Asian Journal of Information Science and Technology (AJIST)
Vol1.No.2 2013 pp 26-30.
available at: www.goniv.com
Paper Received :15-09-2013
Paper Published:28-12-2013
Paper Reviewed by: 1. John Arhter 2. Hendry Goyal
Editor : Prof. P.Muthukumar

Outlier Detection using Granular Box Regression Methods

M.Pandiaraj
Asst. Professor
Department of Computer Science & Applications
Sri Vidya Mandir Arts & Science College
Katteri
pandiya13@gmail.com

C.Prabu
Asst. Professor
Department of Computer Science & Applications
Sri Vidya Mandir Arts & Science College
Katteri
chinnarajuprabu@gmail.com

Abstract: Granular computing (GrC) is an emerging computing paradigm of information processing. It concerns the processing of complex information entities called information granules, which arise in the process of data abstraction and derivation of knowledge from information. Granular computing is more a theoretical perspective, it encourages an approach to data that recognizes and exploits the knowledge present in data at various levels of resolution or scales. Granular computing provides a rich variety of algorithms including methods derived from interval mathematics, fuzzy and rough sets and others. Within this framework granular box regression was proposed recently. The core idea of granular box regression is to determine a fuzzy graph by embedding a given dataset into a predefined number of "boxes". Granular box regression utilizes intervals a challenge is the detection of outliers. In this paper, we propose borderline method and residual method to detect outliers in granular box regression. We also apply these methods to artificial as well as to real data of motor insurance.

Keywords: Granular computing, Granular box regression, Outliers.

I. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions**.**

In applications like fraud detection, medical analysis, intrusion detection, etc, outlier detection is an important problem, since the rare events or exceptional cases are more interesting and useful than the common cases. In the context of outlier detection, what makes the problem more difficult is that not everyone has the same idea of what constitutes an outlier. Intuitively, an object is an outlieror abnormal if it is in some waysignificantly different from itsneighbors. Different answers to what constitutes a neighborhood, how to determine difference and whether it is significant would lead to various sets of objects defined as outliers.

The past decades have been characterized by a dramatically increasing complexity in virtually any real-life domain. Some important drivers behind this increasing complexity are without discussing any dependencies between them or going into any detail.For example the rapid progress in information technology and globalization.

To address these challenges two principle strategies are possible that are completely opposite to each other: first, the advancement of finer-grained algorithms and second, the development of simplified, coarse-grained algorithms.

Granular computing follows the second strategy: As a multi-disciplinary field of research it aims to develop coarse-grained algorithms [2,5]. It has gained increasing attention over the past decade [3,7]. Although a well-accepted definition of granular computing is still missing,

here three issues as a motivation towards granular box regression is discussed briefly.

First, Yao [1] proposes one direction of granular computing as human-inspired problem solving. Humans have the ability to choose an appropriate level of granularity for a problem by ignoring distracting details. For example, recent studies confirm that too much information may lead to an ‘‘information overload’’ that may reduce the quality of human decisions and actions [8]. Hence, granular computing may help to bridge possible gaps between complex machine-like and human-like information processing and representation. This goes along the lines with Zadeh’s proposal of human-level machine intelligence.

Second, granular computing is motivated by the insight that many real life situations are (still) much too complex to be addressed by the most sophisticated algorithms available today. Such algorithms may even show a precision that is, by no means, justified by the real data. Granular approaches take this into account by limiting themselves to coarser approximations rather than heading for (possibly unreachable) exact solutions.

II. Related Work

Data mining, also called Knowledge Discovery is the process of automatically searching large volumes of data for patterns using association rules [20].An outlier is an observation that deviates so much from other observations so as to arouse suspicion that it was generated from different mechanism.Outlier mining is the task to find small groups of data objects that are exceptional when compared with rest large amount of data [18]. To study differences between artificial and real deception, an experiment was performed using deception level and data generation method as factors and directed distance and outlier score as outcome variables [9].Anomaly detection approaches can make use of supervised or unsupervised methods to detect abnormal behaviors in patterns [19]. Certain types of attacks are more harmful than others and their detection is critical to protection of the system [11].

III. Granular Box Regression

Granular box regression was proposed by Peters [10] recently. Its fundamental idea is to approximate a set of objects $o_1 = (y, x_1,x_i. . .x_I)_1$ with $l = 1, . . . , L$ by a predefined number K of ‘‘boxes’’ (hyper-dimensional interval numbers). Granular box regression constitutes an asymmetry problem with I independent dimensions and one dependent dimension: $(x_1. . . x_i, . . . , x_I)\rightarrow y$. This has motivated to label it with the term ‘‘regression’’. Obviously (I + 1)-dimensional boxes to surround the given data set is needed. The goal of granular box regression is to enclose this data set by the boxes as closely as possible.

To map the asymmetry between the independent variables $(x_1. . . x_i, . . . , x_I)$ and the one dependent variable y onto the boxes Peters suggested the following strategy:

Control the one dependent dimension y so that the boxes do not overlap here. Do not control the I independent dimensions $(x_1. . . x_i, . . . , x_I)$ regarding overlapping boxes, i.e. the boxes generally overlap in these dimensions.

The rationale behind this is that one can secure non-overlapping boxes in one dimension only. Hence, an asymmetric box configuration is available with I generally overlapping dimensions and one non-overlapping dimension which is similar to the asymmetry $(x_1. . .x_i . . . , x_I)\rightarrow y$ in the given data set.

A small data set and its surrounding boxes to illustrate this: the boxes overlap in the x-dimension while they do not overlap in the (controlled) y-dimension.

Granular box regression has three major challenges:

1. How shall the number of boxes K be defined
2. How shall the goal, to surround the objects by the boxes as closely as possible, be concretized
3. How can this goal be obtained.

These challenges are addressed briefly in the following paragraphs:

Like in clustering the initial setting of the number of boxes (respectively clusters) is of significant importance. In clustering this challenge has been addressed extensively. In contrast to clustering, the selection of the optimal number of boxes K is still not addressed in detail in granular box regression. However, it might be possible to get some inspiration from clustering to tackle this challenge.

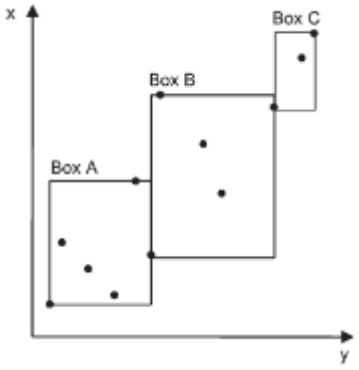

Figure.1: A Possible Box Configuration for a Small Data Set

The goal is well separated in the algorithm. Therefore, it can be easily exchanged. For example, the minimization of the diameters of the boxes or the maximization of the coefficient of determination R is possible goals. In this

paper, the minimization of the sum of the volumes of the boxes is used as objective criterion.

Peters suggested swapping the border objects in the non-overlapping y-dimension to optimize the box configuration. Two border objects are present in this dimension, border object AB separating the boxes A and B and the object BC defining the border between the boxes B and C. The swapping options s1 and s2 for the border object AB and s3 and s4 for border object BC are also depicted. In granular box regression the optimum box configuration is determined by swapping these border objects with their nearest neighbors in the y-dimension until a (local) minimum is obtained. Algorithmically granular box regression is inspired by k-medoid clustering, where medoids and non-medoids are swapped to obtain the optimum cluster configuration. So, granular box regression could be labeled as asymmetric clustering.

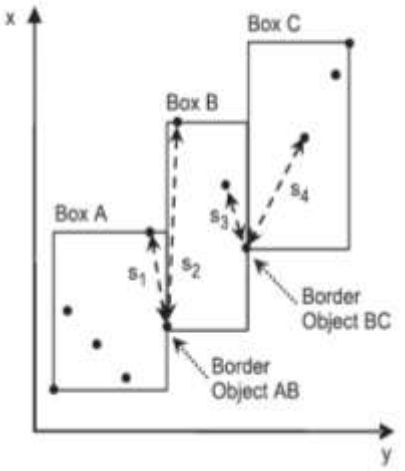

Figure. 2: Swapping Options in Granular Box Regression

Peters [10] discussed two possible interpretations of granular box regression, in the context of regression analysis and for rule induction:

1. Regression analysis: The boxes can be understood as a fuzzy graph, respectively as a generalization of a function. In this context granular box regression enriches the literature on granular regression. In contrast to the already existing approaches that are primarily derived from fuzzy concepts granular box regression constitutes a non-functional relationship via boxes between the independent variables and the one dependent variable.

2. Rule induction: The development of fuzzy rule based systems is an important area in soft computing. Granular box regression contributes to this area since the obtained boxes can be used for rule induction. They provide a basis for the design of linguistic variables as part of IF-THEN rules.

IV. Outlier Detection

The interest in outliers is twofold. On the one hand outliers are considered as perturbing objects which should be deleted (e.g. disturbing noise). On the other hand outliers may be the objects of particular interest, e.g. in crime detection (credit card fraud [4] or network intrusion [12], etc.).

Outlier detection has a long history. Several ideas have been proposed how to discover outliers for different methods and applications [4,6]. Taxonomies of outlier detection methods have also been proposed [6]. Common categorizations for outlier detection are univariate vs. multivariate methods, or parametric (statistical) vs. non-parametric (distance-based) approaches [13].Probably the simplest one is the ‘‘eyeball method’’ applicable for up to 3-dimensional data: an expert tries to identify outliers by investigating a graphical representation of the data. Obviously, higher dimensional data required methods that are independent of our ‘‘eyeballs’’.

A common parametric approach utilizes the standard deviation. Assuming the data are distributed as a Gaussian function, an object may be defined as an outlier when its distance from the mean is greater than 2 or 3 standard deviations [16]. Examples for non-parametric approaches are (i) to define objects that are very far away from the mean of the dependent data as outliers [14] or (ii) in cluster analysis objects of a weakly occupied cluster.This project utilizes some of the well-established concepts of outlier detection and adapted them to outlier detection in granular box regression.

Basic Strategies for Detecting Outlier
Phase 1 – Perform Granular Box Regression

Granular box regression was proposed by Peters[10]recently. Its fundamental idea is to approximate a set of objects $o_1 = (y, x_1 ... x_i ... x_I)l$ with $l = 1...L$ by a predefined number K of ‘‘boxes’’ (hyper-dimensional interval numbers). Granular box regression constitutes an asymmetry problem with I independent dimensions and one dependent dimension: $(x_1 ... x_i ... x_I)y$. This has motivated to label it with the term ‘‘regression’’. Obviously we need (I+1)-dimensional boxes to surround the given data set. The goal of granular box regression is to enclose this data set by the boxes as closely as possible.

Phase 2 – Identify Potential Outliers

In this phase we identify potential outliers. The set $S_k=\{s\}$ contains these objects for box k. For example all objects that lie on the edges of a box should be considered as potential outliers since their elimination would decrease the volume of this box.

Phase 3 – Identify Actual Outliers

In this phase the set of potential outliers is examined for false outliers – we are trying to avoid swamping[17]. We define objects as actual outliers when the elimination of one potential outlier or a combination of potential outliers reduces the sum of the volumes of the boxes significantly. The set $T_k = \{t\}$ contains these actual objects for box k.In principle, granular box regression must be performed after the elimination of the potential outliers since the borders between the boxes possibly change. However, this approach would be rather computationally intensive.

Phase 4 – Update Results of Granular Box Regression

The outliers are eliminated and granular box regression is performed again to obtain the final result.

Application of the Algorithm for Motor Insurance Claims

After the satisfactory testing of the algorithm, it was used for the commercial purpose where the curves are unknown and the relationship is needed to be established. So, the curve fitting is one of the important factors in Insurance sector for the estimation of frequency and severity of the claims to occur. Motor insurance claims data is taken into consideration for this work. The factors which are considered here are:

1. Age of the Vehicle
2. Sex of the driver
3. Age of the Driver
4. Driving Experience
5. Place of Repair
6. Claims

Here, the first 5 factors mentioned above were inputs and claim amount in INR was the output**.**

First the data was collected to cover all ranges of motor vehicle users. This included all vehicle models, geographical regions, vehicle ages etc for example collecting the cases of claims for all ranges of ages starting with ages 18-25, 45-60 and 60 and above. The same was considered for every field. After collecting data rigorously, the next step was to clean the data to avoid any ambiguous data. So, various tests were applied to the excel sheet of data like the age of the driver should not be less than 18years otherwise the claimswill not be paid.

V. Algorithm

Step 1: Initialization.
Let n = 0, a counter for the iterations.
Define the number of boxes 1 < K < L 1(k =1... K), with L the number of objects.
Select the K1 border objects.
Calculate the initial total volumes of all boxes V.
Step 2: Swap Border Objects.
Let n = n +1.
Replace each current border object with its closest right and left neighbor in the y-dimension (a total of m =2(K1) swaps are possible).
Calculate the m sums of the volumes of all boxes V m; n ¼ P K k¼1 V k; m; n
Define $V_n$min ¼min 8 m $V_{m;n}$ as the minimal volume of iteration n.
Step 3: Check for Convergence.
IF V_n min <V_n min
THEN [Replace the corresponding border object. Go to Step 2]
ELSE [Stop].

VI. Experimental Results

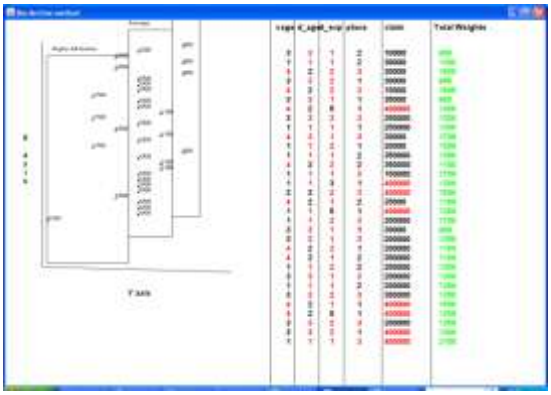

Fig. 3: Granular box regression for borderline method

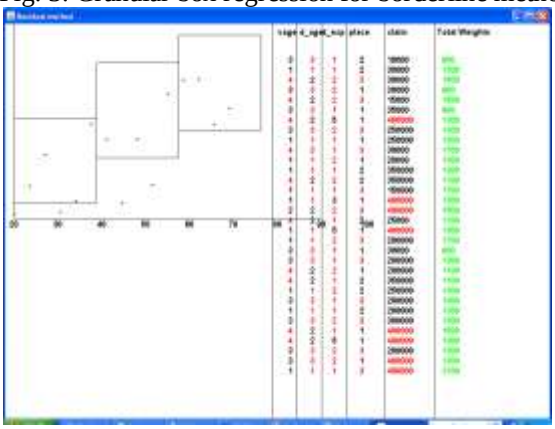

Fig. 4: Granular box regression for residual method

Fig.3 shows the highly risk factors with actual outlier values, potential outlier values and false outlier values for borderline method. Fig.4 shows the highly risk factors with actual outlier values, potential outlier values and false outlier values for residual method.

VII. Conclusion

The risk factors of the insurance data are found using Granular Box Regression. Borderline method and residualmethod are proposed for addressing outliers in granular box regression and applied them to real and artificial data. A main problem, generally immanent in outlier detection, setting the parameters and thereby defining when an object can still is considered as normal and when it should be treated as an outlier, remains a challenge in the presented methods. Setting these parameters needs to be addressed in a context-dependent manner. The borderline and the residual methods mix the concepts of classic regression analysis with granular box regression. The proposed method can become useful to actuaries, as it provides full credibility criteria for Granular box regression, at a time when these are becoming popular in the analysis of insurance and data.

References

[1] Y.Y. Yao, "Human-inspired granular computing", IGI Global, Hershey, PA, USA, 2010, pp. 1–15.

[2] Y.Y. Yao, "Perspectives of granular computing", in: Proceedings 2005 IEEE International Conference on Granular Computing (GrC 2005), vol. 1, pp.85–90.

[3] J.T. Yao, "A ten-year review of granular computing", in: Proceedings 2007 IEEE International Conference on Granular Computing (GrC 2007), pp. 734–739.

[4] D.M. Hawkins, "Identification of Outliers", Chapman and Hall, London, New York, 1980.

[5] L.A. Zadeh, "Toward human level machine intelligence – is it achievable? The need for a paradigm shift", IEEE Computational Intelligence Magazine3(2008) 11–22.

[6] V.J. Hodge, J. Austin, "A survey of outlier detection methodologies", Artificial Intelligence Review 22 (2004) 85–126.

[7] J.T. Yao, "Recent developments in granular computing: a bibliometrics study", in: Proceedings IEEE International Conference on Granular Computing,2008, pp. 74–79.

[8] J. Davis, "Aversion to loss and information overload: an experimental investigation", in: Proceedings International Conference on Information Systems(ICIS 2009), Phoenix, AZ, USA, 2009.

[9] Yanjuan yang, Michael V.Mannino, "An experimental comparison of real and artificial deception using a deception generation model",Information Sciences 212 (2012) 44–56.

[10] M.J. Gacto, R. Alcal, F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures", Information Sciences181 (2011) 4340–4360.

[11] Inhokang, MyongK.Jeong, Dongjoon Kong, "A differentiated one-class classification method with application to intrusion detection", Expert Systems with Applications 39 (2012) 3899–3905.

[12] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection", in: Proceedings of the Third SIAM International Conference on Data Mining, 2003.

[13] E.M. Knorr, R.T. Ng, "Finding intentional knowledge of distance-based outliers", 25th International Conference on Very Large DataBases (VLDB 1999).

[14] A. van Eye, C. Schuste, "Regression Analysis for Social Sciences", Academic Press, San Diego, 1998.

[15] J. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley, New York, USA, 1990.

[16] J. Cohen, P. Cohen, S. West, L. Aiken, "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences", Lawrence Erlbaum Associates,Mahwah, NJ, 2003.

[17] P.J. Rousseeuw, M. Hubert, "Robust statistics for outlier detection", WIREs Data Mining and Knowledge Discovery 1 (2011) 73–79.

[18] Fang Yu Ke, Fu Yan, Zhou Jun Lin, "Research of outlier mining based adaptive intrusion detection techniques", third international conference on knowledge discovery and data mining 2010.

[19] V.Arunkumar, Dr.A.Saradha, "An efficient data retrieval clustering based anomaly intrusion detection system in mining process with time prediction", International journal of communication and engineering, volume 04- no.4, Issue: 02 March2012.

[20] SherishJohri, "Novel method for intrusion detection using data mining", International journal of advanced research in computer science and software engineering, volume 2, issue 4, April 2012.