Open your mind. LUT.
Lappeenranta University of Technology

Lappeenranta University of Technology

Faculty of Technology

Department of Mathematics and Physics

Feature selection using Fuzzy Entropy measures with Yu's Similarity measure

The topic of this Master's thesis was approved by department council of the Department of Mathematics and Physics on 22^{nd} February, 2012.

The supervisors and examiners of the thesis were
PhD Pasi Luukka and Mr. David Koloseni.

Lappeenranta, March 20, 2012.

Cesar Iyakaremye
Liesharjunkatu9 A8
53850 Lappeenranta, FINLAND
+358465959753
Cesar.Iyakaremye@lut.fi

Lappeenranta University of Technology
Department of Mathematics and Physics

Cesar Iyakaremye

Feature selection using fuzzy entropy measures with Yu's similarity measure.
Master Thesis
2012
50 pages, 6 figures, 8 tables

Abstract

In this study, feature selection in classification based problems is highlighted. The role of feature selection methods is to select important features by discarding redundant and irrelevant features in the data set, we investigated this case by using fuzzy entropy measures. We developed fuzzy entropy based feature selection method using Yu's similarity and test this using similarity classifier. As the similarity classifier we used Yu's similarity, we tested our similarity on the real world data set which is dermatological data set. By performing feature selection based on fuzzy entropy measures before classification on our data set the empirical results were very promising, the highest classification accuracy of 98.83% was achieved when testing our similarity measure to the data set. The achieved results were then compared with some other results previously obtained using different similarity classifiers, the obtained results show better accuracy than the one achieved before. The used methods helped to reduce the dimensionality of the used data set, to speed up the computation time of a learning algorithm and therefore have simplified the classification task.

Keywords:Classification, Entropy measure, Feature selection

Acknowledgements

My sincerest thanks to Lappeenranta University of Technology through the Department of Mathematics and Physics for providing financial support and all the efforts to make our studies possible. I highly acknowledge all your deeds. My highest gratitude is addressed to the staff of the department of Applied Mathematics especially Prof. Heikki Haario, Prof. Matti Heiliö, and Prof. Tuomo Kauranne.

I'm particularly indebted to Dr.Pasi Luukka for the professional supervision of this thesis, your excellent and scientific guidance, comments, advice are mostly the fruits of this thesis, thanks. Many thanks are also addressed to my second supervisor Mr. David Koloseni.

My deepest thanks to Matylda for the useful information you provided me when i was filling the master's application form, your kind encouragement in my studies, your assistance in Latex,matlab and especially the time you spent reviewing my thesis are all unforgettable. My highest gratitude goes to Idrissa for your help in matlab and latex, i can't forget the family environment you and your wife provided me with your lovely children . I can't forget your guidance,advises and help in my studies, your best wishes and blessing have accompanied me in all, accept my highest recognition Prof. Verdiana Grace Masanja.

Last but not least, I would like to thank my brothers David Habiyaremye and Kassim Ndindabahizi for the financial and moral support you provided me. Many thanks to my parents, brothers sisters and relatives for your endless love and care. Many thanks to all friends in Finland.

Lappeenranta, March 20, 2012.

Contents

List of Tables

List of Figures

1 Introduction

Many real world problems are characterized by large dimensionality data sets, most of them are burdened with uncertainty of different kinds, the analysis and the learning from them requires a preprocessing method in order to discard uncertainties. In machine learning one of the crucial problem when dealing with big data sets is the selection of the relevant features and elimination of non important features. In addressing this problem different methods of data reduction have been used and managed to eliminate the redundancy and non-important features present in the data sets. Among them feature selection (FS) has been shown to be a powerful approach of dealing with high dimensional data by selecting relevant features from data set at the same time removing irrelevant and (or) redundant (highly correlated with others) features that harm the quality of the results, and therefore build a good learning model [18]. A good feature selection techniques should be able to detect and model the noisy and misleading features from the domain problem and help to get a minimal feature subsets but still keep the important information present in the original data [19].

Feature selection methods have been successfully used in many areas such as machine learning, pattern recognition, systems control, and signal processing [19]. Interesting applications are found in bioinformatics [17] especially in medical diagnosis to reduce the size of features collected during the clinical testings and experiments [2], physicians are confronted with massive data sets which are ranged from simple blood pressure and heart rate to magnetic resonance imaging and electronencephalogram waveforms, they have to deal with them correctly in order to avoid the cost associated with misdiagnosis, failure to diagnose or delayed diagnosis and therefore improve accuracy in well treating and serving the patients.

As far as medical data sets are concerned, in this study dermatological data about erythemato-squamous diseases will be analyzed to test our model. This data set was successively used in classification by different researchers: in 1988 Güvernir developed a classification algorithm VFI5(for Voting Feature Intervals) and apply it to the differential diagnosis of erythemato-squamous diseases, he built up a genetic algorithm which he combined with the VFI5 algorithm to determine the weight of the features in the domain of differential diagnosis of erythemato-squamous, the obtained weights facilitated the VFI5 algorithm in a sense that 99.2% classification accuracy was achieved [14]. Further research in 2005 by Übeyli and Güler showed a new approach based on adaptive neuro-fuzzy inference system (ANFIS) for the detection of erythemato-squamous diseases, the ANFIS model was assessed in terms of training performance and classification accuracies and showed to perform well in detecting erythemato-squamous diseases, the model achieved a total classification accuracy of 95.5% which is concluded to be good in comparison to the one of 85.5% achieved with the stand-alone neural network [9].

In [24] Pasi Luukka and Leppälampi presented a new approach based on similarity classifier with generalized mean and applied it medical data: the presented method managed to detect erythemato-squamous diseases, a good mean classification accuracy of 97.02% was obtained. In recent published study [26] fuzzy entropy measures were used in feature selection. This method successfully managed to discard the non-important features in the data sets, this has positively facilitated the classification task which was done by using the similarity based on Lukasiewicz structure where a mean accuracy of 98.28% was achieved. In this study wrapper feature selection method based on fuzzy entropy measures is performed to get rid of unwanted features present in the data set [27], the use of fuzzy entropy based-feature selection will facilitate our classification task to be performed faster and to increase the classification accuracy. We will test the efficiency of the similarity classifier constructed from Yu's norm [20] [25] on dermatological data set. $Matlab^{TM}$ software will be used for the computation purpose. The structure of this thesis is organized as follows: In Chapter 1 we give general introduction to the feature selection methods, its usefulness in many areas of research especially in medicine. Mathematical background on fuzzy sets theory and fuzzy data analysis methods is presented in Chapter 2. Feature selection methods: filter, wrapper and embedded methods are briefly introduced in Chapter 3. In Chapter 4 we will present the classification and similarity classifier methods. Next comes Chapter 5 which comprises the data sets and its properties, then the obtained results using Matlab software. Finally in Chapter 6, a comparison of our results with some other previously obtained is done, conclusion and suggestions for future work are also presented in this part.

2 Mathematical background

Various mathematical methods were successfully used for year to define, structure and to solve real life problems. Among them, great contribution of statistical and optimization methods in different areas such as medicine, economics etc. In addressing solutions to some problems researchers confronted various challenges whereby some problems presented various types of uncertainties [16] which were coming from many sources, more uncertainty in the problem the less precise was its understanding. The sources of uncertainties can be such as: errors of measurement, deficiency in history and statistical data, insufficient theory, subjectivity and preference of human judgements etc, among different uncertainties we have randomness of occurrence of events, imprecision vagueness and ambiguity. Different types of uncertainties can be categorized as stochastic and fuzziness, stochastic uncertainty are related to the occurrence of event and the stochastic systems related are solved by using probability theory, on the other hand fuzziness uncertainty are originated from vagueness of human language and behavior, impreciseness and ambiguous in the system of data whereby the information could not be well described and defined due to its limited knowledge and deficiency. Lotfi A. Zadeh [31] introduced the Fuzzy sets theory and fuzzy logic which was specifically the mathematical representation of uncertainty and vagueness and provide formalized tools for dealing with the imprecision intrinsic to many problems which are perception based, he allowed uncertainty to exist in the characteristic function, this made fuzzy sets theory to be the extension of the classical set theory where instead of an element in the universe to be a member or non-member he added the degree of membership. Various useful applications and solutions have been provided by fuzzy sets to many real life problems and its usefulness has been growing rapidly since fuzzy sets theory was found to be a good bridge to characterize and quantify the uncertainty within different areas, some of the pertinent applications are found in approximate reasoning, fuzzy pattern recognition, fuzzy modeling, expert system, fuzzy control and fuzzy arithmetic, etc. In his paper [31], Zadeh addressed much on the set membership as key to decision making when faced with uncertainty [29], therefore all the operations on fuzzy sets are defined based on the membership function which is considered as the gradual property for fuzzy set.

2.1 Crisp set versus Fuzzy set

Let A be a crisp set defined over a Universe X, the classical set theory is built on the fundamental concept such that element is either a member A or not, this concept can be clarified using the characteristic function (membership function) $\mu_A(x)$ taking only two values 1 to indicate if an element $x \in X$ is a member of A and 0 otherwise:

$$\mu_A(x) = \begin{cases} 1 & \text{for } x \in A \\ 0 & \text{for } x \notin A \end{cases} \tag{1}$$

In fuzzy set theory this property is generalized by accepting even partial membership of a set, this make the fuzzy set theory to be an extension of the classical (crisp) set theory.

Example 1 *consider the universe* $X = \{a, b, c, d, e, f, j\}$ *and its subsets* $A = \{b, d, e\}$. *Only three elements of six in* X *are members of* A. *We have:*$\mu_A(b) = \mu_A(d) = \mu_A(e) = 1$ $\mu_A(a) = \mu_A(c) = \mu_A(f) = 0$

Example 2 *In a basketball team the coach wants to select the tall players: players with* $2.09m$ *are obviously qualified, in order for the coach to select some other players he needs to base on their degree of tallness, player whose height is* $1.8m$ *is not as tall as well as the one with* $2.03m$ *, as the height* *increases the membership grade increases, this example is presented in Figure 1.*

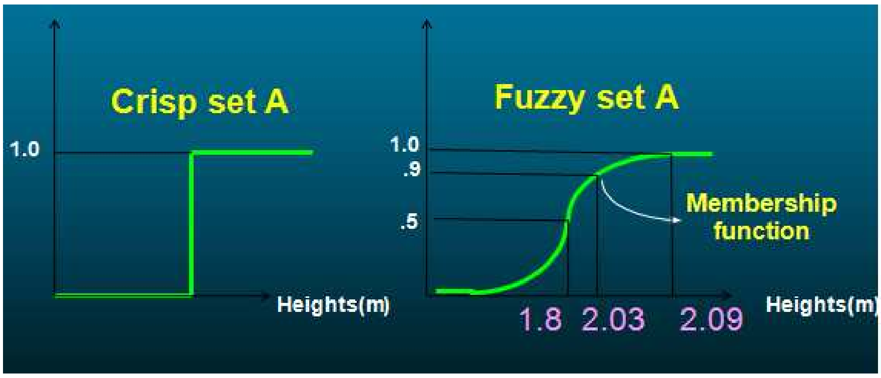


Figure 1: Crisp set, Fuzzy set.

Definition 1 (Fuzzy set) *If we allow our valuation set* $\{0, 1\}$ *to be the real interval* $[0, 1]$ *then* A *is called a Fuzzy set [16], [5] [12].*
The membership function of fuzzy set is denoted by:μ_A*; that is* $\mu_A : X \to [0, 1]$.

$\mu_A(x)$ *is the degree to which* $x \in A$*, the closer the value of the degree of membership* $\mu_A(x)$ *is to* 1*, the more* x *belongs to* A.

Notice that A *is completely determined by the set of ordered pairs:* $A = (x, \mu_A(x)), x \in X$.

Example 3 *Suppose that we want to classify* 5 *people: John, Petter, Mary, Bob and Bill who drink beer using the property " being drunkard". Let A be fuzzy set that describes them,*
$A = \{(John, 1)(Peter, 0.2)(Mary, 0.1)(Bob, 0.8)(Bill, 0.9)\}$. *Obviously, John is more drunkard than everyone because his degree of drunkenness is high i.e his degree of membership is* 1*, whereas Mary is less drunkard with membership degree* 0.1.

Zadeh proposed a more convenient way of notation for a fuzzy set A.

Definition 2 *When X is a finite set $x_1, ..., x_n$, a fuzzy set on X is noted as*

$$A = \mu_A(x_1)/x_1, ..., \mu_A(x_n)/x_n = \sum_{i=1}^{n} \mu_A(x_i)/x_i. \tag{2}$$

When X is not finite, we write

$$A = \int_x \mu_A(x)/x \tag{3}$$

Example 4 *$X = \mathbb{N}$: positive integers.*
Let $A = \{0.1/7 + 0.5/8 + 0.5/9 + 1.0/10 + 0.8/11 + 0.5/12 + 0.1/13\}$.
A is a fuzzy set of integers approximately equal to 10.

Example 5 *$X = \mathbb{R}$: real numbers.*
Let $\mu_A(x) = \frac{1}{1+[\frac{1}{5}(x-10)]^2}$

2.2 Basic properties of fuzzy sets

In the following we will define some basic properties of fuzzy sets theory [31], [4]

Definition 3 (Identity of two fuzzy sets) *Two fuzzy sets A and B are identical, denote $A = B$ iff $\forall x \in X : \mu_A(x) = \mu_B(x)$.*

Definition 4 (Inclusion) *A fuzzy set A is a subset of B, denotes $A \subseteq B$ iff $\forall x \in X : \mu_A(x) \leq \mu_B(x)$*

Definition 5 (Convexity) *A fuzzy setA is convex if the membership function of μ_A is quasi-convex, i.e $\forall x, y \in X$ and $\lambda \in [0, 1]$ the condition $\mu_A(\lambda x + (1 - \lambda)y) \geq \min(\mu_A(x), \mu_A(y))$ is satisfied.*

Definition 6 (Support) *Let A be a fuzzy subset of X, the support of A denoted supp(A) is the crisp subset of X whose elements all have none zero membership grades in A.*
That is, $supp(A) = \{x \in X : \ \mu_A(x) > 0\}$.

Definition 7 (Core) *We define the core of a fuzzy set as the crisp subset of X such that $\mu_A(x) = \ 1$,*
that is, $core(A) = \{x \in X : \ \mu_A(x) = 1\}$.

Definition 8 (Width) *The width of a fuzzy set A is $w(A) = sup[supp(A)] - inf[supp(A)]$*

Definition 9 (Height) *The height of a fuzzy set A is the number $hgt(A) = sup_{x \in X}\{\mu_A(x)\}$*

Definition 10 (Normality) *A fuzzy set A of a classical (crisp) set X is said to be normal if there exists an $x \in X$ such that $\mu_A(x) = \ 1$ or simply $hgt(A) = 1$, otherwise A is subnormal.*

Points $x \in X$ with $\mu_A(x) = \frac{1}{2}$ are called cross-over points. The empty sets $\emptyset$ and the Universe X are incorporated by $\forall x \in X$:$\mu_\emptyset(x) = 0$,$\mu_X(x) = 1$.

Notice that if a fuzzy set is not normal means $hgt(A) < 1$ and $core(A) = \emptyset$, it can be normalized for example by stretching mapping. One such a stretching mapping is

$$x \longmapsto \frac{A(x)}{hgt(A)} \tag{4}$$

The height, the core and support of a fuzzy set are shown in following Figure 2.

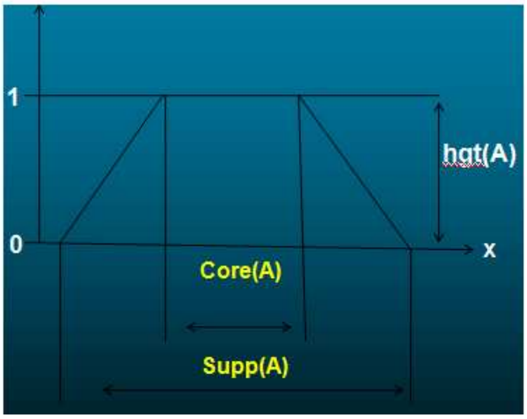


Figure 2: Height, support and core of a fuzzy set A.

Example 6 *To illustrate the previous definitions consider:*
a fuzzy set A such that $A(x) = \frac{x-4}{5}, x \in [4,8]$*and* 0 *elsewhere . what is the core, support and height?*

1. $core(A) = \{x \in X : \ \mu_A(x) = 1\}, core(A) = \emptyset.$

2. $supp(A) = \{x \in X : \ \mu_A(x) > 0\}, supp(A) = (4,8).$

3. $hgt(A) = sup_{x \in X}\{\mu_A(x)\}, hgt(A) = \frac{4}{5}.$
   Since our fuzzy set is not normalized we can normalize it by

4. $\frac{A(x)}{hgt(A)} = \frac{x-4}{4}$

2.3 Union, intersection and complement

Definition 11 *Given three fuzzy sets A and B on the universe X. For a given element x of the universe X, the following function-theoretic operations for the set-theoretic operations of union, intersection, and complement are defined for A, B on X:*

1. $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$*: Union.*
2. $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$*: Intersection.*
3. $\mu_{\overline{A}}(x) = 1 - \mu_A(x)$*: Complement.*

These operations are known as the standard fuzzy operations.

Let us illustrate these operations numerically by an examples:

Example 7 *: Given the universe set $X = \{1, 2, 3, 4, 5\}$, let A and B be two discrete fuzzy sets,*

$A = \frac{1}{2} + \frac{0.5}{3} + \frac{0.3}{4} + \frac{0.2}{5}$ *and* $B = \frac{0.5}{2} + \frac{0.7}{3} + \frac{0.2}{4} + \frac{0.4}{5}$
Using the previous definitions we can calculate:

1. *Complement:* $\overline{A} = \frac{1}{1} + \frac{0}{2} + \frac{0.5}{3} + \frac{0.7}{4} + \frac{0.8}{5}$.
   $\overline{B} = \frac{1}{1} + \frac{0.5}{2} + \frac{0.3}{3} + \frac{0.8}{4} + \frac{0.6}{5}$.
2. *Union:* $A \cup B = \frac{1}{2} + \frac{0.7}{3} + \frac{0.3}{4} + \frac{0.4}{5}$.
3. *Intersection:* $A \cap B = \frac{0.5}{2} + \frac{0.5}{3} + \frac{0.2}{4} + \frac{0.2}{5}$.
4. *Difference:* $A|B = A \cap \overline{B} = \frac{0.5}{2} + \frac{0.3}{3} + \frac{0.3}{4} + \frac{0.2}{3}$.
   $B|A = B \cap \overline{A} = \frac{0}{2} + \frac{0.5}{3} + \frac{0.2}{4} + \frac{0.4}{3}$.

The standard fuzzy sets operations are the same as those for classical sets when the range of membership values is restricted to the unit interval. However, these operations are not the only ones which can be applied to fuzzy sets. These standards operation can be generalized to a broad class of functions whose members can be considered as their fuzzy generalization [29]: these functions are quantified as fuzzy intersection and fuzzy union and are referred in the literature as t-norms and t-conorms (or s-norms).

2.4 t-norm and t-conorm

The triangular norms (t-norm) and triangular conorms (t-conorms), which generalize the form of intersection and union, are next well described and later will be used to construct our similarity measure:

For any x, y, z and $u \in [0, 1]$

Definition 12 (t-norm) *A two-place function $T : [0,1] \times [0,1] \rightarrow [0,1]$ is called t-norm if the following conditions are satisfied:*

1. *$T(x, 1) = x$: one identity;*
2. *$x \leq z, y \leq u \Rightarrow T(x, y) \leq T(z, u)$: monotonicity;*
3. *$T(x, y) = T(y, x)$: commutativity;*
4. *$T(T(x, y), z) = T(x, T(y, z))$: associativity.*

Definition 13 *A t-norm is called Archimedean if and only if T is continuous and $\forall x \in [0, 1] : T(x, x) < x$.*

Example 8 *Find out whether the algebraic product $T(x, y) = xy$ is a t-norm.*

We have to verify if the t-norm conditions are verified by using the given rule:
for any $x, y, z and\ d \in [0, 1]$

1. $T(x, 1) = x1 = x$*: identity;*
2. $if x < z and\ y \geq d$ *then* $T(x, y) = xy \geq zd$ *and* $T(x, y) \geq T(z, d)$*: monotonicity;*
3. $T(x, y) = xy = yx = T(y, x)$*: commutativity;*
4. $T(x, T(y, z)) = xT(y, z) = x(yz) = (xy)z = T((xy)z) = T(T(x, y), z)$*: associativity.*

Definition 14 (t-conorm) *A two-place function* $Sn : [0, 1] \times [0, 1] \rightarrow [0, 1]$ *is called t-conorm if the above conditions are satisfied:*

1. $Sn(x, 0) = x$*: zero identity;*
2. $y \leq z, y \leq u \Rightarrow Sn(x, y) \leq Sn(z, u)$*: monotonicity;*
3. $Sn(x, y) \leq Sn(y, x)$*: commutativity;*
4. $Sn(Sn(x, y), z) = Sn(x, Sn(y, z))$*: associativity.*

Example 9 *:*
Prove whether the given expression $Sn(x, y) = x + y - xy$ *is a t-conorm, it means we have to verify whether the conditions of t-conorm are satisfied:*

1. $Sn(x, 0) = x + 0 - x0 = x$*: zero identity;*
2. $y \leq z, y \leq u \Rightarrow Sn(x, y) \leq Sn(z, u)$*: monotonicity;*
3. $Sn(x, y) = x + y - xy = y + x - yx = Sn(y, x)$*: commutativity;*

4. *$Sn(Sn(x,y),z) = x+y+z-xy-xz-yz+xyz = x+(y+z-yz)-x(y+z-yz) = Sn(x,Sn(y,z))$: associativity.*

Notice that t-norms are functions which are called fuzzy intersections and unions are the common shorthand term for triangular norms, t-norm and t-conorm only differ on their boundary conditions. Some additional properties of t-norm and t-conorm are presented in the following definitions [20].

Definition 15 (Continuity) *T and Sn are continuous functions:*
A t-norm $T : [0,1] \times [0,1] \Rightarrow [0,1]$ is continuous if for all convergent sequences $(x_n)_n \in \mathbb{N}(Y_n)_n \in \mathbb{N} \in [0,1]^n$ we have $T(\lim_{x\to\infty} x_n \lim_{x\to\infty} y_n) = \lim_{x\to\infty} T(x_n, y_n)$

Definition 16 *A continuous t-norm that satisfies this condition:*
$T(x,x) < x$ (for t-norm) or $Sn(x,x) > x$ for t-conorm $\forall x \in [0,1]$ is called an Archimedean t-norm (respectively t-norm or t-conorm).

Definition 17 *A t-norm (t-conorm) is strict [1] if it is continuous on $[0,1]^2$ and strictly increasing in each place on $[0,1]^2$ so that $T(x_1,y) < T(x_2,y)$, whenever $x_1 < x_2, y > 0$, $T(x,y_1) < T(x,y_2)$, whenever $x > 0, y_1 < y_2$.*

Definition 18 (Duality of t-norms) *A function*
$Sn : [0,1] \times [0,1] \rightarrow [0,1]$ is dual t-conorm of t-norm such that for all $x, y \in [0,1]$ both the following equivalent equalities hold

1. *$Sn(x,y) = 1 - T(1-x, 1-y)$*
2. *$T(x,y) = 1 - Sn(1-x, 1-y)$,*
   where $(1-x)$ and $(1-y)$ are respectively complements of x and y.

The duality of t-norms is of a great importance since it combines t-norms together and helps to transform t-norm to t-conorm and vice-versa, also duality helps to change the order in a way that if two t-norms are ordered as $T_1 \leq T_2$ then the corresponding t-conorms are ordered as $S_1 \geq S_2$.

Triangular norms have been of great use and investigation i.e starting from pioneering work of Schweizer/Sklar (1961, 1983) and Ling (1965). We will focus our attention mainly on the properties of t-norms with the idea in mind that the similar results for t-conorm can be obtained by the duality relations [18].

Example 10 *:*

Einstein sum is a t-conorm function, its bivariate form is given by $Sn(x_1, x_2) = \frac{x_1+x_2}{1+x_1x_2}$ *by using the dual form of the t-conorm we get the corresponding t-norm which is*
$T(x_1, x_2) = \frac{x_1x_2}{2-x_1-x_2+x_1x_2}$

Next we present a list of the main well know and most frequently used t-norms [12], [20]:

$$T_{min}(x, y) = min(x, y) : \text{Minimum}; \tag{5}$$

$$T_{prod}(x, y) = xy : \text{Algebraic product}; \tag{6}$$

$$T_{bprod}(x, y) = max(0, x + y - 1) : \text{Bounded product}; \tag{7}$$

$$T_{Ds}(x,y) = \begin{cases} x: & \text{when } y = 0 \\ y: & \text{when } x = 0 \\ 1: & \quad otherwise \end{cases} \text{ : Drastic Sum;} \tag{8}$$

$$T_{\alpha}^{H}(x,y) = \frac{xy}{\alpha + (1-\alpha)x + y - xy}, \alpha \geq 0 : \text{Hamacher's } t\text{-norm ;} \tag{9}$$

$$T_{\beta}^{F}(x,y) = \log_{\beta}(1 + \frac{(\beta^{x}-1)(\beta^{y}-1)}{\beta - 1}), \beta > 0, \beta \neq 1 : \text{Frank's } t\text{-norm;} \tag{10}$$

$$T_{\gamma}^{Y}(x,y) = 1 - min((1-x)^{\gamma} + (1-y)^{\gamma})^{\frac{1}{\gamma}}, 1, \gamma > 0 : \text{Yager's } t\text{-norm;} \tag{11}$$

$$T_{k}^{Do}(x,y) = 1 - \frac{1}{1 + ((\frac{1-x}{x})^{k} + (\frac{1-y}{y})^{k})^{\frac{1}{k}}}, k > 0 : \text{Dombi's } t\text{-norm ;} \tag{12}$$

$$T_{\theta}^{W}(x,y) = max(0, \frac{x + y - 1 + \theta xy}{1 + \theta}), \theta > -1 : \text{Weber's } t\text{-norm ;} \tag{13}$$

$$T_{SS}^{1}(x,y) = max(0, (x^{p} + y^{p} - 1)^{\frac{1}{p}}), p > 0 : Schweizer - \text{Sklar's } t\text{-norm;} \tag{14}$$

$$T_{\lambda}^{Yu}(x,y) = max(0, (1+\lambda)(x+y-1) - \lambda xy), \lambda > -1 : \text{Yu's } t\text{-norm} \tag{15}$$

By using the duality we can easily establish the Yu's t-conorm, which is

$$S_{\lambda}^{Yu}(x,y) = min(1, x + y + \lambda xy), \lambda > -1 \tag{16}$$

2.5 Fuzzy relation

As the fuzzy set is the extension of crisp set the concept and properties of crisp relations are also generalized to fuzzy sets relations [29].

Definition 19 (Fuzzy relation) *Let X and Y be non-empty sets, a fuzzy relation R is a mapping from the cartesian space $X \times Y$ to the interval $[0,1]$, where the 'strength' of the relation is expressed by the membership function $\mu_R(x,y)$ of the relation for ordered pairs (x,y) respectively from the two sets.*

When $X = Y$ our fuzzy relation R is called a binary fuzzy relation.

Definition 20 (Fuzzy Cartesian product) *Let A be a fuzzy set for the universe X and B, a fuzzy sets from the universe Y. The Cartesian product between A and B results in a fuzzy relation R: $A \times B = R \subset X \times Y$, where the fuzzy relation has membership $\mu_R(x,y) = \mu_{A\times B}(x,y) = \min(\mu_A(x), \mu_B(y))$*

Definition 21 (Similarity relation) *A fuzzy relation T on X is called similarity relation on X the following axioms hold:*

1. *Reflexivity: $\forall x_1 \in X : \mu_T(x_1, x_1) = 1$: Every object is completely similar to itself;*
2. *Symmetry: $\forall x_1, x_2 \in X : \mu_T(x_1, x_2) = \mu_T(x_2, x_1)$: the degree in which x_1 is similar to x_2 coincides with the degree in which x_2 is similar to x_1;*
3. *$\forall x_1, x_2, x_3 \in X : \mu_T(x_1, x_3) \geq \max_{x_2}(\min(\mu_T(x_1, x_2), \mu_T(x_2, x_3)))$. If x_1 is similar to x_2 and x_2 is similar to x_3 then x_1 is similar to x_3*

A fuzzy binary relation which is is reflexive, symmetric and transitive is known as a fuzzy equivalence relation or similarity relation [21]. The idea behind fuzzy equivalence relation is to compare elements according to their grade of equivalence and classify them. We can use this knowledge in our classification task to construct classes of objects which are similar or equivalent. Since the theoretical operations of fuzzy set are the base for fuzzy logical operation, we will use fuzzy logic equivalence to establish the law which will help us to compare similar objects in our classification task. In the book [22] Lowen defined different ways classical logic connectives can be extended to fuzzy logic, he showed how the notion of t-norms, t-conorms and negation can be combined to derive the equivalent relation of the form:

$$E(\mu_a(x), \mu_b(x)) = T(S_n(\mu_{\overline{a}}(x), \mu_b(x), S_n(\mu_{\overline{b}}(x), \mu_a(x))) \tag{17}$$

$\forall \mu_a(x), \mu_b(x) \in [0, 1]$, where S_n denotes fuzzy union and T denotes fuzzy intersection, a and b are fuzzy sets, $\overline{a}$ and $\overline{b}$ are respectively the complements of a and b. This equation (17) will be used later to construct our similarity measure.

3 Feature Selection Methods

Feature is any aspect quality or characteristics of any object, usually in big data sets not all the features are important to describe the target concept that is why feature selection method is needed to select a subset of the original feature present in a given data set that provides most useful information. The feature selection task can be formulated as follows: given a feature set $Y = (y_1, y_2, ..., y_n)$ find a subset $Z = (y_1, y_2, ..., y_k)$ of Y with $k < n$, which optimizes an objective function W(Y).

The process of feature selection is very important because it reduces the dimensionality of the data and enables learning algorithms to operate faster (reduction of the computation time) and more efficiently and, therefore, increases the accuracy of the resulting model. Feature selection is performed using feature selection algorithm which in general comprises the following components [32]: search strategy: feature space is searched and a subset select from the candidates: i.e sequential feature selection. The forwards feature selection begins with an empty sets of features (zero features), evaluates all features subsets and select the ones with best performance criteria. The backwards feature selection starts with all features and repeatedly removes features this way improving best the performance criteria. In general, machine learning provides the technical basis of data mining for feature subset selection, which can be grouped as [18]: filters, wrappers, and embedded techniques.

3.1 Filter Methods

In filter techniques, we evaluate the relevant of features based on some discriminating criterion that looks at the general characteristics of the data with the idea of producing the most promising subset before learning commences.

It is a preprocessing step where undesirable features that have little chance to be useful in the analysis of data are filtered out through checking data consistency and elimination of features whose information is represented by others or considered as irrelevant. These methods do not use the learning algorithm or large data set. The results of such method are usually a ranked list of features where at the top of the list are relevant features and at the bottom of the list are not so relevant or totally irrelevant features. Filter methods provide the cheapest approach to the evaluation of feature relevance, moreover filter methods performed before help wrapper and/or embedded methods to be more feasible.

3.2 Wrapper methods

Wrapper techniques incorporate the learner (classifier) in the process of selecting the most relevant subset features, and may improve the overall machine learning algorithm performance. Wrapper methods use the learning machine to measure the quality of subsets of features without incorporating knowledge about specific structure of classification function and can therefore be combined with any learning machine. In wrapper methods we search for an optimal feature subset trough testing the performance of candidates subsets using the learning algorithm. However this process is proved to be slower than the filter methods because the induction algorithm is repeatedly called. Wrapper methods are known to be more accurate than filters due to the fact that they are oriented to the specific interaction between an induction algorithm and its training data but it is more computationally expensive, and do not scale up well high dimension data set.

3.3 Embedded methods

Embedded methods differ from others feature selection method in the way feature selection and the learning interact. Filter methods do not incorporate learning while wrapper methods use a learning machine to measure the quality of subsets of features without incorporating knowledge about the specific of structure of the classification or regression function, and can therefore be combined with any learning machine. In contrast to the above two methods, embedded methods do not separates the learning from the feature selection part, the structure of the class under consideration plays a crucial role.

Notice that every family of feature selection methods (filter, wrapper and embedded) has its own advantages and drawbacks. In general, filter methods are fast since they do not incorporate learning. Most wrapper methods a search for optimal features, the learning algorithm is called repeatedly and this make the wrapper methods to be slower than the filter methods, the embedded methods are faster than the wrapper methods. Embedded methods tend to have higher capacity than filter methods and are therefore more likely to overfit.

We expect filter methods to perform better if only small amounts of training data samples are available or in case of very high dimensional data. With high dimensional data, embedded methods will eventually outperform filter methods. The above three methods where introduced in [13] summarized in Figure 3.

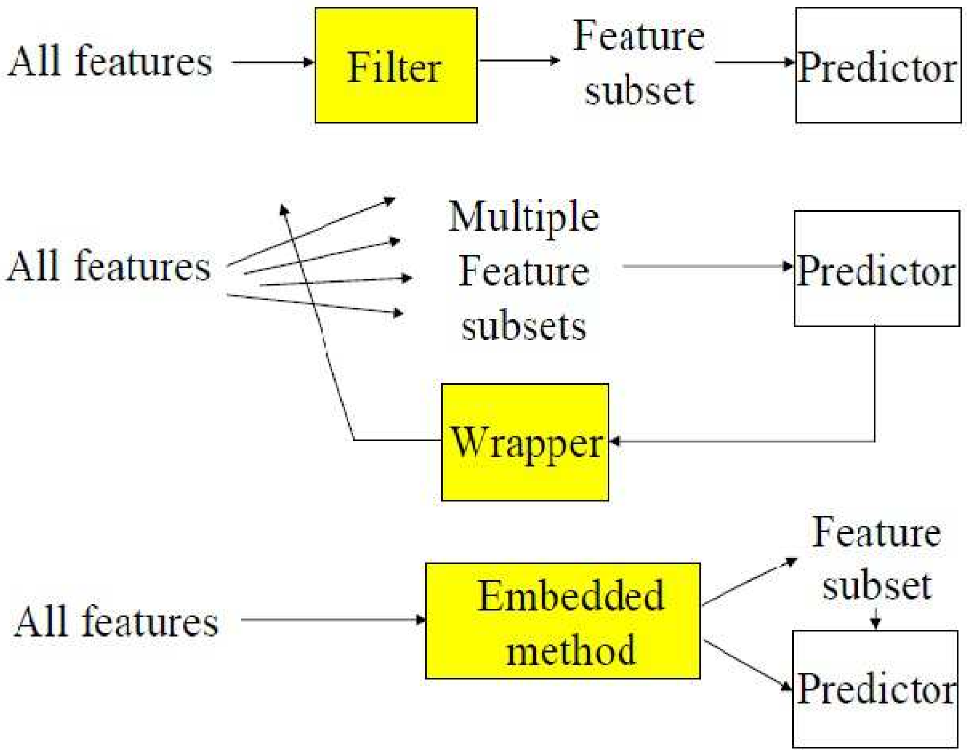


Figure 3: Filter, wrapper and embedded methods.

3.4 Fuzzy feature selection methods

Fuzzy set and fuzzy logic theory provide a way of measuring and reducing uncertainties in data sets through different methods [5], fuzziness measures and fuzzy entropy measures are below defined and discussed.

3.4.1 Fuzziness measure

Fuzziness measure is defined as follows:

Definition 22 *Given a universe set X and a nonempty family $\mathcal{C}$ of a subsets of X, a fuzzy measure on $(X, \mathcal{C})$ is a function $g : \mathcal{C} \to [0, 1]$ that satisfies the following requirements:*

$(R1)$. *$g(\emptyset) = 0$ and $g(X) = 1$, (boundary requirements);*

$(R2)$. *$\forall A, B \in \mathcal{C}$, if $A \subseteq B$, then $g(A) \leq g(B)$, (monotonicity);*

$(R3)$. *For any decreasing sequence $A_1 \subset A_2 \subset \cdots \in \mathcal{C}$ if $\bigcup_{i=1}^{\infty} \in \mathcal{C}$, then*
$\lim_{i \to \infty} g(A_i) = g(\bigcup_{i=1}^{\infty})$
(continuity from below);

$(R4)$. *For any decreasing sequence $A_1 \supset A_2 \supset \cdots \in \mathcal{C}$ if $\bigcap_{i=1}^{\infty} \in \mathcal{C}$, then*
$\lim_{i \to \infty} g(A_i) = g(\bigcap_{i=1}^{\infty})$
(continuity from above).

The boundary requirement $(R1)$ states that since the $\emptyset$ does not contain any element, it cannot contain the element of our interest, on the other hand since the universal set contain all elements under consideration therefore it must contain our elements of interest as well. Requirement $(R2)$ explains that the membership degree of an element in a subset $A \subseteq B$ is smaller than the degree of membership of an element in B. $R3$ and $R4$ are considered in case of infinity universal set.

In many cases we are interested in having suitable measures of impreciseness and vagueness since the uncertainty of data can have different sources [4], this takes us to fuzzy measures. Fuzzy measures give us a good knowledge on of how far a given fuzzy set is from a well defined classical (crisp) reference sets.

3.4.2 Fuzzy entropy measures

Fuzzy entropy represents the fuzziness of a fuzzy set, fuzziness of a fuzzy set is represented through degree of ambiguity, hence the entropy is obtained from fuzzy membership itself.

Definition 23 (Entropy) *Entropy is a measure of the amount of uncertainty in the outcome of a random experiment, or equivalently, a measure of the information obtained when the outcome is observed [7].*

This will play an important role in this research because of its importance in partitioning the input feature space into decision regions and selecting relevant features with good separability for the classification task [7]. Various definitions of fuzzy entropy have been proposed, basically, a well-defined fuzzy entropy measure must satisfy the following four axioms [7], [4]:

1. $E(A) = 0$ iff $A \in 2^X$, where A is a non empty sets and 2^X indicates the power set of A,

2. $E(A) = 1$ iff $\mu_A(x_i) = 1 \forall$ i ,

3. $E(A) \leq E(B)$ if $\mu_A(x) \leq \mu_B(x)$ when $\mu_B(x) \leq 0.5$ and $\mu_A(x) \geq \mu_B(x)$ when $\mu_B(x) \geq 0.5$

4. $E(A) = E(A^c)$

In the following we will define some of well known fuzzy entropy measures [4]. De Luca and Termini suggested that the corresponding to Shannon probabilistic entropy, the measure of fuzzy entropy should be:

$$H_1(A) = -\sum_{j=1}^{n} (\mu_A(x_j) \log \mu_A(x_j) + (1 - \mu_A(x_j)) \log(1 - \mu_A(x_j))) \qquad (18)$$

where $\mu_A(x_j)$ are values in $[0, 1]$.
This fuzzy entropy measure is considered to be a fuzziness measure [4], and it evaluate global deviations from the type of ordinary sets, i.e. any crisp set A_o leads to $H(A_o) = 0$. Note that the fuzzy set A with $\mu_A(x) = 0.5$ plays the role of maximum element of the ordering defined by H.

Newer fuzzy entropy measures were introduced by Parkash [4] and are defined as follows:

$$H_2(A; w) = \sum_{j=1}^{n} w_j (\sin \frac{\pi \mu_A(x_j)}{2} + \sin \frac{\pi(1 - \mu_A(x_j))}{2} - 1) \qquad (19)$$

$$H_3(A; w) = \sum_{j=1}^{n} w_j (\cos \frac{\pi \mu_A(x_j)}{2} + \cos \frac{\pi(1 - \mu_A(x_j))}{2} - 1) \qquad (20)$$

These fuzzy entropy measures(18), (19), (20) will be used in feature selection process to evaluate the relevance of different features in the feature set, this is done by discarding those features with highest fuzzy entropy value in our training set: if the entropy value is high we assume that the feature is not contributing much for the deviation between classes, then it will be removed in our feature set. This process will be repeated for all features in the training set. The classification of the data set becomes relevant after removing irrelevant and redundant features in the data set.

3.5 Feature selection based on entropy measures and Yu's similarity

The notion about the feature selection based on entropy measure and similarity measure was originally introduced in [26] and is extended in this study to cover the similarity based on Yu's norm which is applied to dermatology data set, the main idea is first to create ideal vectors $V_i = (v_i(f_1), ..., v_i(f_N))$, $i = 1, ..., N$ that represent the class i as well as possible. This can be user defined or calculated from some samples set X_i of the vectors $X = (x(f_1), ..., x(f_D))$ which are known to belong to class C_i. Here we use the generalized mean to create these class ideal vectors.

We then calculate the similarities $(S(x, Vi))$ between the samples x and all the ideal vectors V_i. The role of similarity measure is to evaluate the degree of the resemblance or likeness of matched objects, for this purpose we have chosen the similarity measure developed from Yu's norms (15), (16) and it will help to compare similar objects in the classification task. This similarity measure was constructed by replacing Yu's norms (15), (16) in the relation (17) and is below defined:

$$S(x,v) = \max(0, (1+\lambda)(Sn(\overline{x}, v) + Sn(x, \overline{v}) - 1) - \lambda Sn(\overline{x}, v) Sn(x, \overline{v})), \quad (21)$$

where $\lambda > 1$ is a weight parameter,

$$Sn(x,v) = min(1, x + v + \lambda xv) \quad (22)$$

and negation is written as $\overline{x} = 1 - x$ for $x, v \in [0,1]^d$.

In calculating the similarity for our samples vectors and ideal vectors we get j similarities where j is the number of features. We then collected those similarities into one similarity matrix. At this step comes the idea of using the entropy measures to evaluate the relevance of the features, we used the equation (18) with $\mu_A(x_j)$ being similarity values, we remarked that the higher similarity values are, the lower the entropy values are. If the similarity values are close to 0.5 we conclude that we get high entropy values, this underlying idea will help us to identify the features with high entropy values. By summing the entropy values for all the samples in learning set for the feature we get t entropy values for t features, Since hight entropy value corresponds to high uncertainty, we will remove all features with hight entropy values in samples because based on the assumption that they are not contributing much for the deviation between classes, this procedure can be repeated to remove all those non important features. The above method is illustrated in the Figure 4.

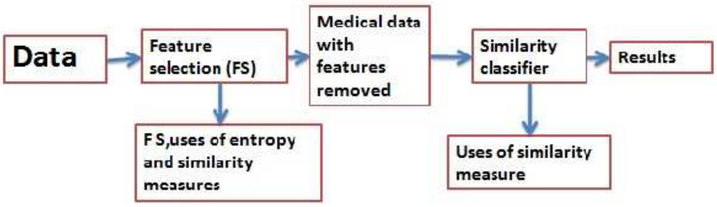

Data
Feature selection (FS)
Medical data with features removed
Similarity classifier
Results
F S,uses of entropy and similarity measures
Uses of similarity measure

Figure 4: Feature selection based on entropy measures and Yu's similarity.

The feature selection based on similarity and fuzzy entropy measure is demonstrated in the following pseudo-code algorithm (1).

Algorithm 1 Pseudo code for feature selection.

```
Require: $idealvec[1, \dots, l]$, $Datalearn[1, \dots, m]$
  for $j = 1$ to $m$ do
    for $i = 1$ to $t$ do
      for $k = 1$ to $l$ do
        $Sn1[i][j][k] = Sn(1 - Datalearn[i][j], idealvec[j][i][k])$
        $Sn2[i][j][k] = Sn(Datalearn[i][j], 1 - idealvec[j][i][k])$
        $sim[i][j][k] = max[0, (1-\lambda)(Sn1[i][j][k] + Sn2[i][j][k] - 1) - \lambda Sn1[i][j][k]Sn2[i][j][k])]$
      end for
    end for
  end for
  Sort similarity values $sim[i][j][k]$ according to feature set $U$
  for $i = 1$ to $t$ do
    $H[i] = -\sum_{x \in U} \mu_i[x] ln\mu_i(x) + (1 - \mu_i(x))ln(1 - \mu_i(x))$
  end for
  $J = argmax_i H[i]$
  Remove $J$:th feature from the data.
```

The algorithm presents m samples, t features, l classes. In the algorithm datalearn stands for the learning set matrix, a typical learning algorithm requires two sets of examples [32]: training sets to produce the learned concept description and test sets to evaluate classification accuracy. The algorithm computes similarity values which are then sorted in one large matrix of $ml \times t$ from which the fuzzy entropy values for each feature summing through ml values can be calculated for each feature. We then find feature with big fuzzy entropy value and remove them in our data set. After feature removal we use the classifier (21) to classify the remaining data.

We illustrate the feature selection based on similarity and fuzzy entropy measure in the following example:

Example 11 (Feature selection with Lukasiewicz measure) *Consider the following data in Table 1.*

Table 1: Feature selection example.

object	*V1*	*V2*	*V3*
O1	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{2}{7}$
O2	$\frac{1}{2}$	$\frac{5}{10}$	$\frac{6}{7}$
O3	*1*	$\frac{6}{10}$	$\frac{1}{7}$
O4	*1*	$\frac{10}{10}$	$\frac{7}{7}$

We know that objects $O1$ and $O2$ belong to class A and objects $O3$ and $O4$ belong to class B, use the (26) and (18) measures to remove variables with highest uncertainty. The mean vector for class A and B are $\mu_A = [0.5, \frac{4}{10}, \frac{4}{7}]$, $\mu_B = [1, \frac{8}{10}, \frac{4}{7}]$, to compute the similarity measure between objects and mean vectors we use the equation (26).

$S(O1, \mu_A) = [1, \frac{9}{10}, \frac{5}{7}]$, $S(O2, \mu_A) = [1, \frac{9}{10}, \frac{5}{7}]$, $S(O3, \mu_B) = [1, \frac{8}{10}, \frac{4}{7}]$, $S(O4, \mu_B) = [1, \frac{8}{10}, \frac{4}{7}]$.

This similarity values are next presented in matrix form

$$\begin{bmatrix} 1 & \frac{9}{10} & \frac{5}{7} \\ 1 & \frac{9}{10} & \frac{5}{7} \\ 1 & \frac{8}{10} & \frac{4}{7} \\ 1 & \frac{8}{10} & \frac{4}{7} \end{bmatrix} \tag{23}$$

Next we calculate fuzzy entropy for our variables $V1$,$V2$,$V3$, and remove variables with highest uncertainty, we use the equation (18) to compute the entropy values, we get $h(V1) = 0$,$h(V2) = 1,65$,$h(V3) = 2.37$.

We decide to remove variable $V3$ because of its highest entropy values.

Example 12 *The same example is again down presented using our similarity measure (21): To compute the similarity measure between objects and* *mean vectors we use the similarity based on Yu's norm (21) when the parameter $\lambda = 1$. These similarity values are presented in Table 2*

Table 2: Similarities in feature selection example.

similarity	V1	V2	V3
$S(O1, \mu_A)$	1	1	0.8367
$S(O2, \mu_A)$	1	1	0.7959
$S(O3, \mu_A)$	0.5	0.96	0.6327
$S(O4, \mu_A)$	0.5	0.4	0.5714
$S(O1, \mu_B)$	0.5	0.56	0.8367
$S(O2, \mu_B)$	0.5	0.8	0.7959
$S(O3, \mu_B)$	1	0.92	0.6327
$S(O4, \mu_B)$	1	0.8	0.5714

Next we calculate fuzzy entropy values for our variables $V1$,$V2$,$V3$, and remove the variable with highest uncertainty. For calculating entropy we use the equation (20, and we get $h(V1) = 1.66$, $h(V2) = 1.50$, $h(V3) = 2.55$. Next we decide to remove variable $V3$ because of its highest entropy values.

4 Similarity classifier

The problem of classification is basically one of partitioning the feature space into regions, one region for each category i.e establish boundaries in feature space. Ideally, one would like to arrange this partitioning so that none of the decisions is ever wrong. When this cannot be done one would like to minimize the probability of error [26]. Simply, the classification task can be understood in this way: assigning objects to classes (groups) on the basis of measurements made on the object. Usually, the problem of classification starts with a vaguer general knowledge about the situation together with a number of designed samples particular representatives of the patterns we want to classify, then our problem will consist in exploiting the given information to design the classifier. Classifiers are divided into categories according to their learning methods: supervised learning and unsupervised learning. In classification, learning means that the algorithm usually learns through samples of how the classification should be done and then it can classify data sets with similar problems. In supervised learning the set of classes is specified in advance and the goal is to decide whether candidate objects belong to those classes. In unsupervised learning the goal is to decide which object should be grouped together, no classes are specified in advance. This classification process is again clarified as follows: given the data to be classified, k samples, we first calculate ideal vectors, secondly we compute the similarity between the samples and ideal vectors, lastly we classify the samples based on the highest similarity value.

In classification we have chosen to use a similarity measure developed from Yu's norm (21)

We would like to classify a set X of objects to N different classes $C_1, ..., C_N$ by their features. Let D be the number of different kinds of features $f_1, ...f_D$.

We assume that the values for the magnitude of each feature is normalized so that it can be presented between $[0, 1]$, this implies that the object we want to classify are vector belong to $[0, 1]^D$.
In the first step we determine the ideal vector $V_i = (v_i(f_1), ..., v_i(f_D))$, $i = 1, ..., N$ that represent the class i as well as possible,this vector can be user defined or calculated from some samples set X_i of the vectors $X = (x(f_1), ..., x(f_D))$ which are known to belong to class C_i. We can actually use the generalized mean to calculate V_i which is.

$$v_i(r) = (\frac{1}{\sharp X_i} \sum_{x \in X_i} x(f_r)^m)^{\frac{1}{m}} \quad \forall r = 1, ..., D \tag{24}$$

where the power value m is fixed for all i and D, and $\sharp X_i$ is the number of samples class i.
In the second step, we have to make a decision to which class an arbitrary chosen $x \in X$ belongs, this can be done by comparing it to the ideal vector by means similarity measure (21). Briefly, the method compares the ideal vector to every sample in the test set using the similarity measure. We decide that $\mathbf{x} \in C_i$ if

$$S\langle \mathbf{x}, \mathbf{v}_i \rangle = \max_{i=1,...,N} S\langle \mathbf{x}, \mathbf{v}_i \rangle \ . \tag{25}$$

In this way, the sample is classified to a class with highest similarity value. Next we demonstrate the above steps with an example:

Example 13 *:*

Given data that presents six samples each having four measured values (features) $(f1, ..., f4)$ *in Table 3 and we know which class they belong, we get a new sample:* $x = [69, 31, 51, 20]$ *without the knowledge to which class it belongs. Compute using Similarity based on Lukasiewicz structure:*

$$S(x, v) = \frac{1}{t}\sum(1 - |x(f_r) - v(f_r)|), for[x, v] \in [0, 1] \tag{26}$$

, to which class the given sample belongs?

Table 3: Similarity classifier problem.

F1	*F2*	*F3*	*F4*	*Classes*
64	*32*	*45*	*15*	*1*
69	*31*	*49*	*15*	*1*
74	*28*	*61*	*19*	*2*
79	*38*	*64*	*20*	*2*

Solution:

All values are positive. The maximum values for Features are: $[79, 38, 64, 20]$

Calculate the mean vectors: v_1, v_2 *for classes:* $v_1 = [\frac{64+69}{2}, \frac{32+31}{2}, \frac{45+49}{2}, \frac{15+15}{2}] = [66.3, 31.5, 46.5, 15]$

$v_2 = [\frac{74+79}{2}, \frac{28+38}{2}, \frac{61+64}{2}, \frac{19+20}{2}] = [76.5, 33, 62.5, 19.5]$.

Next we compute the total similarity values between samples and mean vectors that are representing the classes:

$S(x, v_1) = \frac{1}{4}(1 - |\frac{69-66.5}{79}| + 1 - |\frac{31-31.5}{38}| + 1 - |\frac{51-46.5}{64}| + 1 - |\frac{20-15}{20}|) = 0.9087$

$S(x, v_2) = \frac{1}{4}(1 - |\frac{69-76.5}{79}| + 1 - |\frac{31-33}{38}| + 1 - |\frac{51-62.5}{64}| + 1 - |\frac{20-19.5}{20}|) = 0.9119$

By looking at our similarity values we conclude that our sample belongs to the second class which correspond to the highest similarity value of 0.9119*.*

5 Data set and Results from data set

5.1 Dermatological data set

The data set used in this study are freely available from UCI machine learning data repository [27] and came from Gazi University and Bilkent University: it was donated by N. Ilker and H.A. Guvenir. The fundamental properties of the data set is shown in Table 4.

Table 4: Dermatological data set and its main properties.

Data set	Nb. classes	Nb. features	Nb. cases
Dermatology	6	34	366

The department of dermatology is concerned with the diagnostic of erythemato-squamous diseases which are grouped as follow: soriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris, they all present common clinical features of erythma and scaling with slight variation. Although biopsy is required for the diagnostic of erythemato-squamous diseases it was remarked that these diseases share many histopathological features, a disease may show at the beginning stage the feature of another and may have the characteristic feature at the following stage, this make the problem of erythemato-squamous diseases more seriously and a real concern of the department of dermatology.

The following Table 5 shows different attributes and class distribution of data dermatology data set:

Table 5: Class distribution of dermatology data set.

Class	Attributes clinical	Histopathological
Psoriasi (111)	Att. 1: erythema	Att. 12: melanin incontinence
Seboreic dermatitis(60)	Att.2: scaling	Att. 13: eosinophils in infiltrate
Lichen planus (71)	Att. 3: definite borders	Att. 14: PNL infiltrate
Pityriasis rosea(48)	Att. 4: itching	Att. 15:fibrosis of the papillary dermis
Cronic dermatitis(48)	Att.5:koebner phenomenon	Att. 16:exocytosis
Pityarisis rubra pilaris (20)	Att. 6: polygonal papules	Att.17:acanthosis
	Att.7:follicular papules	Att. 18: hyperkeratosis
	Att.8:oral mucosal involvement	Att. 19: parakeratosis
	Att.9:knee and elbow involvement	Att. 20: clubbing of the rete ridges
	Att.10:scalp involvement	Att. 21: elongation of the rete ridges
	Att.11:family history	Att.22: thinning of the suprapapillary epidermis
	Att.34:age	Att. 23: pongiform pustule
		Att. 24: munro microabscess
		Att. 25: focal hypergranulosis
		Att. 26: disappearance of the granular layer
		Att. 27: vascularization and damage of basal layer
		Att. 28: spongiosis
		Att. 29: saw-tooth appearance of retes
		Att. 30: follicular horn plug
		Att. 31: perifollicular parakeratosis
		Att. 32: inflammatory mononuclear infiltrate
		Att. 33: band-like infiltrate

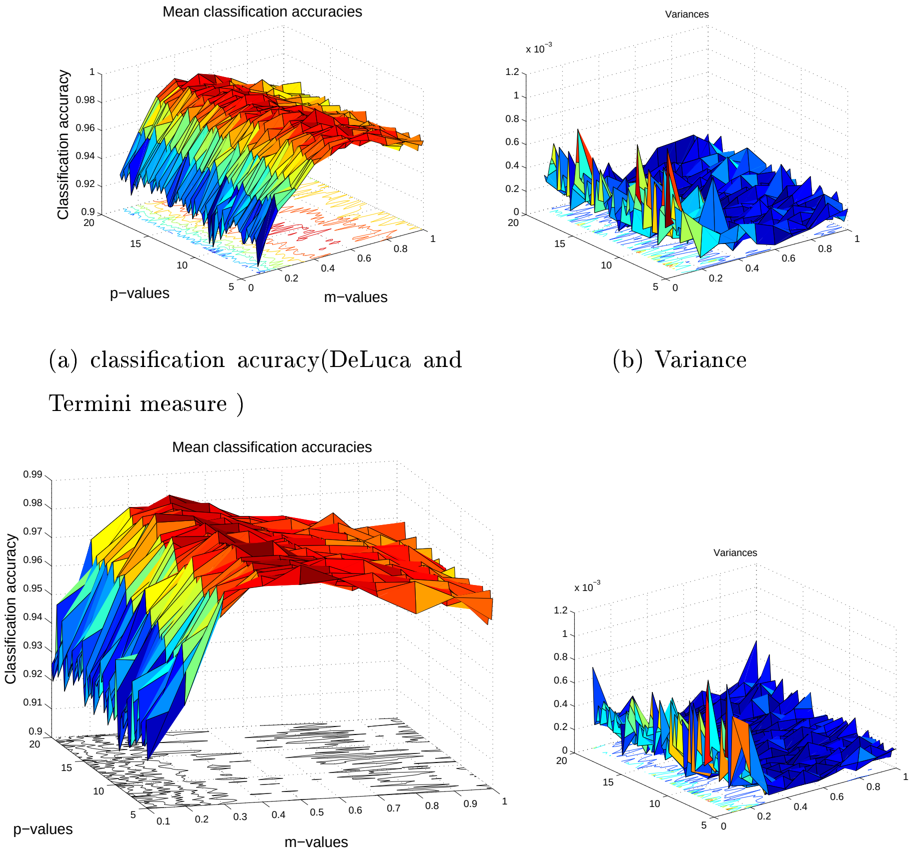
5.2 Results

In this part, we present the obtained result from our data set analysis.
The given data set was divided into two parts: one part was used for training and another for testing and this procedure was repeated randomly 30 times, we computed mean classification accuracies and variances. Also parameter ranges where studied. For the ideal vector computations in the generalized mean, range $m \in (0, 1]$ was shown to provide highest accuracy and for parameter p in similarity measure range $p \in [5, 20]$ seemed to provide the highest results. In the following Table 6 we present how the choice of different p values affects the classification accuracy values. There the first row shows different p values, the second row shows the classification accuracies percentages obtained by using De Luca and Termini fuzzy entropy measure as the feature selection method. In the third row the classification accuracy percentages obtained by using Parkash fuzzy entropy measure are shown.

Table 6: Feature removal using De Luca and Termini, and Parkash measure w.r.t parameter p value.

Parameters	0.1	0.2	0.3	0.4	0.5
Accuracies%(Luca)	98.61	98.83	98.61	98.61	98.39
Accuracies%(Parkash)	98.83	98.72	98.61	98.50	98.39

Both the two feature selection methods De Luca and Termini and Parkash entropy measures applied to the similarity matrix contributed much to achieve high classification accuracy, value 98.83% was obtained. With 99% confidence interval the mean classification accuracy is $98,83 \pm 0.2$ (with Student's t-distribution $\mu \pm t_{1-\frac{\alpha}{2}} S_\mu/\sqrt{n}$). In the following Figure 5, we can see how parameter values from the classifier (generalized mean m and p value from similarity measure) effected the mean classification accuracy and variance. In the first case De Luca and Termini's fuzzy entropy measure was used as feature selection method before actual classification and its performance is seen in Figure 6*a* and 6*b*. Parkash fuzzy entropy measure was used in the second case as feature selection method and again after that classification results were done with similarity classifier, its performance is seen in Figures 6*c* and 6*d*.



(a) classification acuracy(DeLuca and Termini measure)

(b) Variance

(c) classification accuracy(Parkash measure)

(d) Variance

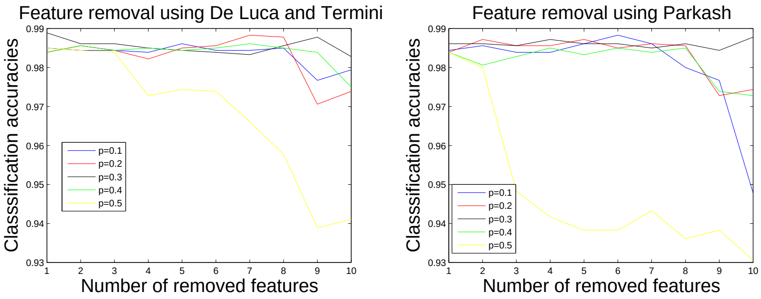
Figure 5: Mean classification results and Variances plotted with respect to parameter p and mean values using entropy measures on Dermatology set.

The results with the highest mean accuracies together with the correct number of the removed features are are presented in the following Table 7:

Table 7: Classification result.

Methods	Mean accuracy(%)	Variance	Dim	Removed features
sim+Luca	98.83	0.0442	33	17
Sim+Parka	98.83	0.0237	33	1

As it can be seen in the Table 7 the use of De Luca's fuzzy entropy measure with 33 features in the data set the highest mean accuracy was obtained. With Parkash's fuzzy entropy measure highest mean accuracy was also achieved after removing one feature in the data set. In Figure 6 one can see how reducing the number of features from the data set effected the classification accuracies for both fuzzy entropy measures. After removing more than 10 features classification accuracies started to deteriorate quite rapidly. In the Figure 6 results are also studied w.r.t. parameter p value in feature selection process to see how it effects the results.



(a) classification accuracy w.r.t feature selection subsets(Luca and Termini)

(b) classification accuracy w.r.t feature selection subsets(Parkash)

Figure 6: Classification accuracies w.r.t. reduced features for data sets (Dermatology).

In the Table 8 we present and compare our results with some others previously obtained using the different similarity classifiers with the feature selection method applied to the same data set.

Table 8: Obtained accuracies with different classifiers.

Classifiers	Mean accuracy%
Sim-using Lukasiewicz	96.04
Sim-using Yu's norm	97.19
FS + Sim-using Lukasiewicz	98.28
FS+Sim-Yu's measure	98.83

In the Table 8 we present the results for similarity classifier using generalized Lukasiewicz similarity which was taken from [26], similarity classifier using Yu's norm taken from [25], in the first two case no feature selection method was done, in the second case we show the results when feature selection method is combined with generalized Lukasiewics similarity [26]. As we can see from the results, the proposed method produced the highest mean classification accuracies and also clearly enhanced the mean accuracy with this data set. Results shown in this study are shortly cited in author's paper [6]

6 Conclusion and future work

In this study, the fuzzy entropy and similarity based feature selection was performed, the used entropy measures (18), (19), (20) managed to discard redundant and irrelevant features in our data set. We reduced the computation time and this has positively impacted our similarity classifier see (4) to achieve highest classification accuracy of 98.83% when testing our similarity measure to real world data set. The results in Table 8 show that the used method managed to achieve the highest accuracy when comparing it with some other results previously obtained, we assure that this is the best accuracy ever obtained with the same data set. we remarked the used method help to reduce the dimensionality of large data sets but also to speed up the computation time of a learning algorithm and therefore simplify our classification task. Notice that over 98% mean accuracy was achieved with the methods after removing 10 features from the data set, which is about 30% reduction of features from this data set.

For future work we acknowledge that these are not the only fuzzy entropy measures that exist and usage of these with different similarity measures is a subject which needs to be thoroughly addressed in the future. Creation of ideal vectors in the feature selection process is one future area of investigation. Also beside these, creation of pool of similarities and proper selection via optimization is also one research area.

References

[1] Alsina Claudi, Berthold Schweizer, Claudi Frank J Maurice, (2006) *Associative Function: Triangural norms and copulas.* World scientific publishing Co, Pte.ltd ISBN 98-256-671-6.

[2] Andrews Chrysostomou Kyriacos (2008) *The Role of Classifiers in Feature selection Selection: Number vs Nature, School of Information Systems.* Computing and Mathematics, Brunel University.

[3] A. Pethalakshmi, K. Thangavel, (2006) *Feature Selection for Medical Database Using Rough System.* AIML Journal. ISBN:978-960-474-41-3

[4] Bandemer Hans and Näther Wolfgang, (1992) *Fuzzy data analysis.* Kluwer Academic.

[5] Bo Yuan, Klir George J., (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall. Publishers, Dordrecht.

[6] Cesar Iyakaremye, Pasi Luukka, David Koloseni(2012) *Feature selection using Yu's similarity measure and fuzzy entropy measures*, Accepted for publication to IEEE International Conference on Fuzzy Systems, 2012.

[7] Chih-Ming Chen, Hahn-Ming Lee, Jyh-Ming Chen, Yu-Lu Jou, (2001) *An Efficient Fuzzy Classifier with Feature Selection Based on Fuzzy Entropy.* IEEE Transactions on Systems, Man, and Cybernetics – Part b: cybernetics, 1083-4419(01)04860-9.

[8] David G.Stork, Duda O. Rishard E.Hart (1973) *Pattern classification and scene analysis.* John Wiley, Sons.

[9] Derya Elif, Übeyli, Güler, I. (2005) *Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems.* Computers in Biology and Medicine, 35, 5, pp. 147-165.

[10] Didier Dubois and Henri, (1980) *Prade Fuzzy Sets and Systems: Theory and Applications.* Academic Press.

[11] Fan Ya-Ju, Wanpracha Art Chaovalitwongse, (2008) *Optimizing feature selection to improve medical diagnosis.* Springer Science and Business Media, vol. 174, no1, pp. 169-183.

[12] Gottwald Siegfried (1993) *Fuzzy sets and fuzzy logic.* Artificial intelligence, ISNB 3-528-05311-9.

[13] Guyon Isabelle (2007) *Introduction to feature selection.* Available from http://videolectures.net (accessed on 20 October 2011).

[14] H. Altay Güvenir, G. Demiroz, Nilsel Ílter (1988) *Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals.* Artificial Intelligennce in Medicine 13 (1998) 147–165.

[15] Huiqing Liu, Jinyan Li, Limsoon Wong, (2002) *A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns.* Genome Informatics journal 13: 51-60.

[16] Ikou Kaku, Jiafu Tang, JianMing Zhu, Yong Yin, (2010) *Data mining: concepts, methods and applications in managements and engineering design.* ISBN 978-1-84996-337-4 DOI 10.1007/978-1-84996-338-1, Springer London Dordrecht, Heidelberg, New York.

[17] Inaki Inza, Pedro Larranaga, Yvan Saeys, (2007) *Review of feature selection techniques in bioinformatics.* Vol.23 no.19, pages 2507–2517 doi:10.1093/bioinformatics/btm344.

[18] Isabelle Guyon, Gunn Steve, Masoud Nikravesh, Lotfi A. Zadeh, (2006) *Feature Extraction: Foundations and Applications.* Springer-Verlag, Berlin Heidelberg.

[19] Jensen Richard, Qiang Sheng, (2008) *Computational intelligence and feature selection: Rough and Fuzzy Approaches.* ISBN:978-0-470-22975-0. IEEE Press Series on Computational Intelligence.

[20] Kalle Saastamoinen (2008) *Many valued algebraic structure as measures of comparison.* PhD thesis, Lappenranta University of Technology.

[21] Luukka Pasi, (2008) *Similarity classifier using similarity based on modified probabilistic equivalence realations.* Elservier journal, Knowledge-Based System 22(2009) 57-62.

[22] Lowen R., (1996) *Fuzzy set theory, Basic concepts, Techniques and Bibliography.* Kluwer Academc Publishers.

[23] Luukka Pasi, (2005) *Similarity measure based classification.* PhD thesis, Lappenranta University of Technology.

[24] Luukka Pasi, Tapio Leppälampi (2006) *Similarity classifier with generalized mean appied to medical data.* Elservier Journal, Computers in Biology and Medicine pp.1026-1040.

[25] Luukka Pasi, (2007) *Similarity classifiers using similarity measure derived from Yu's norm in classification of medical data.* Computers in Biology and Medicine pp. 1133-1140.

[26] Luukka Pasi, (2010) *Feature selection using fuzzy entropy measures with similarity classifier.* Expert Systems with Applications. doi:10.1016,j.eswa.2010.09.133

[27] Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., (2007). *UCI Repository of machine learning databases.* Irvine, CA: University of California, Department of Information and Computer Science.http://www.ics.uci.edu/Emlearn/MLRepository.html

[28] Paavo Kukkurainen and Pasi Luukka, (2006) *New Classifier Based on Fuzzy Level Set Subgrouping.* Lecture Notes in Computer Science, Volume 4253,2006, 383-389.

[29] Timothy J.Ross, (2010) *Fuzzy logic with engineering applications.*3rd edition, ISBN 978-0-470-74376-8. Library of Congress Cataloging-in-Publication Data

[30] Hanss Michael, (2005) *Applied Fuzzy Arithmetic: An Introduction with Engineering Applications.* ISBN 3-540-24201-5. Springer Berlin, Heidelberg New York.

[31] L.Zadeh,(1965). *Fuzzy Sets, Information and Control.* pp. 338-353. Department of Electrical engineering and electronics research laboratory, University of California, Berkeley, California.

[32] Hall A. Mark, (1991) *Correlation-based feature selection for Machine Learning* .PhD thesis, Department of Computer Science, The university of Waikato, Hamilton, NewZealand.