END OF PROJECT REPORT FOR FRGS

VOTE 0736

PROJECT TITLE

SOFT SET APPROACH FOR CATEGORICAL DATA CLUSTERING AND MAXIMAL ASSOCIATION RULES MINING

HEAD OF RESEARCHER

PROF. DR. MUSTAFA BIN MAT DERIS

FRGS GRANT

VOTE 0736

2012

END OF PROJECT REPORT FOR FRGS

Project Title:
Soft Set Approach for Categorical Data Clustering and Maximal Association Rule Mining

FRGS Field:
Technology and Engineering

Project Leader:
Prof. Dr. Mustafa bin Mat Deris

Researchers:
Dr. Rozaida binti Ghazali
Tutut Herawan
Iwan Tri Riyadi Yanto
Bana Handaga (GRA-PhD)

Faculty of Computer Science and Information Technology,
University Tun Hussein Onn Malaysia (UTHM)
86400 Parit Raja, Batu Pahat, Johor

1. Executive Summary

Recent advances in information technology have led to significant changes in today's world; both generating and collecting data have been increasing rapidly. This explosive growth in stored or transient data has generated an urgent need for new techniques that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. Classification is one form of data analysis in data mining, which can be used to extract models describing important data classes. Researchers have proposed many classification methods. An important point is that each technique typically suits some problems better than others do. Thus, there is no universal data-mining method.

In 1999, Molodtsov initiated the concept of soft set theory as a mathematical tool for dealing with uncertainties. The soft set theory has a rich potential for applications in several directions. However, application of soft set theory on data classification still not widely studies. There are few researches of data classification based on soft set theory. Although those methods are quite successful for data classification, however they are still need improvement. This research aim to propose a new approach to classify data based on soft set theory, to improve the accuracy and efficiency. It is called Fuzzy Soft Set Classifier (FSSC)

Soft Set Theory

Definition 1: A soft set F_A over U is a set defined by a function f_A representing a mapping $f_A: E \rightarrow P(U)$ such that $f_A(x) = 0$ if $x \notin A$ and $A \in E$.

Here, f_A is called approximate function of the soft set F_A, and value $f_A(x)$ is a set called *x-element* of the sot set for all $x \in E$. It is worth noting that the sets $f_A(x)$ may be arbitrary, empty, or have nonempty intersection. Clearly, a soft set is not a (crisp) set. Thus, a soft set over U can be represented by the set of ordered pairs

$$F_A = \{(x, f_A(x)): x \in E, f_A(x) \in P(U)\}$$

Fuzzy Soft Set

Definition 2: An fuzzy soft (*fs*)-set Γ_A over U is a set defined by a function γ_A representing a mapping $\gamma_A: E \rightarrow F(U)$ such that $\gamma_A(x) = 0$ if $x \notin A$.

Here, γ_A is called fuzzy approximate function of the *fs*-set Γ_A, and the value $\gamma_A(x)$ is a set called x-element of the *fs*-set for all $x \in E$. Thus, an *fs*-set Γ_A over U can be represented by the set of ordered pairs

$$\Gamma_A = \{(x, \gamma_A(x)): x \in E, \gamma_A(x) \in F(U)\}$$

Similarity Measure between Two Fuzzy Soft Set

Definition 3: Let Γ_A and Γ_B be two *fs*-sets over U. Then the similarity between them, denoted by $S_{A,B}$, is defined by

$$S_{A,B} \equiv \frac{\sum_{i=1}^{n}\{\overrightarrow{\gamma_A}(x_i) \cdot \overrightarrow{\gamma_B}(x_i)\}}{\sum_{i=1}^{n}\left\{\left(\overrightarrow{\gamma_A}(x_i)\right)^2 \vee \left(\overrightarrow{\gamma_B}(x_i)\right)^2\right\}}$$

where $\overrightarrow{\gamma_A}(x_i)$ is fuzzy membership vector of *fs*-sets Γ_A.

Based on tree definition above we construct document classification algorithm based on fuzzy soft set theory.

FSSC Algorithm.

Preprocessing Phase

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

Training Phase

2. Given N samples obtained from the data class w.
3. Calculate the cluster center vector E_w using Equation

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

4. Obtain a fuzzy soft set model for class w, where (Γ_w, E) is a cluster center vector for class w having D features.
5. Repeat steps (2), (3) and (4) for all W classes

Classification Phase

6. Get an unknown class data
7. Obtain a fuzzy soft set model for unknown class data, (Γ_x, E).
8. Compute similarity between (Γ_x, E) and (Γ_w, E) for each w.
9. Assign the unknown data to class w if similarity it reach maximum

$$w = \arg\left[\max_{w=1}^{W} S_{x,w}\right]$$

Comparison tests on seven datasets from UCI Machine Learning Repository have been carried out. It is shown that the proposed (FSSC) algorithm provides better accuracy and efficiency as compared to the baseline algorithm using soft set theory; see Figure-1 and Figure-2.

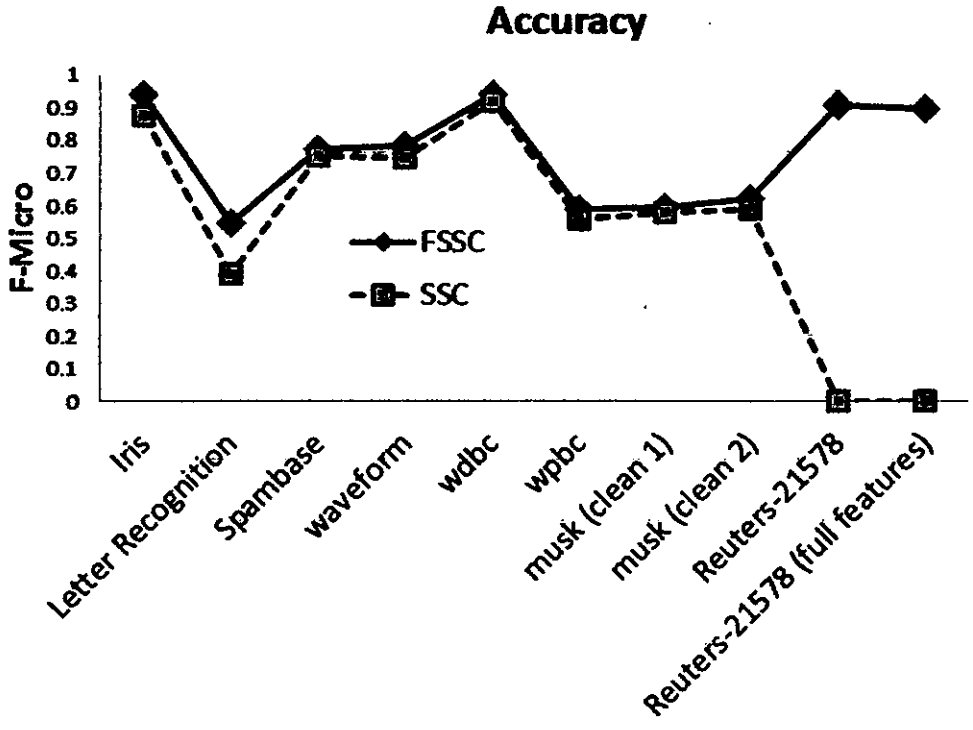

Figure 1. Comparison of accuracy between SSC & FSSC on seven UCI datasets

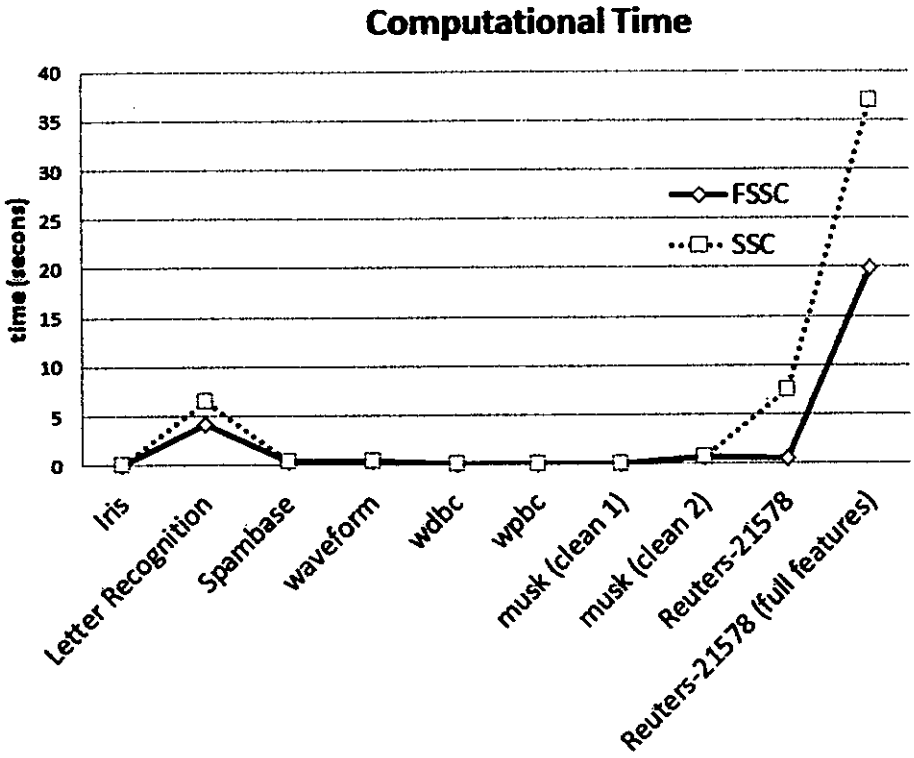

Figure 2. Comparison of efficiency between SSC & FSSC on seven UCI datasets

2. Outcomes/Achievement

Based on the grant, several outcomes have been established as follows:

1. New classification algorithm based on fuzzy soft set theory has been found.

2. Extend the application of soft set theory to classify numerical data.

3. Human Capital Development
    a. One PhD student in data mining

4. Publications at Journal/Conferences:

    a. Handaga, Bana and Mat Deris, Mustafa (2011), Similarity Approach on Fuzzy Soft Set Based Numerical Data Classification, In Zain, JasniMohamad & Wan Mohd, WanMaseribt & El-Qawasmeh, Eyas (Eds.), *Software Engineering and Computer Systems*, Communications in Computer and Information Science, Springer Berlin Heidelberg, 180 (pp. 575-589)

    b. Handaga, B., Herawan, T., & Deris, M. M. (2012). FSSC: An Algorithm for Classifying Numerical Data Using Fuzzy Soft Set Theory. *International Journal of Fuzzy System Applications* (IJFSA), 2(4), 29-46.

    c. Bana Handaga, Mustafa Mat Deris (2011), Similarity Approach on Fuzzy Soft Set Based Numerical Data Classification, In *proceeding of: Software Engineering and Computer Systems - Second International Conference*, ICSECS 2011, Kuantan, Pahang, Malaysia, June 27-29, 2011

Similarity Approach on Fuzzy Soft Set Based Numerical Data Classification

Bana Handaga and Mustafa Mat Deris

University Tun Hussein Onn Malaysia
handaga.bana@gmail.com, mmustafa@uthm.edu.my

Abstract. Application of soft sets theory for classification of natural textures has been successfully carried out by Mushrif et. al.. However the approach can not be applied in a particular classification problem, such as problem of text classification. In this paper, we propose the new numerical data classification based on similarity fuzzy soft sets. In addition can be applied to text classification, this new fuzzy soft sets classifier (FSSC) can also be used in general numerical data classification. As compare to previous soft sets classifier on seven real data sets experiments, the new proposed approach give high degree of accuracy with low computational complexity.

Keywords: fuzzy soft set theory, numerical data, classification.

1 Introduction

Classification, one of the most popular and significant machine learning areas, is particularly important when a data repository contains samples that can be used as the basis for future decision making: for example, medical diagnosis, credit fraud detection or image detection. Machine learning researchers have already proposed many different types of classification algorithms, including nearest-neighbour methods, decision tree induction, error backpropagation, reinforcement learning, lazy learning, rule-based learning and relatively new addition is statistical learning. From amongst this vast and ever increasing array of classification algorithms, it becomes important to ask the question 'which algorithm should be the first choice for my present classification problem?'

To answer this question, Ali and Smith [2] has conducted research comparing the 8 algorithms/classifiers with 100 different classification problems. The relative weighted performance measures showed that there was no single classifier to solve all 100 classification problems with best performance over the experiments. There have been numerous comparisons of the different classification and prediction methods, and the matter remains a research topic. No single method has been found to be superior over all others for all data sets [8]. This is our motivation to proposed a new classification algorithm, based on soft sets theory.

In 1999, D. Molodtsov [16], introduced the notion of a soft set as a collection of approximate descriptions of an object. This initial description of the object has an approximate nature, and we do not need to introduce the notion of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 575–589, 2011.

exact solution. The absence of any restrictions on the approximate description in soft sets make this theory very convenient and easily applicable in practice. Applications of soft sets in areas ranging from decision problems to texture classification, have surged in recent years [5,12,17,21].

The soft sets theory can work well on the parameters that have a binary number, but still difficult to work with parameters that have a real number. There are many problems in the classification involving real numbers. To overcome this problem, Maji et al. [11] have studied a more general concept, namely the theory of fuzzy soft sets, which can be used with the parameters in the form of real numbers. These results further expand the scope of application soft sets theory. There are two important concepts underlying the application of the theory of soft sets in numerical classification problems. First, the concept of decision-making problems based on the theory of fuzzy soft sets, and the second is the concept of measuring similarity between two fuzzy soft sets.

Based on an application of soft sets in a decision making problem presented by [12], Mushrif et al. [17] presented a novel method for classification of natural textures using the notions of soft set theory, all features on the natural textures consist of a numeric (real) data type, have a value between [0,1] and the algorithm used to classify the natural texture is very similar to the algorithms used by Roy and Maji [21] in the decision-making problems with the theory of fuzzy sets softs. In their experiments, Mushrif et al. used 25 texture classes with 14 texture features. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less in case of soft set method in comparison with Bayes classification method. However, soft sets approach proposed by the [17] can not work properly in particular classification problem, such as problem of text classification. This classifier has a very low accuracy, and even failed to classify text data.

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovering task, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. The study of how to measure the similarity between soft sets have been carried out by Majumdar and Samanta [15] and Kharal [9], then Majumdar and Samanta [14] is also expanding his study to measure the similarity of fuzzy soft set and describe how to used that formula in medical diagnosis to detect wheter an ill person to perform certain is suffering from a disease or not.

This paper proposed a new approach of classification based on fuzzy soft set theory, using concept of similarity between two fuzzy soft sets, we call this classifier as fuzzy soft sets classsifier (FSSC). FSSC, first construct model fuzzy soft sets $(\widetilde{F}, E)$ for each class and construct model fuzzy soft sets $(\widetilde{G}, E)$ for data without class labels. Next find the similarity measure, $S(\widetilde{F}, \widetilde{G})$, between $(\widetilde{F}, E)$ and $(\widetilde{G}, E)$. The new data will be given a class label according to the class with the highest similarity. As compare to (fuzzy) soft sets classification of natural

textures by Mushrif et. al. on two real data sets experiments, the new proposed approach give high degree of accuracy with low computational complexity. We use the seven types of data sets from UCI to compare between the proposed fuzzy soft-set classifier to soft set classifier proposed by [17].

The rest of this paper is organized as follows. In the next section, we describes soft set and fuzzy soft sets. In Section 3, describes the concepts of classification base on soft sets theory and we describes our new propose FSSC algorithm in this section. Section 4, describes the classifier evaluation methodology. In Section 5, discuss about experiments to compare between FSSC and soft sets classification algorithm proposed by Mushrif et. al.. In the final section, some concluding comments are presented.

2 Soft Set Theory

In this section, we recall the basic notions of soft sets and fuzzy soft sets. Let U be an initial universe of objects and E_U (simply denoted by E) the set of parameters in relation to objects in U. Parameters are often attributes, characteristics, or properties of objects. Let $P(U)$ denote the power set of U and $A \subseteq E$. Following [16,13], the concept of soft sets is defined as follows.

2.1 Definition of Soft Set

Definition 1. ([16]) Let U be initial universal set and let E be a set of parameters. Let $P(U)$ denote the power set of U. A pair (F, E) is called a soft set over U, if only if F is a mapping given by $F{:}A \rightarrow P(U)$.

By definition, a soft set (F, E) over the universe U can be regarded as a parameterized family of subsets of the universe U, which gives an approximate (soft) description of the objects in U. As pointed in [16], for any parameter E, the subset may be considered as the set of ϵ-approximate elements in the soft set (F, E). It is worth noting that $F(\epsilon)$ may be arbitrary: some of them may be empty, and some may have nonempty intersection [16]. For illustration, Molodtsov considered several examples in [16]. Similar examples were also discussed in [13,1].

From that definition, it is known that fuzzy set introduced by L.A. Zadeh [23] is kind of special soft sets. Let A be a fuzzy set and μ_A be a subject function (μ_A be a mapping of U into [0,1]). To this problem, Maji et al. [11] initiated the study on hybrid structures involving both fuzzy sets and soft sets. They introduced in [11] the notion of fuzzy soft sets, which can be seen as a fuzzy generalization of (crisp) soft sets.

2.2 Definition of Fuzzy Soft Sets

Definition 2. ([11]) Let U be an initial universal set and let E be set of parameters. Let $\widetilde{P}(U)$ denote the power set of all fuzzy subsets of U. Let $A \subset E$. A pair $(\widetilde{F}, E)$ is called a fuzzy soft set over U, where $\widetilde{F}$ is a mapping given by $\widetilde{F}{:}A \rightarrow \widetilde{P}(U)$.

In the above definition, fuzzy subsets in the universe U are used as substitutes for the crisp subsets of U. Hence it is easy to see that every (classical) soft set may be considered as a fuzzy soft set. Generally speaking $\widetilde{F}(\epsilon)$ is a fuzzy subset in U and it is called the fuzzy approximate value set of the parameter ϵ.

It is well known that the notion of fuzzy sets provides a convenient tool for representing vague concepts by allowing partial memberships. In the definition of a fuzzy soft set, fuzzy subsets are used as substitutes for the crisp subsets. Hence every soft set may be considered as a fuzzy soft set. In addition, by analogy with soft sets, one easily sees that every fuzzy soft set can be viewed as an (fuzzy) information system and be represented by a data table with entries belonging to the unit interval [0,1]. For illustration, we consider the following example

Example 3. This example taken from [14], suppose a fuzzy soft set $(\widetilde{F}, E)$ describes attractiveness of the shirts with respect to the given parameters, which the authors are going to wear. $U = \{x_1, x_2, x_3, x_4, x_5\}$ which is the set of all shirts under consideration. Let $\widetilde{P}(U)$ be the collection of all fuzzy subsets of U. Also let E= e_1= "colorful", e_2 = "bright", e_3 = "cheap", e_4 = "warm". Let

$$F(e_1) = \left\{\frac{x_1}{0.5}, \frac{x_2}{0.9}, \frac{x_3}{0}, \frac{x_4}{0}, \frac{x_5}{0}\right\}, \; F(e_2) = \left\{\frac{x_1}{1.0}, \frac{x_2}{0.8}, \frac{x_3}{0.7}, \frac{x_4}{0}, \frac{x_5}{0}\right\}$$

$$F(e_3) = \left\{\frac{x_1}{0}, \frac{x_2}{0}, \frac{x_3}{0}, \frac{x_4}{0.6}, \frac{x_5}{0}\right\}, \; F(e_4) = \left\{\frac{x_1}{0}, \frac{x_2}{1.0}, \frac{x_3}{0}, \frac{x_4}{0}, \frac{x_5}{0.3}\right\}$$

Then the family $\{\widetilde{F}(e_i), i = 1, 2, 3, 4\}$ of $\widetilde{P}(U)$ is a fuzzy soft set $(\widetilde{F}, E)$. Tabular representation for fuzzy soft sets $(\widetilde{F}, E)$ show in Table-1.

Table 1. Tabular representation of fuzzy Soft-Set $(\widetilde{F}, E)$

U	e_1	e_2	e_3	e_4
x_1	0.5	1.0	0	0
x_2	0.9	0.8	0	1.0
x_3	0	0.7	0	0
x_4	0	0	0.6	0
x_5	0.2	0.8	0.8	0.4

There are two important concepts underlying the application of the theory of fuzzy soft sets in numerical classification problems. First, the concept of decision-making problems based on the theory of fuzzy soft sets early proposed by Maji et. al. [21], and the second is the concept of measuring similarity between two fuzzy-soft-sets proposed by Majumdar and Samanta [14]. The following will be explained briefly about these two concepts.

2.3 Fuzzy Soft Set Based Decision Making

We begin this section with a novel algorithm designed for solving fuzzy soft set based decision making problems, which was presented in [21]. We select

this algorithm, because this algorithm have similarity with soft sets classifier algorithm proposed by Mushrif et. al. [17] for classification of natural textures. Next Feng [5] show that this algorithm is actually a useful method for selecting the optimal alternative in decision making problems based on fuzzy soft sets, while the counter example given in [10] is not sufficient for concluding that the Roy-Maji method is incorrect.

Algorithm 4. Fuzzy soft sets decision making problem algorithm by Roy-Maji [21].

1. Input the fuzzy soft sets $(\widetilde{F}, A)$, $(\widetilde{G}, B)$ and $(\widetilde{H}, C)$.
2. Input the parameter set P as observed by the observer.
3. Compute the corresponding resultant fuzzy soft set $(\widetilde{S}, P)$ from the fuzzy soft sets $(\widetilde{F}, A)$, $(\widetilde{G}, B)$, $(\widetilde{H}, C)$ and place it in tabular form.
4. Construct the comparison table of the fuzzy soft set $(\widetilde{S}, P)$ and compute r_i and t_i for o_i, $\forall i$.
5. Compute the score s_i of o_i, $\forall i$.
6. The decision is o_k if $s_k = max_i\ s_i$.
7. If k has more than one value then any one of o_k may be chosen.

Roy and Maji [21] pointed out that the object recognition problem may be viewed as a multi-observer decision making problem, where the final identification of the object is based on the set of inputs from different observers who provide the overall object characterization in terms of diverse sets of parameters. The above Algorithm-4 gives solutions to the recognition problem by means of fuzzy soft sets. This method involves construction of comparison table from the resultant fuzzy soft set and the final optimal decision is taken based on the maximum score computed from the comparison table.

The comparison table is a square table in which rows and columns both are labelled by the object names $o_1, o_2, \ldots, o_n$ of the universe, and the entries c_{ij} indicate the number of parameters for which the membership value of o_i exceeds or equals the membership value of o_j. Clearly, $0 \leq c_{ij} \leq m$, where m is the number of parameters. The row-sum r_i of an object o_i is computed by

$$r_i = \sum_{i=1}^{n} c_{ij} \tag{1}$$

Similarly the column-sum t_j of an object o_j is computed by

$$t_j = \sum_{j=1}^{n} c_{ij} \tag{2}$$

Finally, the score s_i of an object o_i is defined by

$$s_i = r_i - t_i \tag{3}$$

The basic idea of Algorithm-4 was illustrated in [21] by a concrete example (see Section 4 of [21] for details).

2.4 Similarity between Two Soft Sets

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovering task, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. Several researchers have studied the problem of similarity measurement between fuzzy sets, fuzzy numbers and vague sets. Recently Majumdar and Samanta [14,15] have studied the similarity measure of soft sets and fuzzy soft sets. Similarity measures have extensive application in several areas such as pattern recognition, image processing, region extraction, coding theory etc.

In General Fuzzy Soft Set (GFSS) [14] explain similarity between the two GFSS as follow. Let $U = \{x_1, x_2, \ldots, x_n\}$ be the universal set of elements and $E = \{e_1, e_2, \ldots, e_m\}$ be the universal set of parameters. Let F_ρ and G_δ be two GFSS over the parametrized universe (U, E). Hence $F_\rho = \{(F(e_i), \rho(e_i)), i = 1, 2, \ldots, m\}$ and $G_\delta = \{(G(e_i), \delta(e_i)), i = 1, 2, \ldots, m\}$.

Thus $\widetilde{F} = \{F(e_i), i = 1, 2, \ldots, m\}$ and $\widetilde{G} = \{G(e_i), i = 1, 2, \ldots, m\}$ are two families of fuzzy soft sets.

Now the similarity between $\widetilde{F}$ and $\widetilde{G}$ is found first and it is denoted by $M(\widetilde{F}, \widetilde{G})$. Next the similarity between the two fuzzy sets ρ and δ is found and is denoted by $m(\rho, \delta)$. Then the similarity between the two GFSS F_ρ and G_δ is denoted as $\mathcal{S}(F_\rho, G_\delta) = M(\widetilde{F}, \widetilde{G}) \cdot m(\rho, \delta)$. Here $M(\widetilde{F}, \widetilde{G}) = \max_i M_i(\widetilde{F}, \widetilde{G})$, where

$$M_i(\widetilde{F}, \widetilde{G}) = 1 - \frac{\sum_{j=1}^{n} \left| \widetilde{F}_{ij} - \widetilde{G}_{ij} \right|}{\sum_{j=1}^{n} \left(\widetilde{F}_{ij} + \widetilde{G}_{ij} \right)}, \; \widetilde{F}_{ij} = \mu_{\widetilde{F}(e_i)}(x_j) \; and \; \widetilde{G}_{ij} = \mu_{\widetilde{G}(e_i)}(x_j)$$

Also

$$m(\rho, \delta) = 1 - \frac{\sum |\rho_i - \delta_i|}{\sum (\rho_i + \delta_i)}, \; where \; \rho_i = \rho(e_i) \; and \; \delta_i = \delta(e_i).$$

if we used universal fuzzy soft set then $\rho = \delta = 1$ and $m(\rho, \delta) = 1$, now formula for similarity is

$$\mathcal{S}(F_\rho, G_\delta) = M_i(\widetilde{F}, \widetilde{G}) = 1 - \frac{\sum_{j=1}^{n} \left| \widetilde{F}_{ij} - \widetilde{G}_{ij} \right|}{\sum_{j=1}^{n} \left(\widetilde{F}_{ij} + \widetilde{G}_{ij} \right)}, \tag{4}$$

where $\widetilde{F}_{ij} = \mu_{\widetilde{F}(e_i)}(x_j) \; and \; \widetilde{G}_{ij} = \mu_{\widetilde{G}(e_i)}(x_j)$.

Example 5. Consider the following two GFSS where $U = \{x_1, x_2, x_3, x_4\}$ and $E = e_1, e_2, e_3$.

$$F_\rho = \begin{pmatrix} 0.2 & 0.5 & 0.9 & 1.0 & 0.6 \\ 0.1 & 0.2 & 0.6 & 0.5 & 0.8 \\ 0.2 & 0.4 & 0.7 & 0.9 & 0.4 \end{pmatrix} \quad \text{and} \quad G_\delta = \begin{pmatrix} 0.4 & 0.3 & 0.2 & 0.9 & 0.5 \\ 0.6 & 0.5 & 0.2 & 0.1 & 0.7 \\ 0.4 & 0.4 & 0.2 & 0.1 & 0.9 \end{pmatrix}$$

here

$$m(\rho,\delta) = 1 - \frac{\sum|\rho_i - \delta_i|}{\sum|\rho_i + \delta_i|} = 1 - \frac{0.1 + 0.1 + 0.5}{1.1 + 1.5 + 1.3} = 0.82$$

and

$$M_1(\widetilde{F},\widetilde{G}) \cong 0.73, \quad M_2(\widetilde{F},\widetilde{G}) \cong 0.43, \quad M_3(\widetilde{F},\widetilde{G}) \cong 0.50$$

$$\therefore M_1(\widetilde{F},\widetilde{G}) \cong 0.73$$

Hence the similarity between the two GFSS F_ρ and G_δ will be

$$S(F_\rho, G_\delta) = M_1(\widetilde{F},\widetilde{G}) \cdot m(\rho,\delta) = 0.73 \times 0.82 \cong 0.60$$

for universal fuzzy soft sets $\rho = \delta = 1$ and $m(\rho,\delta) = 1$, then similarity $S(F_\rho, G_\delta) = 0.73$.

3 Classification Based on Soft Sets Theory

There are two concepts that underlie the classification algorithm in the soft-sets, namely classification based decision making problem as proposed by Mushrif et. al. [17], and classification algorithms based on the similarity between two fuzzy softs sets, this algorithm is a new classification algorithm proposed in this paper. We'll discuss both in this section.

3.1 Soft Sets Classifier Based on Decision Making Problems

This classifier learns by calculating the average value of each parameters (attributes or features) from all object or instant with the same class label, to construct soft sets model with universe consisting of all of class labels. In other words, an object in the universe represents all data derived from the same class label. Furthermore, to classify the test data or data of unknown class labels. First, construct the soft sets model of data using a certain formula, then construct a comparison-table in the same manner as the preparation of comparison-table in the case of decision making problem. The next step is to calculate the score to determine the class label for the data.

Mushrif et al. using this algorithm to classify the natural texture [17], we called Soft Sets Classifier (SSC). In their experiments, Mushrif et al. used 25 texture classes with 14 texture features. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less in case of soft set method in comparison with Bayes classification method.

We can use the above algorithm to classify numerical data in general. By modifying the second step in both phases of train and stage classification. To classify numerical data with this algorithm, the second step is replaced with fuzzification process, which is like counting the normalization so that all parameters have a

value between 0 to 1. For example, if the classification algorithm is applied to the iris dataset. Fuzzification can be done by dividing each attributes value with the largest value at each attributes, $e_{fi} = e_i/max(e_i)$. Where e_i, $i = 1, 2, \ldots, n$ is the old attribute and e_{fi} is attribute with new value between [0,1]. We can use a different formula in this regard.

3.2 Soft Sets Classifier Based on Similarity between Two Fuzzy Soft Sets

This is a new approach to numerical classification algorithm based on the theory of fuzzy soft sets, which we propose in this paper, we refer to as Fuzzy Softs Set Classifier (FSSC). FSSC have the same learning phase with the previous soft sets classification algorithm, but the FSSC has a different classifier, it uses the similarity between two fuzzy-soft-sets. As described by Majumdar and Samanta [14]. FSSC complete algorithm is as follows.

Algorithm 6. FSSC algorithm

Pre-processing phase

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

Training phase

1. Given N samples obtained from the data class w.
2. Calculate the cluster center vector E_w using Equation-5.

$$E_w = \frac{1}{N} \sum_{i=1}^{N} E_{wi} \tag{5}$$

3. Obtain a fuzzy soft set model for class w, $(\widetilde{F}_w, E)$, is a cluster center vector for class w having D features.
4. Repeat the process for all W classes.

Classification phase

1. Get the one unknown class data.
2. Obtain a fuzzy soft sets model for unknown class data, $(\widetilde{G}, E)$
3. Compute similarity between $(\widetilde{G}, E)$ and $(\widetilde{F}_w, E)$ for each w using Equation (4).
4. Assign the unknown data to class w if similarity is maximum.

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

If FSSC applied to the iris with fuzzification formula and data distribution (train and test) are same with the case of iris data classification in the previous examples. The results are as follows, after learning phase the fuzzy soft sets model for each class are as shown in Table-2 (a), (b), and (c).

Table 2. Tabular representasion of fuzzy soft sets model for iris dataset

(a) Setosa Class, $(\widetilde{F}_{Setosa}, E)$

U	e_1	e_2	e_3	e_4
o_{setosa}	0.99	0.68	0.68	0.64

(b) Versicolor Class, $(\widetilde{F}_{Versicolor}, E)$

U	e_1	e_2	e_3	e_4
o_{setosa}	0.85	0.87	0.82	0.84

(c) Virginia Class, $(\widetilde{F}_{Virginia}, E)$

U	e_1	e_2	e_3	e_4
o_{setosa}	0.84	0.67	0.82	0.79

Table 3. Tabular representasion of fuzzy soft sets model for unknown class, $(\widetilde{G}, E)$

U	e_1	e_2	e_3	e_4
o_{setosa}	0.63	0.52	0.48	0.40

To classify new data with the following features, $e_1 = 0.63$, $e_2 = 0.52$, $e_3 = 0.48$, and $e_4 = 0.40$, the fuzzy soft sets model $(\widetilde{G}, E)$ as shown in Table-3. First, calculate the similarity between $(\widetilde{G}, E)$ and fuzzy soft sets for each class $(\widetilde{F}_{setosa}, E)$, $(\widetilde{F}_{versicolor}, E)$, and $(\widetilde{F}_{virginia}, E)$, using Equation (4) as follows

$$S_1(\widetilde{G}, \widetilde{F}_{setosa}) = 0.78,\ S_2(\widetilde{G}, \widetilde{F}_{versicolor}) = 0.89,\ S_3(\widetilde{G}, \widetilde{F}_{virginia}) = 0.79$$

In this case higest similarity is $S_2(\widetilde{G}, \widetilde{F}_{versicolor}) = 0.89$, so to the new data, we gived class label "versicolor".

The main FSSC advantage compared with previous algorithms are, have a lower complexity in the classification phase. If n is the number of test data, w is the number of class labels, and d is the number of features. So for all the test data n, the pevious algorithms required number of arithmetic operations consists of, (1) wd number of arithmetic operations to compute the soft model sets, (2) w^2d number of arithmetic operations to construct the comparison table, and (3) Finally, w^2 number of arithmetic operations to calculate the score. The total complexity for the previous algorithm is $Q(n[wd + w^2d + w^2])$. The number of arithmetic operations on the new algorithm, FSSC, for all test data n consists of (1) $2wd$ number of arithmetic operations to calculate the similarity between two fuzzy soft sets, and (2) w^2 number of arithmatic operations to seek the highest similarity value. The total complexity to FSSC is $Q(n[2wd + w^2])$. So FSSC has a lower complexity of order $Q(nw^2d)$. This order will be very influential if the number of features becomes more and more, as happened in the case of text data classification. Where the number of features can reach hundreds of thousands.

4 Classifier Evaluation

In classification problems, the primary source of performance measurements is a coincidence matrix or contingency table [18]. Figure 1. shows a coincidence matrix for a two-class classification problem. The equations of most commonly used metrics that can be calculated from the coincidence matrix is also given below

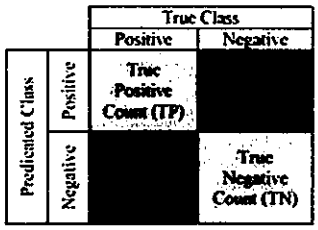

Fig. 1. Contingency table for binary classification

The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples.

When the classification problem is not binary, the coincidence matrix gets a little more complicated (see Fig. 2). In this case the terminology of the performance metrics becomes limited to the "overall classifier accuracy". These formulas in Equation 6.

$$(\text{Overall Classifier Accuracy})_i = \frac{\sum_{i=1}^{n} (True\ Classification)_i}{(Total\ Number\ of\ Cases)_i} \tag{6}$$

		Actual Classification of Classes in the Dataset			
		Class 1	*Class 2*	*Class 3*	
Model Classification	*Class 1*	22	7	2	
	Class 2	5	18	7	
	Class 3	3	5	21	
	Sum	30	30	30	
	Probability	0.33	0.33	0.33	
	Accuracy	0.73	0.60	0.70	0.68

Fig. 2. A sample coincidence matrix for a three class classifier

where i is the class number, n is the total number of classes. To minimize the effect of the imbalance between minority an mayority classes distributions, we used the weight F-measure method [2], which is equal to the harmonic mean of recall (ρ) and precision (π) [22]. Recall and precision are defined as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}, \qquad \rho_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

Here, TP_i (True Positives) is the number of datas assigned correctly to class i: FP_i (False Positives) is the number of datas that do not belong to class i but are assigned to class i incorrectly by the classifier; and FN_i (False Negatives) is the number of datas that are not assigned to class i by the classifier but which actually belong to class i.

The F-measure values are in the interval (0,1) and larger F-measure values correspond to higher classification quality. The overall F-measure score of the entire classification problem can be computed by two different types of average, micro-average and macro-average [22].

Micro-averaged F-Measure. In micro-averaging, F-measure is computed globally over all category decisions. ρ and π are obtained by summing over all individual decisions:

$$\pi = \frac{\sum_{1=i}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FP_i)}, \qquad \rho = \frac{\sum_{1=i}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FN_i)} \tag{8}$$

where M is the number of categories. Micro-averaged F-measure is then computed as:

$$F(\text{micro-averaged}) = \frac{2\pi\rho}{\pi + \rho} \tag{9}$$

Micro-averaged F-measure gives equal weight to each data and is therefore considered as an average over all the class pairs. It tends to be dominated by the classifiers performance on common classes.

Macro-averaged F-Measure. In macro-averaging, F-measure is computed locally over each class first and then the average over all classes is taken. ρ and π are computed for each class as in Equation 7. Then F-measure for each category i is computed and the macro-averaged F-measure is obtained by taking the average of F-measure values for each category as:

$$F_i = \frac{2\pi_i\rho_i}{\pi_i + \rho_i}, \quad F(\text{macro-averaged}) = \frac{\sum_{i=1}^{M} F_i}{M} \tag{10}$$

where M is total number of classes. Macro-averaged F-measure gives equal weight to each class, regardless of its frequency. It is influenced more by the classifier's performance on rare classes. We provide both measurement scores to be more informative.

5 Experimental Results

To compare accuracy and computational time of the two soft sets classifier we perform an experiment to classify some types of classification problems, the source data comes from the UCI datasets, we select the type of classification with a real numerical features, and having multiclass, class labels more than two (binary). Accuracy is calculated using Overall Classifier Accuracy (OCA) and F-measure (micro-average and macro-average). We divide each datasets into two parts, 70% used for the training process and the remaining is for testing, and pre-process (fuzzification) applied to all datasets, to form a fuzzy value between [0,1] for each attribute. At each datasets, experiments performed at least 10 times, and train (70%) and test (30%) data were selected randomly each time we starting the experiment. The experiments are performed on a Core(TM) 2 Duo CPU, 2.1 GHz and 2 GB memory computer using MATLAB version 7.10, 32-bit.

5.1 Data Set

In our experiments, we used seven dataset from [6] i.e. iris, letter-recognition, breast-cancer-wisconsin (wdbc and wpdc), waveform, spambase, musk (clean1 and clean2), and Reuters-21578 document collection which had been prepared in the format matlab by [4]. Special for Reuters-21578 datasets, we used features selection algorithm used to reduce the dimension of data, in this experiment we used information gain algorithm derived from the weka [7] to select 2,784 features from 17,296 features.

5.2 Results and Discussion

From Table 4 we observe that by using seven types of dataset accuracy and time computation of the new classifier (FSSC) are always better compare to soft sets classifier proposed by Mushrif et al. (SSC). Results of experiments on classification problems with 10 types of data sets, show that in general, the classifier both can work well. The highest achievement occurred in the wdbc (Wisconsin Diagnostic Breast Cancer) dataset with accuracy (F-macros) 0.94 (FSSC) and 0.92 (SSC. While the lowest achievement occurred in the letter-recognitions dataset, ie 0.378 (SSC) and 0.55 (FSSC). In this case, the F-macros have the same value with Overall Classifier Accuracy (OCA) and we choose to represent the classifier accuracy.

Special case occurs in Reuters-21578 datasets, for this text data classification problems, with the number of attributes in the thousands and even tens of thousands, but most attribute value is 0. We used two conditions on reuters-21578 datasets, the first by using all its attributes, the number is 17,296 attributes. Second, using the attributes that have been reduced by using information gain that was reduced to 2,784 attributes. We use the five largest classes in these datasets.

Table 4. Classification accuracy and computation time for soft sets classifier proposed by Mushrif et. al. (SSC) and the new **fuzzy soft sets classifier** (FSSC)

a. Iris (i:150; f:5; c:3)[*]

	SSC	FSSC
OCR	0.8756	0.9400
F-Macro	0.8756	0.9400
F-Micro	0.8736	0.9399
Time(s)	0.0090	0.0085

b. Letter recognitions (i: 20000; f:16; c:3)

	SSC	FSSC
OCR	0.3788	0.5527
F-Macro	0.3788	0.5527
F-Micro	0.3883	0.5434
Time(s)	6.4274	4.1135

c. Spambase (i:4601; f:57; c:2)

	SSC	FSSC
OCR	0.7516	0.7743
F-Macro	0.7516	0.7743
F-Micro	0.7510	0.7742
Time(s)	0.3483	0.3267

d. waveform (i: 5000; f:21; c:3)

	SSC	FSSC
OCR	0.7495	0.7967
F-Macro	0.7495	0.7967
F-Micro	0.7430	0.7815
Time(s)	0.3858	0.3586

e. wdbc (i:569; f:30; c:2)

	SSC	FSSC
OCR	0.9263	0.9409
F-Macro	0.9263	0.9409
F-Micro	0.9195	0.9368
Time(s)	0.0268	0.0260

f. wpbc (i: 198; f:33; c:2)

	SSC	FSSC
OCR	0.6153	0.6441
F-Macro	0.6153	0.6441
F-Micro	0.5569	0.5913
Time(s)	0.0108	0.0101

g. musk (clean1) (i:476; f:166; c:2)

	SSC	FSSC
OCR	0.5790	0.5979
F-Macro	0.5790	0.5979
F-Micro	0.5766	0.5960
Time(s)	0.0286	0.0264

h. musk (clean2) (i:6598; f:166; c:2)

	SSC	FSSC
OCR	0.6718	0.7246
F-Macro	0.6718	0.7246
F-Micro	0.5893	0.6199
Time(s)	0.6783	0.6336

i. Reuters-21578 (i:6632; f:2784; c:5)

	SSC	FSSC
OCR	0.5604	0.9364
F-Macro	0.5604	0.9364
F-Micro	0	0.9068
Time(s)	7.48	5.32

j. Reuters-21578 (full features) (i:6632; f:17296; c:5)

	SSC	FSSC
OCR	0.5604	0.9335
F-Macro	0.5604	0.9335
F-Micro	0	0.8966
Time(s)	36.89	19.94

[*] i : instances; f: features; c : classes

To this text datasets the SSC may not work well, all test data is labeled the same class. We still do not know for sure, why the SSC is not able to work well on text data, the most likely is because the text data is very sparse, very few features that have a value not equal to 0. While FSSC can still work well on these datasets, even with relatively good accuracy is 0.93. In addition, FSSC also can work faster, more and more features are used FSSC will work faster than the SSC, for full features reuters-21578 datasets, computation times for SSC is 36.89s, while FSSC only took for 19.94s. This is where one of the weaknesses of the SSC because they have to compare every feature for two different objects, when build the comparison-table, so that if the number of features becomes large, the classification process will more slowly. While FSSC does not need build the comparison-table, so it will work much faster.

6 Conclusion

In this paper we investigate the new classification based on fuzzy soft set theory. We use seven datasets from UCI to test the accuracy and computation time of the fuzzy soft sets classifier (FSSC) compares with the previous soft sets classfier proposed by Mushrif et al (SSC). In general, both can do the classifying numerical data, and both have the highest achievement in wdbc datasets, where for the FSSC accuracy is 0.94 and for SSC accuracy is 0.92. While the lowest achievement in letter-recognition classification problem, where accuracy to FSSC is 0.55 and accuracy for the SSC is 0.38. For all datasets FSSC has higher accuracy and shorter computational time. In addition to the SSC, If the number of features in the dataset the higher, then computational time become longer, also SSC can not work properly on Reuters-21578 dataset (text data). Furthermore, we will study the performance of the FSSC to classify text documents more detail, compared with a text classifier such as support vector machine (SVM), k-NN, and the others.

Acknowledgments. This work was supported by the FRGS under the Grant No. Vote 0736, Ministry of Higher Education, Malaysia.

References

1. Aktas, H., Çagman, N.: Soft sets and soft groups. Inf. Sci., 2726–2735 (2007)
2. Ali, S., Smith, K.: On learning algorithm selection for classification, App. Soft Comp. 6, 119–138 (2006)
3. Chen, D., Tsang, E., Yeung, D., Wang, X.: The parametrization reduction of soft sets and its applications. Comput. Math. Appl. 49, 757–763 (2005)
4. Cai, D., Wang, X., He, X.: Probabilistic Dyadic Data Analysis with Local and Global Consistency. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 105–112 (2009)
5. Feng, F., Jun, Y.B., Liu, X., Li, L.: An adjustable approach to fuzzy soft set based decision making J. Comp. App. Math. 234, 10–20 (2010)

6. Frank, A., Asuncion, A.: UCI Machine Learning Repository University of California, Irvine, School of Information and Computer Sciences (2010)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update (2009)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
9. Khaŕal, A.: Distance and Similarity Measures for Soft Sets. New Math. & Nat. Comp (NMNC) 06, 321–334 (2010)
10. Kong, Z., Gao, L., Wang, L.: Comment on a fuzzy soft set theoretic approach to decision making problems. J. Comput. Appl. Math. 223, 540–542 (2009)
11. Maji, P., Biswas, R., Roy, A.: Fuzzy soft sets. J. Fuzzy Math. 9(3), 589–602 (2001)
12. Maji, P., Roy, A., Biswas, R.: An application of soft sets in a decision making problem. Comput. Math. Appl. 44, 1077–1083 (2002)
13. Maji, P., Biswas, R., Roy, A.: Soft set theory. Comput. Math. Appl. 45, 555–562 (2003)
14. Majumdar, P., Samanta, S.: Generalised fuzzy soft sets. J. Comp. Math. App. 59, 1425–1432 (2010)
15. Majumdar, P., Samanta, S.: Similarity measure of soft sets. NMNC 04, 1–12 (2008)
16. Molodtsov, D.: Soft set theory–First results. Comp. Math. App. 37, 19–31 (1999)
17. Mushrif, M.M., Sengupta, S., Ray, A.K.: Texture classification using a novel, soft-set theory based classification algorithm. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3851, pp. 246–254. Springer, Heidelberg (2006)
18. Olson, D., Delen, D.: Advanced Data Mining Techniques, 1st edn. Springer, Heidelberg (2008)
19. Pawlak, Z.: Rough sets. Int. J. of Inform. Comput. Sci. 11, 341–356 (1982)
20. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic, Boston (1991)
21. Roy, A., Maji, P.: A fuzzy soft set theoretic approach to decision making problems. J. Comp. App. Math. 203, 412–418 (2007)
22. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM, New York (1999)
23. Zadeh, L.: Fuzzy sets. Inform. Control 8, 338–353 (1965)

FSSC: An Algorithm for Classifying Numerical Data using Fuzzy Soft Set Theory

Bana Handaga, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

Tutut Herawan, Departement of Computer Science, Faculty of Computer Systems andSoftware Engineering, Universiti Malaysia Pahang, Malaysia

Mustafa Mat Deris, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malayisa

ABSTRACT

This paper introduces a new algorithm for the classification of numerical data using the theory of fuzzy soft set, named Fuzzy Soft Set Classifier (FSSC). The algorithm use fuzzy approach in the pre-processing stage to obtain features, and similarity concept in the process of classification. It can be applied not only to binary-valued datasets, but also be able to classify the data that consists of real numbers. Comparison tests on seven datasets from UCI Machine Learning Repository have been carried out. It is shown that the proposed algorithm provides better accuracy and higher accuracy as compared to the baseline algorithm using soft set theory.

Keywords: *Fuzzy soft set theory; Numerical data; Classification.*

INTRODUCTION

In 1999, D. Molodtsov introduced the notion of a soft set as a collection of approximate descriptions of an object (Molodtsov, 1999). This initial description of the object has an approximate nature, and we do not need to introduce the notion of exact solution. The absence of any restrictions on the approximate description in soft set makes this theory very convenient and easily applicable in practice [Herawan et al., 2009; Herawan and Deris, 2010; Herawan et al., 2010b; Xiuqin et al., 2011b]. Applications of soft set in areas ranging from decision problems to texture classifications have surged in recent years (Maji et al., 2002; Mushrif et al., 2006; Roy & Maji, 2007; Feng et al., 2010).

The soft set theory can work well on the parameters that have a binary number [Herawan et al., 2010a, Xiuqin et al., 2011a], but still difficult to work with parameters that have real numbers. To overcome this problem, Maji *et al.* have studied a more general concept, namely the theory of fuzzy soft set, which can be used with the parameters in the form of real numbers (Maji *et al.*, 2001). These results further expand the scope of applications of soft set theory [Hongwu et al., 2011]. One of the potential applications of fuzzy soft set theory is numerical data classification. There are two important concepts underlying the application of the theory of soft set in numerical classification problems. Firstly, the concept of decision making problems based on fuzzy soft set theory, and secondly is the concept of measuring similarity between two fuzzy soft set. Based on an application of soft set in a decision making problem presented by (Maji et al., 2002), Mushrif *et al.* (2006) presented a novel method for classification of natural textures using the notions of soft set theory, all features on the natural textures consist of a numeric (real) data type, have a value between [0,1] and the algorithm used to classify the natural texture is very similar to the algorithms used by Roy and Maji (2007) in the decision making problems with the theory of fuzzy soft

set. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less as compared to with Bayes classification method.

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovering task, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. The studies on measuring the similarity between soft set have been carried out (Majumdar & Samanta, 2008; Kharal, 2010). They then extended their research to measure the similarity of fuzzy soft set and describe how it can be applied to medical diagnosis to detect whether a person is suffering from a certain disease (Majumdar & Samanta, 2010).

This paper proposes a new classification approach based on fuzzy soft set theory, using similarity between two soft set. We call this classifier as a Fuzzy Soft Set Classifier (FSSC). The proposed approach results in high degree of accuracy and with low computational complexity as compared to soft set classification based on decision making problem. Seven types of UCI data set have been used have been used for comparing the proposed fuzzy soft set classifier and the soft set classifier suggested by (Mushrif et al., 2006).

The rest of the paper is organized as follows. Soft set and fuzzy soft set theory are introduced. The classification based on soft set theory, the proposed algorithm and the classifier evaluation methodology are discussed. An experiment was done to compare both of classification algorithms. Finally, the concluding remarks and some directions for future research are given.

PRELIMINARY

In this section, we recall the basic notions of soft set and fuzzy soft set. Let U be an initial universe of objects and E_U (simply denoted by E) the set of parameters in relation to objects in U. Parameters are often attributes, characteristics, or properties of objects. Let $P(U)$ denote the power set of U and $A \subseteq E$. Following (Molodtsov, 1999; Maji et al., 2003), the concept of soft set is defined as follows.

Definition of Soft Set

Definition 1 (See Molodtsov (1999)). Let U be initial universal set and $A \subseteq E$ be a subset of parameters. Let $P(U)$ denote the power set of U. A pair (F,A) is called a soft set over U, if and only if F is a mapping given by $F:A \rightarrow P(U)$.

By definition, a soft set (F,E) over the universe U can be regarded as a parameterized family of subsets of the universe U, which gives an approximate (soft) description of the objects in U. As pointed in (Molodtsov, 1999), for any parameter E, the subset may be considered as the set of ε-approximate elements in the soft set (F,E). It is worth noting that $F(\varepsilon)$, for $\varepsilon \in E$ may be arbitrary: some of them may be empty, and some may have non-empty intersection. For illustration, Molodtsov (1999) considered several examples of soft sets. Similar examples were also discussed in (Maji et al., 2003; Aktas & Çagman, 2007).

From the definition, it is known that fuzzy set introduced by L.A. Zadeh (1965) is a kind of a special soft set. Let A be a fuzzy set and μ_A be a subject function (μ_A be a mapping of U into [0,1]). To this problem, Maji *et al.* (2001) initiated the study on hybrid structures involving both fuzzy sets and soft set. They introduced the notion of fuzzy soft set, which can be seen as a fuzzy generalization of (crisp) soft set.

Definition of Fuzzy Soft set

Definition 2 (See Maji et al., (2001)). Let U be an initial universal set and let E be set of parameters. Let $\underline{P}(U)$ denote the power set of all fuzzy subsets of U and $A \subseteq E$. A pair $(\underline{F}, E)$ is called a fuzzy soft set over U, where $\underline{F}$ is a mapping given by $\underline{F}:A \rightarrow \underline{P}(U)$.

In the above definition, fuzzy subsets in the universe U are used as substitutes for the crisp subsets of U. Hence it is easy to see that every (classical) soft set may be considered as a fuzzy soft set. Furthermore, since every fuzzy set is a soft set (Molodtsov, 1999), then it is easily to conclude that every fuzzy set is a fuzzy soft set. Generally speaking $\underline{F}(\varepsilon)$ is a fuzzy sub-set in U and it is called the fuzzy approximate value set of the parameter ε.

It is well known that the notion of fuzzy sets provides a convenient tool for representing vague concepts by allowing partial memberships. It has been widely applied in many fields, including [Chen, 2011a; Chen, 2011b]. In the definition of a fuzzy soft set, fuzzy subsets are used as substitutes for the crisp subsets. Hence every soft set may be considered as a fuzzy soft

set. In addition, by analogy with soft set, one easily sees that every fuzzy soft set can be viewed as an (fuzzy) information system and be represented by a data table with entries belonging to the unit interval [0,1]. For illustration, we consider the following example.

Example 1. This example taken from (Majumdar & Samanta, 2010), suppose a fuzzy soft set $(\underline{F},E)$ describes attract-iveness of the shirts with respect to the given parameters, which the authors are going to wear. $U = \{x_1, x_2, x_3, x_4, x_5\}$ which is the set of all shirts under consideration. Let $\underline{P}(U)$ be the collection of all fuzzy subsets of U. Also let $E = e_1$= "colorful", e_2 = "bright", e_3 = "cheap", e_4 = "warm".

Let

$\underline{F}(e_1) = \{ x_1/0.5, x_2/0.9, x_3/0.0, x_4/0.0, x_5/0.0\}$,
$\underline{F}(e_2) = \{ x_1/1.0, x_2/0.8, x_3/0.7, x_4/0.0, x_5/0.0\}$,
$\underline{F}(e_3) = \{ x_1/0.0, x_2/0.0, x_3/0.0, x_4/0.6, x_5/0.0\}$,
$\underline{F}(e_4) = \{ x_1/0.0, x_2/1.0, x_3/0.0, x_4/0.0, x_5/0.0\}$.

Then the family $\{\underline{F}(e_i); i=1,2,3,4\}$ of $\underline{P}(U)$ is a fuzzy soft set $(\underline{F},E)$. Tabular representation for fuzzy soft set $(\underline{F},E)$ is shown in Table 1.

Table 1. Representation of fuzzy soft set $(\underline{F},E)$

U/E	e_1	e_2	e_3	e_4
x_1	0.5	1.0	0	0
x_2	0.9	0.8	0	1.0
x_3	0	0.7	0	0
x_4	0	0	0.6	0
x_5	0	0	0	0.3

There are two important concepts underlying the application of the theory of fuzzy soft set in numerical classification problems. Firstly, the concept of decision making problems based on the theory of fuzzy soft set proposed by Roy & Maji (2007) and secondly, is the concept of measuring similarity between two fuzzy soft sets proposed by Majumdar & Samanta (2010). The following are brief explanation about these two concepts.

Fuzzy Soft Set-based Decision Making

We begin this section with a novel algorithm designed for solving fuzzy soft set based decision making problems, which was presented by (Roy & Maji, 2007). This algorithm has similarity with soft set classifier algorithm proposed by Mushrif et al. (2006) for classification of natural textures. Also, Feng et al. (2010) shows that this algorithm is a useful method for selecting the optimal alternative in decision making problems based on fuzzy soft set, while the counter example given by Kong et al. (2009) is not sufficient for concluding that the Roy and Maji method is incorrect.

Algorithm 1. Fuzzy soft set decision making problem algorithm by Roy & Maji (2007).

a. Input the fuzzy soft set $(\underline{F},A)$, $(\underline{G},B)$ and $(\underline{H},C)$.
b. Input the parameter set P as observed by the observer.
c. Compute the corresponding resultant fuzzy soft set $(\underline{S},P)$ from the fuzzy soft sets $(\underline{F},A)$, $(\underline{G},B)$, $(\underline{H},C)$ and place it in tabular form.
d. Construct the comparison table of the fuzzy soft set $(\underline{S},P)$ and compute r_i and t_i for o_i, for $i=1,2,\ldots,n$.
e. Compute the score s_i of o_i, for $i=1,2,\ldots,n$.
f. The decision is o_k if $s_k = \max\{s_i\}$.
g. If k has more than one value then any one of o_k may be chosen.

Here, Roy & Maji (2007) pointed out that the object recognition problem may be viewed as a multi observer decision making problem, where the final identification of the object is based on the set of inputs from different observers who provide the overall object characterization in terms of diverse sets of parameters. The Algorithm 1 gives solutions to the recognition problem by means of fuzzy soft set. This method involves construction of comparison table from the resultant fuzzy soft set and the final optimal decision is taken based on the maximum score computed from the comparison table.

The comparison table is a square table in which rows and columns both are labelled by the object names o_1, o_2, ..., o_n of the universe, and the entries c_{ij} indicate the number of parameters for which the membership value of o_i exceeds or equals the membership value of o_j. Clearly, $0 \le c_{ij} \le m$, where m is the number of parameters. The row sum r_i of an object o_i is computed by

$$r_i \equiv \sum_{j=1}^{n} c_{ij} \tag{1}$$

Similarly the column sum t_j of an object o_j is computed by

$$t_j = \sum_{i=1}^{n} c_{ij} \tag{2}$$

Finally, the score s_i of an object o_i is defined by

$$s_i = r_i - t_i \tag{3}$$

for $i = 1,2, \ldots, n$.