Rough Clustering and Regression Analysis

Georg Peters

Munich University of Applied Sciences
Faculty of Computer Science and Mathematics
Lothstrasse34, 80335 Munich, Germany
georg.peters@muas.de

Abstract. Since Pawlak introduced rough set theory in 1982 [1] it has gained increasing attention. Recently several rough clustering algorithms have been suggested and successfully applied to real data. Switching regression is closely related to clustering. The main difference is that the distance of the data objects to regression functions has to be minimized in contrast to the minimization of the distance of the data objects to cluster representatives in k-means and k-medoids. Therefore we will introduce rough switching regression algorithms which utilizes the concepts of rough clustering algorithms as introduced by Lingras at al. [2] and Peters [3].

Keywords: Rough sets, switching regression analysis, clustering.

1 Introduction

The main objective of cluster analysis is to group similar objects together into one cluster while dissimilar objects should be separated by putting them in different clusters.

Besides many classic approaches [4,5] cluster algorithms that utilize soft computing concepts have been suggested, e.g. Bezdek's fuzzy k-means [6] or Krishnapuram and Keller's possibilistic approach [7]. Recently also rough cluster algorithms have gained increasing attention and have been successfully applied to real life data [2,8,9,10,11,12,3,13,14].

Switching regression models [15,16] are closely related to cluster algorithms. However, while cluster algorithms, like the k-means, minimize the cumulated distance between the means and the associated data objects the objective of switching regression analysis is to minimize the cumulated distance between the K regression functions $Y_k \quad (k = 1, ..., K)$ and their associated data objects (Figure 1).

The objective of the paper is to transfer the concepts of rough clustering algorithms to switching regression models and introduce rough versions. We also briefly specify possible areas of applications.

The paper is structured as follows. In the following Section we give a short overview on switching regression models and rough cluster algorithms. In Section 3 we introduce rough switching regression models. In the last Section we give a brief discussion and conclusion.

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 292–299, 2007.

2 Fundamentals: Switching Regression and Rough Clustering

2.1 Switching Regression Models

Switching regression models were introduced in the fifties of the last century [15]. In the meantime these classic, probabilistic based models have been accompanied by switching regression models that utilize soft computing concepts like fuzzy set theory.

Classic Switching Regression Models. Let us consider a simple data constellation as depicted in Figure 1. Obviously two linear regression functions (Y_1 and Y_2) should adequately represent these data:

$$Y_1(x) = a_{10} + a_{11}x \quad \text{and} \quad Y_2(x) = a_{20} + a_{21}x \tag{1}$$

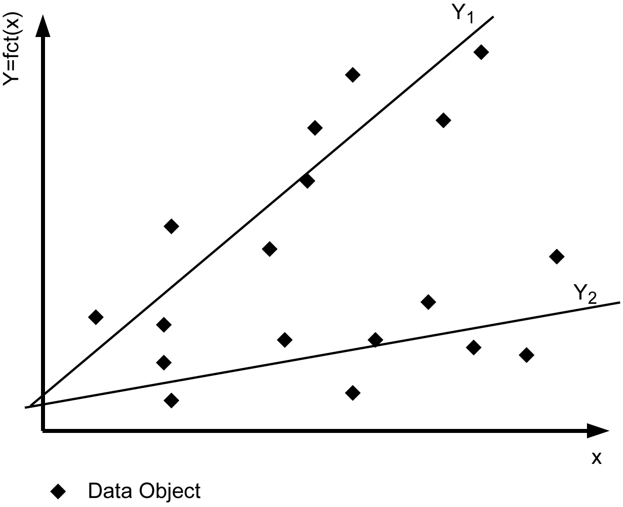

Fig. 1. Switching Regression Analysis

The challenge is to determine which of the two regression functions should represent a certain observation y_i:

$$\begin{gathered} y_i = \widehat{Y_1}(x_i) = a_{10} + a_{11}x_i + \mu_{1i} \quad \text{or} \quad y_i = \widehat{Y_2}(x_i) = a_{20} + a_{21}x_i + \mu_{2i} \\ \text{with} \quad \mu_{1i} \quad \text{and} \quad \mu_{2i} \quad \text{error terms.} \end{gathered} \tag{2}$$

To solve this problem in switching regression analysis - the estimation of the parameters $\boldsymbol{a}$ - one can apply Goldfeld and Quandt's D-method [17].

Fuzzy Switching Regression Models. Besides classic switching regression models Hathaway and Bezdek [18] suggested a fuzzy switching regression model which is closely related to Bezdek's fuzzy k-means [6]. Jajuga [19] also proposed a linear switching regression model that consists of a two step process: (1) the data are clustered with the fuzzy k-means, (2) the obtained membership degrees are used as weights in weighted regression analysis.

2.2 Rough Clustering Algorithms

Lingras' Rough k-Means. Lingras et al. rough clustering algorithm belongs to the branch of rough set theory with a reduced set of properties [20]:

1. A data object belongs to no or one lower approximation.
2. If a data object is no member of any lower approximation it is member of two or more upper approximations.
3. A lower approximation is a subset of its underlying upper approximation.

The part of an upper approximation that is not covered by a lower approximation is called boundary area. The means are computed as weighted sums of the data objects $\boldsymbol{X_n}(n = 1, ..., N)$:

$$\boldsymbol{m_k} = \begin{cases} w_L \sum\limits_{\boldsymbol{X_n} \in \underline{C_k}} \frac{\boldsymbol{X_n}}{|\underline{C_k}|} + w_B \sum\limits_{\boldsymbol{X_n} \in C_k^B} \frac{\boldsymbol{X_n}}{|C_k^B|} & \text{for } C_k^B \neq \emptyset \\ w_L \sum\limits_{\boldsymbol{X_n} \in \underline{C_k}} \frac{\boldsymbol{X_n}}{|\underline{C_k}|} & \text{otherwise} \end{cases} \tag{3}$$

where $|\underline{C_k}|$ is the number of objects in lower approximation and $|C_k^B| = |\overline{C_k} - \underline{C_k}|$ ($\overline{C_k}$ the upper approximation) in the boundary area of cluster k ($k = 1, ..., K$). Then rough cluster algorithm goes as follows:

1. Define the initial parameters: the weights w_L and w_B, the number of clusters K and a threshold ϵ.
2. Randomly assign the data objects to one lower approximation (and per definitionem to the corresponding upper approximation).
3. Calculate the means according to Eq (3).
4. For each data object, determine its closest mean. If other means are not reasonably farer away as the closest mean (defined by ϵ) assign the data object to the upper approximations of these close clusters. Otherwise assign the data object to the lower and the corresponding upper approximation of the cluster of its closest mean (see Figure 2).
5. Check convergence. If converged: STOP otherwise continue with STEP 3.

Extensions and Variations of the Rough k-Means. Lingras rough k-means was refined and extended by an evolutionary component. Peters [12,3] presented a refined version of the rough k-means which improves its performance in the presence of outliers, its compliance to the classic k-means, its numerical stability besides others. To initialize the rough k-means one has to select the weights

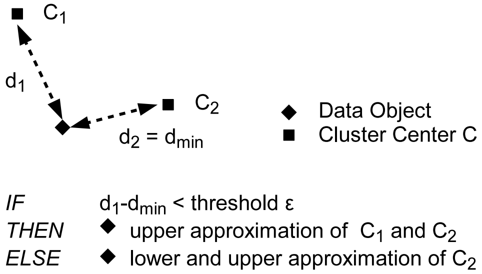
C1
d1
C2
d2 = dmin
◆ Data Object
■ Cluster Center C
IF d1-dmin < threshold ε
THEN ◆ upper approximation of C1 and C2
ELSE ◆ lower and upper approximation of C2

Fig. 2. Lingras' Rough k-Means

of the lower approximation and the boundary area as well as the number of clusters. Mitra [11] argued that a good initial setting of these parameters is one of the main challenges in rough set clustering. Therefore she suggested an evolutionary version of Lingras rough k-means which automates the selection of the initial parameters. And, recently Mitra et al. [10] introduced a collaborative rough-fuzzy k-means.

3 Rough Switching Regression Models

The new rough switching regression models utilize the concepts of rough clustering as suggested by Lingras [8] and Peters [3]. First let us define some terms and abbreviations:

- Data set: $\boldsymbol{S_n} = (y_n, \boldsymbol{x_n}) = (y_n, x_{0n}, ..., x_{Mn})$ for the nth observation and $\boldsymbol{S} = (\boldsymbol{S_1}, ..., \boldsymbol{S_N})^T$ with $n = 1, ..., N$. The variable y_n is endogenous while $\boldsymbol{x_n} = (x_{0n}, ..., x_{Mn})$ with $m = 0, ..., M$ (features) and $x_{0n} := 1$ represent the exogenous variables.
- Y_k the kth function: $\hat{y}_{kn} = Y_k(\boldsymbol{x_n}) = \sum_{m=0}^{M} a_{km} x_{mn}$ for $k = 1, ..., K$.
- Approximations: $\underline{Y_k}$ is the lower approximation corresponding to the regression function Y_k, $\overline{Y_k}$ the upper approximation and $Y_k^B = \overline{Y_k} - \underline{Y_k}$ the boundary area. This implies $\underline{Y_k} \subseteq \overline{Y_k}$.
- The distance in y between the data object $\boldsymbol{S_n}$ and the regression function Y_k is given by $d(\boldsymbol{S_n}, Y_k) = |y_n - \hat{y}_{kn}|$.

3.1 A First Rough Switching Regression Algorithm Based on Lingras' k-Means

First we present a rough switching regression model based on Lingras' k-means.

- **Step 0: Initialization**
  (i) Determine the number K of regression functions.
  (ii) Define the weights for the lower approximations and the boundary areas: w_L and w_B with $w_L + w_B = 1$.

(iii) Randomly assign each data object $\boldsymbol{S_n}$ to one lower approximation $\underline{Y_k}$ of the corresponding regression function Y_k.

- **Step 1: Calculation of the New Regression Coefficients**
  The new regression coefficients a_{km} are calculated using weighted regression analysis with weights defined as follows:

$$w_{kn} = \begin{cases} w_B & \text{for } \boldsymbol{S_n} \in Y_k^B \\ w_L & \text{for } \boldsymbol{S_n} \in \underline{Y_k} \\ 0 & \text{else} \end{cases} \tag{4}$$

- **Step 2: Assignment of the Data Objects to the Approximations**
  (i) For an object $\boldsymbol{S_n}$ determine its closest regression function Y_h (Figure 3):

$$y_{hn}^{min} = d(\boldsymbol{S_n}, Y_h) = \min_k d(\boldsymbol{S_n}, Y_k). \tag{5}$$

  Assign $\boldsymbol{S_n}$ to the upper approximation of the function Y_h: $\boldsymbol{S_n} \in \overline{Y_h}$.
  (ii) Determine the regression functions Y_t that are also close to $\boldsymbol{S_n}$. They are not farther away from $\boldsymbol{S_n}$ than $d(\boldsymbol{S_n}, Y_h) + \epsilon$ with ϵ a given threshold:

$$T = \{t : d(\boldsymbol{S_n}, Y_k) - d(\boldsymbol{S_n}, Y_h) \leq \epsilon \wedge h \neq k\}. \tag{6}$$

  - **If** $T \neq \emptyset$ ($\boldsymbol{S_n}$ is also close to at least one other regression function Y_t besides $\underline{Y_h)}$
    Then $\boldsymbol{S_n} \in \overline{Y_t}, \quad \forall t \in T$.
  - **Else** $\boldsymbol{S_n} \in \underline{Y_h}$.
- **Step 3: Checking the Convergence**
  The algorithms has converged when the assignments of all data objects to the approximations remain unchanged in the latest iteration i in comparison to iteration $i - 1$.
  - **If** the algorithm has not converged **Then** continue with Step 1.
  - **Else** STOP.

However, the algorithm has similar weaknesses as Lingras' k-means (see Peters [3] for a detailed discussion). E.g., please note that the algorithm does not enforce that each regression function has two or more data objects in its lower approximation.

3.2 A Rough Switching Regression Algorithm Based on Peters Rough k-Means

- **Step 0: Initialization**
  (i) Determine the number K of the regression functions which is limited by: $2 \leq K \leq \frac{N}{2}$ since each regression function should be defined by at least two data points.
  (ii) Randomly assign each data object $\boldsymbol{S_n}$ to one and only one lower approximation $\underline{Y_k}$ of the corresponding regression function Y_k so that each regression function has at least two data objects in its lower approximation.

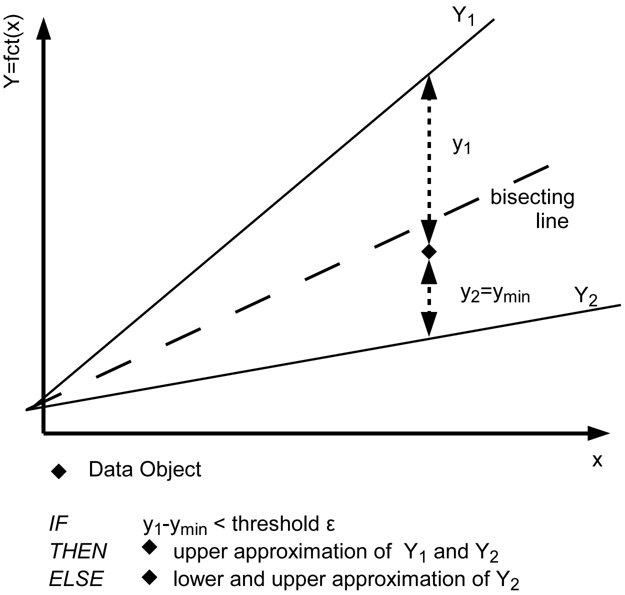
Y=fct(x)
Y1
y1
bisecting line
y2=ymin
Y2
x
◆ Data Object
IF y1-ymin < threshold ε
THEN ◆ upper approximation of Y1 and Y2
ELSE ◆ lower and upper approximation of Y2

Fig. 3. Assignment of the Objects to the Approximations

- **Step 1: Calculation of the New Regression Coefficients**
  The new regression coefficients a_{km} are calculated using weighted regression analysis (see Eq 4). The weights are defined as follows:
  (i) A data object $\boldsymbol{S_n}$ in lower approximations of a regression functions k is weighted by 1: $w_L = 1$.
  (ii) A data object $\boldsymbol{S_n}$ that is member of b boundary areas is weighted by $w_B = \frac{1}{b}$.

  Alternatively the weights of the lower approximation w_L and the boundary area w_B can be determined by the user.
- **Step 2: Assignment of the Data Objects to the Approximations**
  (i) Assign the data object that best represents a regression function to its lower and upper approximation.
  (a) Find the minimal distance between all regression functions Y_k and all data objects $\boldsymbol{S_n}$ and assign this data object $\boldsymbol{S_l}$ to lower and upper approximation of the regression function Y_h:

$$d(\boldsymbol{S_l}, Y_h) = \min_{n,k} d(\boldsymbol{S_n}, Y_k) \Rightarrow \boldsymbol{S_l} \in \underline{Y_k} \wedge \boldsymbol{S_l} \in \overline{Y_k}. \tag{7}$$

  (b) Exclude $\boldsymbol{S_l}$. If this is the second data object that has been assigned to the regression function Y_h exclude Y_h also. If regression functions are left - so far, in Step (a) no data object has been assigned to them - go back to Step (a). Otherwise continue with Step (ii).

(ii) For each remaining data points $\boldsymbol{S'_{n'}}$ ($n' = 1, ..., N'$, with $N' = N - 2K$) determine its closest regression function Y_h:

$$y_{hn'}^{min} = d(\boldsymbol{S'_{n'}}, Y_h) = \min_k d(\boldsymbol{S'_{n'}}, Y_k). \tag{8}$$

Assign $\boldsymbol{S'_{n'}}$ to the upper approximation of the function h: $\boldsymbol{S'_{n'}} \in \overline{Y_h}$.

(iii) Determine the regression functions Y_t that are also close to $\boldsymbol{S'_{n'}}$. Take the relative distance as defined above where ζ is a given relative threshold:

$$T' = \left\{ t : \frac{d(\boldsymbol{S'_{n'}}, Y_k)}{d(\boldsymbol{S'_{n'}}, Y_h)} \leq \zeta \wedge h \neq k \right\}. \tag{9}$$

- **If** $T' \neq \emptyset$ ($\boldsymbol{S'_{n'}}$ is also close to at least one other regression function Y_t besides Y_h)
  Then $\boldsymbol{S'_{n'}} \in \overline{Y_t}, \forall t \in T'$.
- **Else** $\boldsymbol{S'_{n'}} \in \underline{Y_h}$.

– **Step 3: Checking the Convergence**
The algorithms has converged when the assignments of all data objects to the approximations remain unchanged in the latest iteration i in comparison to iteration $i - 1$.
  - **If** the algorithm has not converged **Then** continue with Step 1.
  - **Else** STOP.

4 Discussion and Conclusion

In the paper we proposed rough switching regression models which are based on rough clustering. While classic switching regression models have been extensively applied in economics (e.g. [21,22]) applications to bioinformatics can hardly be found. However Qin et al. [23] suggested the related CORM method (Clustering of Regression Models method) and applied it to gene expression data.

Therefore future work can go in different directions. First, the rough switching regression model should be applied to real life data and compared to classic models, especially in the field of economics. Second, the potential of switching regression (classic, fuzzy, rough) for bioinformatics could be further evaluated.

References

1. Pawlak, Z.: Rough sets. International Journal of Information and Computer Sciences **11** (1982) 145–172
2. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. Technical Report 2002-002, Department of Mathematics and Computer Science, St. Mary's University, Halifax, Canada (2002)
3. Peters, G.: Some refinements of rough k-means clustering. Pattern Recognition **39**(8) (2006) 1481–1491
4. Hartigan, J.: Clustering Algorithms. John Wiley & Sons, Inc., New York, New York (1975)

5. Mirkin, B.: Mathematical Classification and Clustering. Kluwer Academic Publishers, Boston (1996)
6. Bezdek, J.: Pattern Recognition with Fuzzy Objective Algorithms. Plenum Press, New York (1981)
7. Krishnapuram R., K.J.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems **1** (1993) 98–110
8. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems **23** (2004) 5–16
9. Lingras, P., Yan, R., West, C.: Comparison of conventional and rough k-means clustering. In: International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Volume 2639 of LNAI., Berlin, Springer (2003) 130–137
10. Mitra, S., Banka, H., Pedrycz, W.: Rough-fuzzy collaboration clustering. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics **36**(4) (2006) 795–805
11. Mitra, S.: An evolutionary rough partitive clustering. Pattern Recognition Letters **25** (2004) 1439–1449.
12. Peters, G.: Outliers in rough k-means clustering. In: Proceed. First International Conference on Pattern Recognition and Machine Intelligence. Volume 3776 of LNCS., Kolkata, Springer Verlag (2005) 702–707
13. Voges, K., Pope, N., Brown, M.: Cluster Analysis of Marketing Data Examining On-Line Shopping Orientation: A Comparision of k-Means and Rough Clustering Approaches. In: Heuristics and Optimization for Knowledge Discovery. Idea Group Publishing, Hershey PA (2002) 207–224
14. Voges, K., Pope, N., Brown, M.: A rough cluster analysis of shopping orientation data. In: Proceedings Australian and New Zealand Marketing Academy Conference, Adelaide (2003) 1625–1631
15. Page, E.S.: A test for a change in a parameter occurring at an unknown point. Biometrika **42** (1955) 523–527
16. Quandt, R.: The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association **53** (1958) 873–880
17. Goldfeld, S., Quandt, R.: Nonlinear Methods in Econometrics. North-Holland, Amsterdam (1972)
18. Hathaway, R., Bezdek, J.: Switching regression models and fuzzy clustering. IEEE Transactions on Fuzzy Systems **1**(3) (1993) 195–204
19. Jajuga, K.: Linear fuzzy regression. Fuzzy Sets and Systems **20** (1986) 343–353
20. Yao, Y., Li, X., Lin, T., Liu, Q.: Representation and classification of rough set models. In: Proceedings Third International Workshop on Rough Sets and Soft Computing, San Jose, CA (1994) 630–637
21. Fuglie, K., Bosch, D.: Economic and environmental implications of soil nitrogen testing: A switching-regression analysis. American Journal of Agricultural Economics **77**(4) (1995) 891–90
22. Ohtani, K., Kakimoto, S., Abe, K.: A gradual switching regression model with a flexible transition path. Economics Letters **32**(1) (1990) 43–48
23. Qin, L., Self, S.: The clustering of regression model method with applications in gene expression data. Biometrics **62**(2) (2006) 526–533