A NEW CLASIFIER METHOD USING THE ROBBINS-MONRO STOCHASTIC APPROXIMATION ALGORITHM

Jae-Kook Lee, Chun-Taek Ko, Young-Shick Ro, Kab-Ju Hwang and Won-Ho Choi

School of Electrical Engineering, University of Ulsan
San-29 Moogu-2 Dong, Namgu, Ulsan 680-749, Korea

ABSTRACT

This paper presents a new data classification method using the Robbins-Monro stochastic approximation algorithm, k-nearest neighbor algorithm and probability analysis. The centroid of the test data set is decided by k-nearest neighbor algorithm and Robbins-Monro stochastic approximation algorithm. To decide the members of each class, the probability analysis is applied on the basis of the decided centroid point in data set. The proposed classification method is compared to the conventional fuzzy c-mean method, k-nn algorithm and discriminant analysis algorithm. The results show that the proposed method is more accurate than fuzzy c-mean method, k-means algorithm and discriminant analysis algorithm.
Keywords: data classification, Robbins-Monro stochastic approximation algorithm, distribution analysis.

1. INTRODUCTION

Data classification techniques are used to separate data sets in subsets which have the same features for data analysis, pattern recognition, fault detection, reliability analysis and etc. Many classification methods such as fuzzy c-mean algorithm, discriminant analysis and k-nn algorithm are widely used[1][2][3]. Recently, the advanced data classification technique is needed as the sensor technique is developed and as the using of the multivariable sensors is increasing in manufacturing or industrial system.

This paper presents a new data classification method using the Robbins-Monro stochastic approximation algorithm, k-nearest neighbor and distribution analysis.

The RMSA algorithm, originally proposed by Robbins and Monro in 1951 [6], is concerned with the problem of root finding of function $y = R(x)$ which is known or directly observed.

$$x_{n+1} = x_n - a_n(f(x_n) + e_n) \qquad (1)$$

We consider the Robbins-Monro algorithm (eq.1) for finding the zero of a function f where x_n is the estimate for the location of the zero of f, a_n is a sequence of positive constants tending to zero, and e_n represents measurement noise[5][7]. In Figure 1, It is shown the flow chart of sequence of the proposed algorithm. To decide and select centroid of the test data set, we used k-nn algorithm. The k-nn algorithm is a non parametric classification technique which has been shown to be effective in statistical applications. The technique can achieve high classification accuracy in problems which have unknown and non-normal distributions.

However, it is difficult to classify in large number of vectors and high computational complexity[4]. Then in the next step, we calculate the threshold value of the test data set. For more accurate the classification of the test data set, we apply to the probability theory to the data set. In the last step, we apply to the Robbins-Monro stochastic approximation algorithm to the data set.

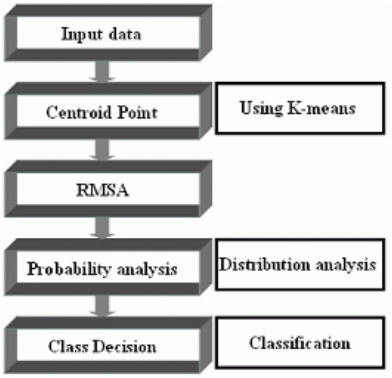


Fig. 1. The flow chart of proposed method

2. Distribution Analysis and Robbins-Monro Stochastic Approximation

2.1. Distribution analysis

The threshold of the outliers is determined by statistical test, assuming the data in each class are Gaussian distributed, the class means and variances for each feature are initially estimated.

We determined an outlier which is used a null hypothesis H_0 and an alternative hypothesis H_1. The significance α is equivalent to the probability distribution $|x_j - \mu_{ij}| \geq Threshold(T)$ is true given H_0:

$$\alpha = \Pr ob(|x_j - \mu_{ij}| > T \mid H_0) \quad (3)$$

The Figure 2 is decision for classifying classes using hypothesis theory.

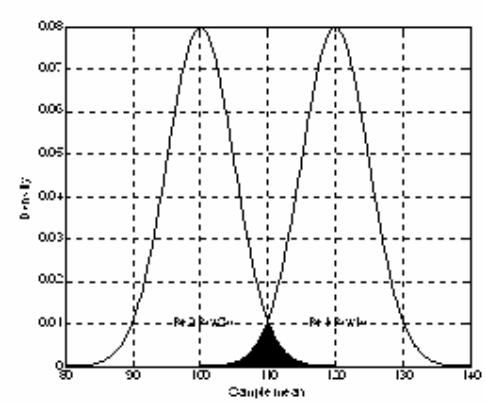


Fig. 2. Threshold decision for classifying

2.2. Robbins-Monro stochastic approximation

Centroid of the data set is used to the RMSA algorithm.

The goal is to estimate the parameter θ from a sequence $\{x_n\}$ of observations. The observations are of the form

$$x_n = \theta + v_n, \qquad n \geq 1,$$

Where the v_n are independently distributed random variables, each with pdf G which is symmetric about zero $(G(v) = 1 - G(-v^-))$. The information available about G is incomplete and is used to define a convex set P of symmetric pdf's, each with zero location parameter, to which G is confined. An estimate T is defined as a sequence $\{T_n\}$ of functions $T_n : R^n \to R$ where R is the real line. If F is in P and T is an estimate for which $T_n(\bar{x}_n) \to \theta$ almost surely or in probability and $n^{1/2} T_n(\bar{x}_n)$ is asymptotically normal when the v_n are distributed as $F(\bar{x}_n = (x_1, x_2, ..., x_n))$, then the asymptotic variance is denoted by V[T. F][8].

$$T_{n+1} = T_n - g_n(u(T_n - x_{n+1}) - p)] \quad (4)$$

Where $u(\bullet)$ is $u(x) = \begin{cases} 1, & x \geq 0 \\ 0, & otherwise \end{cases}$ and $\{g_n\}$ is the sequence of positive numbers.

In paper, the RMSA algorithm is applied to the following steps for decision most suitable centroid point.

Step 1. if data set $\{x_n\}$ is finite, equation $f(\theta) = 0$ is exist.

Step 2. θ is initial. $\theta^{k+1} = \theta^k - w_n Y_n$

Where Y_n is pdf's of $f(\theta^k)$, and w_n is constant : $0 \leq w_n \leq 1$.

Step 3. Iteration for decision $f(\theta) = 0$, Using equation

$$T_{k+1} = T_k - w_k(f(\theta^k) - f(\theta\ ))$$

The minimum θ is used to decision of centroid point in data set.

The initial centroid point is decided by k-means algorithm.

Figure 3 shows the area of RMSA algorithm to decide the centroid point of the iris data set.

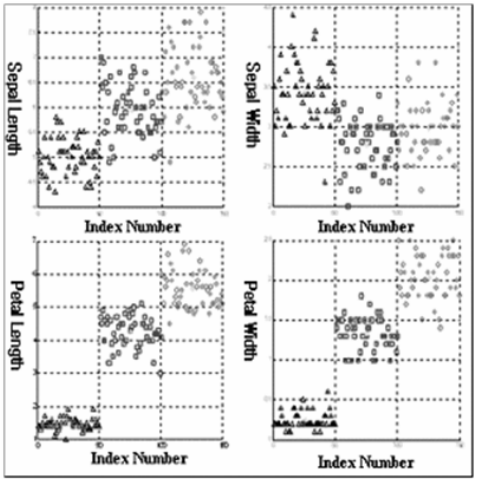


Fig. 3. Classification result of Iris data : setosa(△), versicolor(□), and virginica(O)

The Iris data is a common benchmark in classification and pattern recognition studies[6].

It contains 50 measurements of four features from each of the three classes Sepal length, Sepal width, Petal length, and Petal width.

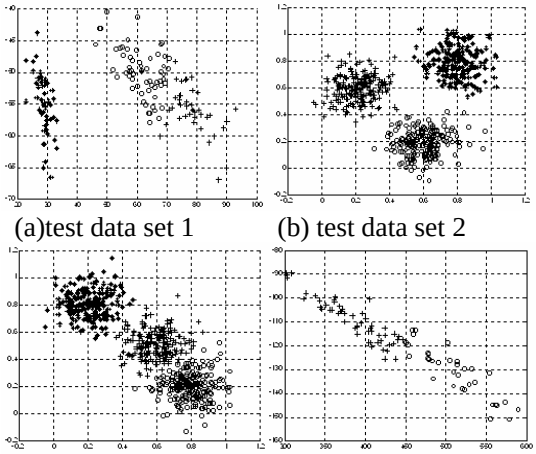
Table 1. Comparison of the centroid point variance of iris data

Centroid Point Comparison [Sepal Length]						
	Index	Setosa	Index	Versicolor	Index	Virginica
K-means	25, 5000	5, 0060	75, 5000	5, 9360	125, 5000	6, 5880
RMSA	25, 0000	5, 0061	75, 0000	5, 9408	125, 0000	6, 6020
Centroid Point Comparison [Sepal Width]						
	Index	Setosa	Index	Versicolor	Index	Virginica
K-means	75, 5000	2, 7700	125, 5000	2, 9740	25, 5000	3, 4280
RMSA	75, 0000	2, 7694	125, 0000	2, 9735	25, 0000	3, 4306
Centroid Point Comparison [Petal Length]						
	Index	Setosa	Index	Versicolor	Index	Virginica
K-means	25, 5000	1, 4620	75, 500	4, 2600	125, 5000	5, 5520
RMSA	25, 0000	1, 4633	75, 000	4, 2633	125, 0000	5, 5612
Centroid Point Comparison [Petal Width]						
	Index	Setosa	Index	Versicolor	Index	Virginica
K-means	25, 5000	0, 2460	75, 5000	1, 3260	125, 5000	2, 0260
RMSA	25, 0000	0, 2469	75, 0000	1, 3265	125, 0000	2, 0306

In table 1, centroid variance is compared to four features which have Sepal length, Sepal width, Petal length, and Petal width using k-means algorithm and proposed algorithm.

3. EXPERIMENTAL RESULTS

The experimental data sets have Iris data set and synthetic data which consist of the three classes and two classes of data. In Figure 4, it shows the test data set which are obtained by random function of MATLAB. To separate each other class, the different shape description is used in the figures of the experimental artificial data sets.



(a)test data set 1 (b) test data set 2

(c)test data set 3 (d) test data set 4

Fig. 4. The synthetic test data set

The test data set 1, 2, and 3 contains 200 measurements of 3 classes and test data set is contains 50 measurements of 2 classes.

In the Table 2, the results of the compared four algorithms are shown. It shows description of the performance rate using the fuzzy c-mean algorithm, k-means algorithm, discriminant analysis algorithm, and the proposed algorithm. In case of test data set 1, the performance rate of the proposed algorithm using RMSA is about 96%, the other algorithm is about 95%. In case of test data set 2, the performance rate of the proposed algorithm is about 99%, its of the fuzzy c-mean algorithm, k-nn algorithm and discriminant analysis is similar to performance rate of the proposed algorithm because the test set is clearly classify the classes. In the point of performance results, the performance of our proposed algorithm is better than the performance using conventional fuzzy c-mean algorithm, k-nn algorithm, and discriminant analysis algorithm.

Table 2. Comparison of the rate of classification

No	Number Of Data	Fuzzy c-mean	K-means	Discrimdnant analysis	RMSA
Test data 1	150	95%	95%	95%	96%
Test data 2	600	99%	99%	99%	99%
Test data 3	600	94%	94%	94%	94%
Test data 4	100	94%	94%	95%	95%

4. CONCRUSION

In this paper, we proposed a new data classification algorithm using the Robbins Monro stochastic approximation algorithm**,** k-means algorithm and distribution analysis to improve the classification performance. Because the centroid point in data set is very important to classify the data, the centroid point of the data set values is determined by RMSA algorithm, and then classification of each class is decided by distribution analysis. From the experimental results, we used the iris data set and synthetic data for classification performance. the proposed data classification algorithm shows better performance than the conventional fuzzy c-mean classification method, k-means classification method, and disciminant analysis and compared the variance centroid point in test iris data set.

For our future works, new classification algorithm will be applied to the real data sets and various data sets.

REFERENCES

1. Ming Chuan Hung. and Dong Lin Yang. "An Efficient Fuzzy C-Means Clustering Algorithm," Data Mining, 2001. ICDM 2001. Proc IEEE international Conf., (2001), pp. 225-232.
2. Pal N.R. and Bezdek J.C., "On Cluster Validity for the Fuzzy C-Means Model," IEEE Trans. on Fuzzy Systems. vol.3(1995) pp. 370-379.
3. Frakt A. B., Karl W.C., and Willsky A.S., "A Multiscale Hypothesis Testing Approach from Noisy Tomographic Data," IEEE Trans. on Image Processing. vol.7(1998) pp. 825-837.
4. Kato Y, Takahashi M and Ohtsuki R.,Yamaguchi., "A Proposal of Fuzzy Test for Statistical Hypothesis," Systems, Man, and Cybernetics, IEEE international Conf(2000). pp. 2929-2934.
5. Setnes, M. and Roubos, H. "GA-fuzzy modeling and classification: complexity and performance," IEEE Trans on. Vol8(2000) pp.509 - 522.
6. Kulkarni, S.R. and Horn, C. "Convergence of the Robbins-Monro algorithm under arbitrary disturbances", IEEE conf., on Decision and Control(1993), pp.537 – 538.
7. Price, E. VandeLinde, V, "Robust estimation using the Robbins-Monro stochastic approximation algorithm", Trans. IEEE, on Information Theory, Vol. 25 (1979) pp. 698 – 704.