AN EFFECT OF DISTANCE MEASURES IN CLASSIFYING LARGE DATASETS

Dr. Sinciya P.O
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
posinciya@amaljyothi.ac.in

ABSTRACT -As digital usage increases day to day, a large voluminous data is accumulated in various applications. Bulky data contain useful facts and information which may go unidentified if not processed. Data mining is a promising technology which processes bulky data to extract facts. Data mining technique is better influenced by the classification work. Effective use of classifier with accurate distance measure calculation helps in extracting minute facts available in the bulky data. In this paper, a set of classifiers such as K nearest neighbor, Support vector machine and centroid are implemented. Then a novel classifier, an improved fuzzy soft classifier is implemented and results are produced using two distance measures viz., Euclidean and Jaccard. The proposed classifier shows better results when Euclidean distance measure is used with continuous data and it shows better results when jaccard distance is used with categorical data.

Keywords – Fuzzy soft set, Euclidean distance, Jaccard distance, Support Vector Machine, K Nearest Neighbour.

1. INTRODUCTION

Classification is the task of assigning class labels to the unlabeled instances. K-nearest neighbor [1], Fuzzy set, Multi-layer perceptron, Centroid based, Rough set, Decision tree and support vector machines are the well-known classification algorithms. Support vector machine performs well in many application because it has the ability to learn independent of the dimensionality of the feature space. Though it has a drawbacks of high algorithmic complexity and memory requirements. Generally, the whole performance of instance based classifiers is good and are very simple and easily understandable.

Some of the instance based classifiers are Centroid based, KNN and naïve Bayes, performs well and takes less execution time and memory requirements than SVM. [3] Proposed a new method of KNN that competes in performance with other classifiers such as SVM. An improved form of Naïve Bayes is suggested in [2], which also outperforms than SVM. Centroid classifier also performs well for small sized data sets.

In this paper, a new approach for real world data classification based on fuzzy soft set theory is proposed.

Fuzzy soft set is a comprehensive form of fuzzy set [4]. It is applied on different problem domains such as game theory, medical diagnosis and operations research etc. This approach is applied for object recognition from imprecise data [5] and in decision making [6]. A new classification algorithm is given by Handaga in [7] for numerical data analysis using fuzzy soft set theory. [8] Proposed a text classification algorithm by considering the corresponding class membership value of each document. A number of researchers studied similarity between two fuzzy sets and also applied in medical diagnosis [8] [9].

Soft set theory is applied for data classification using the comparison table approach for decision-making problems [10] [11]. Modified comparison table approach data classification is introduced in this paper. Performance of classification with this approach is very much better when compared with Bayesian classifier (multinomial). This fuzzy soft set based classifier also has competing performance with the simple and the efficient classifier, the centroid classifier.

The organization of the rest of the chapter is as follows. In Section 2, fuzzy soft set theory concepts are dealt with. In Section 3, proposed classification algorithm and in Section 4, the experimental methodologies and results are discussed and Section 5 summarizes the chapter.

2. FUZZY SOFT SET

Many classification algorithms proved that the use fuzzy set theory is a good choice for dealing with uncertainties [12]. But there is no suitable mechanism to deal with membership function because it may change on the basis of problem domain. In order to avoid these problems, a new mathematical model is introduced by Molodtsov [13] called soft set concept which have necessary parameterization to deal uncertainty problems.

Definition 1: *Soft set.* Suppose U be an initial universal set and T be a set of features. Let P(U) be the set of all subset or power set of U and A ⊂ T. A pair vector (F, A) is termed as soft set over U where F is a mapping given by F: A → P(U).

Objects	Parameters			
	e_1	e_2	e_3	e_4
l_1	0	1	0	0
l_2	0	0	1	0
l_3	1	0	1	0
l_4	0	1	0	1
l_5	1	0	1	0
l_6	0	1	0	1

Table 1 :Tabular representation of soft set-(F, A)

2.1 Example for soft set

As an illustration of soft set, the following example is considered. If U is the universe (set of objects), U = {l_1, l_2, l_3, l_4, l_5, l_6}, which is a set of six diabetes patients under consideration. The parameter set is E = {e_1, e_2, e_3, e_4, e_5, e_6, e_7}, where the parameters e_i, i = 1, 2, 3, 4, 5, 6, 7 refers to the attributes Plasma glucose concentration, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), Hour serum insulin (mu U/ml), Body mass index, Diabetes pedigree function and Age respectively. Let A = {e_1, e_2, e_3, e_4} ⊂ E and let a soft set (F, A) denote whether the patient shows the sign of diabetes. Defining (F, A) involves specifying the sign of diabetes for each parameter. So the mapping F: A → P(U) should be defined.

For this example, F: A → P(U) is given as F(e_1) = {l_3, l_5}, F(e_2) = {l_1, l_4, l_6}, F(e_3) = {l_2, l_3, l_5} F(e_4) = {l_4, l_6}. Thus the soft set (F, A) is a parameterized family {F(e_i), i = 1, 2, 3, 4} of subsets of the universe given by (F, A) = {F(e_1) = {l_3, l_5}, F(e_2) = {l_1, l_4, l_6}, F(e_3) = {l_2, l_3, l_5}, F(e_4) = {l_4, l_6}}. The soft set (F, A) can also be represented in a tabular form with an entry of 1 in each cell of the table if $l_i \in F(e_j)$ and 0 otherwise, i.e if l_{ij} are the entries of the table, l_{ij} = 1 if $l_i \in F(e_j)$ and 0 otherwise. This is given in Table 2.1.The soft set can also be represented by the set of ordered pairs given by (F, A) = {(e_1, F (e_1)), (e_2, F(e_2)), (e_3, F(e_3)), (e_4, F(e_4))} i.e. (F, A) = {(e_1, {l_3, l_5}), (e_2, {l_1, l_4, l_6}), (e_3, {l_2, l_3, l_5}) (e_4, {l_4, l_6})}.

Definition 2: *Fuzzy Soft set.* The family of all fuzzy sets of U is denoted by Ƒ (U). Let $A_i \subseteq E$. Then a pair (F_i, A_i) is called a fuzzy soft set over U, where F_i is a mapping given by F_i: A_i → Ƒ (U).

2.2 Comparison table

Comparison table is a square table with the number of rows equal to the number of columns, and both the rows and the columns are labeled by object names d*1, d2… dn* of the universe. In Roy & Maji (2007), each value in the table (*gij*) = the number of attributes of object *di* that is greater than or equal to the attributes of object *dj* in terms of the membership value. Therefore, $0 \leq gij \leq p$, and $gii = p$, $\forall i, j$, where *p* is the number of parameters in a fuzzy soft set. Thus, *gij* is a numerical measure, which is an integer value, and *di* dominates *dj* in *gij* number of parameters out of *p* parameters.

Row sum for each object *di*, sum of weights of *di* that are greater than other object weights, can be calculated using Equation (1).

$$r_i = \sum_{j=1}^{n} g_{ij} \quad (1)$$

Thus, *ri* indicates the total number of parameters in which *di* dominates all the members of *U.* Column sum of an object *di*, sum of weights of other objects that are greater than *di* can be calculated as below.

$$c_i = \sum_{i=1}^{n} g_{ij} \quad (2)$$

The value *ci* indicates the total number of parameters in which *dj* is dominated by all the members of *U.* Score of an object *i*, difference of row sum and column sum, is given by Equation (5.3).

$$S_i = r_i - c_i, i = 1,2 \ldots n. \quad (3)$$

Based on this score value, the unknown class label of an object can be determined.

3. PROPOSED FUZZY SOFT CLASSIFICATION (FSC) ALGORITHM

1. Calculate the centroid vector *Ci* (for each class *i*), by calculating the average weights of the terms present in the dataset *Di* (set of instances of class *i*) using Equation

$$C_i \rightarrow = \frac{1}{D_i} \sum_{d_j \in D_i} \quad d_j \rightarrow$$

2. Represent the centroid vectors as a table of size *I X N* (*I* classes and N features) which can be considered as a soft set (F, E). An entry in the table is *gin*, *i* = 1, 2, ...,*I* and *n* = 1, 2, ...,*N* .

3. Get a feature vector *Ef* from the unknown dataset.

4. Generate a soft set (F, A) with its entry as *kin*, *i* = 1, 2… *I* and n = 1, 2... *N*, calculated using different distance measures.

a) Using Euclidean distance

$$kin \ = 1 - \sqrt{(gin - Ef)^2}$$

b) Using Jaccard distance

$$K_{in} = 1 - \frac{min\ (g_{in}, E_f)}{max\ (g_{in}, E_f)}$$

5. Obtain a comparison table from (F, A).

6. Determine the score vector *S* = <*s1*,*s2*,..*sK*> for the comparison table as in Section 2.2.

7. Assign the test document to class c, where c is the class for which *sc* > *sv* for all *v* = 1, 2…*K* and c ≠ *v*.

4. EXPERIMENTAL SETUP

4.1 Dataset Description

The performance of our proposed method was evaluated by conducting numerical experiments using three real world cancer datasets with diverse sizes, features, and classes. Tenfold cross validation is adopted to guarantee the unbiased comparison of the classification results and avoid generating random results. The standard data used for experimentation are collected from the UCI repository [14] are used is given in (Table 1). In this work, three real time datasets collected namely vehicle dataset with 846 instances and 19 features, Sonar data with 208 instances and 60 features and diabetes data with 568 instances and 2 features. From this vehicle and Sonar data set includes continuous data values and Diabetes contains categorical values. The selected dataset information are given below.

Table 2: UCI Datasets used in experimentation

Data set	No. of features	No. of samples	No. of classes
Vehicle	19	846	4
Sonar [18]	60	208	2
diabetes	7	568	2

The proposed fuzzy classifier is compared with other well-known classifiers such as Support Vector Machine (SVM), K nearest neighbor (KNN) and Centroid classifier in terms of classification accuracy and F-measure. 10 fold cross validation is applied as an evaluation mechanism.

4.2 Metrics for evaluating performance

It is important to verify the performance of the system after building the classifier model. There are some evaluation metrics to assess the accuracy of the classifier. The different evaluation metrics used are as given below:

Sensitivity

It is also termed as exactness measure. It gives the percentage of tuples categorized as positive is actually positives. It is determined by

$$Sensitivity = \frac{TP}{TP + FP}$$

Specificity

It is also termed as completeness measure. It gives the percentage of tuples categorized as positive from the corrected tuple set. It is determined by

$$Specificity = \frac{TP}{TP + FN}$$

Accuracy

The accuracy measure of classifier gives the percentage of known set of tuples are positively classified. It is also called as recognition rate. It tells how the classifier effectively recognizes the instances of particular classes. It is defined as

$$Accuracy = \frac{(TP + TN)}{(P + N)}$$

Where TP and FP represents the number of true positive and false positive tuples and TN and FN represents the number of True negative and false negative tuples.

4.3. **Experimental Results**

Table 3: Classification results in terms of Accuracy, Sensitivity and Specificity for vehicle Data

Classification Algorithms	Accuracy	Sensitivity	Specificity
FSC (Euclidean)	93.2	90.78	89.4
FSC (Jaccard)	95.3	94.22	96.9
SVM	85.3	90.47	89.67
KNN	92.5	82.33	92.7
Centroid	92.7	89.78	92.89

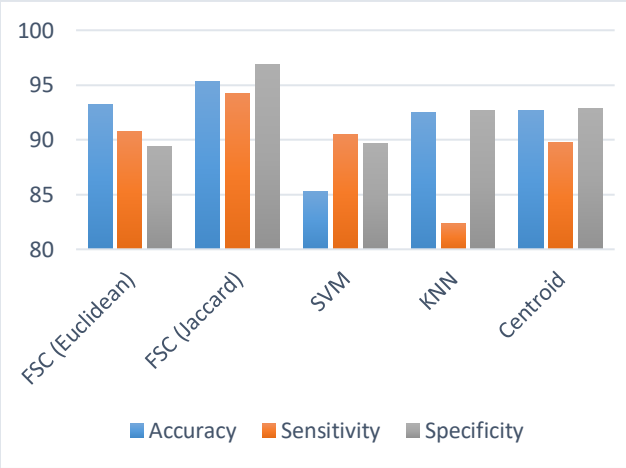

Figure 1. Accuracy results on vehicle data

Table 4: Classification results in terms of Accuracy, Sensitivity and Specificity for Sonar Data

Classification Algorithms	Accuracy	Sensitivity	Specificity
FSC (Euclidean)	98.3	87.22	92.7
FSC (Jaccard)	90.6	83.0	88.8
SVM	93.3	89.4	93.6
KNN	92.5	80.83	89.6
Centroid	92.7	89.7	94.4

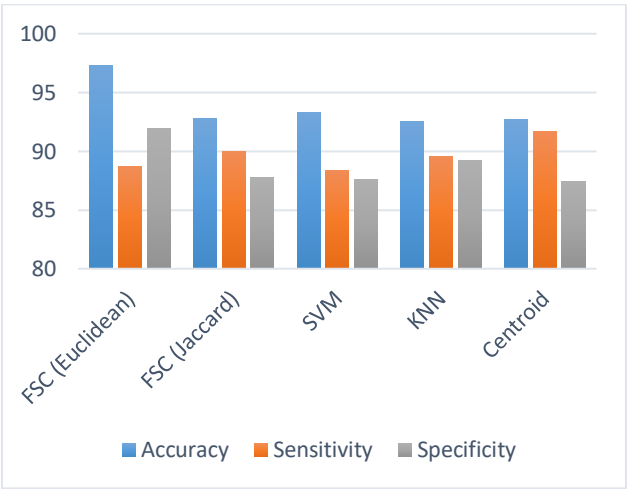

Figure 2. Accuracy results on sonar data

Table 5: Classification results in terms of Accuracy, Sensitivity and Specificity for Diabetes Data

Classification Algorithms	Accuracy	Sensitivity	Specificity
FSC (Euclidean)	97.3	88.72	91.9
FSC (Jaccard)	92.8	90.0	87.8
SVM	93.3	88.4	87.6
KNN	92.5	89.53	89.2
Centroid	92.7	91.7	87.4

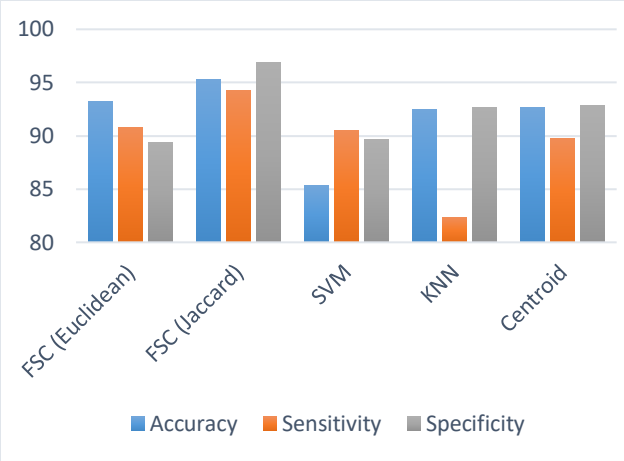

Figure 3. Accuracy results on diabetes data

For experimentation, three real world datasets are considered. From this, Vehicle and Sonar with continuous data values and diabetes dataset with categorical values. In the proposed Fuzzy soft set classifier, two distance measures are considered namely Euclidean and jaccard. The effectiveness of our proposed method was evaluated by comparing its accuracy with the other three popular classification methods, namely, SVM, KNN and Centroid. Based on the experiments, we evaluated the performance of our proposed algorithm in terms of the accuracy rate as tested on all the three datasets. Table3, Table 4 and Table 5 gives the results of FSC, SVM, KNN and Centroid classifiers in terms of accuracy, sensitivity and specificity for Vehicle, Sonar and Diabetes data set. We applied tenfold cross-validation to the datasets and obtained the average classification accuracy, Sensitivity and specificity. The proposed method was superior to all the compared methods. Hence, the utilization of Centroid classifier combined with the fuzzy soft set classifier yielded a higher accuracy compared with a standard SVM classifier. Figure 1, 2 & 3 shows a 97% confidence interval for the mean classification accuracy (with respect to all the continuous and categorical datasets). The figure 1, 2 and 3 clearly shows that Fuzzy soft set classifier (FSC) outperformed the other algorithms. It gives better results for all three data sets in which it performs well for vehicle and sonar data when considering Euclidean distance measure and jaccard distance measure for diabetes data sets. From this we can infer that fuzzy soft set classifier gives better accuracy compared to other classifiers.

CONCLUSION

Fuzzy soft set is an emerging technique which is applied on various application in different fields. So far it is not successfully applied on real world data sets including both categorical and continuous data. In this paper an efficient fuzzy soft set classifier is introduced. It has been tested with four real world data sets including both continuous and categorical data by two distance measures namely Euclidean and jaccard. It is observed that, for continuous data set, with Euclidean distance measure gives better results and at the same time, for categorical, jaccard distance measure performs better. The performance was compared with that of centroid, SVM and KNN classifiers and was found to be interesting.

REFERENCES

1. T.Y. Lin, A set theory for soft computing, a unified view of fuzzy sets via neighbor hoods, In Proceedings of 1996 IEEE International conference on Furzy Systems. New Orleans, LA, September 8-11, pp. 1140-1146, (l996).
2. Kim S. Han, K. Rim, H and Myaeng, SH 2006, “Some effective technique for naïve bayes classification”, IEEE transactions on knowledge and data engineering, vol.18, no.11, pp. 1457 – 1466.
3. Jiang S, Peng G, Wu, M & Kuang, 2012, “An improved k nearest algorithm for data classification”, Expert systems with applications, vol.39, pp.1503-1509.
4. Molodtsov, D 1999, “Soft set theory – first results”, Computers and mathematics with applications, vol.37, pp. 19-31.
5. Mushrif M.M, Sengupta, S & Ray, A.K 2006, “Texture classification using a novel soft set theory based classification algorithm”, ACCV, LNCS, Springer, Heidelberg, vol.3851, pp.246-254.
6. A.R. Roy, P.K. Maji, A fuzzy soft set theoretic approach to decision making problems, J. Comput. Appl. Math. 223 (2007) 540–542.
7. Handaga, B. Herawan, T & Deris, MM 2012, “FSSC: An algorithm for classifying numerical data using fuzzy soft set theory”, IJFSA, vol.2, no.4, pp.29-46.
8. Dusmanta Kumar Sut, ”An Application of Similarity of Fuzzy Soft sets in Decision Making” Int.J.Computer Technology & Applications, Vol 3(2), 742-745.
9. Q.Feng, W.Zheng, “New Similarity Measures of Fuzzy Soft Sets Based on Distance Measures,