Change –Points Detection in Fuzzy Point Data Sets

Hui-hui Wang, and Li-li Wei †,

School of Mathematics and Computer Science, Ningxia University, Yinchuan, P. R. China

Summary

Change-points detection is one of important problems in data analysis. Traditional change-points detection method is based on exact data sets which can't reflect prior information of data. In this paper, a new concept, called "fuzzy point data" which is defined by giving a fuzzy membership to the data in exact data sets, is proposed for helping us handle the confidence of data. We introduce regression-classes mixture decomposition method for Change-points detection in fuzzy point data sets. In the method, different regression classes are mined sequentially in fuzzy point data sets and the estimation of change-points are determined by the two joined regression-classes, the number of the change-points will not be pre-specified. Numerical experiments show that by using fuzzy data point data, important data can make much contribution to mining regression classes. This shows that the change-points we got in fuzzy data point sets are more meaningful than we got in exact data sets.

Key words:

Fuzzy point data, Change-points detection, Robust, Regression-classes

1. Introduction

Detection of change-points in the characteristics of some physical system is one of the most important practical problems arising in signal processing. (speech processing, geophysics, EEG, EMG and ECG analysis, etc., see [1, 2] for several examples of application).

There have been many exciting developments in the theory of change-point detection. New promising directions of research have emerged, and traditional trends have flourished anew. Change-points detection in regression model is a most important branch of change-point detection, and attracts much attention for several decades. Quandt[3] and Kim[4] used the likelihood ratio test for a change in a regression model. Chen[5] used the SIC (Schwarz Information Criterion) to locate and detect a change point in the liner regression models. Hobert[6], Chen Choy and Broemeling[7] adopted some Bayesian methodology. But all the authors mentioned above mainly considered one change point in regression models. Nowadays multiply change-points problem has received much attention. Tang and Fei[8] use SIC to detection change points in polynomial regression models.

All these methods are based on exact data sets, in which all training data are treated uniformly in the detection of change-points. However, in many applications, the effects of the training data are different. Some training data may be more important than others, that is to say, different training data should make different contributions to finding the change. According to these, in these paper, we assign a confidence weight to each training data, and use a highly robust RCMD[9] (Regression-Class Mixture Decomposition) method to detect change-points in the data set. In this method, regression model which has one or more change-points were regard as mixture of different regression-classes[9]. Change-point can be regarded as the demarcative point of two regression-classes. We first mined all of the regression-classes sequentially in the fuzzy data set, and then the number and positions of the change-points will be got simultaneously.

This paper is organized as follows. The basic ideas of fuzzy point data are given in Section 2, in which some possible methods of determining the fuzzy memberships for each train data were discussed. In Section 3, we introduce RCMD method for fuzzy point data set, and discuss the regression change-points detection. To substantiate the theoretical analysis, simulation runs are performed in Section 4 to evaluate the effectiveness of the proposed method. Conclusions are reported in Section 5.

2. The Fuzzy Point Data

Suppose that **x** is *p*-dimensional predicted variable, y is one-dimensional response variable, and $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$ is training sample. Here, each training data was given a different fuzzy membership s_i ($0 < s_i \quad 1$), where s_i can be regarded as importance degree or confidence degree of the corresponding data point towards the population. $((\mathbf{x}_i, y_i), s_i)$ is called a fuzzy point in *p*+1-dimensional space, $(\mathbf{x}_i, y_i)$ and s_i are respectively called support point and height. $\{((\mathbf{x}_i, y_i), s_i), i = 1, 2, ..., n\}$ is called fuzzy point data set.

The fuzzy membership of each training point was determined by the given problem, which has a little subjectivity. So they can reflect the prior information of the data. In general, according to different kind of data, we use different method to compute the fuzzy membership. For example following several cases are given.

(i) Timeliness data: In some application problem such as real-time signal process, data are time-dependent. For example, the data coming late is more important than the early one. In that case, we can denote $s_i = f(t_i)$ (t_i is the time which the point x_i arrived in the system), here $f($✕) is a monotone increasing function.

(ii) Prior information: s_i can indicate the degrees of confidence of training data, and the confidence degrees can be decided by the prior information about training data. For example, if $(\mathbf{x}_i, y_i)$ is very important to analyzer, then a value approximately 1 can be assigned to the confidence degree s_i of this point.

(iii) Repeated measures data: In some research fields, response variable must be measured repeatedly in each carrier. Then the mean of the measured values is regard as observation of this carrier. As different carrier has different times of measure, the fuzzy membership s_i is in contact with the measure frequency of corresponding carrier.

(iv) Categorical data: Usually different categorical has different importance in research. Analyzers always hope more important one can be gotten exactly, while less important one can be allowed something wrong with classification. In this case, we can determine the fuzzy membership of the training data by labeling variable.

(v) Heteroscedasticity data: In heteroscedasticity data, variances of the errors in a model are not always a constant. They change with the different of the independent variables. If we choose appropriate fuzzy weight s_i for each data making the variance of the error be equal, the hypothesis homoscedasticity of variance in classical method can be satisfied.

In a word, determining the fuzzy memberships for training data is not difficulty. In this paper, we suppose $s_i, i = 1,...,n$ have been known.

3. Change-Points Detection in Fuzzy Point Data Set

3.1 RCMD based on Fuzzy Point Data

RCMD is an effect means for data mining. It is a highly robust method which can resist a very large proportion of noisy data[10].

A regression class G_j is defined by the following regression model with random carriers:

$$G_j : y = f_j(\mathbf{x}, b_j) + e_j, \quad j = 1, \quad ,m, \tag{1}$$

where y Î R is the response variable, the explanatory variable that consists of carriers or regressors $\mathbf{x}$ Î R^p is a random(column) vector with probability density function (p.d.f.) $p($✕),the error term e_j is a random variable with a p.d.f. $y(u; s_j)$, s_j is a parameter, $Ee_j = 0$ and $\mathbf{x}$, e_j are independent. Here $f_j($利) is a known regression function, and b_j is an unknown regression parameter (column) vector.

A random vector $(\mathbf{x}, y) \quad G_j$ implies that $(\mathbf{x}, y)$ has a p.d.f.

$$p_j(\mathbf{x}, y; q_j) = p(\mathbf{x})?y(y \quad f_j(\mathbf{x}, b_j); s_j), \quad q_j = (b_j^T, s_j)^T. \tag{2}$$

If the random observations were taken from common mixture distribution population, they obey the regression-class mixture model, the p.d.f. is

$$p_q(\mathbf{x}, y;) = \overset{m}{\underset{j=1}{\text{å}}} p_j p_j(\mathbf{x}, y; q_j). \tag{3}$$

here $q^T = (q_1^T,..., q_m^T)$ .That is, they consist of random observations from m regression-classes with prior probabilities $p_1,..., p_m$ ($p_1 +$ 利? $p_m = 1$, p_i 刈0,1 $i \quad m$).

To given data set $\{(\mathbf{x}_1, y_1),...,(\mathbf{x}_n, y_n)\}$, we assume that there are m regression-classes $G_j, j = 1,...,m$ in data set under study and that m is known in advance (indeed m can be determined at the end of the mining process when all potential regression-classes have been identified). With respect to a particular regression-class G_j, all other regression-classes in the mixture can be readily classified as part of the outlier category in the sense that these other observations obey different statistics. Thus, a mixture density can be viewed as a contaminated density with respect to each cluster in the mixture. According to this idea, the mixture p.d.f. in (3) with respect to G_j can be rewritten as

$$p_e(\mathbf{x}, y; q) = p_j(1 - e_j)p_j(\mathbf{x}, y; q_j) + [1 - p_j(1 - e_j)]g_j(\mathbf{x}, y), \tag{4}$$

where e_j is an unknown fraction of an outlier which present in G_j, and $e = \{e_{1,...,} e_m\}$. Ideally, a sample point $(\mathbf{x}_i, y_i)$ from the above mixture p.d.f. is classified as inliers if it is realized from $p_j(\mathbf{x}, y; q_j)$ or as outliers otherwise (i.e. it comes from the p.d.f. $g_j(\mathbf{x}, y)$).

Assuming that $g_j(\mathbf{x}_1, y_1) = ... = g_j(\mathbf{x}_n, y_n) = d_j$, the above expression is rewritten as

$$p_e(\mathbf{x}, y; q) = p_j(1 - e_j)p_j(\mathbf{x}, y; q_j) + [1 - p_j(1 - e_j)]d_j. \tag{5}$$

After a valid regression-class has been detected, it is extracted from the current data set, and the next regression-class will be identified in the new size-reduces data set. Individual regression-classes continue to be estimated recursively until there are no more valid regression-classes, or the size of the new data set gets to be too small for estimation. Thus the number of the regression-classes can be gotten automatically.

Now each training data is given a fuzzy degree s_i and $e_j = 0, j = 1,\ldots,m$, which means we consider the case in which $p_j(\mathbf{x}, y; q_j)$ is not contaminated. With respect to fuzzy point data set $D = \{((\mathbf{x}_i, y_i), s_i), i = 1,\ldots,n\}$, the log-likelihood function of D can be written as

$$l_j(q_j) = n\log p_j + \mathop{\text{å}}_{i=1}^{n} \log p_j(x_i, y_i, s_i; q_j) + \frac{1-p_j}{p_j} d_j \text{ , (6)}$$

where $p_j(\mathbf{x}_i, y_i, s_i; q_j) = p(\mathbf{x}_i)?y(s_i y_i \quad s_i f_j(\mathbf{x}_i); s_j)$. In order to estimate q_j from D, we need to maximize $l_j(q_j)$ with each d_j subject to $s_j > 0$. Provided that $t_j = (1-p_j)d_j / p_j$, then we can discuss the problem of maximizing $l_j(q_j; t_j)$ inside of maximizing $l_j(q_j)$

$$l_j(q_j; t_j) = \mathop{\text{å}}_{i=1}^{n} \ln p_j(\mathbf{x}_i, y_i, s_i; q_j) + t_j \quad . \qquad (7)$$

In particular, when $\mathbf{x}$ is distributed uniformly (i.e., $p(\mathbf{x}) = c$) and e_j□$N(0, s_j^2)$, the maximization of $\log l(q_j)$ is equivalent to maximizing

$$\bar{l}_j(q_j; \bar{t}_j) = \mathop{\text{å}}_{i=1}^{n} \log[y(s_i y_i - s_i f(b_j, x_i); s_j^2) + \bar{t}_j], \text{ (8)}$$

here $\bar{t}_j = t_j / c$. For simplicity we will still denote $\bar{t}_j$ and $\bar{l}_j$ by t_j and l_j, respectively.

At each selected $t_j^{(v)} (v = 0, \quad ,V)$, we maximize $l_j(q_j; t_j)$ with respect to b_j, s_j by using an iterative algorithm or by using a genetic algorithm (GA). Having solved $\max_{b_j, q_j} l_j(q_j; t_j^{(v)})$ for $\hat{b}_j(t_j^{(v)})$ and $\hat{s}_j(t_j^{(v)})$, the possible regression-class $G_j(\hat{q}_j(t_j^{(v)}))$ can be expressed as

$$G_j(\hat{q}_j(t_j^{(v)})) = \{((\mathbf{x}_i, y_i): \left| y_i - f(\mathbf{x}_i, \hat{b}_j(t_j^{(v)})) \right| \quad 3\hat{s}_j(t_j^{(v)})\}, \quad (9)$$

followed by the test of normality on $G_j(\hat{q}_j(t_j^{(v)}))$. If the test statistic is not significant, then the hypothesis that a valid regression-class $G_j(\hat{q}_j(t_j^{(v)}))$ has been determined, otherwise we proceed to the next partial model until the upper bound $t_j^{(V)}$ has not been reached.

3.3 Change-Points Detection in Fuzzy Point Data Set

To a fuzzy point data set $D = \{((\mathbf{x}_i, y_i), s_i), i = 1,\ldots,n\}$, the detection of change-point can be found by using the following two steps:

(i) Step1: By using RCMD method, extract all of the regression-classes in the fuzzy point data set sequentially;

(ii) Step2: Analyzing two joined regression-classes, and calculating the estimate of the position of change-points. After all of the regression-classes have been considered, the number of the change-points can be gotten.

In step 2, the carriers $\mathbf{x}$ in each regression-class are been arranged in sequence, and the maximum and the minimum of the carriers of each class will be found. By this information we can get the arrangement of the regression–classes, and find which two models are jointed together. Then base on the definition of the change–point, we give the estimation of the change–point as $(\text{Max}_1 + \text{Min}_2) / 2$, here Max_1 is the maximum of carriers in the former regression-class, while Min_2 is the minimum of the carriers in the latter regression-class.

4. Simulations

We give two numerical simulation examples to illustrate the effectiveness and applicability of our method. Example 1 is an application of the method in the problem of linear models. Example 2 deals with structure mining involving the mixture of curve and line.

Example1. This example considers two regression-classes in the data set. In simulation run, 40 data points are generated according to the following models:

$$\begin{cases} \text{reg-class } 1: y = x + e_1, \qquad e_1 \text{□} N(0, 0.25^2), p_1 = 0.5; \\ \text{reg-class } 2: y = 8 - x + e_2, e_2 \text{□} N(0, 0.26^2), p_2 = 0.5. \end{cases}$$

That is, there are approximately 20(40´ 0.5) data point from reg-class1 (regression-class), 20(40´ 0.5) data point from reg-class2, each data point (x_i, y_i), $i = 1,\ldots,40$ was given a fuzzy degree $s_i \quad i = 1,\ldots,40$. The scatter plot of the data set and the fitting result is depicted in Fig.1. The small circles indicate the fuzzy points and each size of these small circles is direct proportion to the correspondence value of fuzzy weight s_i. The dash line in Fig.1 shows the fitting result in the exact data set and the solid line shows the result in the fuzzy data set.

From Fig.1, we can see that bigger weighting fuzzy point can make much contribution to the fitting curve, and the position of change-point is also change accordingly. And by using fuzzy point data sets, the prior information of the data can be considered in analyzing. When s_i º 1, the change-points detection in fuzzy point data sets degenerates to the case in exact point data sets.

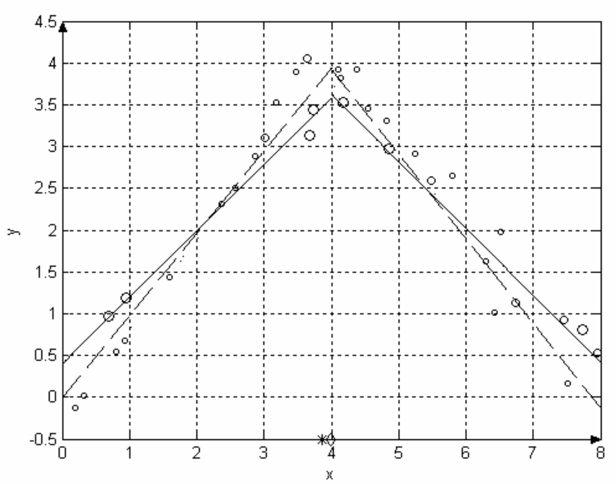

Fig. 1. Scatter plots and line with one change-point, ∗ show the position of the change-point in fuzzy point data, ◇ show the position in the exact data set.

Example2. Detection change-point in nonlinear structure is an important problem of computer vision and pattern recognition. In this example, we use RCMD method to accomplish this work effectively.

The sample data set $\{(x_i, y_i), i = 1,..., 40\}$ is generated according to the following models:

$$\begin{cases} \text{reg-class1:} \quad y = 3x + e_1, \quad e_1 \square N(0, 0.1^2), \quad p_1 = 0.5; \\ \text{reg-class2:} \quad y = 4x - 0.25x^2 + e_2, e_2 \square N(0, 0.2^2), \quad p_2 = 0.5. \end{cases}$$

there are approximately 20(40´ 0.5) data point draw from reg-class 1, 20(40´ 0.5) data point draw from reg-class 2, each data point (x_i, y_i) , $i = 1,..., 40$ was given a fuzzy degree s_i $i = 1,..., 40$. The scatter plot of the data set and the result is showed in Fig.2.The dotted line in Fig.2 is the fitting result of in the exact data set and the solid line shows the result of in the fuzzy data set.

From Fig.2, we can see that the fitting curve is apparently closer to the data which have bigger weight, and the position of change-point also change. That is to say, by using the prior information of the data set, we can get a different value of the change-point which is more suitable than we got form exact data set.

This example also shows that our method can be used to find change-points in nonlinear patterns in fuzzy data sets. Hence, the RCMD method has a great potential in detecting change-points.

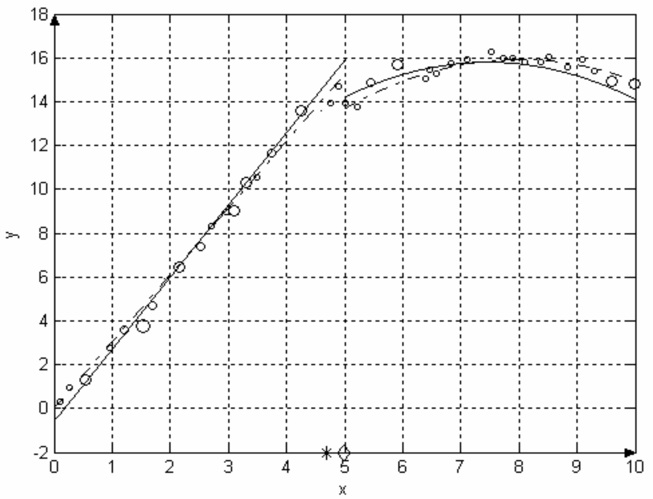

Fig. 2. Detecting change-point in linear and nonlinear structures, ∗ show the position of the change-point in fuzzy point data, ◇show the position in the exact data.

5. Conclusion

In this paper, we introduce the concept of fuzzy point data into change-points detection, so that the prior information of the data which is always ignored can be considered in the method again, and help us to find more reasonable estimation of the change-point. By using RCMD method for change-point detection, we first mining all of the regression-classes sequentially, and then determining the position of the change-point, so the number of the change-point is gotten automatically. Numerical examples have also shown that the method appears to be a promising method to detection change-points and has much potential application in a variety of disciplines such as quality control, seismic signal processing and economy.

In this paper, we only study the change-point detection in 2-dimensional space, how to find change-points in higher space is also a challenging and difficult problem for further research.

Acknowledgments

This work was support by a grant from the National Natural Science Foundation of China (No.60663003), and a grant from the Natural Science Foundation of Ningxia P. R. China (NZ0516).

References

[1] M. Basseville, N. Nikiforov, “The Detection of Abrupt Changes-Theory and Applications,” Information and System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] B. Brodsky, B. Darkhovsky, “Nonparametric Methods in Change-Point Problems,” Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

[3] R.E. Quandt, “Tests of the hypothesis that a linear regression system obeys two separate regimes,” J. Am. Stat. Assoc. vol.55, pp.324-330, 1960.

[4] D. Kim, “Tests for a change-point in linear regression,” IMS Lecture Notes-Monograph Series, vol.23, pp.170-176, 1994.

[5] J. Chen, “Testing for a change point in linear regression models,” Communications in Statistics: Theory and Method, vol.27, pp.2481-2493, 1998.

[6] D. Hobert, “A Bayesian analysis of a switching linear model,” J. of Econo., vol.19, pp.71-78, 1982

[7] J. H. Chin Choy, L. D. Broemling, “Some Bayesian inferences for a changing linear model,” Techno metrics, vol.22, pp.71-78, 1980.

[8] Y. C. Tang, H. L. Fei, “Detecting change points in polynomial regression models with an application to cable data sets,” Acta Math. Appl., Sinica, English Series, vol.20, pp.541-546, 2004.

[9] X. H. Zhuang, Y. Huang and K. Palaniappan, “Gaussian mixture density modeling, decomposition, and applications,” IEEE Trans. on image processing, vol.9, pp.1293–1302, 1996.

[10] J. H. Ma, Y. Leung and J. C. Luo, “A highly robust estimator for regression models,” Pattern Recognition Letters, vol.27, pp. 29–36, 2006.

Wang Hui-hui received the B.S. degree in applied mathematics from Ningxia University in 2004. She is currently pursuing the M.S. degree in applied mathematics from Ningxia University. Her current research interests include applied statistical theory, pattern recognition and data mining.

Wei Li-li received the B.S. degree in mathematics from Northwest Normal University in 1985, and the M.S. degree in mathematics from North-western Polytechnical University in 1988. In 2003, he received the P.H.D. in applied mathematics from Xi’an Jiaotong University. He currently serves associate dean of School of Mathematics and Computer Science and is a professor at Ningxia University. His research interests include applied statistical theory, machine learning and data mining.