로빈스-몬로 확률 근사 알고리즘을 이용한 데이터 분류

이재국, 고춘택, 최원호
울산대학교 전기전자정보시스템공학부

Data Classification Using the Robbins-Monro Stochastic Approximation Algorithm

Jae KooK Lee, Chun Taek Ko and Won Ho Choi
School Of Electrical Engineering, University of Ulsan

ABSTRACT

This paper presents a new data classification method using the Robbins Monro stochastic approximation algorithm, k-nearest neighbor and distribution analysis. To cluster the data set, we decide the centroid of the test data set using k-nearest neighbor algorithm and the local area of data set. To decide each class of the data, the Robbins Monro stochastic approximation algorithm is applied to the decided local area of the data set. To evaluate the performance, the proposed classification method is compared to the conventional fuzzy c-mean method and k-nn algorithm. The simulation results show that the proposed method is more accurate than fuzzy c-mean method, k-nn algorithm and discriminant analysis algorithm.

1. Introduction

Data classification techniques are used to separate data sets in subsets which have the same features. It is applied to the application of 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 which are widely used[1][2][5]. 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 Robbins Monro stochastic approximation algorithm, originally proposed by Robbins and Monro in 1951[4][6], is concerned with the problem of root finding of function $y = R(x)$ which is known or directly observed. We consider the Robbins Monro algorithm

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

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[3].

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 nonnormal 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.

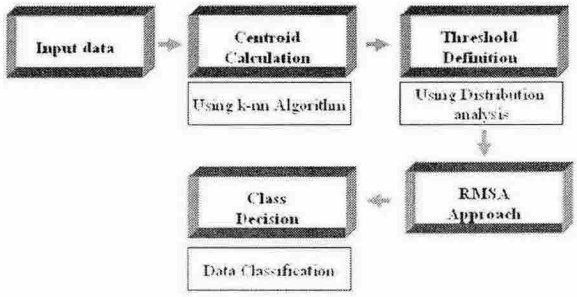

Fig. 1. flow chart of sequence of the proposed algorithm

2. K-NN, Distribution analysis and the Robbins Monro stochastic approximation

A. K-nearest neighbor algorithm[2]

In K-nearest neighbor algorithm, the training dataset is used to classify each member of a target dataset for classification. Generally speaking, K -nearest neighbor algorithm that consists of three steps is as follows:

1. For each row in the target dataset which is the set to be classified, locate the nearest neighbors of the training dataset.

2. An Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.

If input data set is $x_i = (x_{i1}, x_{i2}, ..., x_{id})^T$ and $x_j = (x_{j1}, x_{j2}, ..., x_{jd})^T$, Euclidean Distance is calculated by

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} | x_{ik} - x_{jk} |^2} \quad (2)$$

3. Repeat this procedure for the remaining rows in the target set.

In our proposed algorithm, centroid point is decided by K-nearest neighbor(k-nn) algorithm. Figure 2. is shown the centroid point of the test dataset.

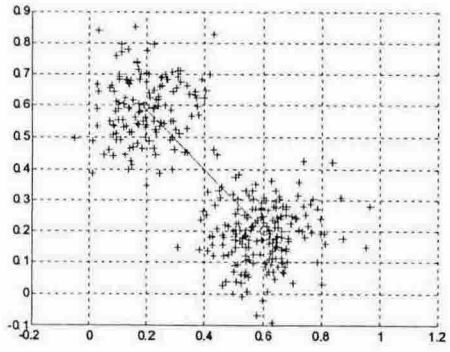

Fig. 2. The centroid point of the test dataset

B. Distribution analysis

The threshold of the outlier is determined by statistically testing, 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 a alternative hypothesis H_1.

The significance α is equivalent to the probability distribution $| x_j - \mu_{ij} | \geq Theshold\ (T)$ is true given H_0:

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

Finally, data set is applied the Robbins Monro stochastic approximation algorithm.

C. Robbins-Monro stochastic approximation

The goal is to estimate the parameter θ from a sequence $\{x_n\}$ of observations. The observations are of the form $x_n = \theta + v_n, \quad n \geq 1,$

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 $T_n : R^n \rightarrow R$ where R is the real line. If F is in P and T is an estimate for which $T_n(\bar{x}_n) \rightarrow \theta$ 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][4].

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

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.

Figure 3. is shown the area of Robbins Monro stochastic approximation algorithm to decide the decided local area of the artificial data set.

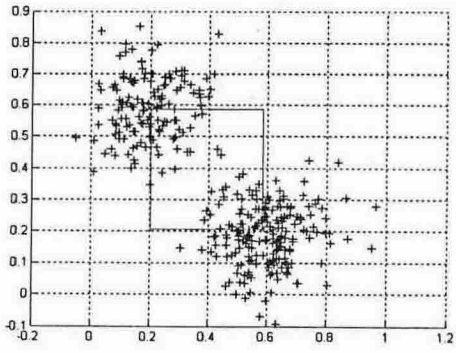

Fig. 3. Decided local area of the artificial data set

3. Experimental results

The experimental data sets have two classes of data. These data are to classify the fault and nonfault. In Figure 4 and Figure 5 the experimental results are shown. To get simplified data sets, they are obtained by random function of MATLAB. To separate each other class, the different color description is used in the figures of the experimental artificial data sets.

The performance of experimental results is compared to the fuzzy c-mean algorithm, k-nn algorithm, discriminant analysis algorithm, and the proposed algorithm.

In the Table 1, the results of the compared four algorithms are shown. It is shown description of the performance rate using the fuzzy c-mean algorithm, k-nn algorithm, discriminant analysis algorithm, and the proposed algorithm. In case of test data set 2, the performance rate of the proposed algorithm using Robbins Monro stochastic approximation is about 96%, one of the fuzzy c-mean algorithm is about 94%, one of the K-nn algorithm is about 94%, and one of the discriminant analysis algorithm is about 95%. In case of test data set 1, the performance rate of the proposed algorithm is about 98%, its of the fuzzy c-mean algorithm, k-nn algorithm and discriminant analysis is similar to performance rate of the proposed algorithm. 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 1 The comparison table of classification rate

No Of Data set	Number Of data	Fuzzy c-mean	K-nn	Discriminant analysis	Proposed algorithm
Test data set 1	400	98%	98%	98%	98%
Test data set 2	100	94%	94%	95%	96%

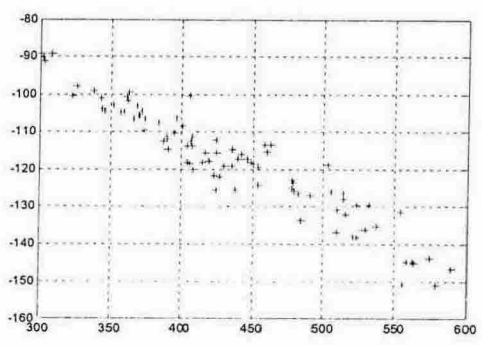

Fig. 5. The artificial test data set 2

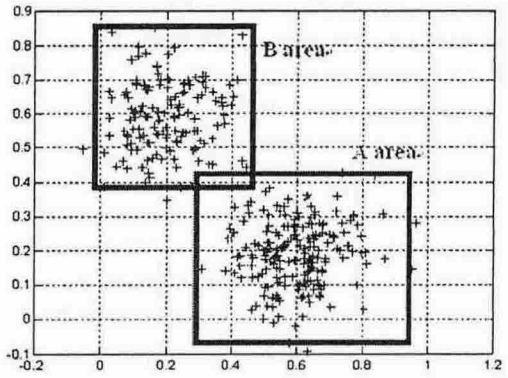

Fig. 4. The artificial test data set 1

In Figure 6 and Figure 7 they are shown the variance of estimate T in the artificial test data 1. Figure 6. shows the variance of estimate from centroid of A area and Figure 7. shows the variance of estimate from centroid of B area.
Because the index of classified data is clearly separated, it can be classified based on the variance of estimate T.
In Figure 8 and Figure 9 they are shown the results of the classification using the proposed algorithm.

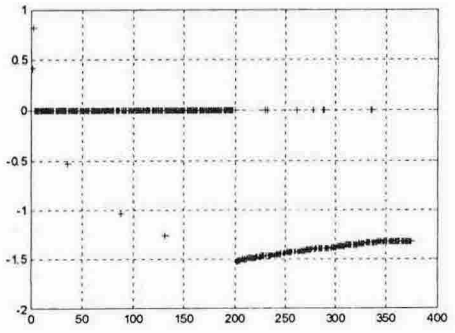

Fig. 6. the variance of estimate T

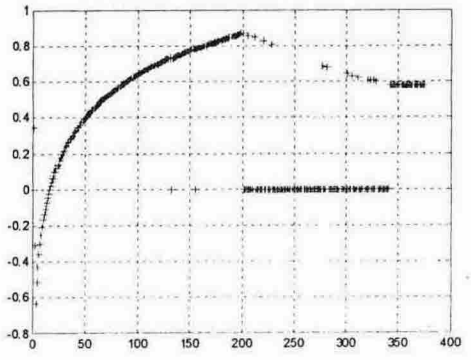

Fig. 7. the variance of estimate T

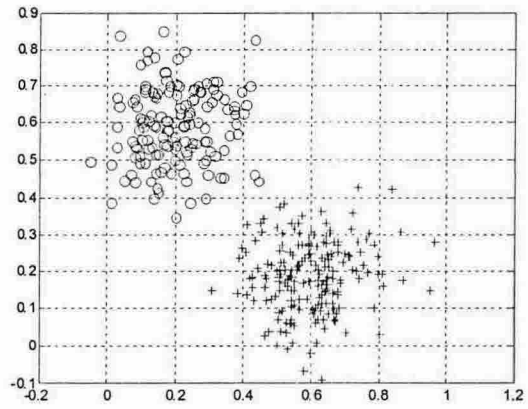

Fig. 8. The results of the test data 1

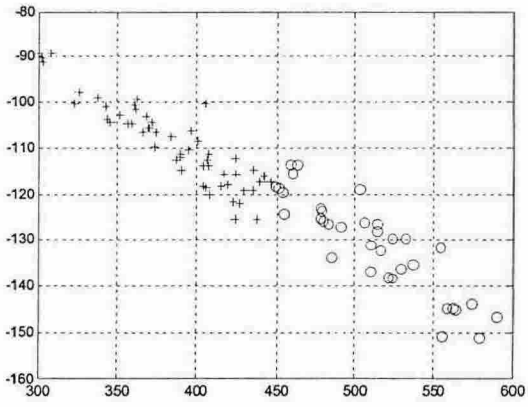

Fig 9. The results of the test data 2

4. Conclusion

In this paper, we proposed a new data classification algorithm using the Robbins Monro stochastic approximation algorithm, k-nearest neighbor and distribution analysis to improve the classification performance. The centroid point of the data set values is determined by k-nn algorithm, and then local area for applying the Robbins Monro Stochastic approximation is decided by distribution analysis. A decision of each class is based on the measured RMSA algorithm.

From the experimental results, the proposed data classification algorithm shows better performance than the conventional fuzzy c-mean classification method, k-nn classification method, and discriminant analysis.

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

이 논문은 NARC와 울산대학교 교비연구비 지원에 의하여 연구되었슴

참 고 문 헌

[1] Ming Chuan Hung., Dong Lin Yang., "An Efficient Fuzzy C-Means Clustering Algorithm," Data Mining, 2001. ICDM 2001. Proc IEEE international Conf., 2001., pp. 225-232, Dec, 2001.

[2] Han J.H. and Kim Y.K., "A Fuzzy K-NN algorithm using weights from the variance of membership values," Computer Vision and Pattern Recognition, 1999. Proc. IEEE Computer Society Conf., pp. 394-399, June, 1999.

[3] Kulkarni, S.R. Horn, C., "Convergence of the Robbins-Monro algorithm under arbitrary disturbances", IEEE conf., on Decision and Control, pp.537 - 538, Dec 1993.

[4] Price, E. VandeLinde, V., "Robust estimation using the Robbins-Monro stochastic approximation algorithm", Trans. IEEE, on Information Theory, , Vol. 25 pp. 698 - 704, Nov 1979.

[5] Parthasarathy, G. Chatterji, B.N., "A class of new KNN methods for low sample problems", Tran. IEEE, on Systems, Man and Cybernetics, Vol 20, pp. 715 - 718, May-June 1990.

[6] Martin, R. Masreliez, C. " Robust estimation via stochastic approximation", Trans. IEEE, on Information Theory, Vol 21, pp. 263 - 271, May 1975.