퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 5, pp. 655-659

Selection of data set with fuzzy entropy function

Sang-Hyuk Lee, Seong-Pyo Cheon and Sungshin Kim

School of Electrical Engineering, Pusan National University
30 Jangjeon-dong, Geumjeong-gu, Busan 609-735, Korea

Abstract

In this literature, the selection of data set among the universe set is carried out with the fuzzy entropy function. By the definition of fuzzy entropy, the fuzzy entropy function is proposed and the proposed fuzzy entropy function is proved through the definition. The proposed fuzzy entropy function calculate the certainty or uncertainty value of data set, hence we can choose the data set that satisfying certain bound or reference. Therefore the reliable data set can be obtained by the proposed fuzzy entropy function. With the simple example we verify that the proposed fuzzy entropy function select reliable data set.

Key words : Fuzzy entropy, distance measure, reliable data set

I. Introduction

Generally for the linear system, reliable input invokes reliable output. Hence the reliable data selection is necessary in the view of reliable result. Previous results concerning this area are related to the pattern recognition and information theory. Pattern recognition is generally used for classifying patterns[1]. Geometrical distance between data sets are mainly used to classify patterns. Also it is well known that the entropy represents the uncertainty of the fact. Hence entropy has been studied in the field of information theory, thermodynamics, or system theory etc.. The results that entropy of a fuzzy set is a measure of fuzziness of the fuzzy set are reported by the numerous researchers[2-9]. the axiomatic definitions of entropy was proposed by Liu. The relation between distance measure and fuzzy entropy was viewed by Kosko. Bhandari and Pal gave a fuzzy information measure for discrimination of a fuzzy set A relative to some other fuzzy set B. Pal and Pal analyzed the classical Shannon information entropy. Also Ghosh used this entropy to neural network. However, these studies are focussed at the design of entropy function and analysis of fuzzy entropy measure, distance measure and similarity measure. Hence we carried out the application of fuzzy entropy to the selection of reliable data set among the universe set.

In this paper, we derived the fuzzy entropy with distance measure. The proposed fuzzy entropy is constructed by the Hamming distance measure, and which has the simple structure compared to the previous proposed entropy[8]. With the proposed entropy, we verify the usefulness through the application of fuzziness measure to the universe data set. We also carried out calculate fuzziness of sample data set.

In the next chapter, definitions of entropy, distance measure and similarity measure of fuzzy sets are introduced and the proof of proposed entropy is discussed. In chapter III, Construction of fuzzy membership function is proposed by the Extension Principle[11]. Also in chapter IV, simple example is carried out. Experiments that choosing middle level 5 students among 65 students are performed, and the chosen sample data are measured by the proposed entropy measure. Finally conclusions are followed in chapter V.

Notations of this paper are used those of Fan and Ma(1999).

II. Fuzzy entropy

In this chapter, we introduce some preliminary results of fuzzy entropy. Measure of fuzziness is an interesting object for the fields of pattern recognition or decision theory. It is well known that the measure of entropy for the fuzzy sets represents the information of uncertainty. Measure of crisp set can be determined by classical mathematical study, whereas the concepts of fuzzy measures and fuzzy integrals had been proposed by Sugeno[12]. Recently, Liu suggested three axiomatic definitions of fuzzy entropy, distance measure and similarity measure as Definition 2.1 3[4]. By these definitions, we can induce entropy which is satisfying definition of fuzzy entropy, and compare it with the result of Fan et al.

접수일자 : 2004년 3월 31일
완료일자 : 2004년 6월 7일
This work has been supported by EESRI (R-2004-B-129), which is funded by MOCIE(Ministry of commerce, industry and energy).

Definition 2.1 [4] A real function $e: F(X)\rightarrow R^+$ is called an entropy on $F(X)$, if e has the following properties:

(E1) $e(D)=0,\ \forall D\in P(X)$

(E2) $e([1/2])=\max_{A\in F(X)}e(A)$

(E3) $e(A^*)\le e(A)$, for any sharpening A^* of A

(E4) $e(A)=e(A^c),\ \forall A\in F(X)$.

Where, $R^+=[0,\infty)$, A^c is the complement of A, and A^* is a sharpening of A. To express entropy function explicitly, distance measure is needed. Next we illustrate definition of distance measure.

Definition 2.2 [4] A real function $d: F^2\rightarrow R^+$ is called a distance measure on $F(X)$ if d satisfies the following properties:

(D1) $d(A,B)=d(B,A),\ \forall A,B\in F(X)$

(D2) $d(A,A)=0\ \ \forall A\in F(X)$

(D3) $d(D,D^c)=\max_{A,B\in F} d(A,B),\qquad \forall D\in P(X),\ \forall A,B\in F(X)$

(D4) $\forall A,B,C\in F(X)$, if $A\subset B\subset C$, then $d(A,B)\le d(A,C)$ and $d(B,C)\le d(A,C)$.

One of well known distance measure is Hamming distance. Similarity measure can be expressed as the complementary meaning of distance measure. Hence the definition is illustrated as follows.

Definition 2.3 [4] A real function $s: F^2\rightarrow R^+$ is called a similarity measure, if s has the following properties:

(S1) $s(A,B)=d(B,A),\ \forall A,B\in F(X)$

(S2) $s(A,A^c)=0\ \ \forall A\in F(X)$

(S3) $s(D,D)=\max_{A,B\in F}s(A,B),\qquad \forall D\in P(X),\ \forall A,B\in F(X)$

(S4) $\forall A,B,C\in F(X)$, if $A\subset B\subset C$, then $s(A,B)\ge s(A,C)$ and $s(B,C)\ge s(A,C)$.

Above definitions are the axiomatic, Liu also pointed out that there is an one-to-one relation between all distance measures and all similarity measures, $d+s=1$. Next, some useful related definitions are listed. If we divide universal set X into two parts D and D^c in $P(X)$, then the fuzzy entropy, fuzzy distance, and similarity are obtained by the following previous results. When we focus interesting area of universal set, then we can extend the theory of entropy, distance measure and similarity measure of fuzzy sets.

Definition 2.4. [7] Let e be an entropy on $F(X)$. Then for any $A\in F(X)$,

$$e(A)=e(A\cap D)\ +\ e(A\cap D^c)$$

is σ-entropy on $F(X)$.

Definition 2.5. [7] Let d be a distance measure on $F(X)$. Then for any $A,B\in F(X)$, and $D\in P(X)$,

$$d(A,B)=d(A\cap D,B\cap D)+d(A\cap D^c,B\cap D^c)$$

be the σ-distance measure on $F(X)$.

Definition 2.6. [7] Let s be a similarity measure on $F(X)$. Then for any $A,B\in F(X)$, and $D\in P(X)$,

$$s(A,B)=s(A\cap D,B\cup D^c)+s(A\cap D^c,B\cup D)$$

be the σ-similarity measure on $F(X)$.

From the properties of Definition 2.5, we can derive the following proposition.

Proposition 2.1 [8] Let d be a σ-distance measure on $F(X)$: then

(i) $d(A,A_{near})\ge d(A^*,A_{near})$

(ii) $d(A,A_{far})\le d(A^*,A_{far})$.

Fan, Ma and Xie also proposed the following theorem[8]. In theorem, they proposed fuzzy entropy function with the distance measure. Proposed entropy contain two crisp set A_{near} and A_{far}

Theorem 2.1[8] Let d be a σ-distance measure on $F(X)$; if d satisfies

(i) $d(\frac{1}{2}D,[0])=d(\frac{1}{2}D,D),\ \forall D\in P(X)$,

(ii) $d(A^c,B^c)=d(A,B),\ A,B\in F(X)$,

then $e(A)=d(A,A_{near})+1-d(A,A_{far})$ is a fuzzy entropy.

Fan and Xie derived new entropy via defined entropy, which is introduces by $e^*=e/(2-e)$, where e is an entropy on $F(X)$. To discriminate between entropies, we give another entropy using Fan's idea.

Theorem 2.2 If e is an entropy on $F(X)$, then $e=e^k$ is also an entropy on $F(X)$, where real number $k\ge 1$.

Proof. It is clear that $0\le e(A)\le 1$ for any $A\in F(X)$, and e satisfy Definition 2.1 as follows

(E1) : $e(D)$ is zero for $\forall D\in P(X)$, hence satisfied.

(E2) : $e([1/2])=\max_{A\in F(X)}e(A)$ is also satisfied.

(E3) : $e(A^*)\le e(A)$ is clear.

(E4) : $e(A)=e(A^c)$ is also easily proved, where $\forall A\in F(X)$.

Q.E.D

Hence the structure of Theorem 2.2 satisfies the entropy which is induced from the another entropy.

It is often required that the reliable data set selection is necessary among many data set. In this chapter, we introduce the relation of fuzzy membership function and the fuzzy entropy. Let X be a space of objects and x be a generic element of X. A classical set A, $A\subseteq X$, is defined as a collection of elements or objects $x\in X$, such that each x can either belong or not belong to the set A. Whereas a fuzzy set A in X is defined as a set of ordered pairs:

$$A = (x, \mu_A(x)) | x \in X$$

where $\mu_A(x)$ is called the membership function for the fuzzy set A. The membership function maps each element of X to a membership grade between 0 and 1.

By the results of Liu, if the fuzzy entropy function expressed by the following Hamming distance measure for the between fuzzy sets A and B, :

$$d(A, B) = \frac{1}{n} \sum_{i=1}^{n} |\mu_A(x_i) - \mu_B(x_i)| \quad (1)$$

where $X = x_1, x_2, \cdots x_n$.

Fuzzy entropy means the uncertainty of the fuzzy set, hence it represents the two times of the shaded area of Fig. 1 [7,8]. In Fig. 1, A_{near} denotes the crisp set of fuzzy set A.

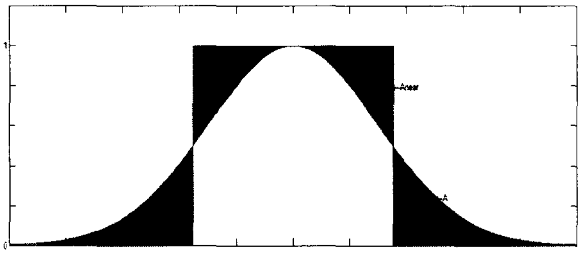


Fig 1. representation of entropy

The more fuzzy set close to the crisp set A_{near}, the more membership function become certain. In the next theorem, we propose fuzzy entropy function with the Hamming distance. Which is different from Theorem of Fan, Ma and Xie[8]. Proposed entropy needs only A_{near} crisp set, and it has the advantage in computation of entropy.

Theorem 2.3 Let d be a σ-distance measure on $F(X)$; if d satisfies $d(A^c, B^c) = d(A, B)$, $A, B \in F(X)$, then

$$e(A) = 2d((A \cap A_{near}), [1]) + 2d((A \cup A_{near}), [0]) - 2 \quad (2)$$

is a fuzzy entropy.

Proofs are carried out by showing whether the (2) satisfy Definition 2.1, and it can be found in [10]. Hence we omit them in here.

Theorem 2.3 uses only A_{near} crisp set, hence we can consider complementary entropy function. Which considers only A_{far}, and it has more compact form than Theorem 2.3.

Theorem 2.4 Let d be a σ-distance measure on $F(X)$; if d satisfies $d(A^c, B^c) = d(A, B), A, B \in F(X)$, then

$$e(A) = 2d((A \cap A_{far}), [0]) + 2d((A \cup A_{far}), [1]) \quad (3)$$

is a fuzzy entropy.

Proofs are similar to those of Theorem 2.3, and it is also found in [10].

Proposed entropies Theorem 2.3 and 2.4 have some advantages to the Liu's, they don't need half part of assumption of Theorem [8] to prove (2) and (3). Furthermore (2) and (3) use only one crisp sets A_{near} and A_{far}, respectively. By the computational results, (2) and (3) are the σ-entropy on $F(X)$. This σ-entropy property has the advantage on the computation burden. It can be shown in the literature[11].

III. Fuzzy membership function design

In this chapter, we introduce fuzzy membership function construction with the Extension Principle[10]. Zadeh had proposed Extension principle for extending nonfuzzy mathematical concepts to fuzzy sets.

Extension Principle

Let $X_1, X_2, \cdots, X_n$, and Y be nonempty crisp sets, be the product set of $X_1, X_2, \cdots, X_n$, and f be mapping from X to Y. Then, for any given n fuzzy sets $A_i \in F(X_i)$, $X = X_1 \times X_2 \times \cdots \times X_n$, $i = 1, 2, \cdots, n$, we can induce a fuzzy set $B \in F(Y)$ through f such that

$$\mu_B(y) = sup_{y = f(x_1, x_2, ..., x_n)} \min[\mu_{A_1}(x_1), \mu_{A_2}(x_2), \ldots \mu_{A_n}(x_n)],$$

where we use the convention

$$sup_{x \in \emptyset} x | x \in [0, \infty] = 0 \text{ if } f^{-1}(y) = \emptyset.$$

As an example, we can obtain a binary operator $*$ on fuzzy sets $A, B \in F(X)$

$$\mu_{A*B}(z) = sup_{x*y=z} [\mu_A(x) \wedge \mu_B(y)], \forall z \in X.$$

Let A and B be fuzzy numbers. Then $A+B$ is defined by

$$\mu_{A+B}(z) = sup_{x+y=z} [\mu_A(x) \wedge \mu_B(y)], \forall z \in X.$$

Then, for any given 2 fuzzy sets $A, B \in F(X_i)$, we can induce a fuzzy set $A+B \in F(Y)$. From this procedure, we can modify fuzzy membership functions for our purpose. Next chapter, we measure the one course of the study uncertainty for one class. One class consist of 65 students. For the simplicity Gaussian distribution function of the scores is modified to the membership function. After obtaining Gaussian function, normalization and truncation of the function is carried out to satisfy membership function.

IV. Illustrative Example

We illustrate the example of reliable data set selection from the universe sets. Statistical mean and variance do not propose the fuzziness or reliability "how much". Hence with the help of the definition of entropy, fuzzy measure has introduced in Chapter II. It is assumed that one class consist of 65 students. Educational level can be classified by the two viewpoints, the one is the heuristic representation and the other is the grade. Mean of 65 students reveals 53.73, and the average level student membership function is shown in Fig. 2. As is explained in Fig. 1, the shaded area of Fig. 2 stands for the uncertainty of average level. The average level students have the grade of B and C, which points are between 37 and 71. In this case, level of chosen 5 students can be measured by the entropy function which are illustrated in (2) and (3). Furthermore, crisp set of grade B and C is also represented in Fig. 2 through rectangle.

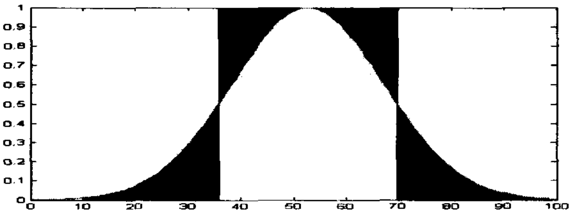


Fig. 2 Average level student membership function and B, C grade

First we choose 5 students randomly. Students points are illustrated by s_1-s_5 in Fig. 3. And its fuzzy entropy value can be calculated with eqs. (2) or (3). Actual values, membership function value and fuzzy entropy values are shown in Table 1. Hence the fuzzy entropy value of the group is 0.2587. Next, repeating this procedure we obtained following results. Additional experiments are carried, and the 5 students are chosen as Fig. 4 and 5. Table 2 and 3 shows the entropy values of the samples. Each time we choose 5 students randomly. Next calculating (2) or (3), then we can calculate student group entropy. As obtained value reaches to zero, it means that student group has higher tendency contained in B, C grade statistically.

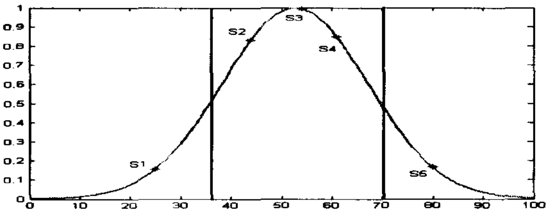


Fig. 3 Selection of 4 students

Table 1 Point, membership value and entropy value of samples(Fig.3 case)

Sample	Point	Membership value	Fuzzy entropy value
S1	25	0.1565	0.3129
S2	44	0.8319	0.3363
S3	54	0.9962	0.0077
S4	61	0.8483	0.3034
S5	80	0.1667	0.3333

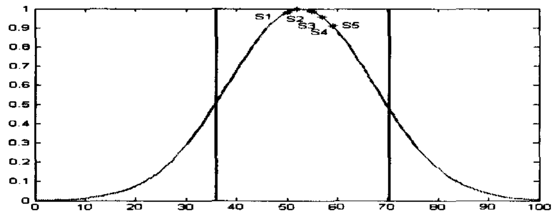


Fig. 4 Selection of 5 students

Table 2 Point, membership value and entropy value of samples(Fig. 4 case)

Sample	Point	Membership value	Fuzzy entropy value
S1	50	0.9821	0.0358
S2	52	0.9987	0.0026
S3	55	0.9877	0.0245
S4	57	0.9572	0.0857
S5	59	0.9098	0.1804

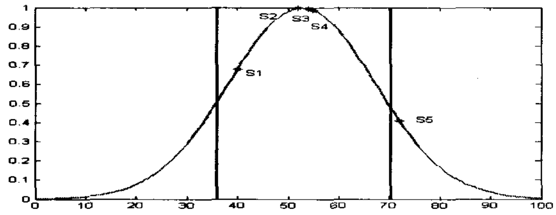


Fig. 5 Selection of 5 students

Table 3 Point, membership value and entropy value of samples(Fig. 5 case)

Sample	Point	Membership value	Fuzzy entropy value
S1	40	0.6762	0.6475
S2	52	0.9987	0.0026
S3	54	0.9962	0.0077
S4	55	0.9877	0.0245
S5	72	0.4088	0.8177

By the Table 1, 2 and 3, the mean values of each trials are 52.8, 54.6, and 54.6 respectively. Statistical results show that sample means of each case are similar to the total average, furthermore 2nd and 3rd trials illustrate same mean values. It is not easy to choose which trial are represents the middle level. Even though 2nd and 3rd trials have the same means, it also represent different meaning in the heuristic viewpoint. Whereas the fuzzy entropy average values of the each groups are 0.2587, 0.0658 and 0.3. By the meaning of fuzzy entropy, it is clear that the 2nd trial is the most reliable. From this experiment, we can offer the metric value to the heuristic viewpoint. As an example, if we refer the reliability level like as 0.2 or *etc.*, then middle level students collection can be possible with some objective guideline. Similarly, it is also possible to collect high level or low level students among 65 students. These results can be extended multi dimensional case with careful consideration. For the multi dimensional case, newly defined membership function will be required.

V. Conclusions

In this literature, we derive the fuzzy entropy with distance measure. The proposed fuzzy entropy is constructed by the Hamming distance measure, and which has the simple structure compared to the previous proposed entropy. With the proposed entropy, the usefulness is verified through the application of measure the fuzziness to the sampled set among universe data set. Through the example, we can verify fuzzy entropy has the different meaning compared to the statistical results. Next, it is valuable that extension of this study to the multi dimensional case or continuous data monitoring.

References

[1] J. C. Bezdek and S.K. Pal, *Fuzzy models for pattern recognition*, IEEE, 1992.

[2] D. Bhandari, N. R. Pal, "Some new information measure of fuzzy sets", Inform. Sci. 67, 209-228, 1993.

[3] A. Ghosh, "Use of fuzziness measure in layered networks for object extraction: a generalization", Fuzzy Sets and Systems 72, 331-348, 1995.

[4] B. Kosko, *Neural Networks and Fuzzy Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[5] Liu Xuecheng, "Entropy, distance measure and similarity measure of fuzzy sets and their relations", Fuzzy Sets and Systems, 52, 305-318, 1992.

[6] N. R. Pal, S. K. Pal, "Object-background segmentation using new definitions of entropy", IEEE Proc. 36, 284-295, 1989.

[7]J. L. Fan, W. X. Xie, "Distance measure and induced fuzzy entropy", Fuzzy Set and Systems, 104, 305-314, 1999.

[8]J. L. Fan, Y. L. Ma, and W. X. Xie, "On some properties of distance measures", Fuzzy Set and Systems, 117, 355-361, 2001.

[9] S. H. Lee and S. S. Kim, "On some properties of distance measures and fuzzy entropy", Proceedings of KFIS Fall Conference 2002, 9-12.

[10] S. H Lee, K. B. Kang and S. S. Kim, "Measure of fuzziness with the fuzzy entropy", to be appeared in Journal of Fuzzy Logic and Intelligent Systems, 2004.

[11] M. Sugeno, Theory of fuzzy integrals and its applications, Ph.D. Dissertation, Tokyo Institute of Technology, 1974.

[12] Z. Wang and G. T. Klir, *Fuzzy measure theory*, New York : Plenum Press, 1992.

[13] *Development of fault detection algorithm with the stator current*, POSCON, 2003.

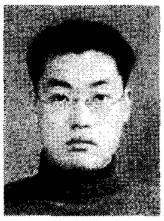
저 자 소 개

이상혁(Sang-Hyuk Lee)

1988년 충북대학교 전기공학과 졸업, 1991년 서울대학교 대학원 전기공학과 졸업(공학석사), 1998년 동대학원 전기공학부 졸업(공학박사), 2003년 충남대학교 대학원 수학과 졸업(이학석사), 1996~1999년 (주)하우 기업부설연구소 책임연구원, 1999~2000년 (주) 지앤티씨 기술이사, 2000년~현재 부산대학교 전자전기정보컴퓨터공학부 기금교수.

관심분야 : 퍼지 이론, 강인제어, 신호처리

E-mail : leehyuk@pusan.ac.kr



천성표(Seong-Pyo Cheon)

1999년 부산대학교 전기공학과 졸업, 2001년 부산대학교 대학원 전기공학과졸업(공학석사), 2001년~2003년 LG CNS 근무, 2004년~현재 동대학원 전기공학과 박사과정

관심분야 : 퍼지 이론, 데이터 마이닝, 지능제어

E-mail : buzz74@pusan.ac.kr

김성신(Sungshin Kim)

1997년 Georgia Institute of Technology 전기공학과졸업(공학박사).

1998년~현재 : 부산대학교 전기공학과 조교수

관심분야 : 지능제어, 웨이브릿, 데이터 마이닝, 공정최적화

E-mail : sskim@pusan.ac.kr