lowest income to highest income or the variable of interest. Then plot the cumulative proportion of the population on the horizontal axis and the cumulative proportion of the variable of interest on the vertical axis (Figure 3.1). The Gini coefficient compares this cumulative frequency curve to the uniform distribution that represents equality. In Figure 3.1 below, the diagonal line represents perfect equality. The greater the deviation of the Lorenz urve from this diagonal line, the greater the inequality. The Gini coefficient is defined as the ratio of the areas in the Lorenz curve. Let the area between the equality line and the Lorenz curve be denoted by A, and the area under the Lorenz curve by B. Then the Gini coefficient is given by A/(A+B). If the area of the rectangle is taken as 1, then A+B = 0.5. This implies that the Gini coefficient is 2A that is double the area between the equality line and the Lorenz curve.

This concept has been extended to classification theory. Suppose there are g classes (populations) which are to be classified into either left subnode or right subnode. Let the equality line represent the case when the right and left subnodes have the same distribution. The greater the impurity between the nodes, the greater the area between the equality line and the Lorenz curve (the Gini coefficient). This is used in the study as the Gini impurity measure. Therefore the greater the Gini index, the better the classification variable. The process is to calculate the Gini index using different variables and choose as a split variable that which maximizes the Gini index.

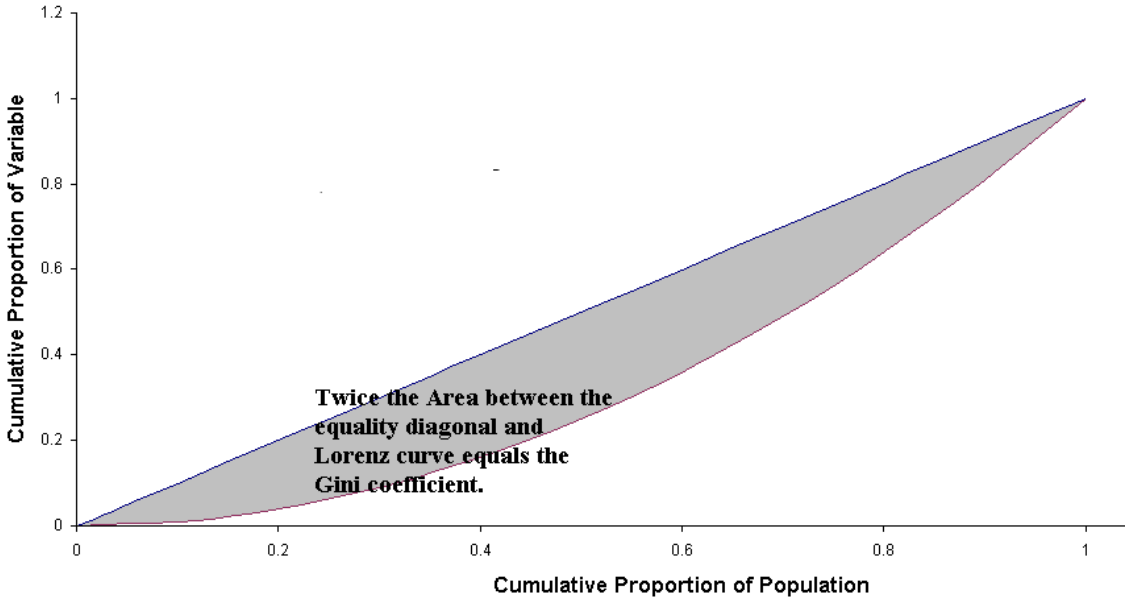

Figure 3.1: Lorenz Curve

Ceriani and Verme (2012).

Let N be the total number of individuals in all the classes, with N_i of these belonging to the i^{th} class, i = 1, 2, . . . , g. The Gini index (also known as Gini impurity measure), at node t denoted by G (t), is given by

$$G(t) = \sum_{i \neq j} p_i p_j$$

Where p_i is the probability that an individual belongs to the i^{th} class.

For binary splits when there are only two subnodes, the Gini impurity measure becomes;

$$
\begin{aligned}
G(t) &= \sum_{i=1}^{g} p_i(1-p_i) \\
&= \sum_{i=1}^{g} p_i - \sum_{i=1}^{g} p_i^2 \\
&= 1 - \sum_{i=1}^{g} p_i^2
\end{aligned}
\tag{3.3}
$$

This is commonly referred to as the Gini index.

Suppose the data is split using a splitting criteria, into two subnode t_1 and t_2 with sizes N_1 and N_2 respectively. The Gini index of the split data is given by;

$$
Gini_{(split)}(t) = \frac{N_1}{N} G(t_1) + \frac{N_2}{N} G(t_2) \tag{3.4}
$$

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

i. Choose a variable and a value and split the node. Calculate the Gin i_{split} (t). Repeat this over all possible variables at the node.

ii. Select the variable and the value with the least $Gini_{spli\ t}$ and use it for splitting. It is known as the splitting variable.

iii. Repeat this process at each node until splitting is completely done.

(Breiman *et al*., 1984)

3.3.2 Pearson's chi –squared statistic

Suppose there are g independent classes. Let N be the total number of individuals with N_i belonging to the i^{th} class, i = 1, 2, . . . , g. The desire is to classify an individual with characteristics X into one of the classes using classification trees.

After a set of splits have been decided, a goodness of fit test is performed to select the best split. A split that makes the two subnodes as pure as possible (most heterogeneous) is the best split. We may define the hypothesis as:

H_0: the population proportions are the same in the right and left subtrees after the split.

H_0 is the most homogeneous tree. However the best split gives the most heterogeneous tree. Therefore, the best split is the one in which Ho is rejected. Therefore a statistic that tests for the homogeneity of a g × 2 contingency table is a suitable test statistic.

The statistic;

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.5)$$

is used to for homogeneity in the contingency tables;

where $O_{ij} = N_{ij}$ are the observed frequencies and $E_{ij} = \frac{N_i N_j}{N}$ are the estimated expected frequencies.

The variable and value that maximizes the χ^2 is the one that gives the most heterogeneous split. Therefore this is the best value and variable to use in the splitting.

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

- Calculate the Pearson's Chi-Squared value among the child branches over all possible points for each variable X_h at each node. h = 1, 2, . . . , k
- Select the variable and the value of that variable with the maximum chi-squared statistic value, denoted by X_{h0} and use it for splitting.

Repeat this process at each node until splitting is completely done. The theoretical basis of the Chi-Square distribution in (3.5) is given in theorem 3.1.

Theorem 3.1

$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ Converges in distribution to the chi-square distribution.

Proof

A sequence of random variables $\{X_n\}$ is said to converge in distribution to a random variable X if

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} P\{X_n \le x\} \\ &= P(X \le x) \\ &= F(x) \end{aligned} \qquad (3.6)$$

at each continuity point x of F(x).

Let $\gamma_\alpha = \sum_{l=1}^{n} X_{l_\alpha}$ be the number of observations in the α^{th} class (the observed frequency in the α^{th} class). Let p_α be the probability of an individual belonging to class α. Then the expected number of individuals in class α is $n\, p_\alpha$.

Let

$$
\begin{aligned}
T_n^2 &= \sum_{\alpha=1}^{k} \frac{(\gamma_\alpha - np_\alpha)^2}{np_\alpha} \\
&= n\left[\sum_{\alpha=1}^{k} \frac{1}{p_\alpha}\left(\frac{\gamma_\alpha}{n} - p_\alpha\right)^2\right] \\
&= n\sum_{\alpha=1}^{k}\left[\frac{1}{\sqrt{p_\alpha}}\left\{\sqrt{n}\left(\frac{\gamma_\alpha}{n} - p_\alpha\right)\right\}\right]^2
\end{aligned}
\tag{3.7}
$$

Define $\sqrt{p_\alpha} = q_\alpha, \quad \alpha = 1,2,\ldots,k$

Let

$$
\Lambda_q = \begin{bmatrix} q_1 & \cdots & 0 \\ 0 & \ddots & \vdots \\ \vdots & \cdots & q_k \end{bmatrix}
$$

Then

$$
\Lambda^{-1}{}_q = \begin{bmatrix} \frac{1}{q_1} & \cdots & 0 \\ 0 & \ddots & \vdots \\ \vdots & \cdots & \frac{1}{q_k} \end{bmatrix} \qquad = \Lambda_{\frac{1}{q}}
\tag{3.8}
$$

Let $\varepsilon_{\alpha_n} = \sqrt{n}\left(\frac{\gamma_\alpha}{n} - p_\alpha\right)$

It follows that,

$$\underline{\varepsilon_n} = \begin{bmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{kn} \end{bmatrix} = \begin{bmatrix} \sqrt{n}\left(\frac{\gamma_1}{n} - p_1\right) \\ \sqrt{n}\left(\frac{\gamma_2}{n} - p_2\right) \\ \vdots \\ \sqrt{n}\left(\frac{\gamma_k}{n} - p_k\right) \end{bmatrix} \tag{3.9}$$

If we let

$$\underline{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix}$$

and define

$$\underline{\xi}_n = \Lambda_{\frac{1}{\underline{q}}} \underline{\varepsilon_n} = \begin{bmatrix} \frac{\varepsilon_{1n}}{q_1} \\ \frac{\varepsilon_{2n}}{q_2} \\ \vdots \\ \frac{\varepsilon_{kn}}{q_k} \end{bmatrix} \tag{3.10}$$

Then,

$$
\begin{aligned}
T_n^2 &= \sum_{\alpha=1}^{k}\left[\frac{\varepsilon_{\alpha n}}{q_\alpha}\right]^2 \\
&= \underline{\varepsilon}_n^{'}\,\underline{\varepsilon}_n \\
&= \left(\Lambda_{\frac{1}{q}}\underline{\varepsilon}_n\right)^{'}\left(\Lambda_{\frac{1}{q}}\underline{\varepsilon}_n\right) \\
&= \underline{\varepsilon}_n^{'}\Lambda_{\frac{1}{q}}^{'}\Lambda_{\frac{1}{q}}\underline{\varepsilon}_n
\end{aligned}
\qquad (3.11)
$$

$\underline{\varepsilon_n}$ is obtained as follows

$$
\begin{aligned}
\underline{\varepsilon_n} &= \begin{bmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{kn} \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{n}\left(\frac{\gamma_1}{n} - p_1\right) \\ \sqrt{n}\left(\frac{\gamma_2}{n} - p_2\right) \\ \vdots \\ \sqrt{n}\left(\frac{\gamma_k}{n} - p_k\right) \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{n}\left(\frac{1}{n}\sum_{l=1}^{n} x_{l1} - p_1\right) \\ \sqrt{n}\left(\frac{1}{n}\sum_{l=1}^{n} x_{l2} - p_2\right) \\ \vdots \\ \sqrt{n}\left(\frac{1}{n}\sum_{l=1}^{n} x_{lk} - p_k\right) \end{bmatrix} \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{l=1}^{n}\begin{bmatrix} x_{l1} \\ x_{l2} \\ \vdots \\ x_{lk} \end{bmatrix} - \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}\right]
\end{aligned}
\qquad (3.12)
$$

Now $\underline{x}_1^{'} = \left(x_{l1}, x_{l2}; \cdots x_{l\alpha}, \cdots x_{lk}\right)$

$$\Pr\{x_{l\alpha}=1\}=p_\alpha,\ \Pr\{x_{l\alpha}=0\}=1-p_\alpha$$

$$E(x_{l\alpha})=p_\alpha \quad E(x_{l\beta}^2)=p_\alpha \quad Var(x_{l\alpha})=p_\alpha(1-p_\alpha)$$

For

$$\alpha\neq\beta, \quad Cov(x_{l\alpha}\ x_{l\beta})=E(x_{l\alpha}\ x_{l\beta})-p_\alpha p_\beta$$
$$=-p_\alpha p_\beta$$

Now

$$\underline{X_l}=\begin{bmatrix} x_{l1} \\ x_{l2} \\ \vdots \\ x_{l\alpha} \\ \vdots \\ x_{lk} \end{bmatrix} \qquad l=1,2,\cdots n$$

Are independently and identically distributed random variables with mean vector

$$\underline{\mu}=\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}$$

and covariance matrix

$$\Sigma=\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 & \cdots & -p_2p_k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & -p_{k-1}p_k & p_k(1-p_k) \end{bmatrix} \qquad (3.13)$$

By the Central Limit Theorem, $L(\underline{\varepsilon}_n) \to L(\underline{\eta})$ where $\underline{\eta}$ has singular normal distribution with mean vector $\underline{0}$ and singular covariance matrix Σ as defined above.

That is $\underline{\eta}$ is degenerate $N_k(0,\Sigma)$. It then follows that $L(\underline{\xi}_n) = L\left(\Lambda_{\frac{1}{q}} \underline{\varepsilon}\right) \to L\left(\Lambda_{\frac{1}{q}} \underline{\eta}\right) = L(\underline{V})$

Where $\underline{V} = \Lambda_{\frac{1}{q}} \underline{\eta}$ is degenerate $N_k\left(0, \Lambda_{\frac{1}{q}} \Sigma \Lambda_{\frac{1}{q}}\right)$.

Therefore,

$L(T_n^2) = L(\underline{\xi}_n' \underline{\xi}_n) \to L(\underline{V}' \underline{V})$ Where $\underline{V}$is degenerate $N_k\left(0, \Lambda_{\frac{1}{q}} \Sigma \Lambda_{\frac{1}{q}}\right)$.

Then,

$$L(T_n^2) \to L(V'\underline{V}) = L\left(\sum_{i=1}^{k} \lambda_i z_i^2\right)$$

(3.14)

Where $z_1, z_2, \ldots, z_k$ are independent normal random variables and $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the characteristic roots of $\left(\Lambda_{\frac{1}{q}} \Sigma \Lambda_{\frac{1}{q}}\right)$. We assume that $\left(\Lambda_{\frac{1}{q}} \Sigma \Lambda_{\frac{1}{q}}\right)$ has characteristic root 1 with multiplicity k-1 and 0 with multiplicity 1. It follows that;

$L(T_n^2) \to L(\chi^2_{(k-1)})$.

3.3.3 Gain ratio

This impurity measure uses the information provided by the attribute. It represents the potential information generated by splitting data into n partitions. Information theory measures information content by use of bits. One bit of information is enough to answer a yes or no question. If the possible answers v_i have probabilities $P(v_i)$, then the information content I (or information gain) of the actual answer is given by;

$$\begin{aligned} \textit{Information Gain}(t) &= I\left(P(v_1), P(v_2), \ldots, P(v_n)\right) \\ &= \sum_{i=1}^{n} - P(v_i)\log_2 P(v_i) \end{aligned} \tag{3.15}$$

The higher the information gain, the better the splitting variable(Shannon and Weaver, 1964).

Suppose we have a set of possible events whose probabilities of occurrence are p_1, p_2, . .., p_n. These probabilities are known but that is all we know concerning which event will occur. The problem is to find a measure of how much "choice" is involved in the selection of the event. This is equivalent to finding how uncertain we are of the outcome. If such a measure exists, say $H(p_I, p_2,.,p_n)$, it has the following properties:

1. H should be continuous in the p_i.

2. If all the p_i are equal, $p_i = 1/n$ then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice can be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The theoretical basis of the distribution in (3.15) is given in theorem 3.2.

Theorem 3.2(Shannon and Weaver, 1964)

The only H satisfying the three assumptions above is of the form:

$$H = -K\sum_i p_i \log_2 p_i$$

Where K is a positive constant.

Proof

$$\text{Let } H\left(\frac{1}{n},\frac{1}{n},\ldots.\frac{1}{n}\right) = A(n)$$

From condition (3) above, we can decompose a choice from S^m equally likely possibilities to a series of m choices from S likely possibilities and obtain

$$A(S^m) = mA(S)$$

Similarly

$$A(t^n) = nA(t)$$

We can choose an n arbitrarily large and find an m to satisfy

$$S^m \le t^n \le S^{m+1}$$

Taking logarithms we get

$$m\log_2 S \le n\log_2 t \le (m+1)\log_2 S$$

Divide throughout by n $\log_2 S$ to obtain;

$$\frac{m}{n} \leq \frac{\log_2 t}{\log_2 S} \leq \frac{m}{n} + \frac{1}{n}$$

This is equivalent to

$$\left(\frac{m}{n} - \frac{\log_2 t}{\log_2 S} \right) < \varepsilon$$

Where ε is an arbitrary small number.

From the monotonic property of A(n), it follows that,

$$A(S^m) \leq A(t^n) \leq A(S^{m+1})$$

That is

$$mA(S) \leq nA(t) \leq (m+1)A(S)$$

Dividing throughout by nA(S) we obtain

$$\frac{m}{n} \leq \frac{A(t)}{A(S)} \leq \frac{m}{n} + \frac{1}{n}$$

Equivalently,

$$\left(\frac{m}{n} - \frac{A(t)}{A(S)} \right) < \varepsilon$$

Therefore,

$$\left(\frac{A(t)}{A(S)} - \frac{\log_2 t}{\log_2 S} \right) < 2\varepsilon$$

Let A(t) = -Klogt (where K is a positive constant chosen to satisfy condition 2).

Now suppose we have a choice from n possibilities with commeasurable probabilities

$p_i = \frac{n_i}{\sum n_i}$ where the n_i are integers. We can break down a choice from $\sum n_i$ possibilities into a choice from n possibilities with probabilities $p_1,..., p_n$. Using condition 3 again, we equate the total choice from $\sum n_i$ as computed by the two methods.

$$K \log_2 \sum n_i = H(p_1, p_2, \ldots, p_n) + K \sum p_i \log_2 n_i$$

This implies that,

$$\begin{aligned} H &= K\left(\sum p_i \log_2 \sum n_i - \sum p_i \log_2 n_i \right) \\ &= -K\left(\sum p_i \log_2 \frac{n_i}{\sum n_i} \right) \\ &= -K\left(\sum p_i \log_2 p_i \right) \end{aligned}$$

Without loss of generality, take K = 1 to obtain,

$$H = -\sum p_i \log_2 p_i$$

Attributes with many different values tend to split into many small classes. This gives the impression of a good split variable. To reduce this anomaly the information gain is divided by the intrinsic value, that is the amount of information contained in the value of the variable.

Let N_i be the number of individuals allocated to branch i and let N denote the total number of individuals. Then the intrinsic information at node t is given by;

$$\text{Intrinsic Information(t)} = \sum -\frac{N_i}{N}\log_2\frac{N_i}{N} \tag{3.16}$$

The gain ratio at node t is given by,

$$Gain\ Ratio(t) = \frac{Information\ Gain(t)}{Intrinsic\ Information(t)} \tag{3.17}$$

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

i. Work out the Information Gain among the child branches over all possible decision points for each variable X_j at each node.
ii. Select the variables and the values of these variables with Information Gain that is neither too small nor too large.
iii. Work out the Gain Ratio among the child branches selected in (ii) above

3.4 Testing for Differences in Performance among the Trees

To compare the performance of the trees, the McNemar's test procedure is used. The McNemar's test procedure is used to determine which of two classifiers, C_1 and C_2 say, has lower error rate (Lindgren, 1993). The procedure is as follows:

Run the two classifiers on data set and record the following information

n_{00} : the number of individuals correctly classified by both.

n_{11}: the number of individuals misclassified by both.

n_{01}: the number of individuals correctly classified by C_1 but misclassified by C_2.

n_{10}: the number of individuals correctly classified by C_2 but misclassified by C_1.

Define the statistic M as follows that,

$$M = \frac{\left(|n_{01} - n_{10}| - 1\right)^2}{n_{01} + n_{10}} \tag{3.18}$$

M has a chi-squared distribution with one degree of freedom. The test is to calculate M and conclude that the two classifiers are different at α level of significance if M is larger than the tabulated value of the Chi-Square distribution on one degree of freedom at α level.

CHAPTER FOUR

COMPARISON OF CRISP AND FUZZY CLASSIFICATION USING DIFFERENT IMPURITY MEASURES

4.1 Introduction

In this chapter, fuzzy and crisp classification trees were compared. The fuzzy tree was constructed using triangular membership function. The probabilities of correct allocation were then calculated. Comparison of the performance of fuzzy tree with crisp tree was carried out using simulated data and then applied to real data. The first set of observations was generated from two 3-variate normal populations with different mean vectors but a common dispersion matrix. The second set of observations was generated from three 4-variate normal populations with different mean vectors but a common dispersion matrix. Simulated data was obtained using R statistical package and implemented on Pentium IV running on Windows 7 environment. The two sets of real data that were used in the study came from UCI machine learning repository.

4.2 Results from 3-variate normal populations based on simulated data

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 trees were generated and tested at varied sample sizes. Samples of sizes 50, 100, 200, 500 and 1000 were used. The splitting variable and its value were obtained using Gini index, Pearson's Chi-square Statistic and the Gain Ratio. After the tree was created, the remaining data were used to test the tree's performance. This was done by calculating the

probabilities of correct allocation, P_{11} and P_{22} for both crisp and fuzzy classification trees. Table 4.1 shows the average probabilities of correct allocation using the three impurity measures that is, Gini Index, Chi-square Statistic and Information Gain.

Table 4.1: Probabilities of Correct Allocation for Gini Index, Pearson's Chi-squared Statistic and Gain Ratio

	Gini Index				Pearson's chi-squared statistic				Gain ratio			
Sample size	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}
50	0.825	0.893	0.822	0.892	0.829	0.893	0.822	0. 892	0.600	0.617	0.615	0.618
100	0.829	0. 895	0.823	0.894	0.831	0.897	0.823	0. 894	0.605	0.618	0.620	0.618
200	0.831	0. 897	0.830	0.896	0.831	0.898	0.826	0. 896	0.605	0.620	0.623	0.620
500	0.832	0. 897	0.826	0.895	0.832	0.898	0.827	0. 896	0.608	0.623	0.626	0.622
1000	0.834	0. 898	0.831	0.896	0.834	0.899	0.831	0. 897	0.610	0.623	0.626	0.622

From table 4.1 it was noted that the average probabilities of correct allocation using fuzzy trees are generally higher than the probabilities of correct allocation when crisp trees are used. This was true for the three impurity measures considered in the study. Correct probabilities calculated for the Gini Index and the Pearson's chi-squared statistic are comparable. However those for gain ratio are much lower than for the other two. Therefore gain ratio did not perform as well as the other two impurity measures for this type of data.

It was also noted that as the sample size increased, the probabilities of correct allocation marginally increased both for the crisp and fuzzy classification trees.

The proportion of times probabilities of correct allocation was higher when using crisp cut points than fuzzy decision points is given in table 4.2.

Table 4.2: Proportion of times crisp probabilities outperforms fuzzy probabilities

	Gini index		Pearson's chi-squared statistic		Gain ratio	
Sample size	P_{11}fuzzy< P_{11}crisp	P_{22}fuzzy< P_{22}crisp	P_{11}fuzzy< P_{11}crisp	P_{22}fuzzy< P_{22}crisp	P_{11}fuzzy< P_{11}crisp	P_{22}fuzzy< P_{22}crisp
50	0.074	0.079	0.073	0.080	0.075	0.080
100	0.013	0.027	0.015	0.029	0.020	0.027
200	0.001	0.002	0.002	0.005	0.003	0.005
500	0	0	0	0	0	0
1000	0	0	0	0	0	0

From table 4.2, we note that the proportion of times the crisp classification tree outperformed the fuzzy classification tree was very low. This is true for the three impurity measures. Therefore one can conclude that the fuzzy classification tree performs better than crisp classification tree for these data.

4.2.1 Testing for difference in performance of the impurity measures for simulated data

The values of n_{11}, n_{01} and $n_{10}$for the three impurity measures are given in Table 4.3, C_1 is taken as fuzzy classification tree and C_2 as crisp classification tree. The McNemar's value in equation 3.10 is calculated and compared to the tabulated chi-squared value at 95% confidence.

Table 4.3:McNemarsValues for Gini Index, Pearson's Chi-squared statistic and Gain Ratio

	n_{11}	n_{01}	n_{10}	M
Gini Index	800	518	215	124.43
Pearson's chi-squared statistic	805	1500	1221	28.40
Gain ratio	822	516	460	3.21

At 95% confidence level and p-value of 0.1056, $\chi^2_{1,0.95}$ = 3.84. Since both 28.40 and 124.43 are higher than 3.84, we conclude that the fuzzy tree performed better than the crisp tree when using the Gini Index and the Pearson's chi-squared statistic. When using the gain ratio, the fuzzy classification tree and crisp classification tree were found not to be significantly different since 3.21< 3.84. The gain ratio was not able to discriminate between the two types of trees.

4.2.2 Different Sample Sizes of population one and population two from the 3-variate normal populations simulated data

In this section we checked if the population proportions had any effect on the probabilities of correct allocation. This was done by using different sample sizes in both populations. Tables 4.4a and 4.4b give the results of both crisp classification tree and fuzzy classification tree.

Table 4.4a: Probability of correct allocation to population one for different combinations sample sizes

Samples from pop 1	Samples from pop 2																			
	10		20		30		40		50		60		70		80		90		100	
	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}	P_{11}^{crisp}	P_{11}^{fuzzy}
10	0.83	0.86	0.82	0.88	0.82	0.88	0.84	0.88	0.83	0.88	0.81	0.89	0.82	0.89	0.82	0.89	0.82	0.89	0.81	0.86
20	0.81	0.88	0.84	0.87	0.84	0.87	0.82	0.88	0.83	0.88	0.83	0.86	0.81	0.87	0.82	0.86	0.81	0.87	0.82	0.87
30	0.84	0.88	0.82	0.88	0.82	0.87	0.81	0.88	0.84	0.88	0.81	0.88	0.84	0.89	0.83	0.89	0.83	0.87	0.84	0.88
40	0.82	0.88	0.83	0.87	0.84	0.89	0.84	0.86	0.84	0.88	0.84	0.88	0.82	0.89	0.83	0.87	0.82	0.89	0.84	0.88
50	0.81	0.88	0.84	0.89	0.84	0.86	0.83	0.88	0.84	0.88	0.83	0.87	0.84	0.87	0.81	0.87	0.82	0.88	0.83	0.87
60	0.84	0.86	0.82	0.87	0.81	0.89	0.82	0.89	0.84	0.88	0.82	0.87	0.81	0.88	0.83	0.87	0.83	0.88	0.84	0.88
70	0.84	0.88	0.84	0.86	0.84	0.88	0.84	0.87	0.84	0.88	0.84	0.87	0.84	0.88	0.84	0.89	0.81	0.87	0.82	0.87
80	0.81	0.88	0.84	0.88	0.82	0.88	0.84	0.87	0.84	0.88	0.81	0.89	0.82	0.87	0.82	0.89	0.83	0.87	0.83	0.87
90	0.83	0.87	0.82	0.88	0.82	0.88	0.81	0.87	0.84	0.88	0.84	0.88	0.84	0.88	0.82	0.87	0.81	0.88	0.83	0.88
100	0.83	0.88	0.83	0.88	0.84	0.88	0.84	0.88	0.84	0.88	0.83	0.88	0.81	0.89	0.84	0.89	0.82	0.89	0.83	0.89

Table 4.4b: Probability of correct allocation to population two for different combinations of sample sizes

Samples from pop 1	Samples from pop 2																			
	10		20		30		40		50		60		70		80		90		100	
	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}	P_{22}^{crisp}	P_{22}^{fuzzy}
10	0.80	0.84	0.78	0.87	0.80	0.86	0.80	0.83	0.79	0.88	0.81	0.84	0.80	0.86	0.82	0.86	0.81	0.86	0.81	0.86
20	0.79	0.84	0.79	0.87	0.79	0.86	0.82	0.86	0.82	0.86	0.82	0.86	0.81	0.88	0.81	0.88	0.82	0.89	0.82	0.87
30	0.81	0.86	0.83	0.84	0.79	0.84	0.81	0.88	0.84	0.89	0.82	0.89	0.83	0.86	0.81	0.86	0.82	0.86	0.84	0.88
40	0.83	0.86	0.82	0.88	0.83	0.88	0.81	0.88	0.84	0.88	0.82	0.87	0.83	0.89	0.83	0.86	0.81	0.86	0.84	0.88
50	0.79	0.86	0.82	0.87	0.83	0.88	0.79	0.84	0.82	0.88	0.81	0.89	0.81	0.86	0.83	0.89	0.83	0.86	0.83	0.87
60	0.83	0.84	0.79	0.87	0.82	0.86	0.79	0.84	0.83	0.89	0.81	0.86	0.84	0.88	0.82	0.86	0.83	0.88	0.82	0.86
70	0.81	0.87	0.82	0.84	0.81	0.86	0.82	0.89	0.82	0.88	0.82	0.89	0.83	0.88	0.83	0.89	0.82	0.88	0.84	0.87
80	0.81	0.86	0.82	0.84	0.83	0.86	0.82	0.84	0.82	0.86	0.83	0.86	0.84	0.86	0.83	0.87	0.84	0.86	0.83	0.88
90	0.82	0.86	0.81	0.88	0.81	0.88	0.82	0.87	0.83	0.87	0.82	0.86	0.81	0.88	0.82	0.87	0.82	0.89	0.83	0.89
100	0.82	0.87	0.81	0.88	0.82	0.88	0.82	0.89	0.83	0.89	0.83	0.86	0.83	0.88	0.84	0.89	0.84	0.88	0.82	0.87

From tables 4.4a and 4.4b it is noted that there is no difference in the values of probabilities of correct allocation even with different sample sizes in the two populations. Hence one can conclude that for both crisp and fuzzy classification trees sample size difference does not affect the performance of either fuzzy or crisp classification trees.

4.3 Results from 3-variate normal populations based on real data

This set of data was for measurements of tortoise shells from UCI machine learning repository. The first population was from 500 female tortoise and the second population was from 500 male tortoise. X_1 denotes the shell length, X_2 the shell width and X_3 the shell height. All the measurements were in millimeters. It was assumed that the populations are normally distributed with different means but same dispersion matrix. From the data, the sample mean vectors and pooled sample variance matrix are given as:

$$\overline{\underline{X}_1} = \begin{bmatrix} 113.4 \\ 88.3 \\ 40.7 \end{bmatrix} \qquad \overline{\underline{X}_2} = \begin{bmatrix} 136.0 \\ 102.6 \\ 52.0 \end{bmatrix}$$

$$S_u = \begin{bmatrix} 282.8 & 167.9 & 98.8 \\ 167.9 & 106.2 & 59.9 \\ 98.8 & 59.9 & 37.3 \end{bmatrix}$$

Where

$\overline{\underline{X}_1}$ is the sample mean vector from population one (female tortoise)

$\overline{\underline{X}_2}$ is the sample mean vector from population two (male tortoise)

S_u is the pooled sample variance matrix from population one and population two.

The splitting value that was selected for each variable was the mean value of the combined samples. Both crisp and fuzzy classification trees were used. To generate the tree one-third of the instances were used and the rest two-thirds were used to test the tree performance. This was done by calculating the probabilities of correct allocation.

Shell Length

For fuzzy cut points, the peak is taken as the mean of the shell length, with left width one standard deviation and right width one standard deviation.

From the sample mean vectors and the combined sample variance matrix above, the mean shell length is 124.7mm and the variance of shell length is 282.8mm^2, giving a standard deviation of 16.8mm.

For fuzzy classification tree, any tortoise with shell length, x, less than 124.7mm was allocated to the left branch. If x is greater than 124.7mm, the tortoise was allocated to the left branch with probability p(x), where;

$$p(x)=\begin{cases} 1-\dfrac{x-124.7}{16.8} & 124.7 \le x \le 141.5 \\ 0, & x > 141.5 \end{cases}$$ (see definition 1.3)

For crisp classification trees, any tortoise with shell length of 124.7mm was allocated to left branch with probability one, otherwise it was allocated to the right branch.

Shell width

For fuzzy cut points, the peak is taken as the mean of the shell width, with left width one standard deviation and right width one standard deviation.

The mean shell width is 95.4mm and the variance of shell width is 106.2 mm, giving a standard deviation of 10.3mm.

For fuzzy classification tree, any tortoise with shell width, x, less than 95.4mm was allocated to the left branch. If x is greater than 95.4mm, the tortoise was allocated to the left branch with probability p(x), where;

$$p(x)=\begin{cases}1-\dfrac{x-95.4}{10.3}, & 95.4 \le x 105.7 \\ 0, & x > 105.7\end{cases}$$ (see definition 1.3)

For crisp classification trees, any tortoise with shell width of 95.4mm was allocated to left branch with probability one, otherwise it was allocated to the right branch.

Shell height

For fuzzy cut points, the peak is taken as the mean of the shell height, with left width one standard deviation and right width one standard deviation.

The mean shell height is 46.35mm and the variance of shell height is 37.3mm, giving a standard deviation of 6.1mm.

For fuzzy classification tree, any tortoise with shell height x, less than 46.35 mm was allocated to the left branch. If x is greater than 46.35mm, the tortoise was allocated to the left branch with probability p(x), where;

$$p(x)=\begin{cases}1-\dfrac{x-46.35}{6.1}, & 46.35 \le x 52.45 \\ 0, & x > 52.45\end{cases} \quad \text{(see definition 1.3)}$$

For crisp classification trees, any tortoise with shell height of 46.35mm was allocated to left branch with probability one, otherwise it was allocated to the right branch.

Using the above allocation criteria, the results in Table 4.5 were obtained.

Table4.5: Instances Individuals Are Allocated to Right and Left Branches

		Crisp results			**Fuzzy results**		
Sample from Population 1 (Female Tortoise)		Shell length	Shell width	Shell height	Shell length	Shell width	Shell height
	left	350	400	440	357	403	446
	right	150	100	60	143	97	54
	total	500	500	500	500	500	500
Sample from Population 2 (Male Tortoise)	left	100	130	54	88	127	54
	right	400	370	446	412	373	446
	total	500	500	500	500	500	500

Using equations 34, 3.10 and 3.17 together with the results in Table 4.5, the Gini index, Pearson's chi-squared statistic and the gain ratio were calculated. This was done for both fuzzy and crisp trees. The results are given Tables 4.6a, 4.6b and 4.6c.

Table 4.6a: Information Gain calculated Values

	Crisp values			Fuzzy values		
	Shell length	Shell width	Shell height	Shell length	Shell width	Shell height
Left branch	0.89	0.64	0.53	0.72	0.80	0.49
Right branch	0.64	0.75	0.49	0.82	0.73	0.49
Total	1.53	1.39	1.02	1.54	1.55	0.98

Table 4.6b:Intrinsic information calculated values

Intrinsic Information	Shell length	Shell width	Shell height
Crisp values	0.992	0.997	1.00
Fuzzy values	0.999	0.997	1.00

Table 4.6c: Gini Index, Pearson's Chi-Squared and Gain Ratio calculated values

	Crisp decision point values			Fuzzy decision point values		
	Shell length	Shell width	Shell height	Shell length	Shell width	Shell height
Gini index	0.360	0.348	0.215	0.353	0.305	0.193
Pearson's Chi-squared statistic	614.7	305.8	293.3	640	355	193
Gain ratio	1.54	1.39	1.02	1.55	1.12	0.98

For crisp decision boundaries, the Gini index for the shell height was 0.215mm (Table 4.6c) and was the least, therefore shell height was used as the splitting variable at 46.35mm. When using fuzzy decision boundaries, the Gini index for the shell height was least (0.193mm)(Table 4.6c),

therefore shell height was used as the splitting variable at about 46.35mm. For both crisp and fuzzy trees while applying the Gin index, shell height was used.

For Pearson's chi-square statistic, shell length has the maximum value both crisp and fuzzy trees. Therefore shell length at 124.7mm was used as the split variable when applying the Pearson's chi-square statistic.

The Gain ratio had maximum value for shell length at 124.7mm. Therefore shell length was used as the split variable.

In classifying allocate left branch to population one (female tortoise) and right branch to population two (male tortoise). Using this allocation rule, classification was done using the rest of the data and the probabilities of correct allocation were calculated by applying the Gini index, Pearson's chi-square statistic and the Gain ratio.

Table 4.7: Probabilities of correct allocation

	Gini Index		**Pearson's chi-squared statistic**		**Gain ratio**	
	P_{11}	P_{22}	P_{11}	P_{22}	P_{11}	P_{22}
Crisp	0.83	0.81	0.89	0.89	0.620	0.621
fuzzy	0.86	0.85	0.9	0.9	0.623	0.626

It was noted that the probabilities of correct allocation when applying the fuzzy classification were higher than the probabilities of correct allocation when applying crisp classification for all the impurity measures.

4.3.1 Testing for difference in performance of the impurity measures for real data

Table 4.8McNemars values for real data

	n_{11}	n_{01}	n_{10}	M
Gini Index	100	8	14	1.14
Pearson's chi-squared statistic	102	10	14	0.375
Gain ratio	105	3	7	0.43

It was found out (Table 4.8) that for the three impurity measures, the calculated M values were smaller than $\chi^2_{1,0.95}$ = 3.84. Therefore at 95% confidence and p-value of 0.112, we conclude that, there was no difference between the crisp classification tree and the fuzzy classification trees.

4.4 Results from 4-variate normal populations based on simulated data

This set of data was from three 4-variate normal populations with different mean vectors and common dispersion matrix. From each of the three 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 probabilities of correct allocation, P_{11}, P_{22} and P_{33} were calculated. This was done for the crisp and fuzzy decision points.

Table 4.9 gives the average probabilities of correct allocation for the different sample sizes using crisp and fuzzy decision points.

Table 4.9: Probabilities of Correct Allocation using simulated data

	Gini index						Pearson's chi-squared statistic						Gain Ratio					
Sample size	P_{11} crisp	$P_{11fuzzy}$	$P_{22crisp}$	P_{22} fuzzy	$P_{33crisp}$	$P_{33fruzzy}$	P_{11} crisp	$P_{11fuzzy}$	$P_{22crisp}$	P_{22} fuzzy	$P_{33crisp}$	$P_{33fruzzy}$	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	0.66	0.70	0.86	0.89	0.79	0.81	0.56	0.58	0.60	0.61	0.58	0.60
100	0.64	0.71	0.90	0.90	0.81	0.82	0.65	0.71	0.88	0.90	0.81	0.82	0.60	0.61	0. 60	062	0.59	0.60
200	0.67	0.72	0.92	0.91	0.79	0.84	0.67	0.72	0.91	0.91	0.79	0.84	0.60	0.61	0.61	0.63	0.59	0.61
500	0.69	0.74	0.93	0.92	0.85	0.86	0.69	0.74	0.93	0.90	0.85	0.85	0.61	0.63	0.62	0.63	0.60	0.63
1000	0.68	0.73	0.93	0.94	0.81	0.86	0.71	0.75	0.94	0.93	0.85	0.86	0.61	0.64	0.62	0.63	0.61	0.63

From table 4.9, the average probabilities of correct allocation are higher using fuzzy tree than when using crisp tree. One can conclude that fuzzy tree performed better than the crisp tree with applying Gini index and Pearson's chi-squared statistic impurity measures.

Table 4.10: McNemar's values for simulated data

	n_{11}	n_{01}	n_{10}	M
Gini Index	10003	1005	901	5.57
Pearson's chi-squared statistic	11006	1024	911	6.48
Gain ratio	11015	1015	993	0.22

4.4.1 Conclusion from simulated data

From the chi-squared Tables $\chi^2_{1,0.95}$ = 3.84 at 5% level of significance and p-value of 0.2110. The computed M values from table 4.10 for Gini index and Pearson's chi-squared statistic were greater than 3.84. We can therefore conclude that for the Gini index and Pearson's chi-squared statistic the fuzzy classification tree and crisp classification tree were significantly different. For the Gain ratio, the computed value (0.22) was less than 3.84. In this case we can conclude that, for the Gain ratio the fuzzy and crisp trees are not significantly different.

Table 4.11 gives the proportion of times probabilities of correct allocation which were higher using crisp cut points than when using fuzzy cut points.

Table 4.11 Proportion of times crisp probabilities outperforms fuzzy probabilities

	Gini index			Pearson's Chi-squared statistic			Gain ratio		
Sample size	P_{11}fuzzy $<P_{11}$crisp	P_{22}fuzzy $<P_{22}$crisp	P_{33}fuzzy $<P_{33}$crisp	P_{11}fuzzy $<P_{11}$crisp	P_{22}fuzzy $<P_{22}$crisp	P_{33}fuzzy $<P_{33}$crisp	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	0.095	0.100	0.100	0.100	0.108	0.110
100	0.050	0.060	0.068	0.050	0.055	0.062	0.060	0.068	0.075
200	0.010	0.015	0.020	0.009	0.012	0.012	0.015	0.016	0.020
500	0.001	0.003	0.008	0	0	0	0.005	0.008	0.009
100	0	0	0	0	0	0	0	0	0

Table 4.11 gives the proportion of times correct allocations for the crisp tree which were higher than those of fuzzy tree when applying the three impurity measures. We see that for all the impurity measures this proportion is close to zero. It is therefore reasonable to conclude, from table 4.11, that fuzzy tree for three populations with four variables performed better than the crisp tree for this type of simulated data.

4.5 Results from 4-variate normal populations based on real data

In this section the iris data from UCI machine learning repository was used. This data set consists of a sample of size 150, 50 form iris Setosa, 50 iris Virginica and 50 from iris Veriscolour. Since there are more than two populations, one split was not enough for classification. After the first split was done the process of selecting a split variable and the split value was repeated for next level of splitting. After the last split, probabilities of correct

allocations were calculated. At each level of splitting, the variable and value having minimum Gini split was used as the splitting variable and value.

Four measurements were recorded for each flower. These were petal length, petal width, sepal length and sepal width. Let X_1 denote sepal length, X_2 denote sepal width, X_3 denote petal length and X_4 denote petal width.

Assuming that the population variance is common for the three types of iris flowers, sample mean vectors and sample pooled variance matrix were calculated and used in the splitting process. The different variables (petal length, petal width, sepal length and sepal width) at different values were tested to determine which variable to use as the splitting variable and at what value. Crisp and fuzzy decision points were used.

The fuzzy values of the different variables were obtained using Definition 1.3 and $\alpha = \beta = 1$ standard deviation.

Splitting the data using the different variables and values gave the results displayed in Tables 4.12-4.15.

4.5.1 Splitting Variables Results

Table 4.12: Allocation of individuals using Sepal length

		Crisp results				Fuzzy results			
		Sample from Pop 1(Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total	Sample from Pop 1(Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total
At 5.02	Left	30	5	8	43	34	1	3	38
	Right	20	45	42	107	16	49	47	112
	Total	50	50	50	150	50	50	50	150
At 6.59	Left	45	25	42	112	50	30	45	125
	Right	5	25	8	38	0	20	5	25
	Total	50	50	50	150	50	50	50	150
At 5.94	Left	48	18	30	96	50	14	35	99
	Right	2	32	20	54	0	36	15	51
	Total	50	50	50	150	50	50	50	150

Table 4.13 Allocation of individuals using Sepal width

		Crisp results				Fuzzy results			
		Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total	Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total
At 3.44	Left	32	45	49	126	35	49	50	134
	Right	18	5	1	24	15	1	0	16
	Total	50	50	50	150	50	50	50	150
At 2.98	Left	15	33	44	92	11	36	44	91
	Right	35	17	6	58	39	14	6	59
	Total	50	50	50	150	50	50	50	150
At 2.77	Left	4	20	34	58	2	19	39	60
	Right	46	30	16	92	48	31	11	90
	Total	50	50	50	150	50	50	50	150

Table 4.14: Allocation of individuals using Petal length

		Crisp results				Fuzzy results			
		Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total	Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total
At 1.47	Left	35	2	3	40	36	0	0	36
	Right	15	48	47	110	14	50	50	114
	Total	50	50	50	150	50	50	50	150
At 5.55	Left	45	30	47	122	50	36	50	136
	Right	5	20	3	28	0	14	0	14
	Total	50	50	50	150	50	50	50	150
At 4.26	Left	48	4	31	83	50	0	31	81
	Right	2	46	19	67	0	50	19	69
	Total	50	50	50	150	50	50	50	150

Table 4.15: Allocation of individuals using Petal width

		Crisp results				Fuzzy results			
		Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total	Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total
At 0.27	Left	40	0	0	40	40	0	0	40
	Right	10	50	50	110	10	50	50	110
	Total	50	50	50	150	50	50	50	150
At 2.03	Left	48	30	47	125	50	32	50	132
	Right	2	20	3	25	0	18	0	18
	Total	50	50	50	150	50	50	50	150
At 1.33	Left	45	6	25	76	50	0	24	74
	Right	5	44	25	74	0	50	26	76
	Total	50	50	50	150	50	50	50	150

Table 4.16:Gini Index, Pearson's chi-squared and Gain ratio values at different points

			Gini Index		Pearson's chi-squared		Gain ratio	
	Variable	**Decision Point**	**crisp**	**fuzzy**	**crisp**	**fuzzy**	**crisp**	**fuzzy**
1	**Sepal Length**	5.02	0.585	0.506	62.9	121.4	2.98	2.47
		6.59	0.612	0.596	43.7	51.1	3.24	3.79
		5.94	0.429	0.537	61.6	77.9	2.80	2.63
2	**Sepal Width**	3.44	0.615	0.714	22.5	11.2	4.02	3.88
		2.98	0.558	0.559	48.7	55.3	2.88	2.72
		2.77	0.418	0.349	60.6	61	2.83	2.55
3	**Petal Length**	1.47	0.506	0.456	94.8	101.6	2.02	1.45
		5.55	0.363	0.598	30.7	28.4	3.91	3.50
		4.26	0.491	0.439	102.6	103.8	2.41	1.99
4	**Petal Width**	0.27	0.309	0.309	109.2	109.2	1.30	1.61
		2.03	0.655	0.576	27.1	38	3.80	2.94
		1.33	0.512	0.444	85	100	2.16	2.00

From Table 4.16, it is observed that the Gini index for petal width at 0.27 when using both crisp and fuzzy trees was the minimum (0.309). Therefore petal width at 0.27 was used for the first level of splitting. After first level split, left branch contains individuals from population one (iris Setosa) only. Individuals from population two (Iris Virginica) and population three (iris Veriscolour) were mixed in the right branch (Table 4.15). Therefore the right branch required further splitting, while the left branch did not require further splitting.

As with the Gini index, it was observed that petal width at 0.27 had the maximum Pearson's chi-squared value. Therefore petal width was used at the first level splitting at the point 0.27.

From table 4.16, we observe that the maximum gain ratio for the crisp tree is 4.02 and that of the fuzzy tree is 3.88. This is the Gain ratio for sepal width at decision point 3.44 for both crisp and fuzzy trees. Looking at table 4.13, we find that at point 3.44, both crisp and fuzzy trees allocation is almost all to the left sub-branch. Using Gain ratio at this point does not discriminate the different flowers but groups them in the same branch. Therefore sepal width cannot be used as the split variable at 3.44. We conclude that Gain ratio is not an appropriate impurity measure for this data. Second level splitting in this case was therefore not done since the first level splitting was not successful.

The process of selecting the splitting variable and variable value was repeated for the right branch so as to separate iris Virginica and Iris Veriscolour in the second level splitting. This was done by applying the Gini index and Pearson's chi-squared statistics impurity measures. Individuals from iris Virginica and Iris Veriscolour only were considered. We allocate Iris Setosa to left Branch. The second level splitting results are in tables 4.17 and 4.18.

Table 4.17: Second level allocation

		Crisp results				Fuzzy results			
		Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total	Sample from Pop 1 (Iris Setosa)	Sample from Pop 2 (Iris Virginica)	Sample from Pop 3 (Iris Veriscolour)	Total
Sepal length	Left	10	38	8	56	10	40	10	60
	Right	0	12	42	54	0	10	40	50
	Total	10	50	50	110	10	50	50	110
Sepal width	Left	0	20	25	45	0	19	26	45
	Right	10	30	25	65	10	31	24	65
	Total	10	50	50	110	10	50	50	110
Petal length	Left	10	5	2	17	10	5	0	15
	Right	0	45	48	93	0	45	50	95
	Total	10	50	50	110	10	50	50	110
Petal width	Left	10	8	42	63	10	5	45	63
	Right	0	42	8	47	0	45	5	47
	Total	10	50	50	110	10	50	50	110

The calculated values for Gini index and Pearson's chi-squared statistic split are given in Table 4.18.

Table 4.18: Gini Index, Pearson's chi-squared and Gain ratio values

	Gini Index		Pearson's Chi-squared statistic	
	Crisp	**Fuzzy**	**Crisp**	**Fuzzy**
Sepal length	0.419	0.418	46.6	45.5
Sepal width	0.429	0.562	8.5	9.64
Petal length	0.508	0.471	54	72
Petal width	0.227	0.228	59	75.8

Table 4.18 shows that the Gini index for the petal width is minimum (0.227 for crisp and 0.228 for fuzzy). Therefore the petal width at 1.68 was used as the splitting variable for the second level splitting. Using Pearson's chi-squared statistic, petal width at 1.68 was used as the splitting variable for the second level splitting. This is because maximum values for both crisp and fuzzy trees are at this point. These are the same variables and values used for splitting when applying the Gini index as the impurity measure. The tree generated when applying the Pearson's chi-squared statistic is the same one generated when using Gini index. The probabilities of correct allocation when using the Gini index and Pearson's chi-squared statistic were the same.

The probabilities of allocation were calculated from the terminal nodes. The probabilities for population one (iris setosa) were calculated after level one splitting. The probabilities of correct allocation for population two (iris virginica) and population three (iris veriscolour) were calculated after level two splitting. The results are given in table 4.19.

Table 4.19: Probabilities of allocation for the iris data

	P_{11}	P_{12}	P_{13}	P_{21}	P_{22}	P_{23}	P_{31}	P_{32}	P_{33}
crisp	0.80	0	0.20	0	0.84	0.16	0	0.14	0.86
fuzzy	0.82	0	0.18	0	0.84	0.16	0	0.10	0.90

These probabilities are for Gini Index Pearson's Chi-squared statistic since the two classification trees are the same.

4.5.2 Conclusion from real data

Again, the values of n_{11}, n_{01} and n_{10} described in subsection 1.3.1 were also noted. These values are:

n_{11}=20, n_{01} =10, n_{10} =10

Using n_{11}=20, n_{01} =10, n_{10} =10, M given by equation 3.10 was calculated and found to be 0.05

From the Chi-squared Tables, $\chi^2_{1,0.95}$ = 3.84. Since the computed value of 0.05 is less than the tabulated value, 3.84, we conclude that in this case crisp tree and fuzzy tree are not significantly different at 5% significant level.

From table 4.19, the probabilities of correct allocation p_{11}, p_{22} and p_{33} are slightly higher for the fuzzy classification tree than for crisp classification tree. However the difference is not significant as shown by the McNemar's test above. Therefore we may conclude that the performance of the crisp and fuzzy classification trees for iris flower data is the same.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMEDATIONS

5.1 Introduction

In this chapter, summary, conclusion from the study and recommendations are given. Section 5.2 gives the summary of the work. In section 5.3 conclusion is given whereas the recommendations are in section 5.4.

5.2 Summary

In this study, the performance of crisp and fuzzy classification trees was compared. In chapter two, a brief literature review was given. In chapter three normally distributed data was simulated using R. Impurity measures used in the study were discussed in chapter three also. The McNemars' statistic used for comparing classification procedures was also given in chapter three.

The results obtained after applying the impurity measures on the simulated and real data were given in chapter four. These results were also discussed in chapter four. Chapter five gives the conclusion and recommendations for further work.

5.3 Conclusion

In this work, the aim was to study the performance of fuzzy classification trees and compare them with crisp classification trees using different impurity measures. The impurity measures used in the study were Gini index, the chi-squared statistic and the gain ratio. As observed in the

study the probabilities of correct allocation using fuzzy classification trees were significantly higher at 5% than those of crisp classification trees for simulated data. This was true for the Gini index, the Pearson's chi-squared statistic and the Gain ratio. Therefore it can be concluded that the fuzzy classification tree performed better than the crisp classification tree for simulated data when applying the Gini index, Pearson's chi-squared statistic and the Gain ratio. In the same sections, it was observed that the probabilities of correct allocation for the simulated data were lower for the Gain ration compared to the other two impurity measures. It can therefore be concluded that the Gain ratio was the least appropriate impurity measure among the three impurity measures used in this study.

In the study, when real data was used the probabilities of correct allocation using fuzzy classification trees were higher than those of crisp classification trees. This was true when applying the Gini index and the Pearson's chi-squared statistic. This difference in probabilities was shown not to be significant at 5% significant level. Therefore it can be concluded that for real data the fuzzy and the crisp classification trees were not significantly different when Gini index and Pearson's chi-squared statistic were used. It was observed that the Gain ratio was unable to discriminate between the different populations. In this case the Gain ratio may not be useful as an impurity measure.

From tables 4.1, 4.7, 4.9 and 4.19 it was noted that all the trees generated using the Gini index, Pearson's the chi-squared statistic and the Gain ratio impurity measures had probabilities of correct allocation greater than 0.5. Due to this, all the trees performed better than a purely random tree(heuristic tree).

The trees generated using Gini index and the Pearson's chi-squared statistic were found not be significantly different for the data used in the study.

It was noted that trees generated using Gain ratio had lower probabilities of correct allocation than trees generated by the Gini index or Pearson's chi-squared statistic.

The probabilities of correct allocation were found not to differ even when the population proportions were different. It is therefore reasonable to conclude that population proportions does not affect the classification tree. Therefore both fuzzy and crisp classification trees may be used to classify cases where one population is dominant over others.

5.4 Recommendations

In this work, it was found out that fuzzy classification trees performed better than crisp classification trees when using probabilities of correct allocation.

In chapter one, one of the issues noted on classification is when to stop splitting. In this work this problem was not an objective of the study, therefore it was not worked on. It is the recommendation of this study that this should be done.

REFERENCES

Abdullah, A., Nureize, A. and **Watada**, J. (2001). A fuzzy regression approach to hierarchical evaluation model for oil palm grading. *Fuzzy Sets and Systems*. **98**: 500-503.

Altman, E.I. (1968). Financial ratios, Discriminant analysis and the prediction of cooperate bankruptcy. *The Journal of Finance*, **23:** 280-301.

Banaś. K, Jasiński. A, Banaś. AM, Gajda. M, Dyduch. G, Pawlicki. B and **Kwiatek. W.M.** (2007). Application of discriminant analysis in postrate cancer research. *American Chemical Society*.**79:** 6670-6674.

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

Chalikias, M, Kaimakamis, G., Adam, M. and **Karadimas, N.** (2009). Discriminant Analysis: A case study of war data set. *International Mathematical Forum*, **4:** 351-357.

Ceriani, L. and **Verme, P.** (2012).The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality***10**(3): 421-443.

Dobra, A.(2002). Classification and regression tree construction. Unpublished doctoral dissertation, Cornell University, New York.

Dorokhov,O. and **Vladimir Chernov,V**. (2011). Application of the fuzzy decision trees for the tasks of alternative choices: *Transport and Telecommunication, **12** (2): 4–10.*

Ducinkas, K.and **Saltyte, J**. (2001). Quadratic discriminant analysis of spatially correlated data, Nonlinear analysis: *Modelling and Control*, 6:15-18.

Elyassami, S. and **Idri, A**. (2012). Investigating effort prediction of software projects on the ISBSG dataset. *Computer and Knowledge.* **4** :1204.2404

Francesco, M. (2010).An Introduction to Fuzzy Sets and Systems. Notes from course by International School on Neural Nets on Computational Intelligence Methods for Data Analysis in Oncology Bioinformatics - IIASS, Vietrisul Mare (Italy) 24-29.

Guo ,Y, Hastie T. and **Tibshirani, R.** (2006*). R*egularized linear discriminant analysis and its applications in micro arrays. *Biostatistics Advance Access, Apil.7.*

Hashemi, S., Kangavari, M. and **Yang, Y.** (2008). Class specific fuzzy decision trees for mining high speed data streams. *Information Sciences,* **17**: 4271-4294

Hastie, T; Buja A. and **Tibshirani, R**. *(*1995*).* Penalized discriminant analysis. *Annals of Statistics, 23:* 73-102.

***Higashi*, M.** and **Klir, G.J**. (1983). On measure of fuzziness and *fuzzy* complements. *Information Sciences,* 5**1**:3427-329.

Lien, C.C., Ho, C.C. and **Tsai, Y.M**. (2011).Applying fuzzy decision tree to infer abnormal accessing of insurance customer data. *Fuzzy Systems and Knowledge*, **6:** 154-160.

Lindgren, L. (1993). Fuzzy Sets and Systems. *Science Direct.* ***115*** *(3), 477-483*

Maimon, O. and **Rokach, L**.(2010).*Data Mining and Knowledge Discovery Handbook.* Springer, US.

Morgan, J. N. and **Sonquist, J. A**. (1963). Problems in the analysis of survey data, *Journal of American Statistical Association* **58**: 415-434.

Pyryt, M.C. (2004). Using discriminant analysis to identify gifted children. *Psychology Sceience* **46**: 342-347.

Qiu, X. and **Wu, L.** *(2005).* Stepwise Nearest Neighbor Discriminant analysis *Journal of Machine Learning, **6**:* 466-474.

Shannon, E.C. and **Weaver,W**. (1964). *The Mathematical Theory of Communication.* The University of Illinois press, Urbana.

Srivastava, S. Gupta,M. and **Frigyik,B.** *(2007).* Bayesian Quadratic Discrimnant analysis. *Journal of Machine Learning, 8:* 1277-1305.

Tibshirani, R., Hastie, T., Narasimhan, B. and **Chu,G.** (2003). Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science,* **18**(1): 104-117.

Venkatesan, E. and **Velmurugan, T**. (2015). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. *Indian Journal of Science and Technology*, **8**(29).

Wakaki, H. *(1992).* Discrininant analysis under elliptical populations. *Hiroshima math Journal.* ***24****:* 257-298.

Yager R. R. and **Zadeh, L. A.** (1992), *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, The Kluwer International Series in Engineering and Computer Science; SECS 165, Boston: Kluwer Academic Publishers.

Yu-Shan, S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing*, **9**: 309-315.

Yu-Shin,S. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis.* **45***: 457-466.*

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*,**8**(3): 338-353.

Zeinalkhani, M. and **Eftekhari,** M.(2011). A new measure for comparing stopping criteria of fuzzy decision tree. *Computer and Knowledge.* **3**: 304-309.

APPENDIX I

IRIS DATA SET

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Creator:

R.A. Fisher

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
```

```
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
```

```
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
```

APPENDIX II

A SAMPLE PROGRAM IN R FOR GENERATING AND CLASSIFYING DATA

{This code generates 1000 random values with mean vector mu and variance matrix sigma. First load the library MASS that implements the multivariate procedure.}

```
Library(MASS)

Sigma<-matrix (c(1.0, 0.5, 0.25, 0.5 ,1.25, 0.35, 0.25, 0.35, 0.45),3,3)

Mu<- c(10,20,30)

Mvn<-mvrnorm(n=1000,mu,sigma)


library(MASS)

sigma<-matrix (c(1.0, 0.5, 0.25, 0.5 ,1.25, 0.35, 0.25, 0.35, 0.45),3,3)

mu<-c(10,20,30)

Mvn<-mvrnorm(n=10,mu,sigma)

y<-round(mvn,1)

y

z<-y[,3]

z

length(z)

n=0

for(i in 1:length(z))

{

if (z[i] <30)

{if (z[i]<26) n=n else n=n+n/2}

else

{if (z[i]>35) n=n else n=n+n/2}

}
```

```
n



library(MASS)
sigma<-matrix (c(1, 0, 0, 0 ,1, 0, 0, 0, 1),3,3)
mu<-c(10,20,30)
mvn<-mvrnorm(n=100,mu,sigma)
y<-round(mvn,1)
z<-y[,1]
n=0
n1=0
for(i in 1:length(z))
{
if (z[i]<11) n=n+1 else n =n
{if (z[i] <11)
n1=n1+1
else
{if (z[i]<12) n1=n1+(1-(z[i]-11)) else n1=n1}
}
}
n
n1
sigma<-matrix (c(1, 0, 0, 0 ,1, 0, 0, 0, 1),3,3)
mu<-c(12,20,30)
mvn2<-mvrnorm(n=100,mu,sigma)
x<-round(mvn2,1)
m <- x[,1]
```

```
k=0

k1=0

for(i in 1:length(m))

{

if (m[i]>11) k=k+1 else k =k

{if (m[i] >11)

k1=k1+1

else

{if (m[i]>10) k1=k1+(1-(11-m[i])) else k1=k1}

}
```