A Numerical Classification Technique Based on Fuzzy Soft Set Using Hamming Distance

Iwan Tri Riyadi Yanto[1(✉)], Rd Rohmat Saedudin[2], Saima Anwar Lashari[3], and Haviluddin[4]

[1] Department of Information Systems, University of Ahmad Dahlan, Kampus III UAD, Jalan Prof. Dr. Soepomo, Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id
[2] School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia
rdrohmat@telkomuniversity.ac.id
[3] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Parit Raja, Johor, Malaysia
[4] Faculty of Computer Science and Information Technology, Mulawarman University, Samarinda, Indonesia
haviluddin@unmul.ac.id

Abstract. In recent decades, fuzzy soft set techniques and approaches have received a great deal of attention from practitioners and soft computing researchers. This article attempts to introduce a classifier for numerical data using similarity measure fuzzy soft set (FSS) based on Hamming distance, named HDFSSC. Dataset have been taken from UCI Machine Learning Repository and MIAS (Mammographic Image Analysis Society). The proposed modeling consists of four phases: data acquisition, feature fuzzification, training phase and testing phase. Later, head to head comparison between state of the art fuzzy soft set classifiers is provided. Experiment results showed that the proposed classifier provides better accuracy when compared to the baseline fuzzy soft set classifiers.

Keywords: Fuzzy soft set (FSS) · Similarity measure · Hamming distance, classification

1 Introduction

In recent years, computers and their peripherals have been made cheaper and more readily available and in line with the development of information technology, various kinds of advanced data mining techniques have hit the market. These new age data mining techniques embrace traditional and more recent sophisticated classification algorithms. Both classification techniques are for handling complex datasets such as multidimensionality, user inference and prior knowledge, web data, spurious data points that cause overfitting of models, improvement in human ability, noisy datasets cleaning, mining multimedia datasets and incremental datasets. Interdisciplinary data mining techniques and approaches can be used for all the above mentioned databases

R. Ghazali et al. (eds.), *Recent Advances on Soft Computing and Data Mining*,
Advances in Intelligent Systems and Computing 700,
https://doi.org/10.1007/978-3-319-72550-5_25

for forecasting the impact and discovering meaningful relationships in the data with the purpose of extracting useful information for knowledge generation [1].

Thus, a variety of models have been fitted in order to determine hidden patterns in the data [2, 3]. The approach that is able to produce the most accurate output and relationships pattern in the observed datasets is considered to be the most efficient in the particular model. Such approach fulfills the objective of data mining. Current data mining practices utilizes a range of model functions including classification, regression, clustering, discovering association rules and sequence analysis [4]. Hence, solutions are needed in order to manage and analyze such as complex, diverse, and huge datasets in a reasonable time complexity and storage capacity for enhanced insight and decision-making.

Molodtsov [5] investigated soft set theory that classifies the objects with help of binary information. In principle, the initial description of any object has an approximate nature and one do not need to introduce the notion of exact solution. Therefore, the problem of membership function setting does not arise in this theory. Currently, soft set theory is being rapidly progressing in several fields of sciences, engineering, economics and medicals sciences. Maji et al. [6] did further exploration and analysis on soft set theory by providing some operations and viability of soft set into decision making problems. Thus, the scope application of fuzzy soft set theory is still available to be expand. The numerical data classification is one of the potential applications of it. A novel classification method using notions on soft set theory has been proposed by Mushrif et al. [7] on natural texture where the type data consist of a numerical value between [0,1].

The measurement of the similarity has an important role in classification using FSS. Currently, some use on measuring similarity have been carried out [8, 9]. Handaga et al. [10] proposed an algorithm namely FSSC which provides high accuracy and the proposed FSSC used general similarity measure of FSS. However, there are various distance measures in mathematics. A new similarity measures of FSS based on different distance measures has been proposed by Feng and Zheng [11]. The similarity measure based on Hamming distance in this paper is more reasonable. Thus, we propose an alternative technique for classification based on FSS similarity measurement using hamming distance which is has good performance in term of accuracy and time responses as compared existing FSS classifiers. Eight dataset have been used.

The rest of the paper is organized as follows: soft set and FSS theory are introduced in Sect. 2. Section 3 discusses similarity measure and distance measure in details. Section 4 demonstrates fuzzy soft set classification using hamming distance. Section 5 exhibits the experimental results and summary of paper are given in Sect. 6.

2 Soft Sets and FSS (Fuzzy Soft Sets)

2.1 Soft Sets

Let U be an initial universe set and E be a set of parameters. Parameters are properties of objects, it ca be known as attribute, factor or characteristics of objects. Let the power set of U as $P(U)$ and $A \subset E$ [3]. A family of subsets parameters of the universe U is

called as soft set and can be defined as A pair (F, A), where F is a mapping given by, $F.A \rightarrow P(U)$.

2.2 Fuzzy Soft Sets

Let U be an initial universe set and E be a set of parameters (which are fuzzy words or sentences involving fuzzy words). Let $P(U)$ denotes the set of all fuzzy sets of U. Let $A \subset E$. A pair (F, A) is called a fuzzy soft set (FSS) over U, where F is a mapping given by $F.A \rightarrow P(U)$.

Example 1 Suppose a FSS (F, E) describes desirability of the gowns with respect to the given parameters, which Mr X going to wear $U = \{g_1, g_2, g_3, g_4, g_5\}$ which is the set of gown under consideration and E is a set of decision parameters $E = \{e_1, e_2, e_3, e_4, e_5\}$. Let $P(U)$ be the collection of all fuzzy subsets of U. Also let E = e1 = "exclusive", e2 = "striking", e3 = "vibrant", e4 = "inexpensive", e5 = "warm".

Meanwhile, mapping models $F : E \rightarrow P(U)$ is given by where (.) is to be filled in by one of parameters $e \in E$. Suppose that;

$$\begin{aligned} F(e_1) &= \{g_2, g_4\} \\ F(e_2) &= \{g_1, g_3\} \\ F(e_3) &= \{g_3, g_4, g_5\} \\ F(e_4) &= \{g_1, g_3, g_5\} \\ F(e_5) &= \{g_1\} \end{aligned}$$

Hence, $F(e_1)$ means expensive shirt whose functional value is the set $\{g_2, g_4\}$. Therefore, soft set (F, E) as a collection of approximates as follows:

$$(F, E) = \left\{ \begin{array}{l} \text{exclusive gown} = \{g_2, g_4\}, \\ \text{striking gown} = \{g_1, g_3\}, \\ \text{vibrant gown} = \{g_3, g_4, g_5\}, \\ \text{inexpensive gown} = \{g_1, g_3, g_5\}, \\ \text{warm gown} = \{g_1\}. \end{array} \right\}$$

Now suppose that
Let

$$\begin{aligned} F(e_1) &= \{y_1/0.2, y_2/0.3, y_3/0.5, y_4/0.5, y_5/0.0\}, \\ F(e_2) &= \{y_1/1.0, y_2/0.6, y_3/0.8, y_4/0.6, y_5/0.0\}, \\ F(e_3) &= \{y_1/0.2, y_2/0.4, y_3/0.4, y_4/0.7, y_5/1.0\}, \\ F(e_4) &= \{y_1/0.3, y_2/1.0, y_3/0.1, y_4/0.3, y_5/0.2\}. \end{aligned}$$

Then a FSS (F, E) represents the family $\{F(e_i); i = 1, 2, 3, 4\}$ of $P(U)$ and the FSS (F, E) can be represented as shown in Table 1.

Table 1. Representation of FSS (F, E)

U/E	e_1	e_2	e_3	e_4
y_1	0.2	1.0	0.2	0.3
y_2	0.3	0.6	0.4	1.0
y_3	0.2	0.8	0.4	0.1
y_4	0.5	0.6	0.7	0.3
y_5	0	0	1.0	0.2

3 Similarity Measure and Distance Measure

A similarity between two entities is one of the measurement models in data grouping and clustering. In this study, the fuzzy soft set were measured based on the normalized Hamming distance [10]. Where, assume that the fuzzy soft set (F, A) and (G, B) have the same parameter set, namely, $A = B$. The normalized Hamming distance and normalize distance in FSS using Eqs. (1) and (2).

$$d_1((F,A),(G,B)) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|F(e_i)(x_j) - G(e_i)(x_j)\right| \tag{1}$$

and

$$d_2((F,A),(G,B)) = \frac{1}{mn}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left|F(e_i)(x_j) - G(e_i)(x_j)\right|^2\right)^{\frac{1}{2}} \tag{2}$$

Example 2 As in [10] let $U = \{u_1, u_2, u_3\}$ be as set with parameters $= \{a_1, a_2, a_3\}$. Given two FSS (G, A) and (H, A) are represented by two tables, Tables 2 and 3.

Table 2. Fuzzy set (G, A)

(G,A)	a_1	a_2	a_3
u_1	0.7	0.8	0.6
u_2	0.6	0.7	0.5
u_3	0.5	0.8	0.8

Table 3. Fuzzy set (H, A)

(H,A)	a_1	a_2	a_3
u_1	0.5	0.6	0.9
u_2	0.7	0.8	0.6
u_3	0.4	0.8	1

From Eqs. (1) and (2), respectively, the distance between (G,A) and (H,A) can be calculated as follows

$$\begin{aligned} & d_1((G,A),(H,A)) \\ & = \frac{1}{3\times 3}\sum_{i=1}^{3}\sum_{j=1}^{3}(0.2+0.1+0.1+0.2+0.1+0+0.3+0.1+0.2) \\ & \approx 0.144 \end{aligned}$$

and

$$\begin{aligned} & d_2((F,E),(G,E)) \\ & = \frac{1}{3\times 3}\sum_{i=1}^{3}\sum_{j=1}^{3}\left(0.2^2+0.1^2+0.1^2+0.2^2+0.1^2+0^2+0.3^2+0.1^2+0.2^2\right)^{\frac{1}{2}} \\ & \approx 0.056 \end{aligned}$$

Feng and Zheng [11] extend Eq. (3) into a generalized normalized distance in FSS by using Eq. (3).

$$d((F,A),(G,B)) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\left|F(e_i)\left(x_j\right) - G(e_i)\left(x_j\right)\right|^p)^{\frac{1}{p}}, (p \in N_+) \tag{3}$$

Clearly, if $p = 1$, then Eq. (3) is reduced to Eq. (4).
From Eq. (4), it can be know that

$$d' = \frac{1}{n}\sum_{j=1}^{n}\left|F(e_i)\left(x_j\right) - G(e_i)\left(x_j\right)\right| \tag{4}$$

Indicate the distance between the ith parameter of (F,A) and (G,B), and $d_1((F,A),(G,B))$ indicates that the distance among all parameters of (F,A) and (G,B).

4 Fuzzy Soft Set Classification Using Hamming Distance (HDFSSC)

As shown in Fig. 1, the proposed modelling comprises of three phases that feature *fuzzyfication*, training phase and testing phase. As, data acquisition is one of the crucial elements to design and develop a successful classifier. Data collections were gathered from University of California at Irvine (UCI) machine learning repository and Mammographic Image Analysis Society (MIAS) datasets. Therefore, Table 4 provides description of these dataset.

Table 4. Dataset description

NO	Dataset	Description
1.	BCWO	Breast cancer Wisconsin (original)
2.	SYM8HARD	Sym8 (Hard threshold) Level 1
3.	DB3ROISOFT	Daub3 (Soft threshold) Level 1
4.	DB3SOFT LEVEL4S1	Daub3 (Soft threshold) Level 4
5.	SYM8HARD LEVEL4	Sym8 (Hard threshold) Level 4
6.	DB3SOFT LEVEL4S2	Daub3 (Soft threshold) Level 4
7.	SYM8HARD LEVEL8	Sym8 (Hard threshold) Level 4
8.	DB3HARD LEVEL8S2	Daub3 (Hard threshold) Level 8

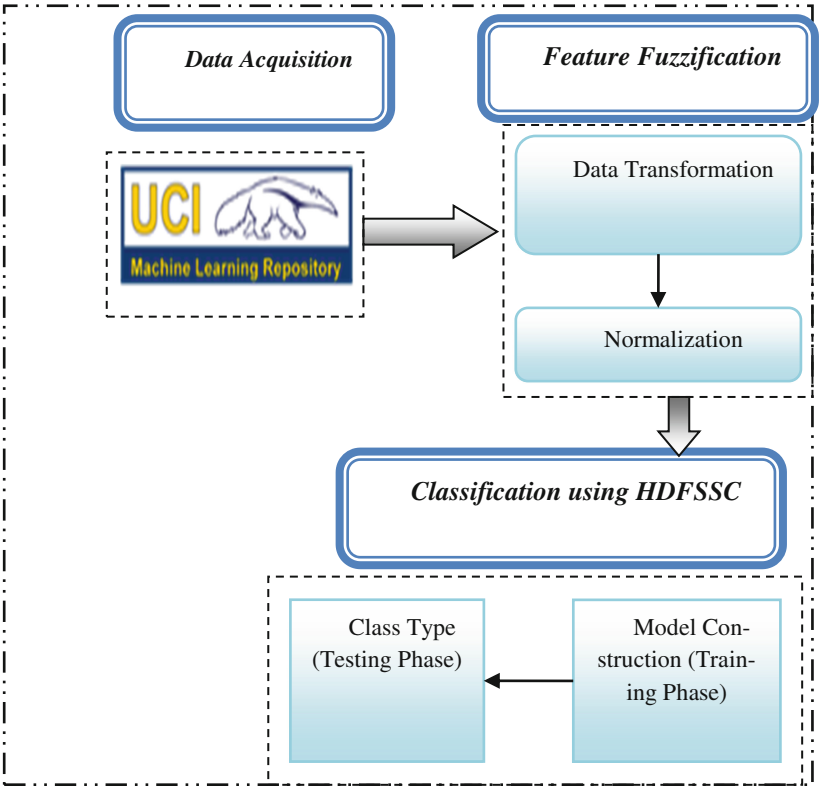

Fig. 1. Proposed modelling for HDFSSC

The HDFSSC algorithm is divided into three phases. The first is feature *fuzzification*. It is to obtain a feature vector for all data including training and testing dataset. The second is training, which to obtain a fuzzy soft is set model for each class. The last is classification, which is to label the unknown data to the target class. The algorithm is shown in the Fig. 2.

Pre-processing phase

a. Feature fuzzification to obtain a feature vector E_{wi}, for i=1,2, …, N For all data, training and testing dataset.

Training phase

b. Given N samples obtained from the data class w.

c. Calculate the cluster center vector E_w using.

$$E_w = \frac{1}{N}\sum_{i=1}^{N} E_{wi}$$

d. Obtain a fuzzy soft set model for class w, where ($\underline{F}_w$, E), is a cluster center vector for class w having D features.

e. Repeat steps (b), (c) and (d) for all W classes.

Classification phase

f. Get an unknown class data.

g. Obtain a fuzzy soft set model for unknown class data, ($\underline{G}$,E).

h. Compute similarity between ($\underline{G}$,E) and ($\underline{F}_w$,E) for each w using hamming distance.

i. Assign the unknown data to class w if similarity it reach maximum.

$$w = \arg\left[\max_{w=1}^{W} S(\underline{G}, \underline{F}_w)\right]$$

Fig. 2. Classification using HDFSSC

5 Experiment Results

The proposed method is compared to baseline algorithms which have been implemented in MATLAB version 8.6.0.267246 (R2015b). The Algorithms were executed on a processor Intel @1.5 GHz (4CPUs) with 2G total memory using Windows 7 Professional 32 bit operating system sequentially. The 80% sample of data is reserved randomly for training and 20% for testing purpose. Table 5 presents the accuracy of proposed method for classification on UCI benchmark datasets. The experimental results strongly suggest that HFSSC has high accuracy even had lower complexity in the testing phase. Figure 3 illustrate the time response of BCWO data set, for the

Table 5. The comparisons in term of accuracy

Dataset	FSSC	FussCyier	HDFSSC
BCWO	0.9258	0.9439	0.9578
SYM8HARD	0.5918	0.6563	0.6603
DB3ROISOFT	0.6485	0.7002	0.7695
DB3SOFT LEVEL4S1	0.7458	0.7976	0.8181
SYM8HARD LEVEL4	0.6859	0.6292	0.7035
DB3SOFT LEVEL4S2	0.6413	0.6345	0.6765
SYM8HARD LEVEL8	0.7295	0.6750	0.7504
DB3HARD LEVEL8S2	0.7283	0.6679	0.7598

HDFSSC can reduce the time response up to 13.09% and 72.53% comparing to the *FussCyier* and FSSC, respectively.

Figure 4 illustrate the data set number 2–8. Based on the Fig. 4, the HDFSSC can reduce the time response up to 4.91% and 61.72% in average comparing to the *FussCyier* and FSSC, respectively

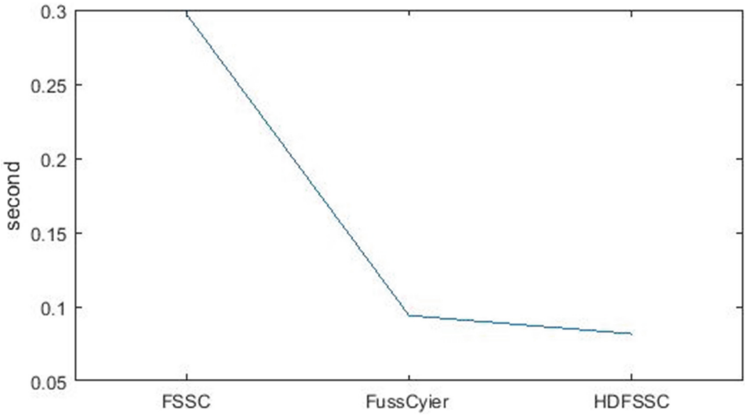

Fig. 3. Time response of BCWO

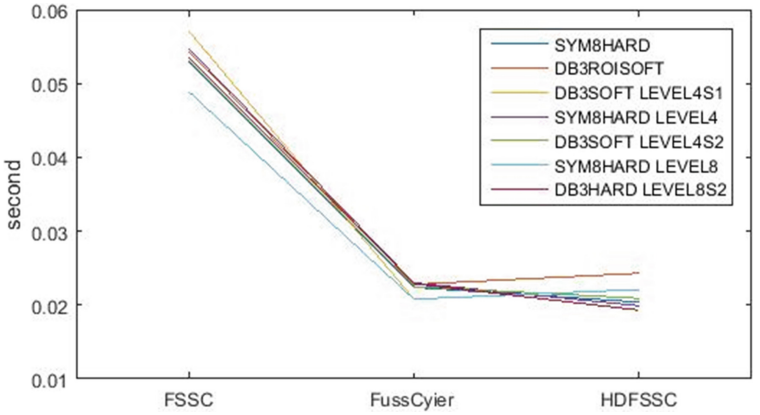

Fig. 4. Time response of dataset number 2–8

6 Conclusion

In this paper, a frameworks of hamming distance approach have been proposed, in order to obtain a balanced solution of a FSS based decision making problem. Moreover, we have justified and compared with the existing FSS classification methods i.e.,

FussCyier and FSSC by using medical datasets. In other words, we confidence that these theories have a lot of future and may serve to solve many decision-making problems.

In the current implementation of this research, the experimental setup involved UCI benchmark datasets, however, for the future works mammogram images classification will be taken into consideration which can be improved by incorporating feature selection phase before classification and the classification results can be more compact and precise.

References

1. Lashari, S.A., Ibrahim, R., Senan, N., Yanto, I.T.R., Herawan, T.: Application of wavelet de-noising filters in mammogram images classification using fuzzy soft set. In: 2016 International Conference on Soft Computing and Data Mining, pp. 529–537 (2016)
2. Yanto, I.T.R., Ismail, M.A., Herawan, T.: A modified fuzzy k-partition based on indiscernibility relation for categorical data clustering. Eng. Appl. Artif. Intell. **53,** 41–52 (2016)
3. Purnawansyah, Haviluddin: K-Means clustering implementation in network traffic activities. In: 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia, pp. 51–54 (2016)
4. Beniwal, S., Arora, J.: Classification and feature selection techniques in data mining. Int. J. Eng. Res. Technol. **1**(6) (2012)
5. Molodtsov, D.: Soft set theory—first results. Comput. Math. Appl. **37**(4–5), 19–31 (1999)
6. Maji, P.K., Roy, A.R., Biswas, R.: An application of soft sets in a decision making problem. Comput. Math. Appl. **44**(8–9), 1077–1083 (2002)
7. Mushrif, M., Sengupta, S., Ray, A.: Texture classification using a novel, soft-set theory based classification algorithm. In: Computer Vision–ACCV 2006, pp. 246–254 (2006)
8. Roy, A.R., Maji, P.K.: A fuzzy soft set theoretic approach to decision making problems. J. Comput. Appl. Math. **203**(2), 412–418 (2007)
9. Kharal, A.: Distance and similarity measures for soft sets. New Math. Nat. Comput. **6**(3), 321–334 (2010)
10. Handaga, B., Herawan, T., Deris, M.M.: FSSC: an algorithm for classifying numerical data using fuzzy soft set theory. Int. J. Fuzzy Syst. Appl. **2**(4), 29–46 (2012)
11. Feng, Q., Zheng, W.: New similarity measures of fuzzy soft sets based on distance measures. Ann. Fuzzy Math. Inf. **7**(4), 669–686 (2014)