Volume 2, No. 1, Jan-Feb 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

A Novel Classification Method Aided SAW

Emad Roghanian
Department of Industrial Engineering
K.N.Toosi, University of Technology, Tehran, Iran
E_roghanian@Kntu.ac.ir

Azadeh Bazleh
Department of Industrial Engineering
Islamic Azad University, Arak, Iran
A.bazleh@gmail.com

Peyman Gholami
Department of Industrial Engineering
Islamic Azad University, Arak, Iran
Peyman711@yahoo.com

Mansour Ahmadi*
Department of Software Engineering
Islamic Azad University, Arak, Iran
Mansourweb@gmail.com

Abstract: The Simple Additive Weighting (SAW) method is one of the easiest techniques for compensation multiple criteria decision making (MCDM). This method used for solving multiple criteria decision problems. On the other hand, data mining is one of the 10 global science-growing and classification algorithms are the most important data mining method used for prediction. Most of these algorithms are complex and time consuming. In this paper, we have proposed a new method to classify data using the SAW method aided Fisher Score for scoring the features. The main advantages of this method are its simplicity, effectiveness and efficiency. We have used fuzzy functions for improving the method accuracy. Based on experimental results on two Datasets, high accuracy is obtained comparing to the most classification algorithms.

Keywords: Data mining; Classification; Simple Additive Weighting; Multiple criteria decision; Fisher Score; Fuzzy

I. INTRODAUCTION

SAW method is one of the compensatory techniques to solve multiple criteria decision making, which was introduced by Fishburn in 1967 [1]. The most important feature of this technique is its simplicity without considering the number of features and samples. The Saw method works based on rating a sample.

Multiple criteria decision making techniques are actually highly regarded in recent artificial intelligence, machine learning and data mining research. There is much attention [2, 3] to the classification process for predicting the class labels of new samples. Classification is a process to find a model that describes the data and for predicting the uncertain class labels. Based on analysis a train dataset, the model is built. There are some famous classification algorithms such as SVM, Naïve Bayesian, and KNN and Decision trees [4]. These algorithms often have large calculations and these issues causes increasing the time order thus, when the number of samples and features is high, too much time are spent on analyzing data and building models. Therefore, in this article we have proposed a new classification method that has all simplicity, effectiveness and efficiency advantages.

Our proposed method consists of three major steps as follows:

A. At first, we would allocate scores to all features in train dataset using Fisher scores that indicate the importance of them.
B. Then, compute the SAW values for all samples in dataset and determine the mean, initial and final points of each class.
C. Finally, assign the ranges of SAW values to the existing classes; furthermore, we have used fuzzy functions for the common intervals.

The proposed method has tested on some standard numeric datasets, and admissible results were obtained. Rest of the paper as following:

In section two we describe the related works. Then we would present the proposed method in section three and subsequent, the results of experiments to be announced and finally conclusions and future work would be presented.

II. RELATED WORKS

SAW method is still the simplest and was widely used in MADM problems [5-6] and recent researches [7] have represented the efficiency of SAW method more than other methods like TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method. Some researchers used SAW method for rating the samples but there aren't used in data mining approach for classification. So in order to increase the efficiency of our proposed method, we use SAW method.

Just as mentioned, another popular method of multiple criteria decision making is TOPSIS that is used somewhere for classification in data mining [8-9]. In these methods, they have used decision trees in some overlapped intervals, and this has caused a classification method is not independent. Therefore, mentioned dependency caused high dispersion so that operations are being time-consuming.
To solve this problem, we have been used fuzzy functions that easy to compute even by hand. Another problem is weighting to features. Based on our results, the factor of Fisher Score [10] is more efficient than the correlation coefficients that are used in mentioned articles.

Furthermore, most of the existing classification algorithms are too complex. For example, SVM model [11] considers the samples as points in space and then regarding to the dimension, the samples are separated with a line,

page, etc. Increasing dimension can be increased the accuracy.

III. PROPOSED METHOD

Our proposed classification method has three major steps that have shown in Figure 1.

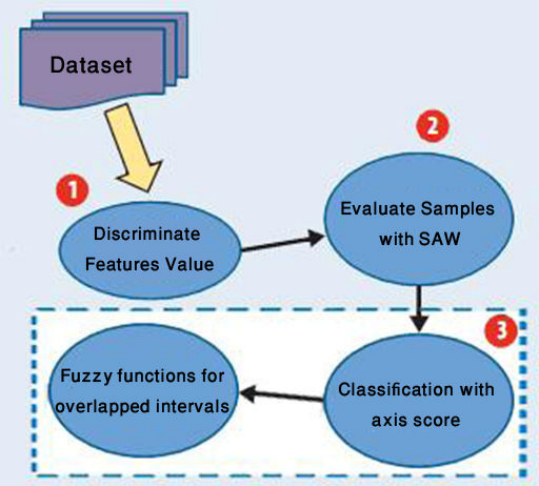

Figure1. Our Method

A. *Rating Features*

To evaluate the discrimination power of each feature, we have been using the statistical criteria of Fisher scores that are defined as follows:

$$Fr = \frac{\sum_{i=1}^{c} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \qquad (1)$$

Where n_i is the number of samples in i_{th} class, μ_i is the mean values of a feature in i_{th} class, $\sigma_i$is the variance values of a feature in i_{th} class, μ is the mean values of a feature in total samples.

Suppose x_{ij} is the values of j_{th} feature in i_{th} class, then μ, μ_i, σ_i are defined as following:

$$\mu = \frac{\sum_i \sum_j x_{ij}}{\sum_i n_i} \qquad (2)$$

$$\mu_i = \frac{\sum_j x_{ij}}{n_i} \qquad (3)$$

$$\sigma_i = \sqrt{\frac{\sum_j (x_{ij} - \mu_i)^2}{n_i - 1}} \qquad (4)$$

When the difference between μ value and μ_i value is high or σ_i value is very small, the Fisher score would be great. If a feature has similar property values in the same class and has very different values in other classes, the Fisher score would be very large. In this case, the features for discriminating samples from different classes are very distinct and use the scores for weighing the features would be very useful.

B. *Evaluate Samples*

In the second step, a score is given to any of the samples in the dataset by SAW method. This technique is one of the methods of multiple criteria decision making has been used for analysis of multiple criteria decision making for the long time. In this method, a number is giving to each sample that is obtained from the following formula:

$$Score(A_i) = \sum_{j=1}^{n} w_j \times x_{ij} \qquad (5)$$

In above relation, w_j is the weight of j_{th} feature equal to Fisher score, A_i is a dataset sample, x_{ij} is the value of j_{th} feature in i_{th} sample and n is the number of features in dataset. Of course in using this technique, you should be noting two points that 1) features should be benefit values and if they are cost values, we have to use techniques to convert them into benefit values and 2) using normalization before calculation of SAW number. There are many techniques for normalization that dividing each feature value to max of feature values is the simplest one.

C. *Classification*

Now we use the SAW score axis to identify new sample class. New samples should be normalized before classification step.

We consider the classes range such as the score axis that these are achieved as follows:

We consider the minimum and maximum SAW values in each class, respectively, as the beginning and the end of the range.

Thus, one of three states occurs for a new sample as following:

[a] The new sample is placed only in one interval.
[b] The new sample would not occur in any intervals.
[c] The new sample is jointly placed in two or more intervals.

There is no doubt for the first case but to resolve the ambiguity in the second or third case, do the following:

In the second case, assume 'l' for the length of the regions that isn't exposure in any intervals (figure 2), the length should be divided into a certain ratio between the two intervals that are located on either side of this region.

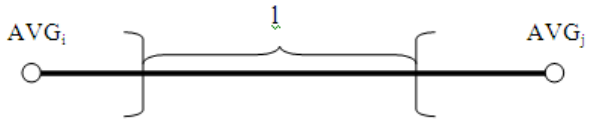

Figure2. The scores axis for second case

In Figure (2), AVG $_i$ is the average of SAW value in i_{th} class and AVG $_j$ is the average of SAW value in j_{th} class. Suppose 'm' for the distance between AVG_j and minimum scores that SAW is devoted to the elements in the class and 'n' for the distance between AVG_j and maximum scores that SAW is devoted to the elements in the class. Now the region l is broken into two parts in ratio $\frac{n}{n+m}$ and $\frac{m}{m+n}$ that is respectively related to i and j class.

In order to resolve ambiguity in third case, we operate as following:

For each class in the overlapped region, we formed a trapezoidal fuzzy membership function (figure 3).

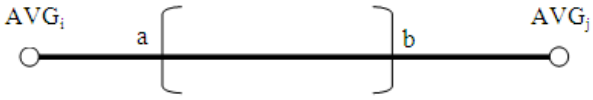

Figure3. The scores axis for third case

Here we describe it for two classes that can also be extended for more classes. For the class in the right interval, we will form the left leg of the trapezoid so that the highest degree of membership (1) is at the point that the biggest score of the left class has gotten from SAW and the lowest degree of membership (0) is at the point that the smallest points of the right class have gotten from saw. Now we define fuzzy functions in the following form:

$$\mu_{class\ i}(x) = \frac{x-a}{b-a}, \qquad a \le x \le b \tag{6}$$

$$\mu_{class\ j}(x) = \frac{b-x}{b-a}, \qquad a \le x \le b \tag{7}$$

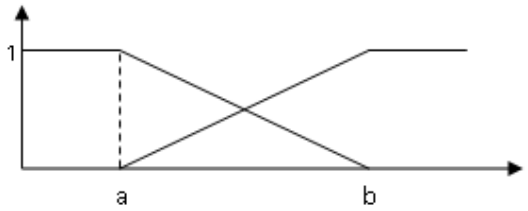

Figure4. Trapezoidal fuzzy function

Point **a** is the minimum score of right class and point **b** is the maximum score of left class. On the other hand, the number of left and right class points may be different from each other that have to be applied to these differences. As follows that if a class has more points, the degree of fuzzy membership function rises, and therefore, it is necessary to multiple each of the functions in the fuzzy relative frequency in the range of overlapping region, For example, suppose If the right class has n points in the overlapped interval and left class has m points in the overlapped interval then the fuzzy functions are defined as follows:

$$\mu^*_{class\ i}(x) = \mu_{class\ i} \tag{8}$$

$$\mu^*_{class\ j}(x) = \mu_{class\ j} \tag{9}$$

If $\mu^*_{class\ i}(x) > \mu^*_{class\ j}(x)$, then x belong to i_{th} class and otherwise belong to j_{th} class. Furthermore, we can have a good approximate of probability of membership of new sample in the new class aided fuzzy functions.

IV. EXPERIMENTS AND RESULTS

The above proposed method has been tested on two standard databases and fully implemented. Characteristics of databases are given in Table (1).

Table1. profiles of database

Breast Cancer			Iris		
Number of features	Number of classes	Number of samples	Number of features	Number of classes	Number of samples
9	2	699	4	3	150

Fisher score of each feature for both of the database is given in table (2)

Table2. Fisher score of each features.

	Iris	Breast Cancer
Fisher Score	Sepallength = 1.59	Clump Thickness = 1.05
	Sepalwidth = 0.63	Uniformity of Cell Size = 2.01
	Petallength = 15.72	Uniformity of Cell Shape = 2.02
	Petalwidth = 12.79	Marginal Adhesion = 0.94
		Single Epithelial Cell Size = 0.8
		Bare Nuclei = 1.90
		Bland Chromatin = 1.33
		Normal Nucleoli = 1.02
		Mitoses = 0.22

The minimum, the maximum and the average of SAW numbers in different classes is given in Table (3).

Table3. SAW score

	Classes	Min Saw	Max Saw	Avg Saw
Iris	setosa	4.31	8.22	6.08
	versicolor	13.85	21.87	18.08
	virginica	20.29	29.41	24.77
Breast Cancer	benign	1.14	7.15	1.89
	malignant	2.54	11.08	7.29

To obtain the accuracy, we have used ten cross fold validation for evaluating our methodology so that each time, nine of them are considered as the train and one of them as the test. The obtained results are very good compared with other famous classification algorithms that are obviously shown in chart (1) and (2).

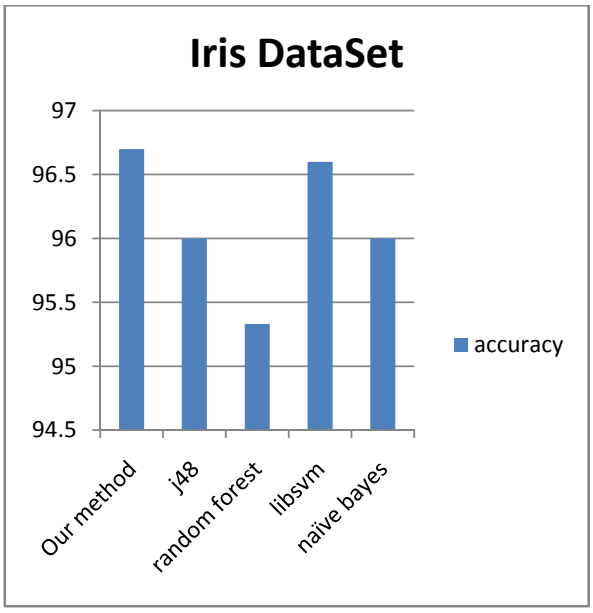

Chart1. Accuracy comparison between classifiers in Iris dataset

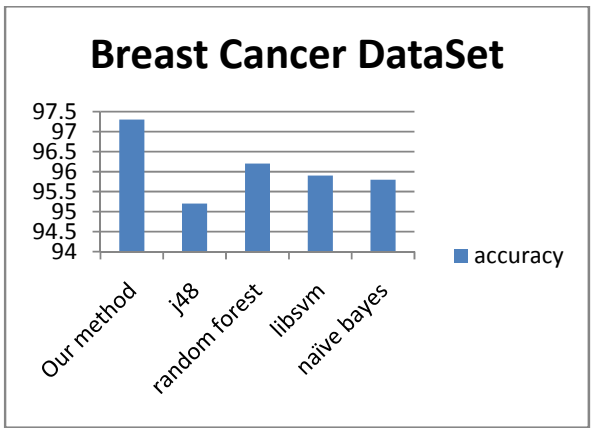

Chart2. Accuracy comparison between classifiers in breast cancer dataset

V. CONCLUSIONS AND FUTURE WORK

Classification is one of the most important processes of data mining. Existing algorithms in this field are often time consuming and complex. In this paper we have presented one of the methods based on multiple criteria decision making (SAW) which the most advantages are its simplicity and speed.

Results of experiments express that the proposed method is high accurate in addition of speed.

In order to complete our proposed method to support all datasets, we plan to convert qualitative features into quantitative. Furthermore, to improve the present approach, we intend to use another criterion and methods of multi criteria decision making for scoring features and samples.

VI. REFERENCES

[1] D.M.J. Asgharpoor, "Multiple criteria decision" 7 ed : Tehran University, 2009.

[2] C. Spathis, Doumpos, M., & Zopounidis, C., "Detecting falsified financial statements: A comparative study using multicriteria analysis and multivariate statistical techniques," European Account-ing Review, Taylor and Francis Journals, vol. 11(3), pp. 509–535, 2002.

[3] L. S. Arie B., "Generating rules from examples of human multiattribute decision making should be simple," Expert Systems with Applications, 2005.

[4] J. a. K. Han, Micheline, Data Mining:Concepts and Techniques, 2 ed.: Morgan Kaufmann, 2006.

[5] S. S.-N. M. MODARRES "FUZZY SIMPLE ADDITIVE WEIGHTING METHOD BY PREFERENCE RATIO," Intelligent Automation and Soft Computing, 2005.

[6] Y.-H. C. Shuo-Yan Chou a, Chun-Ying Shen, "A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes," European Journal of Operational Research, 2008.

[7] L. U. Ruta Simanaviciene, "Sensitivity Analysis for Multiple Criteria Decision Making Methods: TOPSIS and SAW," Procedia Social and Behavioral Sciences, 2010.

[8] D. L. O. Desheng Wu, "A TOPSIS Data Mining Demonstration and Application to Credit Scoring," International Journal of Data Warehousing & Mining, 2006.

[9] D. L. Olson, Wu, D, "Decision making with uncertainty and data mining," Lecture notes in artificial intelligence 2005.

[10] P. H. R. Duda, and D. Stork, Pattern Classification, 2nd edition ed.: Wiley Interscience, 2000.

[11] C.-C .a. L. Chang, Chih-Jen. (2001, LIBSVM -- A Library for Support Vector Machines. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/