A FUZZY HYPERCUBE MODEL FOR CLASSIFICATION PROBLEMS

GEOFFREY ONKUNDI BARINI
MC400-0002/2015

A THESIS SUBMITTED TO PAN AFRICAN UNIVERSITY INSTITUTE FOR BASIC SCIENCES, TECHNOLOGY AND INNOVATION IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN MATHEMATICS (COMPUTATIONAL OPTION)

2018

DECLARATION

This research is my original work and has not been presented for a degree in any other University.

Signature..... **Date**.....
Geoffrey Onkundi Barini

This thesis has been submitted for examination with our approval as University Supervisors.

Signature..... **Date**.....

Professor Livingstone M. Ngoo
Multimedia University of Kenya, Kenya

Signature..... **Date**.....

Professor Ronald M. Waweru
Jomo Kenyatta University of Agriculture and Technology, Kenya.

DEDICATION

This thesis is dedicated to my dear mother, Esther, for her constant encouragement and prayers.

ACKNOWLEDGMENT

Admittedly, undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Special mention goes to my enthusiastic supervisors, Prof. Livingstone Ngoo and Prof Ronald Waweru for the tremendous academic support and encouragement they gave me. Without their guidance and constant feedback this PhD would not have been achievable.

I thank Prof. George Orwa, and Prof. Mathew Kinyanjui wholeheartedly, not only for their persistent mentorship, but also for their constant faith in my research ability.

Similar profound gratitude goes to mum and my siblings for almost unbelievable support. They are the most important people in my world.

Finally I thank my God, my good Father, for letting me through all the difficulties. I have experienced Your guidance day by day. You are the one who let me finish my degree. I will keep on trusting You for my future. Thank you, Lord.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

ABSTRACT

Classification problems are naturally characterized by uncertainty, subjectivity, imprecision and ambiguity. Thus, in designing classification models, mathematical methods that are able to satisfactorily deal with uncertainty, ambiguity and subjectivity are essential. Although fuzzy set theory is a very convenient mathematical tool for treating vagueness and ambiguity due to redundant and irrelevant features, and poor class definition, the existing fuzzy entropy based techniques can not effectively deal with ambiguity. This thesis presents a geometrical fuzzy similarity classifier which allows us not only reduce complexity of classification problems by removing redundant and irrelevant features, but also estimate ambiguity in class assignment using measures of fuzzy specificity. The model was tested using 4 benchmark datasets from University of California Irvine (UCI) machine learning repository yielding very attractive results. With Dermatology data set, a mean classification accuracy of 98.21% was obtained with only 24 features as compared with 97.82% with 34 features. Uncertainty associated with class assignment is also reported. Miss-classified samples display high average uncertainty as compared to those correctly classified.

CHAPTER ONE

INTRODUCTION

1.1 Overview

Unlike crisp sets where an element can be either a member of a set or not, fuzzy sets allow partial membership (Zedeh, 1965). The theory of fuzzy sets was first introduced by Zedeh (1965) as a generalization of the traditional logic. Since then, it has provided a fertile ground for useful researches in a number of fields. In particular, fuzzy mathematics has continued to play a leading role in modeling systems possessing non-statistical uncertainty (Zimmermann, 2011). Fuzzy mathematic has been successfully applied in areas such as: artificial intelligence, industrial control, expert systems, decision analysis, economics, medicine and many others (Zimmermann, 2011).

Classification problems lies at the heart of decision analysis (Duda et al., 2012). Basically, classification problems involve assigning entities to classes defined by essential features shared by these entities (Duda et al., 2012). In practical settings, these classes are characterized by vagueness due to lack of clear cut boundaries and thus can not be meaningfully represented using methods based on crisp logic. Fuzzy methods provide a convenient alternative mathematical framework for modeling problems characterized by this kind of uncertainty (Pedrycz, 1991).

In this chapter the basic concepts of the theory of fuzzy sets are discussed. Towards the end of the chapter, the statement of the research problem, objectives and justification are precisely stated.

1.2 Theoretical Background

Definition 1.2.1. (Zedeh, 1965). Let ζ represent a universe of discourse (finite or infinite). A fuzzy set A in ζ is a set that is characterized by a membership function,

$$m_A : \zeta \to [0,1] \tag{1.2.1}$$

which associates each $x_i \in \zeta$ the degree of membership $m_A(x_i) = a_i$ $(i = 1,2,\ldots,n)$ in A. Fuzzy sets allow partial membership as opposed to crisp sets whose membership values take 0 or 1. A constant fuzzy set A denoted $[a]$ is such $m_A(x_i) = a \forall x \in \zeta$.

Observe that the universe ζ is always a crisp set. For a finite universe ζ, A is expressed as,

$$A = \frac{m_A(x_1)}{x_1} + \frac{m_A(x_2)}{x_2} + \cdots + \frac{m_A(x_n)}{x_n} = \sum_{i=1}^{n} \frac{m_A(x_i)}{x_i} \tag{1.2.2}$$

If the universe is an inifinite set, then fuzzy set A on ζ is expressed as

$$A = \int_A \frac{m_A(x)}{x} \tag{1.2.3}$$

The empty set $\emptyset$ is such that $m_\emptyset(x) = 0$ for all x in ζ. Clearly for any $x \in \zeta$ we have $m_\zeta(x) = 1$. The sets of all crisp and fuzzy sets of ζ will be denoted by 2^ζ and $F\left(2^\zeta\right)$ respectively.

Definition 1.2.2. (Zimmermann, 2011). The support of a fuzzy set A constitute all elements of the universe ζ that have membership greater than zero in A

$$Support(A) = \{x \in \zeta | m_A(x) > 0\} \tag{1.2.4}$$

Definition 1.2.3. (Zimmermann, 2011). The crisp set

$$Support(A) = \{x \in \zeta | m_A(x) \geq \alpha\}, \alpha \in [0,1] \tag{1.2.5}$$

is called an alpha cut of A, denoted A_α.

Definition 1.2.4. (Zimmermann, 2011).The core of a fuzzy set A constitute all ele-

ments of the universe ζ that have full membership in A

$$Core\,(A) = \{x \in \zeta \,|m_A\,(x) = 1\} \tag{1.2.6}$$

Definition 1.2.5. (Zimmermann, 2011).The height of a fuzzy set A, $h\,(A)$ is the largest membership grade in A. If $h\,(A) = 1$, then A is said to normal.

Membership functions assigns to every element of ζ, membership values to fuzzy set A. Membership functions for fuzzy sets can be defined in any number of ways as long as they follow the rules of the definition of a fuzzy set. The shape of the membership function used defines the fuzzy set and so the decision on which type to use is dependant on the purpose. The membership function choice is the subjective aspect of fuzzy sets, it allows the desired values to be interpereted appropriately. The most common membership functions include triangular, trapezoidal, sigmoidal and gaussian functions. The following figures show examples of trapezoidal and triangular membership functions

$$m_A\,(x) = \begin{cases} 0, & \text{if} \quad x < 0 \\ \frac{x}{3}, & \text{if} \quad 0 \leq x \leq 3 \\ 1, & \text{if} \quad 3 \leq x \leq 9 \\ \frac{12-x}{3}, & \text{if} \quad 9 \leq x \leq 12 \\ 0, & \text{if} \quad x > 12 \end{cases} \tag{1.2.7}$$

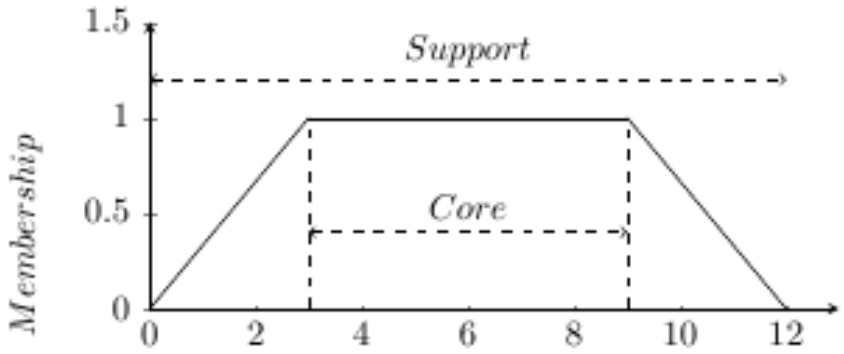

Figure 1.2.1: Trapezoidal membership function

$$m_A(x) = \begin{cases} 0, & \text{if} \quad x < 0 \\ \frac{x}{4}, & \text{if} \quad 0 \leq x \leq 4 \\ \frac{8-x}{4}, & \text{if} \quad 4 \leq x \leq 8 \\ 0, & \text{if} \quad x > 8 \end{cases} \tag{1.2.8}$$

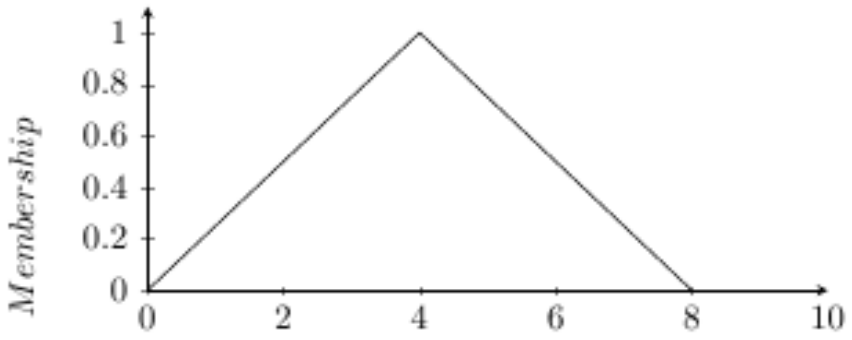

Figure 1.2.2: Triangular membership function

1.2.1 Geometry of Fuzzy Sets

Kosko (1990), introduced a very useful geometrical representation of fuzzy sets. He viewed a fuzzy set $A = ((m_A(x_1), a_1), (m_A(x_2), a_2), \ldots, (m_A(x_n), a_n))$ as a point $(a_1, a_2, \ldots, a_n)$

or fit vector in an n-dimensional fuzzy unit hypercube $[0,1] \times [0,1] \times \cdots \times [0,1] = I^n$. The fit value $m_A(x_i) = a_i$ of A corresponds to the membership value in th i^{th} dimension. Vertices of I^n are the non-fuzzy sets. Thus all the 2^n crisp sets is the power set 2^{ζ}. On the other hand, all fuzzy sets occupy the surface and interior of I^n. Hypercubes of dimensions $1,2$ and 3 are shown in figure (1.2.3) below

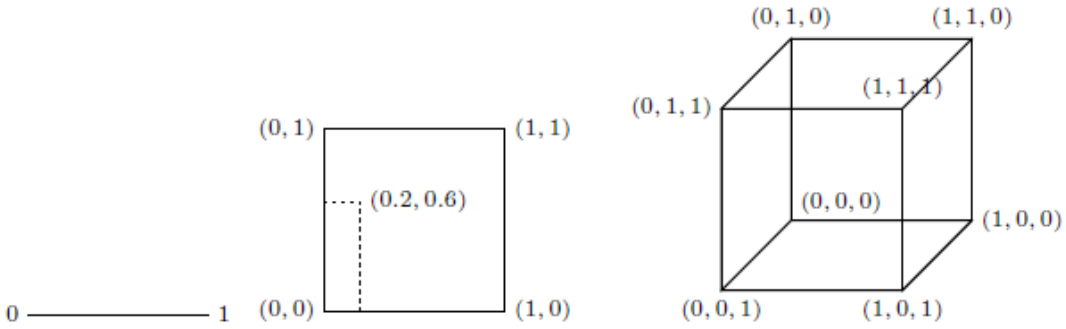

Figure 1.2.3: Fuzzy hypercubes of dimensions n=1,2 and 3 respectively

The centre of the hypercube is the constant fuzzy set $\left[\frac{1}{2}\right]$. The origin $(0,0,0,\ldots,0)$ and the point $(1,1,1,\ldots,1)$ correspond respectively, to the empty set $\emptyset$ and the universal set ζ.

1.2.2 Operations on Fuzzy Sets

The operations on fuzzy sets are extension of the most commonly used operations on crisp sets. This extension imposes a prime condition that all the fuzzy operations which are extensions of crisp concepts must reduce to their usual meaning when the fuzzy sets reduce themselves to ordinary sets, that is, when they have only 0 and 1 as membership values (Zimmermann, 2011).

Definition 1.2.6. (Zedeh, 1965). The complement $\overline{A}$ of A is fuzzy set with membership function

$$m_{\overline{A}}(x) = 1 - m_A(x) \tag{1.2.9}$$

Example 1.2.1. If $A = (0.4,1,0.7,0.5)$, then $\overline{A} = (0.6,0,0.3,0.5)$

Definition 1.2.7. (Zedeh, 1965). The membership function of the intersection of fuzzy sets A and B is given by:

$$m_{A\cap B}(x) = \text{Min}\left(m_A(x), m_B(x)\right) \forall x \in \zeta \tag{1.2.10}$$

Definition 1.2.8. (Zedeh, 1965). The membership function of the union of fuzzy sets A and B is given by:

$$m_{A\cup B}(x) = \text{Max}\left(m_A(x), m_B(x)\right) \forall x \in \zeta \tag{1.2.11}$$

Example 1.2.2. If $A = (0.4, 1, 0.7, 0.5)$ and $B = (0.1, 0, 1, 0.9)$, then $A \cap B = (0.1, 0, 0.7, 0.5)$ and $A \cup B = (0.4, 1, 1, 0.9)$

Definition 1.2.9. (Zedeh, 1965). Two fuzzy sets A and B are said to be equal if and only if
$m_A(x) = m_B(x) \forall x \in \zeta$

Definition 1.2.10. (Zedeh, 1965). The cardinality of A is a non-negative real scalar

$$c(A) = \sum_{i=1}^{n} m_A(x_i) \tag{1.2.12}$$

In example(1.2.1) we have $c(A) = 0.4 + 1 + 0.7 + 0.5 = 2.6$

Definition 1.2.11. (Weber, 1983). A T-norm is a bivalent function

$$T : [0,1] \times [0,1] \to [0,1] \tag{1.2.13}$$

satisfying the following $\forall x, y, z, w \in [0,1]$:

i. $T(0,0) = 0$

ii. $T(x,1) = x$

iii. $T(x,y) = T(y,x)$

iv. $T(x,y) \leq T(w,z)$ if $x \leq w$ and $y \leq z$

v. $T(x,T(y,z)) = T(T(x,y),z)$

This definition allows to combine two membership functions to find the membership function of $A \cap B$. For the union $A \cup B$, we have correspondingly the definition of the T-conorm of S-norm as

Definition 1.2.12. (Weber, 1983). A T_c-conorm is a bivalent function

$$T_c : [0,1] \times [0,1] \to [0,1] \tag{1.2.14}$$

satisfying the following $\forall x,y,z,w \in [0,1]$:

i. $T_c(1,1) = 1$

ii. $T_c(x,0) = x$

iii. $T_c(x,y) = T_c(y,x)$

iv. $T_c(x,y) \leq T_c(w,z)$ if $x \leq w$ and $y \leq z$

v. $T_c(x,T_c(y,z)) = T_c(T_c(x,y),z)$

From these definitions we have:

$$m_{A\cap B}(x) = T(m_A(x), m_B(x)) \tag{1.2.15}$$

and

$$m_{A\cup B}(x) = T_c(m_A(x), m_B(x)) \tag{1.2.16}$$

Definition 1.2.13. (Zedeh, 1965). Fuzzy set A is said to be a subset of fuzzy set B if and only if $m_A(x) \leq m_B(x) \, \forall x \in \zeta$.

Definition 1.2.14. (Boicescu et al., 1991). Lukasiewicz t-norm or conjunction has the form

$$x \odot y = \max\{x + y - 1, 0\} \tag{1.2.17}$$

Definition 1.2.15. (Boicescu et al., 1991). The normal Lukasiewicz structure is defined by

$$x \odot y = \max\{x + y - 1, 0\}, x \to y = \max\{1, 1 - x + y\} \tag{1.2.18}$$

Definition 1.2.16. (Boicescu et al., 1991; Klement et al., 2003). In the normal Lukasiewicz structure the equivalence relation $x \leftrightarrow y$ is defined as

$$x \leftrightarrow y = 1 - |x - y| \tag{1.2.19}$$

Definition 1.2.17. (Boicescu et al., 1991; Klement et al., 2003). In the generalized Lukasiewicz struacture the above equivalence relation becomes

$$x \leftrightarrow y = (1 - |x^p - y^p|)^{\frac{1}{p}}, p > 0 \tag{1.2.20}$$

Definition 1.2.18. (Zadeh, 1975). A linguistic variable is a variable whose values are words that we use in every day communication. For instance, temperature is a lingustic variable whose values are *warm* and $cold$.

1.2.3 Fuzzy Logic

Fuzzy logic is a form of multi-valued logic derived from fuzzy set theory to deal with reasoning that is approximate rather than precise in nature (Zedeh, 1965). Just as in fuzzy set theory the set membership values can range (inclusively) between 0 and 1, in fuzzy logic the degree of truth of a statement can range between 0 and 1 and is not constrained to the two truth values true, false as in classic logic. And when linguistic variables are used, these degrees may be managed by specific functions, as discussed

below.

Both fuzzy degrees of truth and probabilities range between 0 and 1 and hence may seem similar at first. However, they are distinct conceptually; fuzzy truth represents membership in vaguely defined sets, not likelihood of some event or condition as in probability theory. For example, if a 100-ml glass contains 30 ml of water, then, for two fuzzy sets, Empty and Full, one might define the glass as being 0.7 empty and 0.3 full. Note that the concept of emptiness would be subjective and thus would depend on the observer or designer. Another designer might equally well design a set membership function where the glass would be considered full for all values down to 50 ml. A probabilistic setting would first define a scalar variable for the fullness of the glass, and second, conditional distributions describing the probability that someone would call the glass full given a specific fullness level. Note that the conditioning can be achieved by having a specific observer that randomly selects the label for the glass, a distribution over deterministic observers, or both. While fuzzy logic avoids talking about randomness in this context, this simplification at the same time obscures what is exactly meant by the statement the 'glass is 0.3 full'.

A basic application might characterize subranges of a continuous variable. For instance, a temperature measurement for anti-lock brakes might have several separate membership functions defining particular temperature ranges needed to control the brakes properly. Each function maps the same temperature value to a truth value in the 0 to 1 range. These truth values can then be used to determine how the brakes should be controlled.

1.2.4 Subsethood Measure

A measure of subsethood $SB(A,B)$ gives the degree to which a fuzzy set A is a subset of set B. A real function $SB : F\left(2^{\zeta}\right) \times F\left(2^{\zeta}\right) \rightarrow [0,]$ is called a subsethood measure if it satisfies the following properties (Fan et al., 1999):

$(S1)$ $SB(A,B) = 1$ iff $m_A(x) \leq m_B(x) \forall x \in \zeta$

$(S2)$ If $A \subseteq \left[\frac{1}{2}\right]$, then $SB\left(A,\overline{A}\right) = 0$ iff $A = \zeta$.

$(S3)$ If $A \subseteq B \subseteq C$, then $SB(C,A) \leq SB(B,A)$ and $SB(C,A) \leq SB(C,B)$

1.2.5 Fuzzy Distance Measure

A measure of distance between two fuzzy sets is one of the fundamental concepts associated with fuzzy sets, both in theory and applications (Cheng, 1998; Bloch, 1999; Saha and Wehrli, 2004; Guha and Chakraborty, 2010).

Definition 1.2.19. Formally, the function $d : F\left(2^{\zeta}\right) \times F\left(2^{\zeta}\right) \to [0,+\infty)$ is fuzzy distance measure on $F\left(2^{\zeta}\right)$ if $\forall A,B,D \in F\left(2^{\zeta}\right)$ we have (Rosenfeld, 1985):

$(D1)$ $d(A,B) \geq 0$

$(D2)$ $d(A,B) = 0$ iff $A = B$

$(D3)$ $d(A,B) = d(B,A)$

$(D4)$ $d(A,B) \leq d(A,D) + d(D,B)$

While, there exist a variety of functions for distance measure convenient for various applications, the Minkowski class of metrics,

$$d_r(A,B) = \left(\sum_{i=1}^{n} |m_A(x_i) - m_B(x_i)|^r\right)^{\frac{1}{r}}, 1 \leq r \leq \infty \tag{1.2.21}$$

provides a prominent function for distance. For $r = 1$ and $r = 2$, the Minkowski class of metrics collapses to the fuzzy Hamming and the Euclidean distances,

$$d_1(A,B) = \sum_{i=1}^{n} |m_A(x_i) - m_B(x_i)|, \quad d_2(A,B) = \left(\sum_{i=1}^{n} |m_A(x_i) - m_B(x_i)|^2\right)^{\frac{1}{2}} \tag{1.2.22}$$

respectively. They are sometimes normalized as,

$$\widehat{d_1}(A,B) = \frac{1}{n}\sum_{i=1}^{n}|m_A(x_i) - m_B(x_i)|, \quad \widehat{d_2}(A,B) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}|m_A(x_i) - m_B(x_i)|^2\right)^{\frac{1}{2}} \tag{1.2.23}$$

The centre of the hypercube, $\left[\frac{1}{2}\right]$ is such that,

$$d_r\left(A, \left[\frac{1}{2}\right]\right) = d_r\left(A^c, \left[\frac{1}{2}\right]\right) = \left(\sum_{i=1}^{n}|m_A(x_i) - \frac{1}{2}|^r\right)^{\frac{1}{r}} \leq \frac{n^{\frac{1}{r}}}{2} \tag{1.2.24}$$

Distance measure between fuzzy sets gives us the ability to define a concept of fuzzy similarity, which plays an important role in comparing fuzzy concepts.

1.2.6 Fuzzy Similarity

Fuzzy similarity gives a quantitative measure of the resemblance between two fuzzy sets. Most prominent definitions of similarity are based on operations on fuzzy sets and distance between fuzzy sets. Fuzzy similarity is a dual concept of fuzzy distance measure.

Definition 1.2.20. A real valued function:

$$S : F\left(2^{\zeta}\right) \times F\left(2^{\zeta}\right) \to [0,1] \tag{1.2.25}$$

is called a measure of similarity on $F\left(2^{\zeta}\right)$ if it satisfies the following properties (Wang et al., 2008):

$\forall A,B,C \in F\left(2^{\zeta}\right)$

$(S1)$ $S(A,B) = S(B,A)$

$(S2)$ $0 \leq S(A,B) \leq 1$

$(S3)$ $S(A,B) = 1$ if and only if $A = B$

$(S4)$ if $A \subseteq B \subseteq C$ then $S(A,B) \geq S(A,C)$ and $S(B,C) \geq S(A,C)$

Examples of similarity measures based on set operations include (Baccour et al., 2014):

$$S_1(A,B) = \frac{c(A \cap B)}{c(A \cup B)} \tag{1.2.26}$$

$$S_2(A,B) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{min(m_A(x_i), m_B(x_i))}{max(m_A(x_i), m_B(x_i))}\right) \tag{1.2.27}$$

$$S_3(A,B) = \frac{c(\overline{A} \cap \overline{B})}{c(\overline{A} \cup \overline{B})} \tag{1.2.28}$$

$$S_4(A,B) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2min(m_A(x_i), m_B(x_i))}{m_A(x_i) + m_B(x_i)}\right) \tag{1.2.29}$$

Note that if the denominator equals zero in equations (1.1.27) and (1.1.29), then we trivially have $S_2 = S_4 = 1$. Distance based measures of similarity include (Baccour et al., 2014):

$$S_5(A,B) = 1 - \frac{\sum_{i=1}^{n}|m_A(x_i) - m_B(x_i)|}{n} \tag{1.2.30}$$

$$S_5(A,B) = \frac{1}{n}\sum_{i=1}^{n}(1 - |m_A(x_i) - m_B(x_i)|) \tag{1.2.31}$$

$$S_6(A,B) = 1 - \frac{1}{n^{\frac{1}{r}}}\left(\sum_{i=1}^{n}|m_A(x_i) - m_B(x_i)|^r\right)^{\frac{1}{r}} \tag{1.2.32}$$

A similarity measure based on the generalized Lakasiewicz structure has the form (Luukka and Leppälampi, 2006),

$$S_7(A,B) = \frac{1}{n}\left(\sum_{i=1}^{n}(1 - |(m_A(x_i))^p - (m_B(x_i))^p|)^{\frac{m}{p}}\right)^{\frac{1}{m}} \tag{1.2.33}$$

which is based on the equivalence relation,

$$m_A(x) \longleftrightarrow m_B(x) = 1 - |m_A(x) - m_B(x)| \tag{1.2.34}$$

The parameters p and m come from the generalized Lukasiewicz structure and and generalized mean respectively.

Definition 1.2.21. (Qing and Li, 2004). For any $A \in F\left(2^{\zeta}\right)$, the nearest A_{near} and farthest A_{far} crisp sets are such that

$$A_{near} = \begin{cases} 1, & if\ m_A(x) \geq \frac{1}{2} \\ 0, & if\ m_A(x) < \frac{1}{2} \end{cases}, \quad A_{far} = \begin{cases} 0, & if\ m_A(x) \geq \frac{1}{2} \\ 1, & if\ m_A(x) < \frac{1}{2} \end{cases} \tag{1.2.35}$$

Definition 1.2.22. (De Luca and Termini, 1972). A fuzzy set A^* is said to be a sharpened form of A if $m_A(x) \geq \frac{1}{2}$, then $m_{A^*}(x) \geq m_A(x)$ and if $m_A(x) \leq \frac{1}{2}$, then $m_{A^*}(x) \leq m_A(x)$.

1.3 Measures of Fuzzy Uncertainty

Uncertainty may refer to vague, doubtful or ambiguous (Klir, 1987). Two main forms of uncertainty are easily recognizable in fuzzy sets: ambiguity and fuzziness (or vagueness)(Klir, 1987; Zimmermann, 2011). Ambiguity refers to one-to-many relations, that is, situations with more than two or more alternatives that are left unspecified (Klir, 1987; Zimmermann, 2011). Fuzziness on the other hand refers to lack of sharp or precise boundaries (Klir, 1987; Zimmermann, 2011)

1.3.1 Fuzziness and fuzzy entropy

Fuzziness refers to the amount of difficult involved deciding whether an element belongs to a fuzzy set or not. Fuzziness is maximal if membership value of each element

is $\frac{1}{2}$ and zero for a crisp set as there exist no uncertainty as to whether or not an element belongs to the set. It is measured using fuzzy entropy measure (De Luca and Termini, 1972; Zadeh, 1968).

In 1972, Deluca and Termin introduced the now widely adopted axiomatic definition of fuzziness (De Luca and Termini, 1972). In their view, fuzziness of set A can be measured by a set to point mapping (De Luca and Termini, 1972).

$$E : F\left(2^{\zeta}\right) \rightarrow [0,1] \tag{1.3.1}$$

given by

$$E(A) = -k\left(m_A(x_i)\log m_A(x_i) + (1 - m_A(x_i))\log(1 - m_A(x_i))\right) \tag{1.3.2}$$

where $k \geq 0$ is a normalizing constant. Equation (1.3.2) is a measure of fuzzy entropy if it satisfies the following axioms:

$(A1)$ $E(A) = 0 \forall A \in 2^{\zeta}$

$(A2)$ $E(A)$ attains its maximum at the centre of the fuzzy hypercube, $[\frac{1}{2}]$

$(A3)$ if A^* is a sharpened version of any $A \in F\left(2^{\zeta}\right)$ then $E(A^*) \leq E(A)$

$(A4)$ For any $A \in F\left(2^{\zeta}\right)$, $E(A) = E\left(\overline{A}\right)$

Several researchers have successfully formulated new fuzzy entropy measures satisfying the above postulates. They include:

Kaufman (1975):

$$E_{ka}(A) = \frac{2}{n^{\frac{1}{r}}}\left(\sum_{i=1}^{n}|m_A(x_i) - m_{A_{near}}(x_i)|^r\right)^{\frac{1}{r}} \tag{1.3.3}$$

Yager (1979):

$$E_y(A) = 1 - \frac{\gamma_r\left(A,\overline{A}\right)}{n^{\frac{1}{r}}} \tag{1.3.4}$$

(Kosko, 1990):

$$E_{k1}(A) = \frac{d_r(A, A_{near})}{d_r(A, A_{far})} \tag{1.3.5}$$

Kosko (1990):

$$E_{k2}(A) = \frac{c\left(A \cap \overline{A}\right)}{c\left(A \cup \overline{A}\right)} \tag{1.3.6}$$

1.3.2 Measures of Specificity

Lack of ambiguity in a fuzzy set is measured using specificity(Klir, 1987; Yager, 1982). Specificity is a quantitative measure of the extent to which a fuzzy set restricts a variable to a small number of membership values(Yager, 1982). It is a measure of exact knowledge we have regarding a system. Yager (1982), proposed the following definition for specificity of fuzzy sets.

Definition 1.3.1. Yager (1998) Specificity $Sp(A)$ of fuzzy set A is a mapping $Sp : F\left(2^{\zeta}\right) \to [0,1]$ satisfying the following axioms: $\forall A, B \in F\left(2^{\zeta}\right)$

$(S1)$ $Sp(\emptyset) = 0$

$(S2)$ $Sp(A) = 1$ iff A is a singleton set.

$(S3)$ $Sp(A)$ increases as the largest membership value in A increases and decreases as the non-maximal membership values increases.

In addition a measure of specificity is said to be regular if for any $A \in F\left(2^{\zeta}\right)$ such that $m_A(x_i) = a \forall i$ we have $Sp(A) = 0$, where a is constant membership value (Yager, 1998).

Common examples of measures of specificity include:

Yager (1982):

$$Sp(A) = \int_0^{h(A)} \frac{1}{c(A_\alpha)} d\alpha \tag{1.3.7}$$

Where A is a finite set and $h(A)$ is the height of A and A_α is an alpha cut of fuzzy set A. For normal fuzzy sets the above formula becomes

$$Sp(A) = \int_0^1 \frac{1}{c(A_\alpha)} d\alpha \tag{1.3.8}$$

Yager (1995):

$$Sp(A) = a_1 - \frac{1}{n-1} \sum_{i=2}^{n} a_i \tag{1.3.9}$$

where membership values of A be ordered such that $a_1 \geq a_2 \geq \cdots \geq a_n$. This measure is known as the linear measure of specificity.

Dubois and Prade (1985):

$$Sp(A) = \sum_{i=1}^{n} \frac{a_i - a_{i+1}}{i} \tag{1.3.10}$$

The complement of specificity is called non-specificity (Dubois and Prade, 2012) or anxiety, NSp, and is given by

$$NSp(A) = 1 - Sp(A) \tag{1.3.11}$$

1.4 Classfification Problems

Classification problem involves grouping of entities into a set of classes based on similarity among the entities (Duda et al., 2012). Similar entities are put in one class. Classification problem can be viewed as a task of approximating a mapping function f from input variables y to discrete output variables z.The input variables are called attributes or features while the output variables are often called labels, categories or classes. Features can discrete or continuous. Entities to classified are also called instances, examples or samples.

Classification problems are further divided into binary and multi-class classifica-

tion depending on the number of the underlying classes.Binary classification involves classifying entities into either of the two classes. Binary classification problems arise disease diagnosis (Huang et al., 2007; Nahar et al., 2013) and malware detection (Firdausi et al., 2010). Multi-class classification involves classifying the input entities into more than two classes. Multi-class classification problems arise in risk classification (Zabalgoitia et al., 1998), character recognition (Cecotti and Vajda, 2013), biometric identification and security (Bailey et al., 2014), face recognition (Parveen and Thuraisingham, 2006).

Traditional classification methods are based on crisp sets in which each entity can belong to only one class. Accordingly, in crisp classification, class membership is binary, that is, an entity is a member of a class or not. Crisp class membership values can be either 1 when the entity is a prototype of the class and 0 for all other classes. On the other hand, in fuzzy classification, an entity can have membership in many different classes to different degrees at the same time. Fuzzy classes are appropriate for continuous and imprecise data that does not fall neatly into discrete classes, such as data medical applications (Ali et al., 2011; Nauck and Kruse, 1999), image processing (Bezdek et al., 2006), control engineering (Zhang and Liu, 2006), soil classification (McBratney and Odeh, 1997) and many others.

Classification problems where data grouped together based on predetermined characteristics is called supervised learning. On the other hand, if predetermined characteristics are not provided we have unsupervised learning (Marsland, 2011).

An algorithm that performs classification is called a classifier. A classifier should be fast and accurate. Classification algorithms tend to be affected by noise in data. Noise should be reduced as much as possible in order to avoid unnecessary complexity in the inferred models and improve the effectiveness of the algorithm (Strong et al., 1997; Wang and Strong, 1996). Noise in data can be due to (Zhu and Wu, 2004): feature noise or class noise. Attribute noise is caused by errors in the feature values (wrongly measured variables, missing values), irrelevant and redundant features,

while class noise is caused by samples that are labelled to belong in more than one class and misclassifcations (Hira and Gillies, 2015). The presence of irrelevant and redundant features increases the model's computational cost, usually exponentially (Hira and Gillies, 2015). To overcome this problem it is necessary to find a way to reduce the number of features by discarding irrelevant and redundant features (Luukka, 2011; Hira and Gillies, 2015).

1.5 Statement of the Problem

Uncertainty is a natural phenomenon in all classification problems and often leads to loss of information. Main forms of uncertainty that arises in fuzzy classification include: fuzziness (not sharp, unclear, imprecise and approximate) and ambiguity (not specific) (Booker and Ross, 2011). Fuzzy entropy is frequently used to measure uncertainty arising in classification problems particularly due to redundant and irrelevant features. However, Pal and Pal (1992) have shown that this measure is inadequate in feature selection because it fails to detect ambiguity. They instead proposes the use of fuzzy entropy and higher order fuzzy entropy. Nevertheless, it is computationally expensive to implement these measures which makes this approach highly impractical. This thesis aims to address the inadequacy of fuzzy entropy based feature selection method by designing a geometrical fuzzy hypercube classification model that estimates ambiguity in features using measures of specificity. In addition, this measures of ambiguity will allows us estimate uncertainty when entities are assigned to classes. This form of uncertainty, which is rarely reported, will give us insights into how much information we lose during class assignment.

1.6 Justification

To handle larger and more complex classification tasks, the problem of focusing on the most relevant information in a potentially overwhelming quantity of data has become

increasingly important. In fuzzy classification problems, algorithms that are efficient and require less computational effort to implement during feature selection will prove very useful. Such algorithms should be to detect ambiguity in attributes and classification classes.

1.7 Objectives

1.7.1 General Objective

The main objective of this study is to design a fuzzy set theoretic classification model based on the geometry of fuzzy sets.

1.7.2 Specific Objective

Objectives of this study is to:

1. Construct a fuzzy hypercube similarity classifier based on the Lukasiewicz generalized structure.

2. Develop feature selection scheme using the similarity measure based on the generalized Lukasiewicz structure.

3. Derive a measure of uncertainty using Yager's linear measure of specificity.

4. Evaluate the performance of the proposed model using standard bench mark data sets from UCI machine learning repository.

5. Evaluate uncertainty in the proposed model due to ambiguity in class assignment using the measure of uncertainty obtained in $(1.)$ above.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

In this chapter we provide a survey on studies that have considered classification problems within the framework of fuzzy set theory. In particular, we look at studies that have demonstrated how the geometry of fuzzy sets can be applied to pattern recognition. Literature on fuzzy set based feature selection techniques is covered as well.

2.2 Application of Fuzzy Sets in Classification Problems

Fuzzy set based approaches have been successfully applied in classification problems across several fields of study. This has led to well established fuzzy classification algorithms such neural-fuzzy, fuzzy genetic algorithms among others. Most of these algorithms are rule based, implying that their classification accuracy depends on how efficient they generate and manipulate the fuzzy if-then rules. Some of the fuzzy logic rule based classifiers include, fuzzy image processing classifiers (Fageth et al., 1996; Moore et al., 2001) and fuzzy genetic classifiers (Yuan and Zhuang, 1996; Krömer et al., 2011).

Other than the rule based techniques, approaches based on fuzzy sets as basic mathematical structures for representing and quantifying various aspects of non-probabilistic uncertainty associated with real-world problems have been investigated (Sadegh-Zadeh, 1999; Nieto and Torres, 2003; Nieto et al., 2006). Such representation does not only allow the use of fuzzy measures of similarity and distance to efficiently represent entities of a physical problem, but also provide a framework for interpretation of such problems afforded by a fuzzy unit hypercube. We now look at a survey of some studies that have demonstrated how these basic elements can be used in classification problems.

Kang and Vachtsevanos (1993) have developed a tool for intelligent control and

identification. A robust and reliable learning and reasoning mechanism is addressed based on fuzzy set theory. The mechanism stores a priori an initial knowledge base via approximate learning and utilizes this information for identification and control via fuzzy inferencing. This processor is called a fuzzy hypercube. Fuzzy hypercubes can be applied to a class of complex and highly nonlinear systems which suffer from vagueness uncertainty. Evidential aspects of a fuzzy hypercube are treated to assess the degree of certainty or reliability. The implementation issue using fuzzy hypercubes is raised, and a fuzzy hypercube is applied to fuzzy linguistic control.

Other applications of fuzzy hypercube includes disease casuality (Helgason and Jobe, 1998). They argue that diseases such as stroke which involve multiple concomitant causal factors that are difficult to represent using conventional statistical methods. Their work illustrates how fuzzy sets and fuzzy logic form the best paradigm for representing complex multi-causal clinical phenomenon in stroke. In their view, this representation is generalizable to all of clinical science since multiple concomitant causal factors are involved in nearly all known pathological processes. Furthermore, Helgason (2007) how complex interaction between variables associated with stroke can be represented using fuzzy unit hypercube.

Sadegh-Zadeh (1999) has used the geometrical representation of fuzzy sets to introduce the concept of nosology within fuzzy set theoretic frame work. In this framework, the vague notions of health, disease, and nosology are analyzed. It is shown that health and disease as generic concepts, and also individual disease entities are best understood as fuzzy sets. Clinical language and linguistics, nosology and diagnosis may thus become directly amenable to fuzzy theory.

Landscapes are normally classified into three main categories, urban, semi-urban and rural. However, these categories are characterized by vague boundaries which can not be captured using conventional approaches. One of the studies that has taken advantage of the flexibility of fuzzy sets in characterizing urban landscapes can be found in (Heikkila et al., 2003). In this study, the fuzzy unit hypercube has been

used to develop three metrics to measure extent of urbanization, level of fuzziness, and degree of entropy, to characterize levels of urban membership typical in cities of China, and other Asian countries. This is made possible by representing the entire area of study as a point within a fuzzy hypercube. Using these three dichotomies with fuzzy set interpretations, they exploited the geometrical interpretations of fuzzy sets afforded by the fuzzy unit hypercube to show how a study area could be located within a three dimensional fuzzy hypercube.

Similarly, Torres and Nieto (2003) have used a 12-dimensional fuzzy hypercube of to represent triplet codon. They have also illustrated how dissimilitudes between polynucleotides can be measured using geometrical properties of fuzzy sets. Such geomtrical representation allows calculation of frequencies of the nucleotides at the three base sites of a codon in the coding sequences of Escherichia coli K-12 and Mycobacterium tuberculosis H37Rv, when considered as points in the fuzzy space.

On the other hand, Nieto and Torres (2003) have used the fuzzy unit hypercube to define concept of midpoint of fuzzy sets. This concept together with distance measure then forms a basis for representing fuzzy degree of two concurrent food and drug addictions, and a fuzzy representation of concomitant causal mechanisms of stroke.

In (Nieto et al., 2006) a fuzzy unit hypercube is applied in the study of polynucleotides. They have used the concept of a metric space to investigate differences between polynucleotides. The hypercube allows definition of distances between nucleotides and some complete genomes using several metrics within the hypercube. In addition, results on the notions of similarity and equality between polynucleotides are presented.

2.3 Fuzzy Feature Selection Based Techniques

Feature selection seeks to address the issue of geometrical complexity in classification problems. Feature selection plays an important role in classification for several

reasons. First it can simplify the model and this way computational cost is reduced. Secondly, by removing insignificant features from the dataset makes the model more transparent and more comprehensible, providing better explanation, which is an important requirement in applications. Feature selection process can also reduce noise which enhances the classification accuracy.

In quest for higher classification accuracies, feature subset selection has been used for data reduction in areas characterized by high dimensionality due to the large numbers of available features, for example in seismic data processing (Hoffman et al., 1998), remote sensing (Yu et al., 2000), drug design , speech recognition (Abdulla and Kasabov, 2003). Feature selection is expected to improve classification performance, particularly in situations characterized by the high data dimensionality problem caused by relatively few training examples compared to a large number of measured features.

Even if no significant improvements in classification accuracy are achieved, reducing the number of features still has many advantages. These are e.g. reducing the number of measurements required, shortening training and execution times, and improving model compactness, transparency, and interpretability (Luukka, 2011).

Most of the prominent feature selection approaches have been investigated within the proabilistic framework. These techniques can be grouped into three broad categories: embedded, filter and wrapper techniques (Blum and Langley, 1997).

Embedded techniques involve procedures that perform feature selection as part of a classification algorithm. Such methods are common in machine learning. They include ID3(Quinlan, 1986), C4.5 (Salzberg, 1994), rough sets(Pawlak, 1982), classification and regression trees(Breiman, 2017). Filter schemes on the other hand, do not interact with the classifier during feature selection. Within this context, feature selection is performed as a preprocessing stage prior to model development in order to filter out the irrelevant or redundant features from the analysis. Thus, the selection of the most important features is not related to the classification method that is used to build the model. This is the main disadvantage of such algorithms, since the characteristics of

the method are ignored during the feature selection process. However, filter techniques are quite popular mainly owing to their computational efficiency even for large data sets. Some common filter feature selection algorithms include the RELIEF algorithm (Kira and Rendell, 1992), the FOCUS algorithm (Kira and Rendell, 1992), sequential forward and backward generation algorithms(Liu and Yu, 2005).

Feature selection approaches utilizing tools of fuzzy set theory are fairly recent. Most of these approaches use the concept of fuzzy entropy to determine relevant features.

A feature selection approach proposed by Lee et al. (2001) is such that the feature space is partitioned into nonoverlapping decision regions. They then apply fuzzy entropy measure to select important features. This approach was then tested using Iris and Breast cancer datasets from UCI machine learning repository achieving good classification rates.

In (Luukka, 2011) a feature selection method based on fuzzy entropy measures with similarity classifier is introduced. Model was tested with four medical data sets which were, dermatology, Pima-Indian diabetes, breast cancer and Parkinsons data set. With all the four data sets, he managed to get quite good results by using fewer features that in the original data sets. Also with Parkinsons and dermatology data sets, classification accuracy was enhanced significantly. Mean classification accuracy with Parkinsons data set being 85.03% with only two features from original 22. With dermatology data set, mean accuracy of 98.28% was achieved using 29 features instead of 34 original features. Iyakaremye et al. (2012) have proposed a similar approach but instead they have used Yu's similarity measure.

Mitra et al. (2002) have introduced a feature selection algorithm suitable for large data sets. Their method is based on measuring similarity between features whereby redundancy therein is removed. This does not need any search and, therefore, is fast. They used a feature similarity measure, called maximum information compression index. The algorithm is generic in nature and has the capability of multiscale representa-

tion of data sets. The superiority of the algorithm, in terms of speed and performance, is established extensively over various real-life data sets of different sizes and dimensions. It is also demonstrated how redundancy and information loss in feature selection can be quantified with an entropy measure.

Jaganathan and Kuppuchamy (2013) have presented an approach for measurement of feature relevance based on fuzzy entropy for a medical database classification. Three feature selection strategies are devised to obtain the valuable subset of relevant features. Five benchmarked datasets, which are available in the UCI Machine Learning Repository have been used in this work. The classification accuracy shows that the proposed method is capable of producing good results with fewer features than the original datasets.

CHAPTER THREE

Methodology

3.1 Overview

In this chapter, we first outline the key components of the proposed fuzzy hypercube classification model. These include the similarity classifier and the proposed feature selection scheme. We conclude with discussion on how the model can be validated and estimation of uncertainty due to class ambiguity.

3.2 Fuzzy Similarity Classifier

Suppose a set X of samples is classified into classes $C_1, C_2, \ldots, C_n$ based on a set of feature which are shared by these samples to the greatest extent. Let $F = \{f_1, f_2, \ldots, f_q\}$ be a set of features measured for each of the classes. The values of these features are fuzzified by normalizing them so that they lie in the unit interval $[0,1]$. Thus, the samples $\mathbf{x}(i) \in X, 1 \leq i \leq N$ to be classified can be viewed as q-dimensional fit vectors,

$$\hat{\mathbf{x}}(i) = \left(\hat{x}_1(i), \hat{x}_2(i), \ldots, \hat{x}_q(i)\right) \tag{3.2.1}$$

in a fuzzy unit hypercube

$$[0,1] \times [0,1] \times \cdots \times [0,1] = [0,1]^q \tag{3.2.2}$$

In this case, the j^{th} feature corresponds to the j^{th} dimension of the fuzzy unit hypercube. For a classification problem with three features we have a three dimensional hypercube space of features (attributes) shown in the figure below. Classification is performed in this multidimensional feature space by first determining a vector

$$V_k = \left(v_1(k), v_2(k), \ldots, v_q(k)\right), (1 \leq k \leq n) \tag{3.2.3}$$

that carries properties of the k^{th} class to the greatest extent. Pasi Luukka calls this vector an ideal vector for the k^{th} class. It can be defined by an expert or computed using the generalized mean as follows. Suppose the sample set X_k^t of fit vectors $\hat{\mathbf{x}}_k^t(i) = \left(\hat{x}_1^t(i,k), \hat{x}_2^t(i,k), \ldots, \hat{x}_q^t(i,k)\right)$ in the training set X^t is known to belong to the k^{th} class. Then,

$$v_j(k) = \left(\frac{1}{|X_k^t|}\sum\left(\hat{x}_j^t(i,k)\right)^m\right)^{\frac{1}{m}}, 1 \leq j \leq q \tag{3.2.4}$$

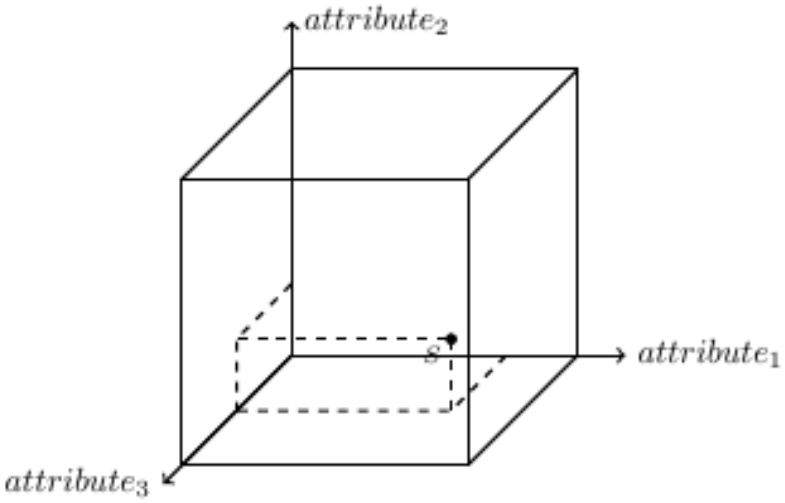

Figure 3.2.1: Attribute Space

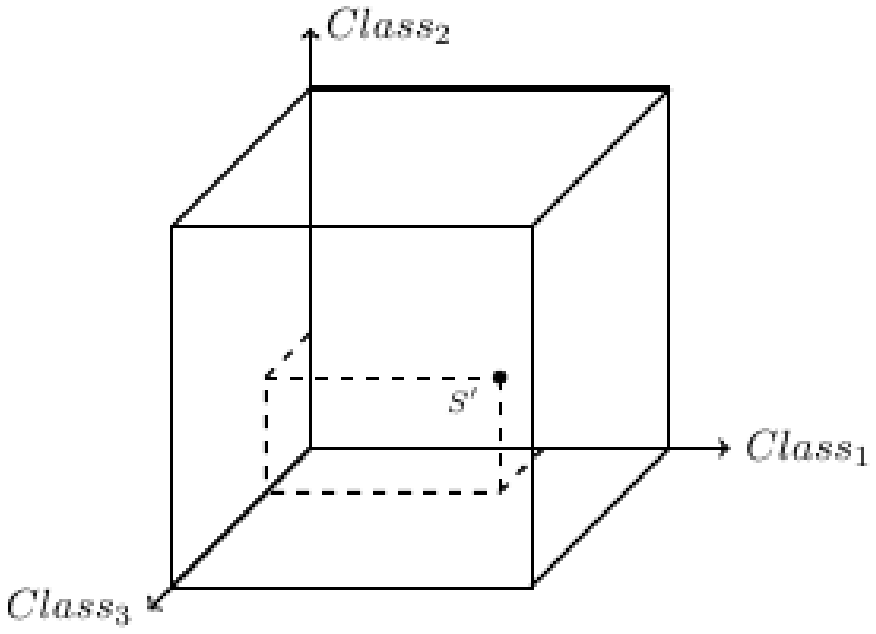

Figure 3.2.2: Space for Classes

Where the sum of the fit values $\hat{x}_j^t(i,k)$ is performed over all the i^{th} training samples in the k^{th} class for each j and $|X_k^t|$ is the number of samples in the k^{th} class of the training set and m is the parameter for the generalized mean. In this study we have used the arithmetic mean, $m = 1$.

Consider artificial training samples given in the table (3.1) below

Table 3.1: Artificial data

Sample	Feature1	Feature2	Feature3	Classes
1	0.6	0.8	0.7	C1
2	0.6	0.5	0.4	C2
3	0.4	0.3	0.3	C2
4	0.8	0.5	0.6	C1
5	0.1	024	0.5	C2
6	0.5	0.4	0.9	C1

In this case the number of classes is $n = 2$ and $|X_k^t|$=3 for both classes, i.e, we have three instances which are known to belong to each of the classes. Ideal vector for each

class is obtain by getting the mean of the values of features for samples belonging to the class. If we use the arithmetic mean, that is, $m = 1$ we have

$$V_1 = \left(\frac{1}{3}(0.6+0.8+0.5), \frac{1}{3}(0.8+0.5+0.4), \frac{1}{3}(0.7+0.6+0.9)\right) = (0.63, 0.57, 0.73) \tag{3.2.5}$$

and

$$V_2 = \left(\frac{1}{3}(0.6+0.4+0.1), \frac{1}{3}(0.5+0.3+0.2), \frac{1}{3}(0.4+0.3+0.5)\right) = (0.37, 0.0.33, 0.4) \tag{3.2.6}$$

To classify an arbitrary sample $x(i)$, we compare its fit vector with the ideal vectors V_k using the similarity measure

$$\begin{aligned} S_k &= S(\hat{\mathbf{x}}(i), V_k) \\ &= \left[\frac{1}{n}\sum_{j=1}^{q} w_j \left(1 - |\left(\hat{x}_j(i)\right)^p - \left(v_j(k)\right)^p|\right)^{\frac{m}{p}}\right]^{\frac{1}{m}} \end{aligned} \tag{3.2.7}$$

where

$$\sum_{j=1}^{q} w_j = 1, w_j \in [0,1] \tag{3.2.8}$$

is a subjective weight reflecting the relative importance of the j^{th} feature. In this study we set $w_j = 1 \forall j$. The parameter p is chosen from the generalized Lukasiewicz structure and m is for the generalized mean. In particular for $p = 1$ we have the normal Lukasiewicz structure.

The sample is then assigned class k^* if it bears the highest similarity with this class, that is,

$$S_{k^*} = \text{Max}_k \quad \{S(\hat{\mathbf{x}}(i), V_k)\} \tag{3.2.9}$$

with $S_{k^*} = 1$ if $x(i)$ coincides with the ideal vector for class k^* and $S_{k^*} = 0$ if it bears no resemblance with the k^{th} class.

Since in the hypercube space $[0,1]^q$, we have a continuum of fuzzy sets, it follows

that $V_k \in [0,1]^q$ is the centre of the k^{th} class. Thus, the similarity classifier can be interpreted as a between cube mapping given by the equation below

$$S_k : [0,1]^q \to \mathscr{C} \tag{3.2.10}$$

where $\mathscr{C}$ is the space of all classification classes. In this space, each point is a similarity vector which can be understood as an ordered fuzzy set.

3.2.1 Feature selection

Consider the ideal vectors

$$\begin{aligned} V_1 &= \left(v_1(1), v_2(1), \ldots, v_q(1)\right) \\ V_2 &= \left(v_1(2), v_2(2), \ldots, v_q(2)\right) \\ &\ldots \\ V_k &= \left(v_1(k), v_2(k), \ldots, v_q(k)\right) \end{aligned} \tag{3.2.11}$$

corresponding to the n classification classes. In a perfect situation the n ideal vectors correspond to the single element sets $(1,0,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,0,0,\ldots,1)$. These are precisely vertices of a fuzzy unit hypercube. Now, assume we want to classify sample $x(i)$ using the j^{th} feature. In the normal Lukasiewicz structure, each of the n similarity values $S_1(j), S_2(j), \ldots, S_n(j)$ is obtained from the equivalence relation

$$S_k(j) = \hat{x}_j(i,k) \longleftrightarrow v_j(k) = 1 - |\hat{x}_j(i,k) - v_j(k)| \tag{3.2.12}$$

The sample is then assigned to the class that gives maximum value of similarity. However, if it has same maximum value of similarity to more than one class, then we are faced with difficult in deciding which class to assign this sample. In other words, the classes are indistinguishable based on the j^{th} feature. We can measure this by computing the distance between each sample $x(i)\,(1 \le i \le N)$ and the ideal vectors.

Observe that negating the equivalence relation given in equation (3.12) we obtain the distance D between sample $x(i)$ and the k^{th} class,

$$D = \left(\hat{x}_j(i,k) \longleftrightarrow v_j(k)\right)^* = |\hat{x}_j(i,k) - v_j(k)| \tag{3.2.13}$$

If $x(i)$ coincides with class say k' then we must have

$$D = |\hat{x}_j\left(i,k'\right) - v_j\left(k'\right)| = 0 \tag{3.2.14}$$

and

$$D = |\hat{x}_j\left(i,k'\right) - v_j(k)| \neq 0 \forall k \neq k' \tag{3.2.15}$$

In the generalized Lukasiewicz structure with parameter p we have

$$D_p = \left(\hat{x}_j(i,k) \longleftrightarrow v_j(k)\right)^* = |\left(\hat{x}_j(i,k)\right)^p - \left(v_j(k)\right)^p|^{\frac{1}{p}} \tag{3.2.16}$$

The distance measure D can be computed for all samples in the training set X^t and all classification classes. The average of this measure can be viewed as a measure of separation between classes. Intuitively, features that do separate classes well are expected to give small average values of D. These features are then successfully removed by setting a minimum threshold value of this measure till no further improvement in classification rate is achieved. In this work, D is computed in the normal Lukasiewicz structure. Because in the normal Lukasiewicz structure D is linear, we see that the suggested measure requires less computational effort to implement as compared to entropy based approach.

3.3 Uncertainty in Class Assignment

If the similarity vector for a given sample is such that the it has same values of similarity to all classes then we are faced with uncertainty as to which class to assign this

sample. This form of uncertainty is called ambiguity. Ambiguity exists when selection must be made from two or more classification classes and attributes do not offer sufficient information to discriminate the classes and determine the appropriate class to assign a sample. For example for a space consisting of three classes $\mathscr{C} = \{C_1, C_2, C_3\}$, the similarity vector $(0.7, 0.4, 0.1)$ indicates a fairly good degree of compatibility between a sample and the first class, while $(0.6, 0.6, 0.2)$ shows ambiguity between the first and the second classes. Further, by assigning a sample to the first class in the former, we ignore the fact that this sample carries properties of the second and the third classes to degrees 0.4 and 0.1 respectively. Using Yager's linear measure of specificity, this form of uncertainty is given as

$$U_i = 1 - Sp(\Omega_i) \tag{3.3.1}$$

Where $\Omega_i = (S_1, S_2, \ldots, S_n)$ is the vector of similarity values for the i^{th} sample.

3.4 Numerical Experiment

3.4.1 Model Validation

After developing the model, the next step is to find out how effective is the model based on some performance measures using data. For purposes of comparing our results with those obtained using similar models, we will use four widely used benchmark datasets from UCI machine learning repository (Blake and Merz, 1998). Each data set is divided into 50% for model training and 50% for model testing. This process is repeated randomly 10 times for each fixed value of parameters in the similarity classifier. We report the following measures of performance, highest mean classification accuracies and variances, specificity and sensitivity. Specificity and sensitivity are computed us-

ing the following formulae

$$Sensitivity = \frac{TP}{TP+FN} \tag{3.4.1}$$

and

$$Selectivity = \frac{TN}{TN+FP} \tag{3.4.2}$$

where TP of C_k is all C_k samples that are classified as C_k.

TN of C_k is all non-C_k samples that are not classified as C_k.

FP of C_k is all non-C_k samples that are classified as C_k.

FN of C_k is all C_k samples that are not classified as C_k. and C_k is the k^{th} classification class.

3.4.2 Characteristics of Datasets

Table 3.2: Validation data and properties

Validation data	Number of classes	Dimension	Number of observations
Dermatology	6	34	366
Pima-Indians	2	8	768
Parkinsons	2	22	197
Thyroid	3	5	215

(i) **Dermatology Dataset.**

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy

is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

(*iii*) **Pima Indian Dataset**

The Pima-Indian data set concerns the presence or absence of diabetes among Pima-Indian women living near Phoenix, Arizona.

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m))
7. Diabetes pedigree function
8. Age (years)

(ii) **Parkinsons Dataset**

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ('name' column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

(iii) **Thyroid Dataset**

This data set contains 3 classes and 215 samples. These classes correspond to to the hyper, hypo and normal function of the thyroid gland. The followings give the 5 tests which are applied to patients to measure the thyroid functions.

1. T3-resin uptake test (as a percentage).
2. Total Serum thyroxin as measured by the isotopic displacement method.
3. Total Serum triiodothyronine as measured by radioimmuno assay
4. Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay
5. Maximal absolute difference of TSH value after injection of 200 micro-grams of thyrotropin-releasing hormone as compared to the basal value.

CHAPTER FOUR

Results and Discussion

4.1 Theoretical results

In this section we investigate Yager's linear measure of specificity (Yager, 1998) within the geometrical setting of fuzzy sets. Thereafter, we demonstrate how this measure can be used to measure uncertainty in classification problems.

4.1.1 Geometrical Measure of Specificity

We now turn to specificity as a measure of certainty. In this discussion we require the concept of the principal diagonal of a fuzzy hypercube. This is the diagonal joining the points $(0,0,0,\ldots,0)$ and $(1,1,1,\ldots,1)$.

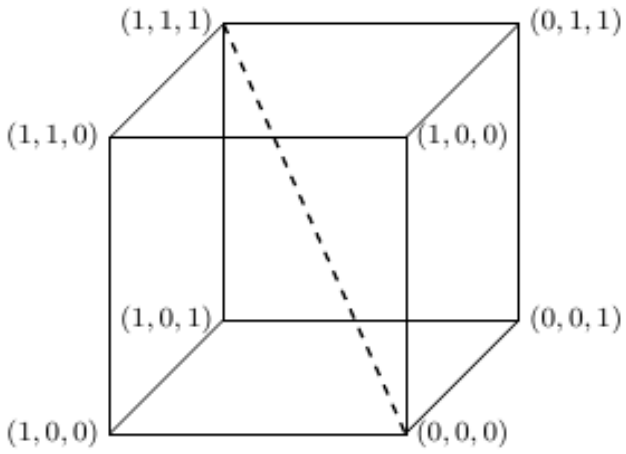

Figure 4.1.1: Principal diagonal

Observe that any point on the principal diagonal is a constant fuzzy set, $[a]$. Moreover, it follows that if $A_{s(j)}$ a singleton subset of X with 1 in the j^{th} position then $d_1\left([a],A_{s(1)}\right) = d_1\left([a],A_{s(2)}\right) = \cdots = d_1\left([a],A_{s(n)}\right)$. The subsets $A_{(j)}\ \forall j$ can be regarded as basis vectors (Yager, 1998). Basis vectors have maximum specificity.

Observe that if A is a singleton set then we must have $A = A_{(j)}$ and hence $d_r\left(A, A_{(j)}\right) = 0$.

Lemma 4.1.1.

$$d_r\left(A, A_{(j)}\right) = (n-1)^{\frac{1}{r}} \text{ if and only if } \quad A = X. \tag{4.1.1}$$

Proof. The result follows from the fact that the whole space X is farthest from any basis vactor A_j. □

Yagers linear measure

$$Sp(A) = a_1 - \frac{1}{n-1}\sum_{r=2}^{n} a_r \tag{4.1.2}$$

can be expressed as,

$$Sp(A) = \frac{1}{n-1}\left(a_1(n-1) - \sum_{r=2}^{n} a_r\right) \tag{4.1.3}$$

Which can be checked to be the same as,

$$Sp(A) = \frac{1}{n-1}\left((a_1 - a_1) + (a_1 - a_2) + (a_1 - a_2) + \cdots + (a_1 - a_n)\right) \tag{4.1.4}$$

Intuitively, this represents the fuzzy Hamming distance between set A and the point $[a_1]$ on the principal diagonal,

$$Sp(A) = \frac{1}{n-1} d_1\left([a_1], A\right) \tag{4.1.5}$$

Let us consider the a more general case of this measure of expressed as the normalized distance between $[a_1]$ and A

Theorem 4.1.1.

$$Sp(A) = \frac{1}{(n-1)^{\frac{1}{r}}} d_r\left(\left[a_j\right], A\right) \tag{4.1.6}$$

Proof. This formula is a measure of specificity if it satisfies satisfies the basic properties of specificity. First, observe that if $A = A_{(j)}$, then $\left[a_j\right] = X$ implying $d_r\left(X, A_{(j)}\right) = (n-1)^{\frac{1}{r}}$ and thus $Sp(A) = 1$. On the other hand, if $A = \emptyset$, then $A = [0]$ and hence $Sp(A) = 0$. Geometrically, specificity decreases as A approaches the principal diagonal- in other words the nearer A gets to the principal diagonal, the more evenly distributed the membership values become and thereby reducing precision of A. □

Further, observe that this measure of specificity is regular since for any constant fuzzy set we have $\left[a_j\right] = [a] = A$.

4.1.2 Estimation of Uncertainty in Classification Problems Using Specificity

We now show how this measure can be used to estimate uncertainty in classification problems. A consider for instance the lenght of a train. If it known this lenght is between $100ft$ to $200ft$ then there if no fuzziness in information being conveyed but rather, there is lack of specificity. We are not told what is the exact value of the length of the train. This form of uncertainty is prominent in classification problems where an entity might posses properties of several classes to same extent.

In the previous chapter, we proposed a feature selection scheme based on some measure of separation between classes. Now, our first assignment is to show that this measure is indeed a measure of uncertainty based on Yager's linear measure of specificity.

Consider a fuzzy similarity vector for the i^{th} sample based on the j^{th} feature

$$\Omega_i(j) = (S_1(j), S_2(j), \ldots, S_n(j)) \tag{4.1.7}$$

If this vector is a single element set say

$$\Omega_i(j) = (1,0,0,\ldots,0) \tag{4.1.8}$$

then this sample precisely belongs to the first class. This is, this sample is unambiguously classified using the j^{th} feature. This lack ambiguity can be easily determined by computing specificity of $\Omega_i(j)$. Because specificity is a measure of certainty, we see that uncertainty in classification of sample i using feature j is given by

$$U_i(j) = 1 - Sp(\Omega_i(j)) \tag{4.1.9}$$

with $U_i(j) = 0$ if $U_i(j)$ is a single element set and $U_i(j) = 1$ if $U_i(j) = [S_k(j)] \forall k$. Consider an ideal situation where ideal vectors correspond to vertices of a fuzzy hypercube. If sample i is assigned class k, then the sum of D over all classes takes a maximum value of $n-1$ and minimum value of 0 when

$$S_1(j) = S_2(j) = \cdots = S_n(j) \tag{4.1.10}$$

This is the same as computing specificity of $\Omega_i(j)$. We wrap this up by considering a simple numerical example. Suppose we have a classification problem with 2 classes, C_1, C_2 and and 3 features say f_1, f_2, f_3. If sample i gives similarity vectors $(0.5,0.4), (0.4,0.4)$ and $(0.9,0.9)$ using features f_1, f_2 and f_3, respectively, then it is difficult to decide which sample assign this sample using features f_2 and f_3. Using specificity measure, we obtain maximum uncertainty with these features. On the other hand, fuzzy entropy based feature selection method will give large uncertainty for f_1 because it is closer to the centre of the hypercube as compared to f_2 and f_3. Thus, using fuzzy entropy, classification using feature f_2 and f_3 are less uncertain. Clearly, this is not intuitive.

Now, after classifying all the samples we would like to determine how ambiguously

they are classified. This is a measure of uncertainty in class assignement. Given a fuzzy similarity vector

$$\Omega_i = (S_1(i), S_2(i), \ldots, S_n(i)) \tag{4.1.11}$$

Then

$$Sp(\Omega_i) = S_{k^*}(i) - \frac{1}{n-1}\sum_{k^* \neq k} S_k^* = \frac{1}{n-1} d_1([S_{k^*}], \Omega_i) \tag{4.1.12}$$

Where $S_{k^*}(i)$ is the largest similarity value. Therefore a sample that is most ambiguously classified is such that its corresponding similarity vector lies on the principal diagonal of the hypercube space for classes.

4.2 Experimental Results

In this section, we present classification results from the four UCI machine learning datasets. Classification rates, specificity and sensitivity values are reported. Plots of classification rates vs parameters p and m are given as well. These results are obtained using MATLAB software, version 17.

(i) Dermatology

The mean classification accuracies of 98.34% and 98.21% are achieved with 28 and 24 features respectively as compared with 97.82% without feature removal as shown in table 4.1. Plots for $34, 29$ and 24 features as shown in figures (4.2.1-4.2.3). The corresponding values of specificity and sensitivity are shown in table (4.2).

The first column in table (4.1) corresponds to method used, similarity classifier (S) or similarity classifier with feature selection (S & F).

Features:

(1) Erythema.

(2) Scaling.

Table 4.1: Classification rate with dermatology data

Method	Mean Accuracy (%)	Variance	Dimension
S	97.82	0.0001	34
S & F1	98.34	0.0000	28
S & F2	98.21	0.0000	24

Table 4.2: Specificity and sensitivity with dermatology data

Dimension	Specificity (%)	Sensitivity (%)
34	99.57	97.70
28	99.69	98.11
24	99.65	98.16

(3) Definite borders.

(4) Itching.

(5) Koebner Phenomenon.

(6) Polgonal popules.

(7) Follicular papules

(8) Oral mucosal involvement.

(9) Knee and elbow involvement.

(10) Scalp involvement.

(11) Family history.

(12) Melanin incontinence.

(13) Eosinophils in the infiltrate.

(14) PNL infitrate.

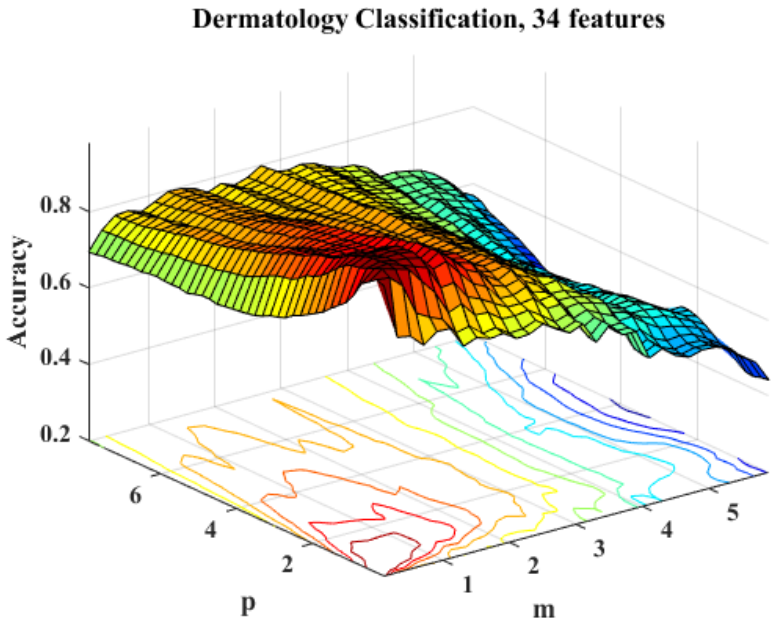

Figure 4.2.1: Dermatology classification using similarity classifier without feature selection

(15) Fibrosis of the papillary dermis.

(16) Exocytosis.

(17) Acantosis.

(18) Hyperkeratosis.

(19) Parakeratosis.

(20) Clubbing of the rete ridges.

(21) Elongation of the rete ridges.

(22) Thinning of the suprapapillary.

(23) pongiform pustule.

(24) Munro microabcess.

(25) Focal hypergranuosis.

(26) Disappearance of granular layer.

(27) Vascularization and damage of basal layer.

(28) Spongosis.

(29) Saw-tooth appearance of retes.

(30) Follicular horn plug.

(31) Perifollicular parakeratosis.

(32) Inflamatory mononuclear infiltrate.

(33) Band-like infiltrate.

(34) Age.

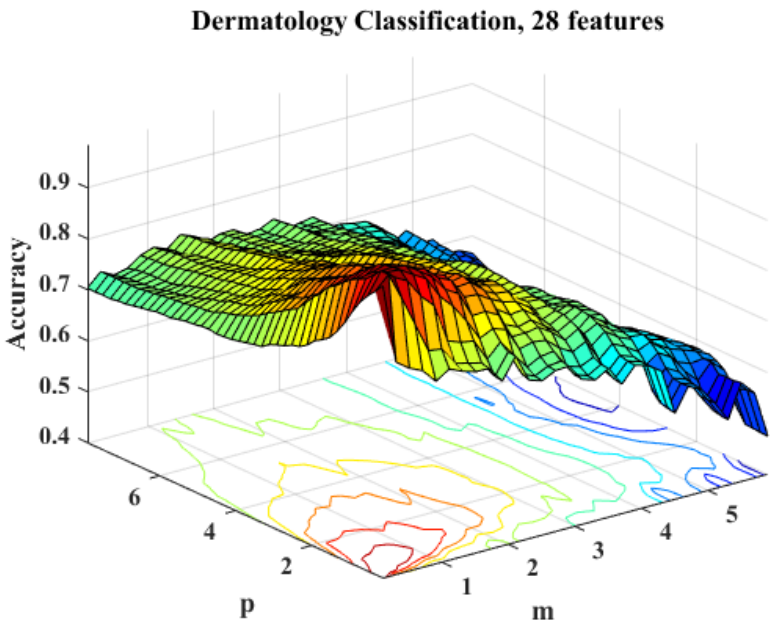

Figure 4.2.2: Dermatology classification using similarity classifier with 28 features

Removed features:

(1) Erythema.

(2) Scaling.

(17) Acantosis.

(13) Eosinophils in the infiltrate.

(32) Inflamatory mononuclear infiltrate.

(34) Age.

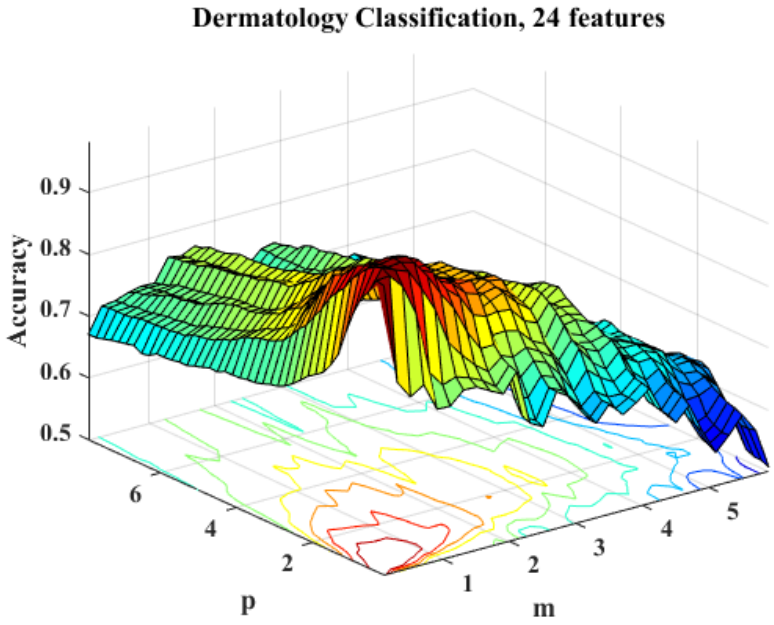

Figure 4.2.3: Dermatology classification using similarity classifier with 24 features

Removed features:

(1) Erythema.

(2) Scaling.

(3) Definite borders.

(13) Eosinophils in the infiltrate.

(17) Acantosis.

(19) Parakeratosis.

(23) pongiform pustule.

(30) Follicular horn plug.

(32) Inflamatory mononuclear infiltrate.

(34) Age.

(*ii*) **PIMA-Indian Diabetes**

With the Pima Indian data set, we only require 1 feature to obtain an accuracy of 77.73% as shown in table (4.3). For these data set, the best classification rate is obtained with 1 feature. The plots of accuracies with respect to these parameters indicate a significant difference with 8 as compared to 3 and 1 features as shown in figures 4.2.4-4.2.6. This dataset gives low values of specificity and sensitivity. From values of specificity and sensitivity shown in table (4.4), the classifier performs fairly well with 5 features.

Table 4.3: Classification rate with Pima Diabetes data

Method	Mean Accuracy (%)	Variance	Dimension
S	77.21	0.0000	8
S & F1	77.47	0.0012	3
S & F2	77.73	0.0008	1

Table 4.4: Specificity and sensitivity Pima Diabetes data

Dimension	Specificity (%)	Sensitivity (%)
8	72.88	78.03
3	70.30	75.70
1	70.45	75.72

Features:

(1) Number of times pregnant.

(2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

(3) Diastolic blood pressure (mm Hg).

(4) Triceps skin fold thickness (mm).

(5) 2-Hour serum insulin (mu U/ml).

(6) Body mass index (weight in kg/(height in m)).

(7) Diabetes pedigree function.

(8) Age (years).

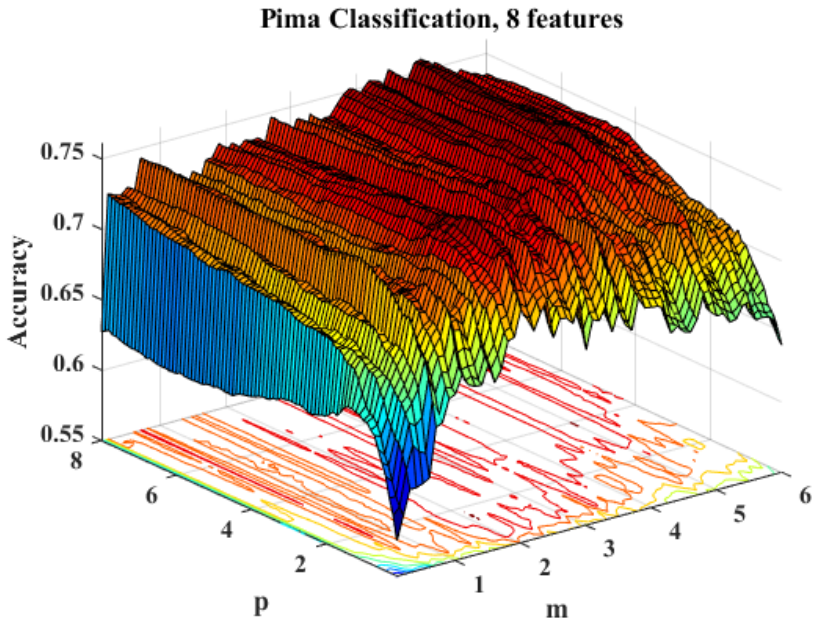

Figure 4.2.4: Classification of PIMA-Indian diabetes data with 8 features

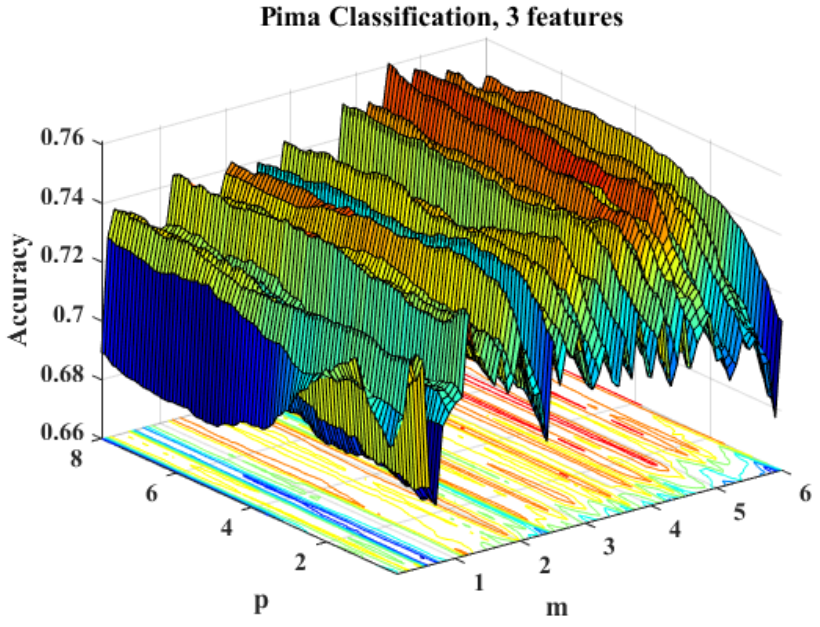

Figure 4.2.5: Classification of PIMA-Indian diabetes data with 3 features

Selected features:

(2) Plasma glucose concentration at 2 hours in an oral glucose tolerance test.

(3) Diastolic blood pressure (mm Hg).

(5) 2-Hour serum insulin (mu U/ml).

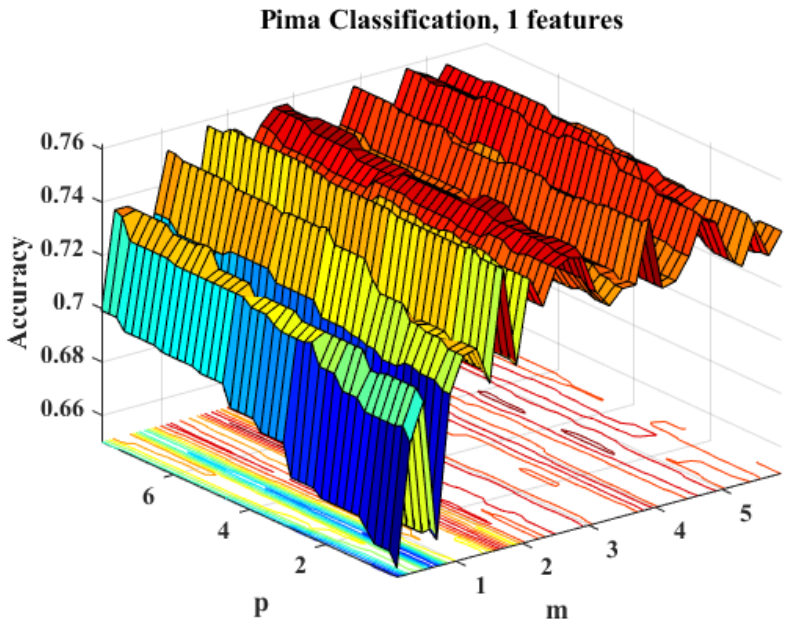

Figure 4.2.6: Classification of PIMA-Indian diabetes data with 1 feature

Selected feature:

(2) Plasma glucose concentration at 2 hours in an oral glucose tolerance test.

(*iii*) **Parkinsons**

The mean classification accuracy attained with similarity classifier alone in 84.69%. With feature selection, we managed to obtain a mean classification accuracy of 87.24% and 87.76% with 3 features and 1 feature respectively. Plots for 22, 3 and 1 features as shown in fig 4.2.7-4.2.9. With 1 feature we have the best values of specificity and sensitivity as shown in table (4.6).

Table 4.5: Classification results with Parkinsons data

Method	Mean Accuracy (%)	Variance	Dimension
S	83.22	0.0000	22
S & F1	87.24	0.0013	3
S & F2	87.76	0.0002	1

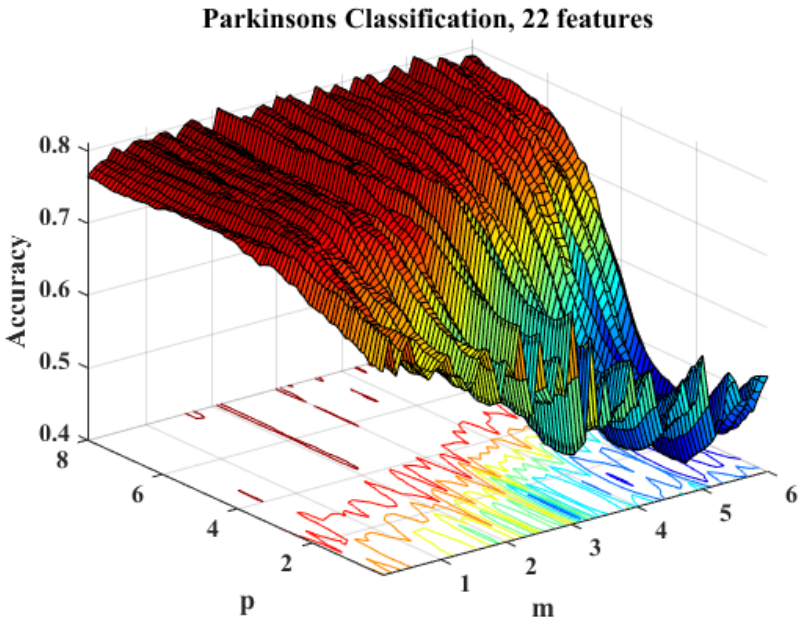

Figure 4.2.7: Classification of Parkinsons data with 22 features

Features:

(1).MDVP:Fo(Hz) Average vocal fundamental frequency,(2) MDVP:Fhi(Hz) Maximum vocal fundamental frequency, (3).MDVP:Flo(Hz) Minimum vocal fundamental frequency, (4).MDVP:Jitter, (5).MDVP:Jitter(Abs),(6).MDVP:RAP, (7). MDVP:PPQ, (8).Jitter:DDP - Several measures of variation in fundamental frequency, (9).MDVP:Shimmer, (10).MDVP:Shimmer(dB), (11). Shimmer:APQ3, (12). Shimmer:APQ5, (13).MDVP:APQ, (14).Shimmer:DDA - Several measures of variation in amplitude (15) NHR, (16).HNR, (17).RPDE, (18).D2, (19). DFA, (20). Exponent spread1 (21).Spread2, (22).PPE

Table 4.6: Specificity and sensitivity for parkinsons data

Dimension	Specificity (%)	Sensitivity (%)
22	78.44	82.75
3	79.98	84.49
1	83.71	86.57

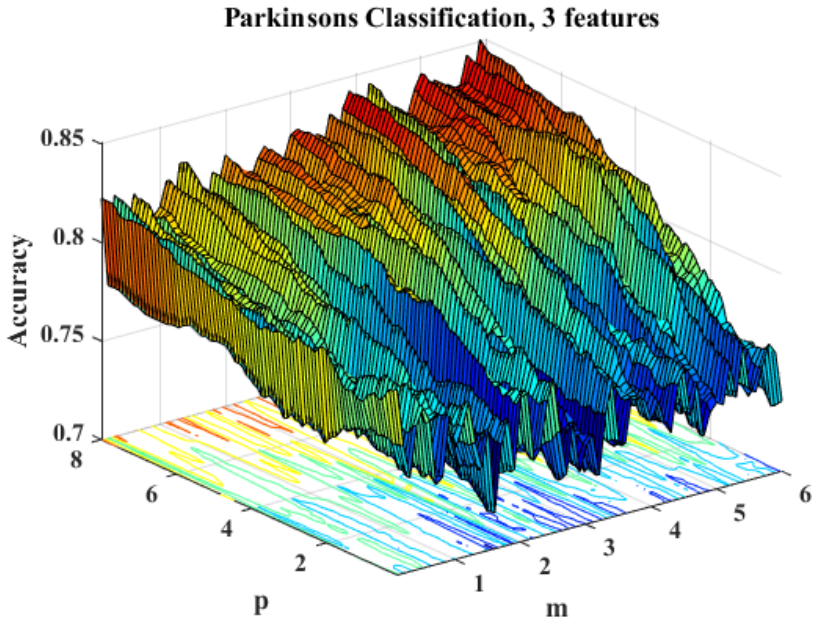

Figure 4.2.8: Classification of Parkinsons data with 3 features

Selected features:

(3) MDVP:Flo(Hz) Minimum vocal fundamental frequency.

(19) DFA.

(20) Exponent spread1.

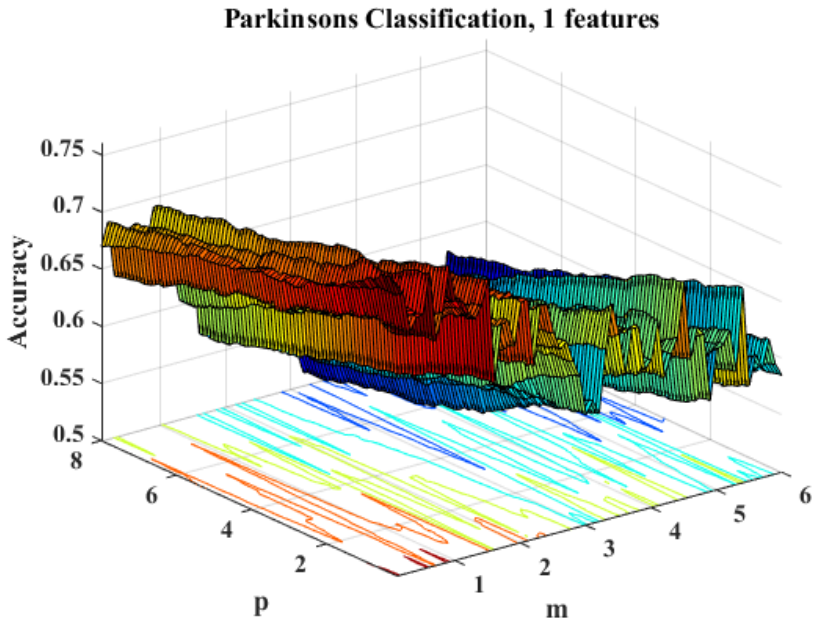

Figure 4.2.9: Classification of Parkinsons data with 1 feature

Selected feature:

(19) DFA.

(*iv*) **Thyroid data set**

The similarity classifier achieved high classification accuracy with this data set. With 5 features we have a mean classification accuracy of 97.69%. On the other hand, with 4 featurs we managed to obtain a classification accuracy of 98.61%. The required plots with and without feature selection are shown in figures 4.2.10-4.2.11.

Table 4.7: Classification rate with thyroid data

Method	Mean Accuracy (%)	Variance	Dimension
S	97.69	0.0000	5
S & F	98.61	0.0000	4

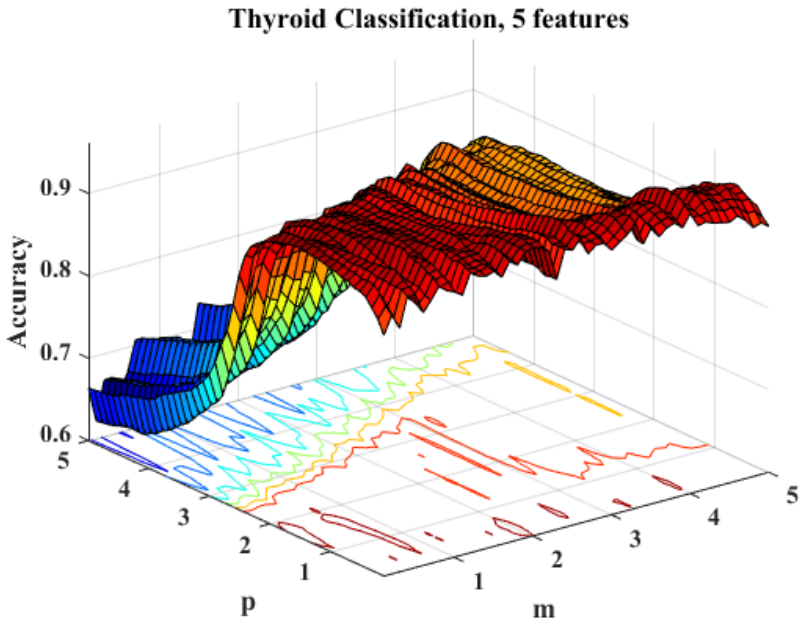

Figure 4.2.10: Classification of Thyroid data with all features

Table 4.8: Specificity and sensitivity with thyroid data

Dimension	Specificity (%)	Sensitivity (%)
5	97.73	95.26
4	98.84	97.82

Features:

(1) T3-resin uptake test (as a percentage).

(2) Total Serum thyroxin as measured by the isotopic displacement method.

(3) Total Serum triiodothyronine as measured by radioimmuno assay.

(4) Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay.

(5) Maximal absolute difference of TSH value after injection of 200 micro-grams of thyrotropin-releasing hormone as compared to the basal value.

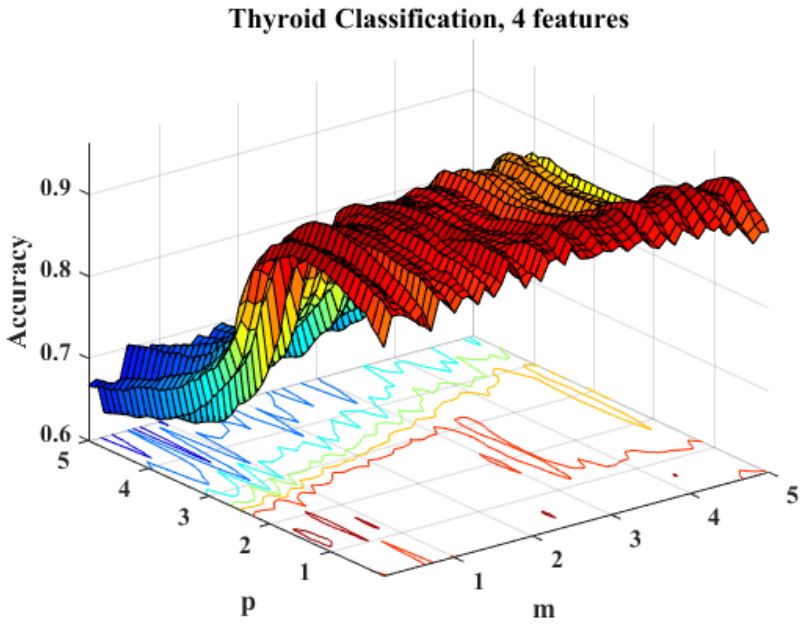

Figure 4.2.11: Classification of Thyroid data with 4 features

Selected features:

(1) T3-resin uptake test. (as a percentage)

(2) Total Serum thyroxin as measured by the isotopic displacement method.

(3) Total serum triiodothyronine as measured by radioimmuno assay.

(5) Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

4.2.1 Disussion

In tables (4.9-4.12), a comparison of the proposed feature selection method to fuzzy entropy based approach by Luukka (2011) and Luukka and Leppälampi (2006) is given. From these results, it can be seen that the proposed method yields very impressive results with some data sets. In particular, with the dermatology data set, we have an

average classification accuracy of 98.21% with only 24 features. Similarly, we report a slightly improved classification rate with the thyroid data set. The similarity classifier gives a classification accuracy of 98.84% with 4 features as compared to 95.45% reported in (Luukka and Leppälampi, 2006). In addition, these two data sets display the best values of specificity and sensitivity. For PIMA Indian and Parkinsons data set, we have fairly good classification results. However, the similarity classifier gives lowest values of specificity and sensitivity with these two data sets.

Table 4.9: Feature Selection Comparison for Dermatology

Method	% Accuaracy (Removed Features)	% Accuaracy (Removed Features)
Pasi Luukka (2011)	98.28 (13,23,24,30,31)	98.15 (31,33)
Proposed Method	98.38 (1,2,13,17,32,34)	98.21 (1,2,3,13,17,19,23,30,32,34)

Table 4.10: Feature Selection Comparison for PIMA

Method	% Accuaracy (Selected Features)	% Accuaracy (Selected Features)
Pasi Luukka(2011)	75.84 (1,2,3,4,5,7,8)	75.97 (1,2)
Proposed Method	77.47 (2,3,8)	77.73 (2)

Table 4.11: Feature Selection Comparison for Parkinsons

Method	% Accuaracy (Selected Features)	% Accuaracy (Selected Features)
Pasi Luukka(2011)	85.03 (19,20)	84.52 (19)
Proposed Method	87.24 (3,19,20)	87.76 (1)

Table 4.12: Feature Selection Comparison for Thyroid

Method	% Accuaracy (Selected Features)
Pasi Luukka(2006)	95.45 (1,2,3,4,5)
Proposed Method	96.21 (1,2,3,5)

It is observed that for most of the data sets, some of the features used for classification are the same for both methods. This is an important observation because features which give fuzzy similarity vectors with maximum fuzziness are equally most ambiguous and therefore not selected for classification by both methods. However, most ambiguous fuzzy similarity vectors are not necessarily most fuzzy. This provides a convincing justification why the proposed feature selection method performs better than entropy based approach.

4.3 Uncertainty in Class Assignment

We now report uncertainty due to ambiguity in class assignment. This form of uncertainty is important since it tells us how much information we lose and where. Since the similarity classifier discussed here assigns samples based on the degree of similarity of features of samples with typical attributes representing classification classes, it follows that samples with high similarity values with two or more classification classes are likely to be mis-classified. Average uncertainty for both correctly and wrongly classified samples are presented in tables (4.13-4.22).

(i) Dermatology Dataset

For this dataset, with 28 features, samples correctly classified display the lowest uncertainty than those correctly classified with both 34 and 24 features. Samples correctly and wrongly classified with 34 features display highest uncertainty as compared to the other two cases.

Table 4.13: Uncertainty for dermatology with 34 features

Parameter	Correct classification	Wrong classification
Average number of samples	175.1	3.9
Mean uncertainty	0.6755	0.7884
Variance	0.0091	0.0009

Table 4.14: Uncertainty for dermatology with 28 features

Parameter	Correct classification	Wrong classification
Average number of samples	175.8	3.2
Mean uncertainty	0.6102	0.7607
Variance	0.0116	0.0047

Table 4.15: Uncertainty for dermatology with 24 features

Parameter	Correct classification	Wrong classification
Average number of samples	176.1	2.9
Mean uncertainty	0.6336	0.7486
Variance	0.0112	0.0012

(ii) **PIMA Dataset**

This dataset displays very high values of average uncertainty for samples (correctly and wrongly classified). With 8 mean uncertainty for samples correctly and wrongly classified is higher than for 5 features and 3 features. Observe that with 8 features and 3 features, we have the same value of uncertainty, which is consistent with classification results obtained for this data set.

Table 4.16: Uncertainty with 8 features

Parameter	Correct classification	Wrong classification
Average number of samples	296.5	87.5
Mean uncertainty	0.9799	0.9890
Variance	0.0001	0.0000

Table 4.17: Uncertainty for PIMA with 3 features

Parameter	Correct classification	Wrong classification
Average number of samples	297.5	86.5
Mean uncertainty	0.9667	0.9689
Variance	0.0003	0.0002

Table 4.18: Uncertainty for PIMA with 1 feature

Parameter	Correct classification	Wrong classification
Average number of samples	298.5	85.5
Mean uncertainty	0.8904	0.9016
Variance	0.0011	0.0016

(iii) **Parkinsons Dataset**

The results obtained for this dataset are similar to those for PIMA dataset. However, we see a slight reduction in uncertainty for both correct and wrong classifications when 1 feature is used. Note that lower uncertainty for 1 feature implies smaller degree of overlap among classes based on the feature. This explains why we have better classification results with 1 feature for this data set.

Table 4.19: Uncertainty for Parkinsons with 22 features

Parameter	Correct classification	False classification
Average number of samples	83	15
Mean uncertainty	0.9927	0.9970
Variance	0.0000	0.0000

Table 4.20: Uncertainty for Parkinsons with 3 features

Parameter	Correct classification	False classification
Average number of samples	85.5	12.5
Mean uncertainty	0.9840	0.9900
Variance	0.0000	0.0000

Table 4.21: Uncertainty for Parkinsons with 1 feature

Parameter	Correct classification	False classification
Average number of samples	86	12
Mean uncertainty	0.9503	0.9721
Variance	0.0004	0.0004

(iv) **Thyroid Dataset**

As already seen, this dataset gives the best classification results. However, we still have relatively high uncertainty in class assignment. In fact, samples correctly wrongly classified display higher class ambiguity than those for dermatology dataset.

Table 4.22: Uncertainty for thyroid with 5 features

Parameter	Correct classification	Wrong classification
Average number of samples	105.5	2.5
Mean uncertainty	0.7730	0.8584
Variance	0.0036	0.0005

Table 4.23: Uncertainty for Thyroid with 4 features

Parameter	Correct classification	False classification
Average number of samples	106.5	1.1
Mean uncertainty	0.7666	0.8548
Variance	0.0038	0.0004

4.3.1 Discussion

The uncertainty reported in tables correspond to classification results discussed in previous previously. As desired, all datasets display high mean uncertainty in misclassified samples as compared to those correctly classified. Also, all the samples correctly classified posses class ambiguity of greater than 0.5. This is an important observation because it tells us that these observations carry properties of more than one class to a fairly good extent. have the lowest class ambiguity with dermatology data set, while the parkinsons and PIMA data sets display the highest class ambiguity. This result is consistent with classification results previously obtained. Further, we observe that these two data sets have the lowest variances in uncertainty. This further explains why classification rates for these two data sets were much lower as compared to the rest of the data sets.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This thesis presented a classification model based on the fuzzy unit hypercube. First, we have developed measures of uncertainty associated with classification problems using fuzzy specificity. Using the similarity classifier based on the generalized Lukasiecz structure, we have demonstrated how these measures can be used to determine uncertainty in class assignment. In addition, an efficient feature selection method has been proposed as well.

The model has been validated using medical data sets from UCI machine learning repository. We managed to obtain exceptionally good results with all data sets. Classification accuracies were better than those obtained in previous studies (Luukka and Leppälampi, 2006; Luukka, 2011). For example, with Dermatology data set we were able to significantly reduce features from 34 to 24 obtaining a classification rate of 98.15% as compared with 29 features with a classification rate of 98.29% given in(Luukka, 2011). For the Pima dataset, we obtained a classification accuracy of 76.32% with 3 features. Even though some data sets presented only a slight improvement in classification rate, the reduced number of features significantly reduces computational time and greatly enhances the model by reducing the number of measurements required. This makes disease screening faster, more convenient and less costly. In addition, models with fewer measurements are more transparent and more comprehensible, providing better explanations of suggested diagnosis, which is important in medical applications.

In addition to classification rate, we have reported uncertainty in class assignment due to ambiguity. All the data sets displayed greater values of average ambiguity for misclassified observations as compared to those correctly classified. This is true because ambiguous class assignments means observations are similar to an equal extent

to two or more classes and are therefore, likely to be misclassified. With PIMA and Parkinsons data sets, we have the highest values of uncertainty. This is consistent with classification results for the two data sets. Uncertainty of this nature gives us some insights on quality of the data sets.

5.2 Recommendations

The model proposed in this thesis uses a similarity classifier and feature selection, both of which are based on the generalized Lukasiewicz structure producing very impressive results. However, it is imperative to explore how other similarity classifiers compare with this classifier in terms of classification rate and computational cost. In addition, further research can include other forms of uncertainty such as discord, dissonance and probabilistic. The model can also be automated for use by physicians in clinical decision making.

REFERENCES

Abdulla, W. H. and Kasabov, N. (2003). Reduced feature-set based parallel chmm speech recognition systems. *Information Sciences*, 156(1):21–38.

Ali, A., Shamsuddin, S. M., Ralescu, A. L., and Visa, S. (2011). Fuzzy classifier for classification of medical data. In *Hybrid Intelligent Systems (HIS), 2011 11th International Conference on*, pages 173–178. IEEE.

Baccour, L., Alimi, A. M., and John, R. I. (2014). Some notes on fuzzy similarity measures and application to classification of shapes, recognition of arabic sentences and mosaic. *IAENG International Journal of Computer Science*, 41(2):81–90.

Bailey, K. O., Okolica, J. S., and Peterson, G. L. (2014). User identification and authentication using multi-modal behavioral biometrics. *Computers & Security*, 43:77–89.

Bezdek, J. C., Keller, J., Krisnapuram, R., and Pal, N. (2006). *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4. Springer Science & Business Media.

Blake, C. and Merz, C. (1998). Uci repository of machine learning databases [http://www. ics. uci. edu/ mlearn/mlrepository. html], department of information and computer science. *University of California, Irvine, CA*, 55.

Bloch, I. (1999). On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 32(11):1873–1895.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271.

Boicescu, V., Filipoiu, A., Georgescu, G., and Rudeanu, S. (1991). *Lukasiewicz-Moisil Algebras*, volume 49. Elsevier.

Booker, J. and Ross, T. (2011). An evolution of uncertainty assessment and quantification. *Scientia Iranica*, 18(3):669–676.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Cecotti, H. and Vajda, S. (2013). Rejection schemes in multi-class classification–application to handwritten character recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 445–449. IEEE.

Cheng, C.-H. (1998). A new approach for ranking fuzzy numbers by distance method. *Fuzzy sets and systems*, 95(3):307–317.

De Luca, A. and Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and control*, 20(4):301–312.

Dubois, D. and Prade, H. (1985). A note on measures of specificity for fuzzy sets. *International Journal of General System*, 10(4):279–283.

Dubois, D. and Prade, H. (2012). *Fundamentals of fuzzy sets*, volume 7. Springer Science & Business Media.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Fageth, R., Allen, W. G., and Jäger, U. (1996). Fuzzy logic classification in image processing. *Fuzzy sets and systems*, 82(3):265–278.

Fan, J., Xie, W., and Pei, J. (1999). Subsethood measure: new definitions. *Fuzzy Sets and Systems*, 106(2):201–209.

Firdausi, I., Erwin, A., Nugroho, A. S., et al. (2010). Analysis of machine learning techniques used in behavior-based malware detection. In *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, pages 201–203. IEEE.

Guha, D. and Chakraborty, D. (2010). A new approach to fuzzy distance measure and similarity measure between two generalized fuzzy numbers. *Applied Soft Computing*, 10(1):90–99.

Heikkila, E. J., Shen, T.-y., and Yang, K.-z. (2003). Fuzzy urban sets: theory and application to desakota regions in china. *Environment and Planning B: Planning and Design*, 30(2):239–254.

Helgason, C. M. (2007). The difference between a dynamic and mechanical approach to stroke treatment. *current treatment options in cardiovascular medicine*, 9(3):213–220.

Helgason, C. M. and Jobe, T. H. (1998). The fuzzy cube and causal efficacy: representation of concomitant mechanisms in stroke. *Neural Networks*, 11(3):549–555.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015.

Hoffman, A., Hoogenboezem, C., Van der Merwe, N., and Tollig, C. (1998). Seismic buffer recognition using mutual information for selecting wavelet based features. In *Industrial Electronics, 1998. Proceedings. ISIE'98. IEEE International Symposium on*, volume 2, pages 663–667. IEEE.

Huang, M.-J., Chen, M.-Y., and Lee, S.-C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert systems with applications*, 32(3):856–867.

Iyakaremye, C., Luukka, P., and Koloseni, D. (2012). Feature selection using yu's similarity measure and fuzzy entropy measures. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, pages 1–6. IEEE.

Jaganathan, P. and Kuppuchamy, R. (2013). A threshold fuzzy entropy based feature

selection for medical database classification. *Computers in Biology and Medicine*, 43(12):2222–2229.

Kang, H. and Vachtsevanos, G. (1993). Fuzzy hypercubes: Linguistic learning/reasoning systems for intelligent control and identification. *Journal of Intelligent and Robotic Systems*, 7(2):215–232.

Kaufman, A. (1975). Introduction to the theory of fuzzy subsets. vol. 1. fundamental theoretical elements. In *New York: Academic Press*.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier.

Klement, E. P., Mesiar, R., and Pap, E. (2003). Book review:" triangular norms". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(02):257–259.

Klir, G. J. (1987). Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like? *Fuzzy sets and systems*, 24(2):141–160.

Kosko, B. (1990). Fuzziness vs. probability. *International Journal of General System*, 17(2-3):211–240.

Krömer, P., Platoš, J., Snášel, V., and Abraham, A. (2011). Fuzzy classification by evolutionary algorithms. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 313–318. IEEE.

Lee, H.-M., Chen, C.-M., Chen, J.-M., and Jou, Y.-L. (2001). An efficient fuzzy classifier with feature selection based on fuzzy entropy. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(3):426–432.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502.

Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, 38(4):4600–4607.

Luukka, P. and Leppälampi, T. (2006). Similarity classifier with generalized mean applied to medical data. *Computers in biology and medicine*, 36(9):1026–1040.

Marsland, S. (2011). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.

McBratney, A. B. and Odeh, I. O. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, 77(2-4):85–113.

Mitra, P., Murthy, C., and Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312.

Moore, T. S., Campbell, J. W., and Feng, H. (2001). A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. *IEEE Transactions on Geoscience and Remote sensing*, 39(8):1764–1776.

Nahar, J., Imam, T., Tickle, K. S., and Chen, Y.-P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1):96–104.

Nauck, D. and Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine*, 16(2):149–169.

Nieto, J. J. and Torres, A. (2003). Midpoints for fuzzy sets and their application in medicine. *Artificial Intelligence in Medicine*, 27(1):81–101.

Nieto, J. J., Torres, A., Georgiou, D., and Karakasidis, T. (2006). Fuzzy polynucleotide spaces and metrics. *Bulletin of mathematical biology*, 68(3):703–725.

Pal, N. R. and Pal, S. K. (1992). Higher order fuzzy entropy and hybrid entropy of a set. *Information Sciences*, 61(3):211–231.

Parveen, P. and Thuraisingham, B. (2006). Face recognition using multiple classifiers. In *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*, pages 179–186. IEEE.

Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences*, 11(5):341–356.

Pedrycz, W. (1991). Fuzzy logic in development of fundamentals of pattern recognition. *International Journal of Approximate Reasoning*, 5(3):251–264.

Qing, M. and Li, T.-R. (2004). Some properties and new formulae of fuzzy entropy. In *Networking, Sensing and Control, 2004 IEEE International Conference on*, volume 1, pages 401–406. IEEE.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rosenfeld, A. (1985). Distances between fuzzy sets. *Pattern Recognition Letters*, 3(4):229–233.

Sadegh-Zadeh, K. (1999). Fundamentals of clinical methodology: 3. nosology. *Artificial intelligence in medicine*, 17(1):87–108.

Saha, P. K. and Wehrli, F. W. (2004). Measurement of trabecular bone thickness in the limited resolution regime of in vivo mri by fuzzy distance transform. *IEEE transactions on medical imaging*, 23(1):53–62.

Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240.

Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, 40(5):103–110.

Torres, A. and Nieto, J. J. (2003). The fuzzy polynucleotide space: basic properties. *Bioinformatics*, 19(5):587–592.

Wang, D.-G., Meng, Y.-P., and Li, H.-X. (2008). A fuzzy similarity inference method for fuzzy reasoning. *Computers & Mathematics with Applications*, 56(10):2445–2454.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.

Weber, S. (1983). A general concept of fuzzy connectives, negations and implications based on t-norms and t-conorms. *Fuzzy sets and systems*, 11(1-3):115–134.

Yager, R. R. (1979). On the measure of fuzziness and negation part i: membership in the unit interval. *International Journal of General Systems*, 5(4):221–229.

Yager, R. R. (1982). Measuring tranquility and anxiety in decision making: An application of fuzzy sets. *International Journal of General Systems*, 8(3):139–146.

Yager, R. R. (1995). On a measure of ambiguity. *International journal of intelligent systems*, 10(11):1001–1019.

Yager, R. R. (1998). Measures of specificity. *Computational intelligence: Soft computing and fuzzy-neuro integration with applications*, pages 94–113.

Yu, S., De Backer, S., and Scheunders, P. (2000). Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for high-dimensional remote sensing data. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 3, pages 1912–1916. IEEE.

Yuan, Y. and Zhuang, H. (1996). A genetic algorithm for generating fuzzy classification rules. *Fuzzy sets and systems*, 84(1):1–19.

Zabalgoitia, M., Halperin, J. L., Pearce, L. A., Blackshear, J. L., Asinger, R. W., Hart, R. G., in Atrial Fibrillation III Investigators, S. P., et al. (1998). Transesophageal echocardiographic correlates of clinical risk of thromboembolism in nonvalvular atrial fibrillation. *Journal of the American College of Cardiology*, 31(7):1622–1626.

Zadeh, L. A. (1968). Probability measures of fuzzy events. *Journal of mathematical analysis and applications*, 23(2):421–427.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information sciences*, 8(3):199–249.

Zedeh, L. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Zhang, H. and Liu, D. (2006). *Fuzzy modeling and fuzzy control*. Springer Science & Business Media.

Zhu, X. and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210.

Zimmermann, H.-J. (2011). *Fuzzy set theory and its applications*. Springer Science & Business Media.