A New Method to Measure the Similarity between Features in Machine Learning using the Triangular Fuzzy Number

Hassan Nosrati Nahook, Mahdi Eftekhari

Abstract — In this paper, we present a new method to measure the similarity between features using fuzzy numbers. The proposed method uses the concept of geometry to calculate the degree of similarity between triangular fuzzy numbers defined on the features. We also prove some properties of the proposed similarity measure and use different data sets to compare the proposed method with existing methods. In the feature selection methods, the proposed similarity measure compared with other fuzzy similarity measures can be more efficient.

Keywords: Similarity Measure, Symmetrical or Asymmetrical Triangular Fuzzy Number, Features.

I. INTRODUCTION

While similarity is an essential concept in human reasoning and plays a fundamental role in theories of knowledge, there is no unique and general-purposed definition of similarity. The reason for this lack of a definition comes from the fact that one can find practical cases where similarity properties are not satisfied (e.g., symmetry, indiscernibility, or transitivity; [1]). Indeed, several studies ([2] and [3]) have shown that similarity measures do not necessarily have to be transitive, implying a contradiction with the most usual approach of comparison, based on geometrical assumptions in the feature space.

Fuzzy set theory provides a consistent basis for information processing and an elegant, mathematically well-founded, representation of the uncertainty in the data. Since the data that are to be processed are often imprecise, using fuzzy set theory or its derivatives (e.g., possibility theory or belief function theory) has become a common approach in recent years [4]. In this paper, similarity measures are defined by the use of membership function that is derived from fuzzy residual implications. We present a new method to calculate the similarity between features based on triangular fuzzy numbers (TFN).

This paper is organized as follows. Section 2 describes the related terms. Section 3 explains the proposed fuzzy similarity measure for evaluate similarity between features. In Section 4, the experimental results of the proposed method are presented. The last section summarises and conclusion related work.

Manuscript received January 01, 2013.

Hassan Nosrati Nahook, Student of MSc Computers - AI, Computer Engineering Department, Science and Research Branch, Islamic Azad University, Kerman, Iran.

Mahdi Eftekhari, Assistant Professor, Computer Engineering Department, Science and Research Branch, Islamic Azad University, Kerman, Iran.

II. RELATED TERMS

A. Fuzzy Sets

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, As shown in equation 1.

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

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. If we allow our valuation set $\{0,1\}$ to be the real interval $[0,1]$ then A is called a Fuzzy set [5]. 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: $A=(x, \mu_A(x))$, $x \in X$.

B. Membership Function

Fuzziness in a fuzzy set is characterized by its membership functions. A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The input space is sometimes referred to as the universe of discourse, a fancy name for a simple concept. It classifies the element in the set, whether it is discrete or continuous. The graphical representations may include different shapes. There are certain restrictions regarding the shapes used. The "shape" of the membership function is an important criterion that has to be considered. There are different methods to form membership functions. Zadeh proposed a series of membership functions that could be classified into two groups: those made up of straight lines, or "linear" and Gaussian forms, or "curved" [6]. Based on this criterion the membership function can be of the following types [7].

(1) Triangular Fuzzy Number Defined by its lower limit a, its upper limit b, and the modal value m, so that $a < m < b$. We call the value $b - m$ margin when it is equal to the value $m - a$.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

As shown in equation (2). Figures 2 and 3 show the triangular fuzzy number.

$$A(x) = \begin{cases} 0 & x \le a \text{ or } x \ge b \\ \frac{(x-a)}{(m-a)} & x \in (a,m) \\ \frac{(b-x)}{(b-m)} & x \in (m,b) \end{cases} \quad (2)$$

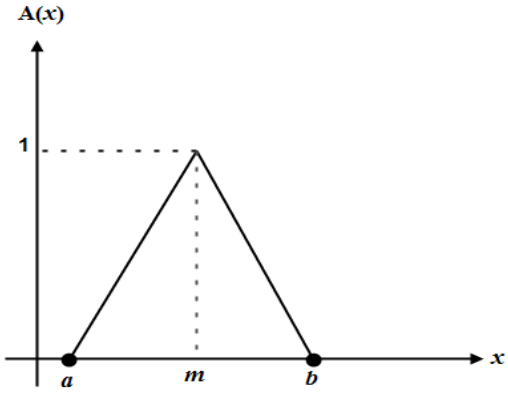

Fig 2: Triangular Fuzzy Symmetrical.

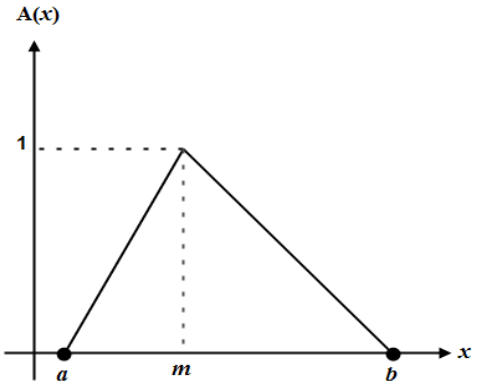

Fig 3: Triangular Fuzzy Asymmetrical.

(2) Trapezoidal Fuzzy Number

Defined by its lower limit a, its upper limit d, and the lower and upper limits of its nucleus or Kernel b and c respectively. As shown in equation (3). Figure 4 show the Trapezoidal fuzzy number.

$$T(x) = \begin{cases} 0 & x \le a \text{ or } x \ge d \\ \frac{(x-a)}{(b-a)} & x \in (a,b) \\ 1 & x \in (b,c) \\ \frac{(d-x)}{(d-c)} & x \in (c,d) \end{cases} \quad (3)$$

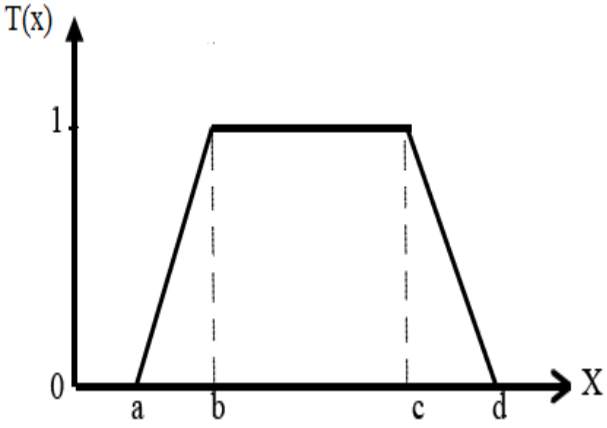

Fig 4: Trapezoidal Fuzzy Number.

(3) Gaussian Bell shape, other MFs and basic properties of fuzzy sets [8].

C. T – norm and T – conorm

The triangular norms (t-norm), 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$\in[0,1]$.

T – norm: A two-place function $T:[0,1]\times[0,1]\to[0,1]$ is called t – norm if the following conditions are satisfied:

$1.T(x,1) = x$: one identity;

$2.x \le z,\ y \le u \Rightarrow T(x,y) \le T(z,u)$: monotonicity ;

$3.T(x,y) = T(y,x)$: commutivity;

$4.T(T(x,y),z) = T(x,T(y,z))$: associativity;

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

T – conorm: A two-place function $Sn:[0,1]\times[0,1]\to[0,1]$ is called t – conorm if the following conditions are satisfied:

$1.Sn(x,0) = x$: zero identity;

$2.x \le z,\ y \le u \Rightarrow Sn(x,y) \le Sn(z,u)$: monotonicity ;

$3.Sn(x,y) = Sn(y,x)$: commutivity;

$4.Sn(Sn(x,y),z) = Sn(x,Sn(y,z))$: 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 [9].

A function $Sn:[0,1]\times[0,1]\to[0,1]$ is dual t-conorm of t-norm such that for all x, y$\in[0,1]$ both the following equivalent equalities hold, $Sn(x,y) = 1 - T(1-x,1-y)$ and $T(x,y) = 1 - Sn(1-x,1-y)$, where (1 – x) and (1 – y) are respectively complements of x and y.

Next we present a list of the main well know and most frequently used t – norms [9], [10]:

$$T_{min}(x,y) = min(x,y): \text{Minimum;} \quad (4)$$

$$T_{prod}(x,y) = xy: \text{Algebraic product;} \quad (5)$$

$$T_{bprod}(x,y) = max(0,x+y-1): \text{Bounded product;} \quad (6)$$

$$T_{Ds}(x,y) = \begin{cases} x & ;y=0 \\ y & ;x=0 \\ 1 & otherwise \end{cases} : \text{Drastic sum;} \quad (7)$$

$$T_{\alpha}^{H}(x,y) = \frac{xy}{\alpha+(1-\alpha)x+y-xy}, \alpha \ge 0: \text{Hamacher's t-norm;} \quad (8)$$

$$T_{\beta}^{F}(x,y) = log_{\beta}\left(1+\frac{(\beta^{x}-1)(\beta^{y}-1)}{\beta-1}\right), \beta > 0, \beta \ne 1: \text{Frank's t-norm;} \quad (9)$$

$$T_{\gamma}^{Y}(x,y) = 1 - min(1,(1-x)^{\gamma}+(1-y)^{\gamma})^{\frac{1}{\gamma}}, \gamma > 0: \text{Yager's t-norm;} \quad (10)$$

$$T_{k}^{Db}(x,y) = 1 - \frac{1}{1+\left(\left(\frac{1-x}{x}\right)^{k}+\left(\frac{1-y}{y}\right)^{k}\right)^{\frac{1}{k}}}, k > 0: \text{Dombi's t-norm;} \quad (11)$$

$$T_{\theta}^{W}(x,y) = max\left(0,\frac{x+y-1+\theta xy}{1+\theta}\right), \theta > -1: \text{Weber's t-norm;} \quad (12)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$T_{SS}^{1}(x,y) = max\left(0,(x^p+y^p-1)^{\frac{1}{p}}\right), p>0: \text{Schweizer Sklar's t-norm;} \quad (13)$$

$$T_{\lambda}^{Yu}(x,y) = max(0,(1+\lambda)(x+y-1)-\lambda xy), \lambda>-1: \text{Yu's t-norm;} \quad (14)$$

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

$$Sn_{\lambda}^{Yu}(x,y) = min(1,x+y+\lambda xy), \lambda>-1 \quad (15)$$

D. Similarity measures for Fuzzy sets

In this section we present a brief review of similarity measures for fuzzy sets and their axiomatic basis. Since the concept of similarity has a wide range of applications, there are different approaches present in literature as axioms for degree or measure of similarity. These axioms have differences and similarities depending upon the contexts in which they are constructed. At first hand, a similarity measure for fuzzy sets is expected to be a *T*- equivalence on $F(X)$, which is later realized to be a very unrealistic requirement. Some other lists of properties are also found in literature that a reasonable similarity measure must satisfy. We shall suffice to present a set of axioms formulated by Bustince [11] for an interval valued similarity measure.

A function $\varsigma: F(X) \times F(X) \to [0,1]$ is called a normal interval valued similarity measure, if ς satisfies following properties for all A, B, C $\in F(X)$:

I. $\varsigma(A,B) = \varsigma(B,A)$,
II. $\varsigma(A,A^c) = [0,0]$,
III. $\varsigma(A,A) = [1,1]$,
IV. Monotonic

if $A \subseteq B \subseteq C$, then $\varsigma(A,B) \geq \varsigma(A,C)$ and $\varsigma(B,C) \geq \varsigma(A,C)$.

Distance based similarity measures for Fuzzy sets. The most obvious way of calculating similarity of fuzzy sets is based on their distance. This calculation is in two step: in first part the distance between two fuzzy sets is obtained by a distance measure and in the second part one of the relationships between similarity and distance comes into play to reach at the degree of similarity.

Various distance measures d are present in literature. The most commonly employed distance measures are:

1. The Hamming distance

$$d_H(A,B) = \sum_{i=1}^{n} |A(x_i) - B(x_i)| \quad (16)$$

2. The normalized Hamming distance

$$d_{nH}(A,B) = \frac{1}{n}\sum_{i=1}^{n} |A(x_i) - B(x_i)| \quad (17)$$

3. The Euclidean distance

$$d_E(A,B) = \sqrt{\sum_{i=1}^{n} \left(A(x_i) - B(x_i)\right)^2} \quad (18)$$

4. In general

$$d_r(A,B) = \left[\sum_{i=1}^{n} |A(x_i) - B(x_i)|^r\right]^{\frac{1}{r}}, r \geq 1 \quad (19)$$

5. The sup distance

$$d_{\infty}(A,B) = \sup_i |A(x_i) - B(x_i)| \quad (20)$$

Where measures 1 – 4 are constructed for finite universe. The relationship between the notions of similarity and distance is expressed in several ways some of which are as follows: If d is the distance measure between two fuzzy sets A and B on a universe X, then following measures of similarity are presented in [12], [13] and [14] respectively:

1. The distance based assessment proposed by Koczy:

$$S(A,B) = \frac{1}{1+d(A,B)} \quad (21)$$

2. The distance based assessment proposed by Williams and Steele:

$$S(A,B) = e^{-\alpha d(A,B)} \quad (22)$$

Where α is the steepness measure.

3. Family of distance based similarity measures presented by Sanitini:

$$S(A,B) = 1 - d_r(A,B), r = 1,2,\infty. \quad (23)$$

E. Fuzzy similarity measures

(1) Simple fuzzy similarity measures

As definition of the cardinality of a fuzzy set A in X we consider the usual sigma-count of A:

$$\#A = \sum_{i=1}^{n} A(x_i) \quad (24)$$

Furthermore, the complement A^c of A is defined by:

$$A^c(x_i) = 1 - A(x_i) \quad (25)$$

and therefore $\#A^c = n - \#A$.

We have expressed T – norms in Section 2.3. In this paper, only the T – norm of equation (4) we use. Consider two fuzzy sets A and B in X and let $a_i = A(x_i)$ and $b_i = B(x_i)$, then we define:

$$A \cap B(x_i) = T(a_i,b_i) \quad (26)$$

$$A \cup B(x_i) = Sn(a_i,b_i) \quad (27)$$

Where T is an arbitrary t-norm, and Sn denotes its dual t-conorm: $Sn(x,y) = 1\text{-}T(1\text{-}x,1\text{-}y)$. We further restrict the t-norm T to the family of Minimum t-norms, namely the t-norms characterized by the functional equation:

$$Sn(x,y) + T(x,y) = x + y \quad (28)$$

Hence, fuzzification equation (27) for set union can be restated in the alternative form:

$$A \cup B(x_i) = a_i + b_i - T(a_i,b_i) \quad (29)$$

Notice that rules (24), (25) and (29) are such that both the expressions $\#(A \cup B) + \#(A \cap B)$ and $\#A + \#B$ are fuzzified to the same expression.

Equations (24) and (25) leads to the fuzzy similarity measures listed in Table (1).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table (1): Simple fuzzy similarity measures.

S	Expression
S_{11}	$\frac{\min(\sum a_i, \sum b_i)}{\max(\sum a_i, \sum b_i)}$
(Complement S_{11})S_{12}	$\frac{n - \max(\sum a_i, \sum b_i)}{n - \min(\sum a_i, \sum b_i)}$
S_{13}	$\frac{\min(\sum a_i, \sum b_i)}{n}$
(Complement S_{13})$S1_4$	$\frac{n - \max(\sum a_i, \sum b_i)}{n}$

(2) ∩ - based fuzzy similarity measures

From Table (1) we see that are candidate for fuzzification by means of equations ((24), (25) and (29). In Table (2) are shown the expressions of the corresponding fuzzy similarity measures.

Table (2): ∩ - based fuzzy similarity measures.

S	Expression
S_{21}	$\frac{\sum T(a_i, b_i)}{\max(\sum a_i, \sum b_i)}$
(Complement S_{21})S_{22}	$\frac{n - \sum a_i - \sum b_i + \sum T(a_i, b_i)}{n - \min(\sum a_i, \sum b_i)}$
S_{23}	$\frac{\sum T(a_i, b_i)}{\sum a_i + \sum b_i - \sum T(a_i, b_i)}$
(Complement S_{23})S_{24}	$\frac{n - \sum a_i - \sum b_i + \sum T(a_i, b_i)}{n - \sum T(a_i, b_i)}$
S_{25}	$\frac{\min(\sum a_i, \sum b_i)}{\sum a_i + \sum b_i - \sum T(a_i, b_i)}$
(Complement S_{25})S_{26}	$\frac{n - \max(\sum a_i, \sum b_i)}{n - \sum T(a_i, b_i)}$
S_{27}	$\frac{\sum T(a_i, b_i)}{n}$
(Complement S_{27})S_{28}	$\frac{n - \sum a_i - \sum b_i + \sum T(a_i, b_i)}{n}$

III. THE PROPOSED FUZZY SIMILARITY MEASURE FOR EVALUATE SIMILARITY BETWEEN FEATURES

In this method, for features of a standard data set, we define a triangular fuzzy number. In this case, the minimum and maximum values in each feature is defined respectively the lower and upper fuzzy numbers. Each feature a triangular fuzzy number will vary according to the center. Whatever triangular fuzzy number related to more asymmetric features, the degree of similarity between two features is greater. That degree of similarity between two feature is calculated as follows:

We Use Triangular Fuzzy number $A(x)$, equation (2) which is defined as Follows:

Let (m, 0) divides, internally, the base of the triangle in ratio $P : 1$, where P is real positive number ($P \in \mathbb{R}^+$).

$$\frac{P}{1} = \frac{m - \alpha}{\beta - m} \Rightarrow m = \frac{\alpha + P\beta}{P + 1} \quad (30)$$

Where α=minimum value of a feature and β=maximum value of a feature.

In [15], Hsieh and Chen proposed a similarity measure using the "graded mean integration representation distance", where the degree of similarity $S(A,B)$ between fuzzy numbers A and B can be calculated with equation (21).

Where

$$d(A,B) = |P(A) - P(B)| \quad (31)$$

$P(A)$ and $P(B)$ are the graded mean integration representations of A and B, respectively. If A and B are triangular fuzzy numbers, where $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$, then the graded mean integration representations $P(A)$ and $P(B)$ of A and B, respectively, are defined as follows [15], [16]:

$$P(A) = \frac{a_1 + 4\,a_2 + a_3}{6} \quad (32)$$

$$P(B) = \frac{b_1 + 4\,b_2 + b_3}{6} \quad (33)$$

It is obvious that the larger the value of $S(A,B)$, the more the similarity between the fuzzy numbers A and B. Finally, we average degree of similarity between features for a data set using equation (34) are calculated.

$$\overline{S_{ff}} = \frac{\sum_{i=0}^{n-1} S(f_i, f_{i+1})}{n} \quad (34)$$

IV. EXPERIMENTS

In this paper, the proposed fuzzy similarity measure of the features on our four data sets taken from the UCI were tested [17]. Characteristics of the data sets in come Table (3).

Table (3): Description of the used data sets.

No.	Data sets	Features	Samples
D_1	CNAE – 9	857	1080
D_2	Semeion Handwritten Digit	266	1593
D_3	Madelon	500	1800
D_4	Dbworld_bodies_stemmed	3721	64

The proposed fuzzy similarity measure for features of a data set, first for each feature we define a triangular fuzzy number. Then using equations (21) and (31), we compute the similarity features. Finally, average degree of similarity of features in a data set to get and based on these parameter we compare the proposed similarity measure with ∩ - based fuzzy similarity measures. Comparing the results come in Table (4).

Table (4): Calculated $\overline{S_{ff}}$ to compare the degree of similarity of features.

Data sets	Similarity average feature – feature ($\overline{S_{ff}}$)		
	proposed fuzzy similarity measure	P	∩ - based fuzzy similarity measures
D_1	0.1958	9	0.2730
D_2	0.22	0.5	0.2641
D_3	0.4715	10	0.9565
D_4	0.033	0.125	0.543

In the feature selection algorithms of machine learning, the average correlation feature – feature less is better. In table (4), we see that the average degree of similarity between the features of the new method is better than the other fuzzy similarity measures. We were able to change parameter P until a more asymmetrical triangular fuzzy number for each feature is created. This change center triangular fuzzy number led that to obtain the best $\overline{S_{ff}}$.

V. CONCLUSION

In this paper, we have presented a new similarity measure to calculate the degree of similarity between features using triangular fuzzy numbers. Firstly, we use the concept of triangular fuzzy number to determine fuzzy number for each feature and then to calculate the degree of similarity between features. Proposed fuzzy similarity measure between features improved, as the performance of this method is shown in Table (4). Of this average the degree of similarity between features improvement can be used in feature selection methods.

REFERENCES

1. A. Tversky and I. Gati, "Similarity, separability, and the triangle inequality," *Psychol. Rev.*, vol. 89, no. 2, 1982, pp. 123–154.
2. M. De Cock and E. Kerre, "On (un)suitable fuzzy relations to model approximate equality," *Fuzzy Sets Syst.*, vol. 133, no. 2, 2003, pp. 137–153.
3. F. Klawonn, "Should fuzzy equality and similarity satisfy transitivity?" *Fuzzy Sets Syst.*, vol. 133, no. 2, 2003, pp. 175–180.
4. E. Hullermeier, "Fuzzy methods in machine learning and data mining: Status and prospects," *Fuzzy Sets Syst.*, vol. 156, no. 3, 2005, pp. 387–407.
5. Ikou Kaku, Jiafu Tang, JianMing Zhu, Yong Yin, (2010) Data mining: concepts, methods and applications in managements and engineering de-
6. sign. ISBN 978-1-84996-337-4 DOI 10.1007/978-1-84996-338-1, Springer London Dordrecht, Heidelberg, New York.
7. Zadeh, L.A., Fuzzy sets, Info and Control, 8, 1965, pp. 338 – 353.
8. Jose Galindo".Handbook of Research in Fuzzy Information Processing in Databases", Information science Reference, 2008.
9. Luukka, P. , Koloseni, D. , Feature selection using Fuzzy Entropy measures with Yu's Similarity measure, Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on. 2012, pp. 1 – 6.
10. Kalle Saastamoinen (2008) many valued algebraic structure as measures of comparison. PhD thesis, Lappenranta University of Technology.
11. Gottwald Siegfried (1993) Fuzzy sets and fuzzy logic. Arti_cial intelligence, ISNB 3-528-05311-9.
12. Bustince, H. Indicator of inclusion grade for interval valued fuzzy sets: Application to approximate reasoning based on interval-valued fuzzy sets, Int. J. Approx. Reasoning, V.23, 2000, pp.137-209.
13. Laszlo Koczy, T., Domonkos, T. Fuzzy rendszerek, Typotex, 2000.
14. Williams, J., Steele, N. Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes, Fuzzy Sets and Systems, V.131, 2002, pp.35-46.
15. Santini, S., Jain, R. Similarity is a geometer, Multimedia Tools and Applications, V.5, N.3, 1997, pp.277-306.
16. C. H. Hsieh and S. H. Chen, "Similarity of generalized fuzzy numbers with graded mean integration representation", in Proceedings of the Eighth International Fuzzy Systems Association World Congress, vol. 2, Taipei, Taiwan, Republic of China, 1999, pp. 551 – 555.
17. Bernard De Baets UGent and Hans De Meyer UGent (2001) EUSFLAT Conference, 2nd, Proceedings. pp.249 – 252.
18. Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 2007.