Information Sciences 537 (2020) 246–260

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Granular regression with a gradient descent method

Yumin Chen [a,*], Duoqian Miao [b]

[a] College of Computer & Information Engineering, Xiamen University of Technology, Xiamen 361024, China
[b] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 22 April 2019
Received in revised form 20 May 2020
Accepted 26 May 2020
Available online 4 June 2020

Keywords:
Granular computing
Regression
Information granules
Gradient descent
Granular regression

ABSTRACT

The regression is one of classical models in machine learning. Traditional regression algorithms involve operations of real values, which are difficult to handle the discrete or set data in information systems. Granules are structural objects on which agents perform complex computations. The structural objects are forms of sets that can measure the uncertainty of data. In order to deal with uncertain and vague data in the real world, we propose a set-based regression model: granular regression. Granules are constructed by introducing a distance metric on single-atom features. Meanwhile, we establish conditional granular vectors, weight granular vectors and decision granules. The operations among them induce a granular regression model. Furthermore, we propose a gradient descent method for the granular regression model, and the optimal solution of granular regression is achieved. We prove the convergence of granular regression and design a gradient descent algorithm. Finally, several UCI data sets are used to test and verify the granular regression model. We compare our proposed model with popular regression models from three aspects of convergence, fitting and prediction. The results show that the granular regression model is valid and effective.

1. Introduction

In 1997, American scientist Zadeh [44] originally put forward the concept of information granule. He believed that granulation, organization and causation are the fundamental stones of human cognition. Agents perform computations on complex objects including behavioral patterns, classifiers, clusters, sets of rules, aggregation operations, approximate reasoning schemes [31]. All such structural objects are called granules [31]. Polkowski presented adaptive calculus of granules in [27]. In 1999, granular computing was first presented by Lin [16] and successfully applied in data mining [17]. Skowron and Nguyen [2,20] proposed the granular computing from rough set theory [23]. Yao analyzed granular computing from the perspectives of philosophy, application and computation. He held that the granular computing is a method of structured thinking, problem-solving and information processing [41]. Canadian academician Pedrycz pointed out that the construction of information granules is the key to granular computing. He proposed an interval-based information granule [24] that is applied in schemes of knowledge management. He also constructed some information granules from the perspective of fuzzy sets and used them for clustering [25,48] and further designed a variety of granular classifiers [1,26,30], which achieved good results. Skowron [32] defined information granules in terms of semantics and grammar, and presented the structure and calculation of granules. Liu [18] proved a principle of granular resolving from rough logic. Yao proposed a neighborhood relation

* Corresponding author.
E-mail address: cym0620@163.com (Y. Chen).

https://doi.org/10.1016/j.ins.2020.05.101
0020-0255/

[42] and developed the neighborhood granule computing [43]. Hu analyzed the neighborhood reduction [10] and designed some neighborhood classifiers [11,47]. Wang discussed a big data processing method using granules [35,38]. Liang and Qian proposed some fusing and processing methods for multi-granularity data [15,28]. We analyzed structures of granules from the views of sets and formal concept analysis [12], further described a three-level granular structure in a neighborhood system, and analyzed some uncertainty and distance measures of granules [4]. Granules and granulation are the significant characteristics of human cognition, which play an important role in modeling complex data and have been widely used in many fields [6,8,29,36,37,39,49].

Models of machine learning are divided into unsupervised learning and supervised learning according to whether or not they have labels. Unsupervised learning mainly includes: clustering (such as k-means) [9], dimensionality reduction (such as PCA) [14] and so on [33]. Supervised learning is related to classification and regression. Classification mainly includes: linear classifiers [3], Support Vector Machine (SVM) [13], Naive Bayesian (NB) [40], K-Nearest Neighbor (KNN) [45], Decision Tree (DT) [7], Convolutional Neural Network (CNN) [21], integrated models and so on [22]. The models of regression are: linear regression [19], ridge regression [5], lasso regression [34] and elastic net regression [46]. These regression models have the advantages of simplicity and high computational efficiency. They are also the basis of building stones of deep learning.

However, most of these models are weak in dealing with uncertain and vague data that can be represented by sets. These models are difficult to tackle set-based data since their operations involve in real values. The structure of a granule is essentially a set, and the operations of granules must be a form of set operations. As methods and techniques of information granulation spring up, many methods of clustering and classification of granules are proposed under the characteristics of set or aggregation of granules. Since regression has an ability of handling continuous real numbers, it is difficult to operate granules by a regression process. From a new angle, we propose a data representation based on set theory and single feature granulation, and define the concepts of granule, granular vector and granular matrix. We put forward some related operations of granule, granular vector and granular matrix, which induce a regression model of granules. Furthermore, we proved the convergence of granular regression and propose a gradient descent method. Finally, we design a gradient descent algorithm of granular regression, and successfully achieve good results of fitting and predicting with the form of information granules. Several UCI data sets are used for experimental analysis, and the results show that the granular regression is valid and effective.

2. Granules and granular vectors

Information systems are widely used in the machine learning field. We can acquire a granule according to distances between samples on a single-atom feature in an information system. A sample forms different granules on different single-atom features, and these granules constitute a granular vector of the sample.

Definition 1. An information system is represented as $S = (X, C \cup Y)$ (decision table) [23], where $X = \{x_1, x_2, \ldots, x_n\}$ is a set of samples, $C = \{c_1, c_2, \ldots, c_n\}$ is a set of features, and Y is a decision attribute or a label.

Definition 2. Let $S = (X, C \cup Y)$ be an information system. For any two samples $x_1, x_2 \in X$ and each single-atom feature $c \in C$, a distance between the samples x_1, x_2 on the feature c is defined as:

$$s_c(x_1, x_2) = |v(x_1, c) - v(x_2, c)|,$$

where $v(x, c) \in [0, 1]$ is a normalized value of sample x on c.

This is a Manhattan distance that measures the similarity between samples. If $0 < s_c(x_1, x_2) < 1$, then x_1 and x_2 satisfy a similar relation. While $s_c(x_1, x_2) = 0$ or $s_c(x_1, x_2) = 1$, the x_1 and x_2 are equivalent or distinct. Therefore, the distance metric is not only suitable to numerical data but also to symbolic data.

Definition 3. [44] Let $S = (X, C \cup Y)$ be an information system. For any sample $x \in X$ and each single-atom feature $c \in C$, a granule of x on c is defined as:

$$g_c(x) = \sum_{j=1}^{n} g_c(x)_j = \sum_{j=1}^{n} \frac{r_j}{x} = \frac{r_1}{x} + \frac{r_2}{x} + \ldots + \frac{r_n}{x} = (r_1, r_2, \ldots, r_n),$$

where $r_j = s_c(x, x_j)$ is a distance between x and x_j, $|X| = n$, '+' represents a union of elements, and '−' is a splitter.

A granule is also named as an atom granule, while $g_c(x)_j$ is the j_{th} nucleus of a granule. A granule is a set of ordered granular nuclei induced by distances between samples. If $\forall r_j = 1$, then $one = (1, 1, \ldots, 1)$ is called as one-granule; if $\forall r_j = 0$, then $zero = (0, 0, \ldots, 0)$ is called as zero-granule. For the decision attribute $y \in Y$ and any sample $x \in X$, it is granulated as a decision granule, represented as $g_y(x) = \sum_{j=1}^{n} \frac{r_j}{x}$, where $r_j = s_y(x, x_j)$.

Definition 4. Let $S = (X, C \cup Y)$ be an information system. For any sample $x \in X$ and a feature $c \in C$, the size and the norms of a granule $g_c(x)$ are defined as follows:

(1) Size of a granule: $Size(g_c(x)) = |g_c(x)| = \sum_{j=1}^{n} r_j$;
(2) Norm-1 of a granule: $Norm\text{-}1(g_c(x)) = ||g_c(x)||_1 = ||g_c(x)|| = \sum_{j=1}^{n} |r_j|$;
(3) Norm-2 of a granule: $Norm\text{-}2(g_c(x)) = ||g_c(x)||_2 = \sqrt{\sum_{j=1}^{n} r_j^2}$.

The size of a granule is an intrinsic characteristic. In order to avoid over-fitting problem, the norms of granules are used for regularization in a machine learning process.

Property 1. The size and norm-1 of a granule are satisfied: $0 \leqslant |g_a(x)| = ||g_a(x)||_1 \leqslant n$.

Proof. From the definition of a granule, we know $r_j = s_a(x, x_j) = |v(x, a) - v(x_j, a)|$. Since $v(x, a) \in [0, 1]$, then $r_j \in [0, 1]$ is achieved. According to definitions of the size and norm-1 of the granule, we achieve $|g_a(x)| \geqslant 0$ and $|g_a(x)| = \sum_{j=1}^{n} r_j = \sum_{j=1}^{n} |r_j| = ||g_a(x)||_1 \leqslant n$. So, the property is proved.

Property 2. The norm-1 and norm-2 of a granule are satisfied: $0 \leqslant ||g_a(x)||_2 \leqslant ||g_a(x)||_1 \leqslant n$.

Proof. From the definition of norm-1 of a granule, we know $||g_a(x)||_1^2 = \left(\sum_{j=1}^{n} |r_j|\right)^2 = (r_1 + r_2 + \ldots + r_n)^2$. According to the definition of norm-2 of the granule, we obtain $||g_a(x)||_2^2 = \left(\sqrt{\sum_{j=1}^{n} r_j^2}\right)^2 = r_1^2 + r_2^2 + \ldots + r_n^2$. Since $r_j \geqslant 0$, then $(r_1 + r_2 + \ldots + r_n)^2 \geqslant r_1^2 + r_2^2 + \ldots + r_n^2$ is obtained. Therefore, $0 \leqslant ||g_a(x)||_2 \leqslant ||g_a(x)||_1 \leqslant n$ is found.

Definition 5. Let $S = (X, C \cup Y)$ be an information system. For any sample $x \in X$ and any feature subset $P \subseteq C$, suppose $P = \{c_1, c_2, \ldots, c_m\}$, then the granular vector of x on P is defined as follows:

$$F_P(x) = (g_{c_1}(x), g_{c_2}(x), \ldots, g_{c_m}(x))^T = \begin{bmatrix} g_{c_1}(x) \\ g_{c_2}(x) \\ \ldots \\ g_{c_m}(x) \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} g_{c_1}(x)_j \\ g_{c_2}(x)_j \\ \ldots \\ g_{c_m}(x)_j \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} r_{1j} \\ r_{2j} \\ \ldots \\ r_{mj} \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} r_1 \\ r_2 \\ \ldots \\ r_m \end{bmatrix}_j,$$

where T is a transpose, 'Σ' represents a union of granular nucleus vectors, and $r_{mj} = g_{c_m}(x)_j$ indicates the j_{th} granular nucleus of $g_{c_m}(x)$.

The granular vector is expressed by a list of granules. The elements of a granular vector are granules, and a granule is composed of granular nuclei. Therefore, the granular vector can be formed by the union of granular nucleus vectors.

Definition 6. Let $S = (X, C \cup Y)$ be an information system. For any sample $x \in X$ and any feature subset $P \subseteq C$, suppose $P = \{c_1, c_2, \ldots, c_m\}$, then the size of a granular vector $F_P(x)$ is defined as:

$$|F_P(x)| = \sum_{i=1}^{m} |g_{c_i}(x)| = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij},$$

where $r_{ij} = s_{c_i}(x, x_j)$, $|P| = m$ and $|X| = n$.

The size of a granular vector is also called the modulus of the granular vector. It is easy to know that the size of a granular vector satisfies: $0 \leqslant |F_P(x)| \leqslant m * n$.

Theorem 1. *Let $S = (X, C \cup Y)$ be an information system. For any sample $x \in X$ and any feature subsets $P, Q \subseteq C$, $F_P(x), F_Q(x)$ are two granular vectors on P, Q. If $P \subseteq Q$, then $|F_P(x)| \leqslant |F_Q(x)|$.*

Proof. For $P \subseteq Q$, suppose $P = \{a_1, a_2, \ldots, a_m\}$, then $Q = P \cup B = \{a_1, a_2, \ldots, a_m\} \cup B$, where B may be an empty set, and suppose $B = \{b_1, b_2, \ldots, b_s\}$. From the Definition 6, we know $|F_P(x)| = \sum_{i=1}^{m} |g_{a_i}(x)| = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}$, then $|F_Q(x)| = \sum_{i=1}^{m} |g_{a_i}(x)| + \sum_{k=1}^{s} |g_{b_k}(x)| = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} + \sum_{k=1}^{s} \sum_{j=1}^{n} r_{kj} = |F_P(x)| + \sum_{k=1}^{s} \sum_{j=1}^{n} r_{kj}$. If B is an empty set, then s is 0, so $\sum_{k=1}^{s} \sum_{j=1}^{n} r_{kj} = 0$; otherwise, $\sum_{k=1}^{s} \sum_{j=1}^{n} r_{kj} > 0$. Therefore, $|F_P(x)| \leqslant |F_Q(x)|$ is obtained.

Example 1. An information system $S = (X, C \cup Y)$ is shown in Table 1. Suppose $X = \{x_1, x_2, x_3, x_4\}$ is a sample set, $C = \{a, b, c\}$ is a feature set, and $Y = \{0, 0, 1, 1\}$ is a decision set.

For the sample set $X = \{x_1, x_2, x_3, x_4\}$, if the granulation is performed on the single-atom feature a, some granules are constructed as follows:

Table 1
An information system.

X	a	b	c	$\Rightarrow Y$
x_1	0.2	0.1	0.2	0
x_2	0.1	0.6	0.1	0
x_3	0.4	0.2	0.5	1
x_4	0.7	0.1	0.5	1

$g_1 = g_a(x_1) = \frac{0}{x_1} + \frac{0.1}{x_1} + \frac{0.2}{x_1} + \frac{0.5}{x_1} = (0, 0.1, 0.2, 0.5)$,
$g_2 = g_a(x_2) = \frac{0.1}{x_2} + \frac{0}{x_2} + \frac{0.3}{x_2} + \frac{0.6}{x_2} = (0.1, 0, 0.3, 0.6)$,
$g_3 = g_a(x_3) = \frac{0.2}{x_3} + \frac{0.3}{x_3} + \frac{0}{x_3} + \frac{0.3}{x_3} = (0.2, 0.3, 0, 0.3)$, and
$g_4 = g_a(x_4) = \frac{0.5}{x_4} + \frac{0.6}{x_4} + \frac{0.3}{x_4} + \frac{0}{x_4} = (0.5, 0.6, 0.3, 0)$.

If the granulation is performed on feature b, the granules are:

$g_5 = g_b(x_1) = \frac{0}{x_1} + \frac{0.5}{x_1} + \frac{0.1}{x_1} + \frac{0}{x_1} = (0, 0.5, 0.1, 0)$,
$g_6 = g_b(x_2) = \frac{0.5}{x_2} + \frac{0}{x_2} + \frac{0.4}{x_2} + \frac{0.5}{x_2} = (0.5, 0, 0.4, 0.5)$,
$g_7 = g_b(x_3) = \frac{0.1}{x_3} + \frac{0.4}{x_3} + \frac{0}{x_3} + \frac{0.1}{x_3} = (0.1, 0.4, 0, 0.1)$, and
$g_8 = g_b(x_4) = \frac{0}{x_4} + \frac{0.5}{x_4} + \frac{0.1}{x_4} + \frac{0}{x_4} = (0, 0.5, 0.1, 0)$.

If the granulation is performed on feature c, the granules are:

$g_9 = g_c(x_1) = \frac{0}{x_1} + \frac{0.1}{x_1} + \frac{0.3}{x_1} + \frac{0.3}{x_1} = (0, 0.1, 0.3, 0.3)$,
$g_{10} = g_c(x_2) = \frac{0.1}{x_2} + \frac{0}{x_2} + \frac{0.4}{x_2} + \frac{0.4}{x_2} = (0.1, 0, 0.4, 0.4)$,
$g_{11} = g_c(x_3) = \frac{0.3}{x_3} + \frac{0.4}{x_3} + \frac{0}{x_3} + \frac{0}{x_3} = (0.3, 0.4, 0, 0)$, and
$g_{12} = g_c(x_4) = \frac{0.3}{x_4} + \frac{0.4}{x_4} + \frac{0}{x_4} + \frac{0}{x_4} = (0.3, 0.4, 0, 0)$.

If the granulation is performed on label Y, the decision granules are:

$d_1 = g_Y(x_1) = \frac{0}{x_1} + \frac{0}{x_1} + \frac{1}{x_1} + \frac{1}{x_1} = (0, 0, 1, 1)$,
$d_2 = g_Y(x_2) = \frac{0}{x_2} + \frac{0}{x_2} + \frac{1}{x_2} + \frac{1}{x_2} = (0, 0, 1, 1)$,
$d_3 = g_Y(x_3) = \frac{1}{x_3} + \frac{1}{x_3} + \frac{0}{x_3} + \frac{0}{x_3} = (1, 1, 0, 0)$, and
$d_4 = g_Y(x_4) = \frac{1}{x_4} + \frac{1}{x_4} + \frac{0}{x_4} + \frac{0}{x_4} = (1, 1, 0, 0)$.

The size, norm-1 and norm-2 of granule g_1 are:

$Size(g_1) = |g_a(x_1)| = 0 + 0.1 + 0.2 + 0.5 = 0.8$;
$Norm\text{-}1(g_1) = ||g_a(x_1)||_1 = 0 + 0.1 + 0.2 + 0.5 = 0.8$; and
$Norm\text{-}2(g_1) = ||g_a(x_1)||_2 = \sqrt{0 + 0.1^2 + 0.2^2 + 0.5^2} = 0.5477$.

If $C = \{a, b, c\}$, then the granular vector of x_1 on C is:

$$F_C(x_1) = (g_1, g_5, g_9)^T = (g_a(x_1), g_b(x_1), g_c(x_1))^T = ((0, 0.1, 0.2, 0.5), (0, 0.5, 0.1, 0), (0, 0.1, 0.3, 0.3))^T$$
$$= \begin{bmatrix} 0 \\ 0.1 \\ 0.2 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.5 \\ 0.1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.1 \\ 0.3 \\ 0.3 \end{bmatrix}.$$

The size of the granular vector $F_C(x_1)$ is:

$$|F_C(x_1)| = (0 + 0.1 + 0.2 + 0.5) + (0 + 0.5 + 0.1 + 0) + (0 + 0.1 + 0.3 + 0.3) = 2.1.$$

3. Operations of granules and granular matrices

The granules are forms of sets with structural data. A granular vector is an ordered array composed by granules that are results of granulating on several single-atom features. Therefore, we can construct granular matrices based on these granular vectors. Furthermore, we induce some operations on these granules and granular matrices.

Definition 7. Let $S = (X, C \cup Y)$ be an information system. For any $x \in X$ and $\forall a, b \in C$, suppose $s = g_a(x) = \sum_{j=1}^{n} \frac{s_j}{x}$, $t = g_b(x) = \sum_{j=1}^{n} \frac{t_j}{x}$ are two granules of x on a and b, then operations of addition, subtraction, multiplication and division among the two granules are defined as follows:

$$s + t = g_a(x) + g_b(x) = \sum_{j=1}^{n} \frac{s_j + t_j}{x} = (s_1 + t_1, s_2 + t_2, \ldots, s_n + t_n),$$

$$s - t = g_a(x) - g_b(x) = \sum_{j=1}^{n} \frac{s_j - t_j}{x} = (s_1 - t_1, s_2 - t_2, \ldots, s_n - t_n),$$

$$s * t = g_a(x) * g_b(x) = \sum_{j=1}^{n} \frac{s_j * t_j}{x} = (s_1 * t_1, s_2 * t_2, \ldots, s_n * t_n),$$

$$s/t = g_a(x)/g_b(x) = \sum_{j=1}^{n} \frac{s_j/t_j}{x} = (s_1/t_1, s_2/t_2, \ldots, s_n/t_n),$$

where 'Σ' represents a union, and '$-$' is a splitter. In division operation, if $t_j = 0$, then it sets a tiny number approximated to zero.

Definition 8. Let $S = (X, C \cup Y)$ be an information system, where $C = \{c_1, c_2, \ldots, c_m\}$. For any sample $x \in X$, suppose $s = g_{c_i}(x)$ is a granule of the sample x on c_i. For the sample set X and the feature set C, a granular matrix is defined as follows:

$$F_C(X) = (F_{c_1}(X), F_{c_2}(X), \ldots, F_{c_m}(X)) = \begin{bmatrix} g_{c_1}(x_1), g_{c_2}(x_1), \ldots, g_{c_m}(x_1) \\ g_{c_1}(x_2), g_{c_2}(x_2), \ldots, g_{c_m}(x_2) \\ \cdots \\ g_{c_1}(x_n), g_{c_2}(x_n), \ldots, g_{c_m}(x_n) \end{bmatrix} = \begin{bmatrix} F_C(x_1)^T \\ F_C(x_2)^T \\ \cdots \\ F_C(x_n)^T \end{bmatrix} = (F_C(x_1), F_C(x_2), \ldots, F_C(x_n))^T$$

Since a granule is composed of granular nuclei, we can express the granular matrix by a union of granular nucleus matrices, which is showed in the follows:

$$F_C(X) = \begin{bmatrix} g_{c_1}(x_1), g_{c_2}(x_1), \ldots, g_{c_m}(x_1) \\ g_{c_1}(x_2), g_{c_2}(x_2), \ldots, g_{c_m}(x_2) \\ \cdots \\ g_{c_1}(x_n), g_{c_2}(x_n), \ldots, g_{c_m}(x_n) \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} g_{c_1}(x_1)_j, g_{c_2}(x_1)_j, \ldots, g_{c_m}(x_1)_j \\ g_{c_1}(x_2)_j, g_{c_2}(x_2)_j, \ldots, g_{c_m}(x_2)_j \\ \cdots \\ g_{c_1}(x_n)_j, g_{c_2}(x_n)_j, \ldots, g_{c_m}(x_n)_j \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} r_{11j}, r_{12j}, \ldots, r_{1mj} \\ r_{21j}, r_{22j}, \ldots, r_{2mj} \\ \cdots \\ r_{n1j}, r_{n2j}, \ldots, r_{nmj} \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \cdots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j,$$

where $r_{nmj} = g_{c_m}(x_n)_j$ is the j_{th} granular nucleus of the granule $g_{c_m}(x_n)$.

Definition 9. Given an information system $S = (X, C \cup Y)$, there are two feature subsets $P, Q \subseteq C$, where $P = \{p_1, p_2, \ldots, p_k\}$ and $Q = \{q_1, q_2, \ldots, q_m\}$. For two sample subsets $S, T \subseteq X$, they are $S = \{s_1, s_2, \ldots, s_n\}$ and $T = \{t_1, t_2, \ldots, t_k\}$. Let $g_{p_i}(s)$ be a granule of sample $s \in S$ on p_i and $g_{q_i}(t)$ be a granule of sample $t \in T$ on q_i, their cardinalities are equal, which are $card(g_{p_i}(s)) = card(g_{q_i}(t)) = n$. For the sample set S, a granular matrix on P is

$$F_P(S) = (F_{p_1}(S), F_{p_2}(S), \ldots, F_{p_k}(S)) = \begin{bmatrix} g_{p_1}(s_1), g_{p_2}(s_1), \ldots, g_{p_k}(s_1) \\ g_{p_1}(s_2), g_{p_2}(s_2), \ldots, g_{p_k}(s_2) \\ \cdots \\ g_{p_1}(s_n), g_{p_2}(s_n), \ldots, g_{p_k}(s_n) \end{bmatrix}.$$

For the sample set T, a granular matrix on Q is

$$F_Q(T) = (F_{q_1}(T), F_{q_2}(T), \ldots, F_{q_m}(T)) = \begin{bmatrix} g_{q_1}(t_1), g_{q_2}(t_1), \ldots, g_{q_m}(t_1) \\ g_{q_1}(t_2), g_{q_2}(t_2), \ldots, g_{q_m}(t_2) \\ \cdots \\ g_{q_1}(t_k), g_{q_2}(t_k), \ldots, g_{q_m}(t_k) \end{bmatrix}.$$

The multiplication of $F_P(S)$ and $F_Q(T)$ is defined as

$$F_P(S) * F_Q(T) = \begin{bmatrix} g_{c_1}(x_1), g_{c_2}(x_1), \ldots, g_{c_m}(x_1) \\ g_{c_1}(x_2), g_{c_2}(x_2), \ldots, g_{c_m}(x_2) \\ \cdots \\ g_{c_1}(x_n), g_{c_2}(x_n), \ldots, g_{c_m}(x_n) \end{bmatrix},$$

where $g_{c_m}(x_n) = g_{p_1}(s_n) * g_{q_m}(t_1) + g_{p_2}(s_n) * g_{q_m}(t_2) + \ldots + g_{p_k}(s_n) * g_{q_m}(t_k)$.

The foregoing definition of multiplication of granular matrices is a general form. We can give another form based on granular nucleus, which is illustrated in the below.

Definition 10. According to the definition of the granular matrix with granular nuclei, suppose that two granular matrices are $S = \sum_{j=1}^{n} \begin{bmatrix} s_{11}, s_{12}, \ldots, s_{1k} \\ s_{21}, s_{22}, \ldots, s_{2k} \\ \cdots \\ s_{n1}, s_{n2}, \ldots, s_{nk} \end{bmatrix}_j$ and $T = \sum_{j=1}^{n} \begin{bmatrix} t_{11}, t_{12}, \ldots, t_{1m} \\ t_{21}, t_{22}, \ldots, t_{2m} \\ \cdots \\ t_{k1}, t_{k2}, \ldots, t_{km} \end{bmatrix}_j$, then the multiplication of two granular matrices is defined as

$$S * T = \sum_{j=1}^{n} \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \cdots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j,$$

where $r_{nm} = s_{n1} * t_{1m} + s_{n2} * t_{2m} + \ldots + s_{nk} * t_{km}$.

Theorem 2. *The multiplication of granular matrices by a general form is equivalent to that by a granular nucleus form.*

Proof. According to the Definition 7, Definition 8 and Definition 9, it is easy to be proved.

Example 2. The following granules are obtained from Example 1. We have $g_1 = \frac{0}{x_1} + \frac{0.1}{x_1} + \frac{0.2}{x_1} + \frac{0.5}{x_1} = (0, 0.1, 0.2, 0.5)$ and $g_5 = \frac{0}{x_1} + \frac{0.5}{x_1} + \frac{0.1}{x_1} + \frac{0}{x_1} = (0, 0.5, 0.1, 0)$. Then, $g_1 + g_5 = \frac{0+0}{x_1} + \frac{0.1+0.5}{x_1} + \frac{0.2+0.1}{x_1} + \frac{0.5+0}{x_1} = (0, 0.6, 0.3, 0.5)$.

For the sample set $X = \{x_1, x_2, x_3, x_4\}$ and the feature set $C = \{a, b, c\}$, the granular matrix is:

$$F_C(X) = \begin{bmatrix} F_C(x_1)^T \\ F_C(x_2)^T \\ F_C(x_3)^T \\ F_C(x_4)^T \end{bmatrix} = \begin{bmatrix} ((0, 0.1, 0.2, 0.5), (0, 0.5, 0.1, 0), (0, 0.1, 0.3, 0.3)) \\ ((0.1, 0, 0.3, 0.6), (0.5, 0, 0.4, 0.5), (0.1, 0, 0.4, 0.4)) \\ ((0.2, 0.3, 0, 0.3), (0.1, 0.4, 0, 0.1), (0.3, 0.4, 0, 0)) \\ ((0.5, 0.6, 0.3, 0), (0, 0.5, 0.1, 0), (0.3, 0.4, 0, 0)) \end{bmatrix}$$

Suppose a granular matrix $S = \begin{bmatrix} ((0, 0.1, 0.2, 0.5), (0, 0.5, 0.1, 0)) \\ ((0.1, 0, 0.3, 0.6), (0.5, 0, 0.4, 0.5)) \\ ((0.2, 0.3, 0, 0.3), (0.1, 0.4, 0, 0.1)) \\ ((0.5, 0.6, 0.3, 0), (0, 0.5, 0.1, 0)) \end{bmatrix}$ and a granular matrix $T = \begin{bmatrix} ((0, 0.1, 0.2, 0.5), (0, 0.5, 0.1, 0), (0, 0.1, 0.3, 0.3)) \\ ((0.1, 0, 0.3, 0.6), (0.5, 0, 0.4, 0.5), (0.1, 0, 0.4, 0.4)) \end{bmatrix}$. Then $S * T = \begin{bmatrix} ((0, 0.01, 0.07, 0.25), (0, 0.05, 0.06, 0), (0, 0.01, 0.1, 0.15)) \\ ((0.05, 0, 0.18, 0.6), (0.25, 0, 0.19, 0.25), (0.05, 0, 0.25, 0.38)) \\ ((0.01, 0.03, 0, 0.21), (0.05, 0.15, 0, 0.05), (0.01, 0.03, 0, 0.13)) \\ ((0, 0.06, 0.09, 0), (0, 0.3, 0.07, 0), (0, 0.06, 0.13, 0)) \end{bmatrix}$.

4. The granular regression with a gradient descent method

According to different features, a sample can be granulated into some granules, which compose a granular vector. At the same time, decision values of the samples are granulated into decision granules. For these aggregate granules or granular vectors, we propose a granular regression model, provide an optimized solution, and design a gradient descent algorithm.

4.1. The model of granular regression

Definition 11. Let $S = (X, C \cup Y)$ be an information system, where the sample set is $X = \{x_1, x_2, \ldots, x_n\}$ and the feature set is $C = \{c_1, c_2, \ldots, c_m\}$. For $\forall x \in X$, suppose that a granular vector over C is $F_C(x) = (g_{c_1}(x), g_{c_2}(x), \ldots, g_{c_m}(x))^T$, and a decision granule over Y is $g_Y(x)$. Given a shared weight granular vector $W_C = (w_{c_1}, w_{c_2}, \ldots, w_{c_m})^T$, where w_{c_i} is a weight granule, then a granular regression model for the sample x is

$$f(x) = F_C(x)^T W_C - g_Y(x) = (g_{c_1}(x) * \ w_{c_1} + g_{c_2}(x) * \ w_{c_2} + \ldots + g_{c_m}(x) * \ w_{c_m}) - g_Y(x).$$

As for all the sample set in X, its granular regression model is:

$$\begin{aligned}
&f(x_1) = F_C(x_1)^T W_C - g_Y(x_1) = (g_{c_1}(x_1) * \ w_{c_1} + g_{c_2}(x_1) * \ w_{c_2} + \ldots + g_{c_m}(x_1) * \ w_{c_m}) - g_Y(x_1),\\
&f(x_2) = F_C(x_2)^T W_C - g_Y(x_2) = (g_{c_1}(x_2) * \ w_{c_1} + g_{c_2}(x_2) * \ w_{c_2} + \ldots + g_{c_m}(x_2) * \ w_{c_m}) - g_Y(x_2),\\
&\ldots\\
&f(x_n) = F_C(x_n)^T W_C - g_Y(x_n) = (g_{c_1}(x_n) * \ w_{c_1} + g_{c_2}(x_n) * \ w_{c_2} + \ldots + g_{c_m}(x_n) * \ w_{c_m}) - g_Y(x_n).
\end{aligned}$$

For convenience, the above formulas are expressed as
$f(X) = F_C(X)W_C - g_Y(X)$, where $F_C(X)$ is a granular matrix.

Definition 12. Give an information system $S = (X, C \cup Y)$, for $\forall x \in X$, the decision granule is $g_Y(x)$, then the loss function of granular regression for x is

$$||g_e(x)||_2 = ||F_C(x)^T W_C - g_Y(x)||_2.$$

As for the whole sample set, it has n samples and m features that can be represented by a matrix, noted as X. The loss function for X is

$$||g_e(X)||_2 = ||F_C(X)W_C - g_Y(X)||_2.$$

$F_C(x)^T W_C$ is a product of two granular vectors. The result of $F_C(x)^T W_C - g_Y(x)$ is a granule. It can be seen that the loss function is the norm-2 of a granule.

4.2. The optimization of granular regression

From the analysis in the previous subsection, we can see that a granular regression model is that the decision granular vector subtracts the product of granular matrix and weight granular vector. In order to get the best weight granular vector, we minimize the loss function of granular regression. The formula is expressed as follows:

$$W_C = \underset{W_C}{argmin} \frac{1}{2}||F_C(X)W_C - g_Y(X)||_2^2.$$

Here is the least square of the loss function. To facilitate calculation, a constant of 1/2 is added. Since $\frac{1}{2}||F_C(X)W_C - g_Y(X)||_2^2 = \frac{1}{2}(F_C(X)W_C - g_Y(X))^T(F_C(X)W_C - g_Y(X))$, then the loss function of granular regression is represented as $J(W_C) = \frac{1}{2}(F_C(X)W_C - g_Y(X))^T(F_C(X)W_C - g_Y(X))$. The derivative of the loss function of granular regression is proved in detail below.

Theorem 3. *If the loss function of granular regression is $J(W_C) = \frac{1}{2}(F_C(X)W_C - g_Y(X))^T(F_C(X)W_C - g_Y(X))$, then its derivative is $F_C(X)^T(F_C(X)W_C - g_Y(X))$.*

Proof. Suppose $F_C(X) = \sum_{j=1}^n \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j$, $W_C = \begin{bmatrix} w_{c_1} \\ w_{c_2} \\ \ldots \\ w_{c_m} \end{bmatrix} = \sum_{j=1}^n \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix}_j$, $g_Y(X) = \sum_{j=1}^n \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}_j$, then

$$J(W_C) = \frac{1}{2}(F_C(X)W_C - g_Y(X))^T(F_C(X)W_C - g_Y(X)) = \frac{1}{2}\left\{\sum_{j=1}^n \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j \sum_{j=1}^n \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix}_j - \sum_{j=1}^n \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}_j\right\}^T *$$

$$\left\{\sum_{j=1}^n \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j \sum_{j=1}^n \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix}_j - \sum_{j=1}^n \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}_j\right\} = \frac{1}{2}\sum_{j=1}^n \left\{\begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix} - \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}\right\}_j^T \left\{\begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix} - \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}\right\}_j. \quad \square$$

Set $X_j = \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j$, $W_j = \begin{bmatrix} s_1 \\ s_2 \\ \ldots \\ s_m \end{bmatrix}_j$, $Y_j = \begin{bmatrix} t_1 \\ t_2 \\ \ldots \\ t_n \end{bmatrix}_j$, then $J(W_C) = \frac{1}{2}\sum_{j=1}^n((XW - Y)^T(XW - Y))_j$. Since $f(W) = (XW - Y)^T(XW - Y) = (W^T X^T - Y^T)(XW - Y) = W^T X^T XW - W^T X^T Y - Y^T XW + Y^T Y$, so $\frac{\partial f(W)}{\partial W} = \frac{\partial(W^T X^T XW - W^T X^T Y - Y^T XW + Y^T Y)}{\partial W} =$

$2X^TXW - X^TY - (Y^TX)^T = 2(X^TXW - X^TY)$. Since $W_C = \begin{bmatrix} w_{c_1} \\ w_{c_2} \\ \cdots \\ w_{c_m} \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} s_1 \\ s_2 \\ \cdots \\ s_m \end{bmatrix}_j$, so $\frac{\partial J(W_C)}{\partial W_C} = \frac{\partial J(W_C)}{\partial \left(\sum_{j=1}^{n} W_j\right)} = \frac{\partial \frac{1}{2}\sum_{j=1}^{n}((XW-Y)^T(XW-Y))_j}{\partial \left(\sum_{j=1}^{n} W_j\right)} = \partial(\frac{1}{2}(X_1W_1 - Y_1)^T(X_1W_1 - Y_1) + \frac{1}{2}(X_2W_2 - Y_2)^T$

$\frac{(X_2W_2-Y_2)+\ldots+\frac{1}{2}(X_nW_n-Y_n)^T(X_nW_n-Y_n))}{\partial(W_1,W_2,\ldots,W_n)=(X_1^TX_1W_1-X_1^TY_1,X_2^TX_2W_2-X_2^TY_2,\ldots,X_n^TX_nW_n-X_n^TY_n)=\sum_{j=1}^{n}(X^TXW-X^TY)_j}$. Since $F_C(X) = \sum_{j=1}^{n} \begin{bmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \cdots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{bmatrix}_j = \sum_{j=1}^{n}(X)_j, W_C = \begin{bmatrix} w_{c_1} \\ w_{c_2} \\ \cdots \\ w_{c_m} \end{bmatrix} = \sum_{j=1}^{n} \begin{bmatrix} s_1 \\ s_2 \\ \cdots \\ s_m \end{bmatrix}_j = \sum_{j=1}^{n}(W)_j, g_Y(X) = \sum_{j=1}^{n} \begin{bmatrix} t_1 \\ t_2 \\ \cdots \\ t_n \end{bmatrix}_j = \sum_{j=1}^{n}(Y)_j$, thus $\frac{\partial J(W_C)}{\partial W_C} = \sum_{j=1}^{n}(X^TXW - X^TY)_j = F_C(X)^T(F_C(X)W_C - g_Y(X))$. So, the theorem is proved.

According to the proof of derivative of the loss function, we can use a gradient descent method to solve the approximate optimal solution of the granular regression model.

The gradient descent formula of the granular regression model is represented as follows:

$$W_C^{t+1} = W_C^t - \alpha F_C(X)^T(F_C(X)W_C^t - g_Y(X)),$$

where α is a learning rate. $F_C(X)W_C^t - g_Y(X)$ is a granular residual. If the granular residual sets E, then the formula is abbreviated as: $W_C^{t+1} = W_C^t - \alpha F_C(X)^T E$.

4.3. The gradient descent algorithm for granular regression

In the previous chapter, we give the derivative of the loss function of granular regression and prove it. Therefore, it ensures the feasibility of the gradient descent method for solving the optimization problem of a granular regression model. We use the gradient descent with sample by sample. After a sample is granulated into granules, a granular vector is constructed. The granular vector is convoluted with a shared weight granular vector to produce a new feature granule. The produced granule is compared with the decision granule to obtain a granular residual, then the residual propagates back to correct the shared weight granular vector. All samples are computed once, called a round of iteration, and then the error has accumulated. When the error is small enough, the iteration terminates. The total error after each iteration is called granular error. There are two forms of granular error, which are expressed as follows:

$$GMSE = \frac{1}{n^2}\sum_{i=1}^{n}||F_C(x_i)^TW_C^t - g_Y(x_i)||_2,$$

$$GMAE = \frac{1}{n^2}\sum_{i=1}^{n}||F_C(x_i)^TW_C^t - g_Y(x_i)||.$$

Suppose the training set has a total of n samples, and x_i represents the i_{th} sample. $F_C(x_i)$ is a granular vector of x_i on the feature set C, and $g_Y(x_i)$ represents a decision granule of x_i on a decision. t denotes the number of iterations. The granular weight shared by all samples is updated as follows:

$$\begin{aligned} &W_C^2 = W_C^1 - \alpha F_C(x_1)(F_C(x_1)^TW_C^1 - g_Y(x_1)),\\ &W_C^3 = W_C^2 - \alpha F_C(x_2)(F_C(x_2)^TW_C^2 - g_Y(x_2)),\\ &\ldots,\\ &W_C^{t+1} = W_C^t - \alpha F_C(x_n)(F_C(x_n)^TW_C^t - g_Y(x_n)). \end{aligned}$$

Repeat the above steps until *GMSE* or *GMAE* converges.

The gradient descent principle of granular regression is given in detail, so we can design the gradient descent learning algorithm of granular regression, which is described as follows.

Algorithm: Gradient Descent Algorithm of Granular Regression (GDAGR) (01) $GMSE^{(0)} = +\infty, \Delta GMSE = GMSE^{(0)}$; (02) While $\Delta GMSE > \epsilon$ Do (03) $t = 1$; (04) For $i = 1, 2, \ldots, n$ (05) $Residual(i) = \sum_{j=1}^{m} g_{c_j}(x_i) * w_{c_j}^i - y_Y(i)$; (06) For $j = 1, 2, \ldots, m$ (07) $Partial(j) = g_{c_j}(x_i) * Residual(i)$; (08) End For (09) $W_C^{i+1} = W_C^i - \alpha * Partial$; (10) End For (11) $GMSE^{(t)} = \frac{1}{n^2}\sum_{i=1}^{n}||Residual(i)||_2$; (12) $\Delta GMSE = GMSE^{(t-1)} - GMSE^{(t)}$; (13) $t = t + 1$; (14) End While

In the gradient descent algorithm, it is an iterative and convergent process. The update process is sample by sample with constantly changing weights. The parameter α is a learning rate for tuning the learning speed.

5. Experimental analysis

In the following experiments, we use UCI data sets Slump, Concrete and QSAR to test several regression models. Slump has seven conditional features and three decision features. In order to reduce the regression error, we expand the conditional features. Each feature multiplies oneself or another to induce a new feature, then we get 49 features, and plus the original 7 features, so a total of 56 features are obtained. For the three decision features (SLUMP, SLOW, CS), we randomly select the SLOW as a decision feature in these experiments. Similarly, Concrete and QSAR have eight conditional features expanded to 72 features respectively. We compare our proposed model with traditional regression models from the convergence and regression error on Slump data set, and further analyze the learning rate of granular regression on different data sets.

In the granular regression model, the weight granular vector is initial a random value and changed by a training process. The test sample is granulated into a granular vector, which is calculated with the weight granular vector, and then a predictive result is obtained. It is a granule, not a real number. Therefore, it is necessary to optimize the predictive result to achieve a real value. Let there be n training samples with a label set $Y = \{y_1, y_2, \ldots, y_n\}$. After the granulation and prediction of the test sample t, a predictive granule is obtained. Suppose it is $g_t = (r_1, r_2, \ldots, r_n)$, how to get the predictive real value? In fact, it is an optimization problem. The formula is expressed as follows:

$$y' = \underset{y'}{argmin} \sum_{i=1}^{n} \left((y' - y_i)^2 - r_i^2\right)^2.$$

Similarly, we derive the function, use a traditional gradient descent algorithm to solve it, then we obtain the predictive value. In this way, we can compare our proposed granular regression with the classical regression models.

5.1. Convergence analysis of granular regression

In this subsection, we use all samples for granulation and training to check whether the regression error converges or not, and compare the convergence effect with the traditional linear regression and ridge regression. The Mean Absolute Error (MAE) is used for evaluating regression models. It is expressed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} ||y_i' - y_i||$$

The experimental results are shown in Figs. 1 and 2. In Fig. 1, the horizontal axis represents the number of iterations and the vertical axis represents the regression error. In Fig. 2, the horizontal axis shows the number of iterations and the vertical axis represents the subtraction of errors between adjacent iterations.

Fig. 1 shows that the fitting errors of the three regressions decrease monotonously with the increasing of iterations. The fitting error of linear regression is close to that of ridge regression. The fitting error of linear regression is smaller than that of ridge regression. The fitting error of granular regression is smaller than those of linear regression and ridge regression.

Fig. 2 illustrates that the subtraction of adjacent fitting errors decreases with the increasing of iterations. When the number of iterations reaches 100,000, the subtraction of adjacent fitting errors is close to zero. Therefore, granular regression, linear regression and ridge regression converge. The convergence speeds of linear regression and ridge regression are very close. The convergence speed of granular regression is faster than those of linear regression and ridge regression.

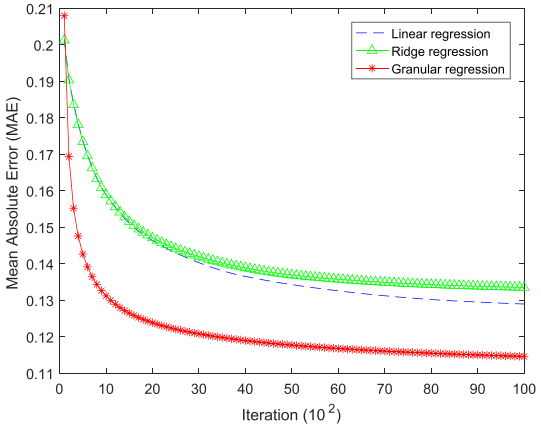

Fig. 1. Fitting errors of all samples.

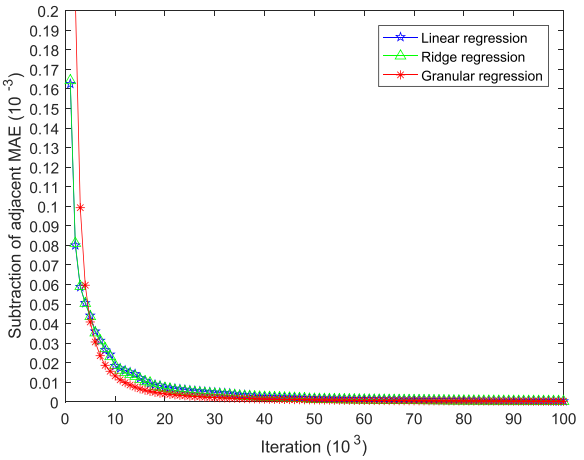

Fig. 2. Subtraction of adjacent fitting errors for all samples.

5.2. Comparisons of granular regression and classical regressions

The classical regression models include linear regression and ridge regression. We use MAE to evaluate the prediction performance of regression algorithms. Eighty percent of the data set is used for training and twenty percent for predicting. For the linear, ridge and granular regressions, we use gradient descent methods. The number of iterations is related to the prediction error. Therefore, we draw a comparison in Figs. 3–10. In Figs. 3 and 7, the horizontal axis is the number of iterations and the vertical axis represents the regression evaluation index MAE. In Figs. 4–6, the horizontal axis is the number of samples of the training set and the vertical axis shows the fitting values of the training set with fifty thousand iterations. In Figs. 8–10, the horizontal axis is the number of samples of the test set and the vertical axis represents the predictive values of the test set with fifty thousand iterations.

Fig. 3 shows that the fitting error of the training set decreases monotonously with the increasing of iterations, and the fitting error of granular regression is smaller than that of linear regression, and the fitting error of linear regression is smaller than that of ridge regression.

From Figs. 4–6, we can see that the fitting curves of linear regression and ridge regression are appropriate, and the fitting effect of granular regression is evident.

In Fig. 7, the predictive error of granular regression decreases monotonously with the increasing of iterations. The predictive errors of linear regression and ridge regression are not monotonic. When the number of iterations is small, the predictive error of ridge regression is smaller than that of linear regression; and when the number of iterations is large, the predictive error of linear regression is smaller than that of ridge regression. Moreover, the predictive error of granular regression is smaller than those of linear regression and ridge regression.

From Figs. 8–10, we can see that the linear regression, ridge regression and granular regression have similar predictive effects. Because the predictive errors of the three methods are large, ranging from 0.17 to 0.2, the prediction effects of

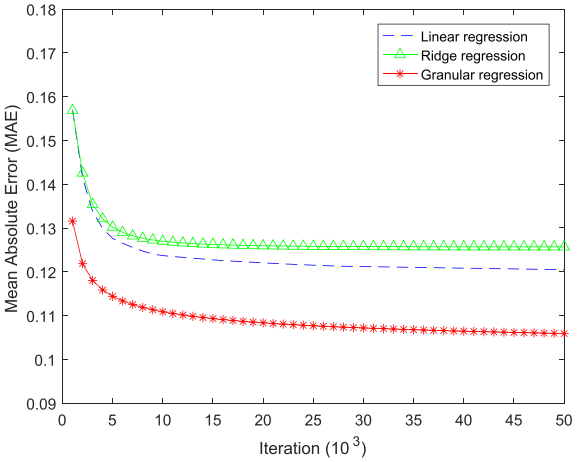

Fig. 3. Fitting errors of the training set.

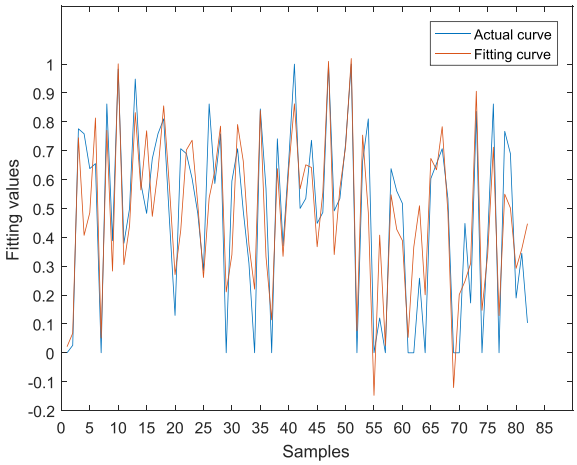

Fig. 4. Fitting curves of the training set for linear regression.

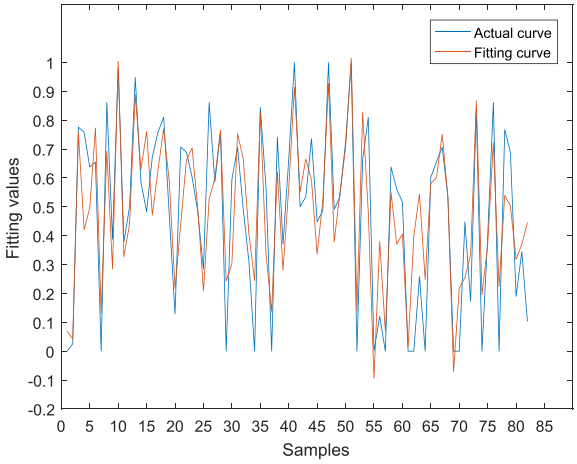

Fig. 5. Fitting curves of the training set for ridge regression.

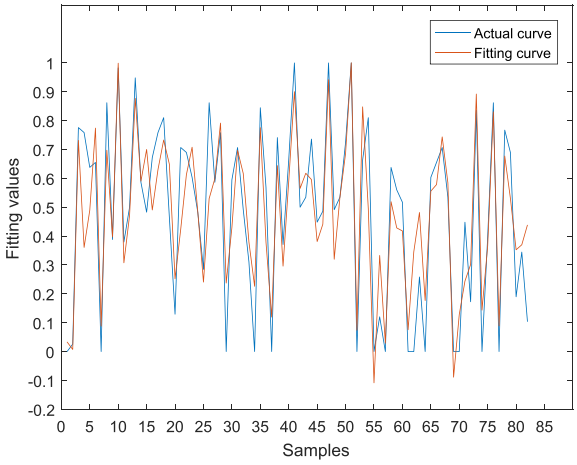

Fig. 6. Fitting curves of the training set for granular regression.

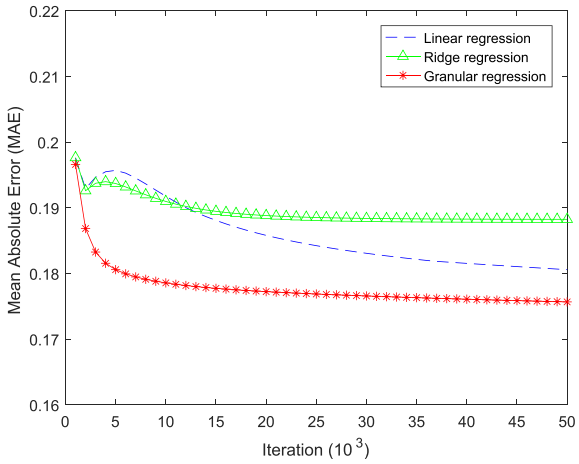

Fig. 7. Predictive errors of the test set.

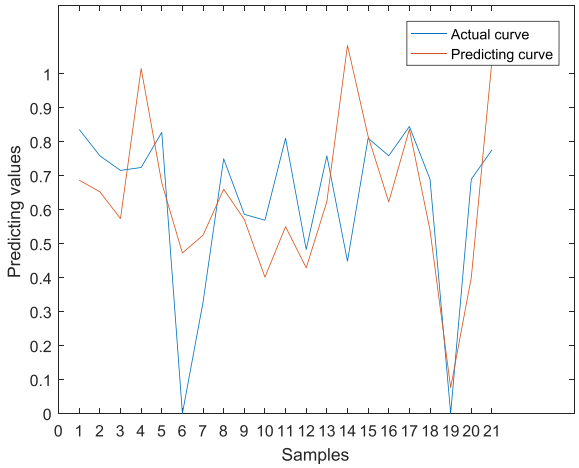

Fig. 8. Predictive curves of the test set for linear regression.

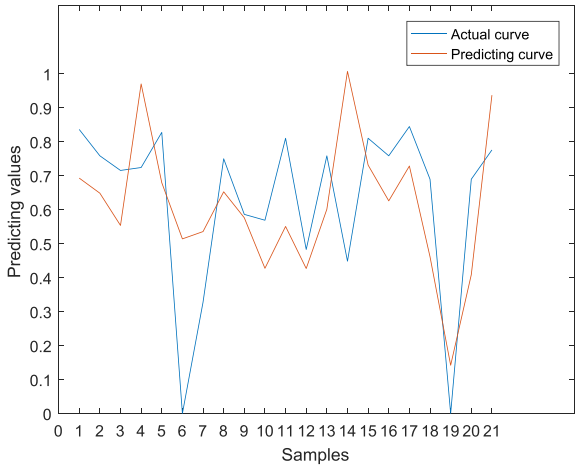

Fig. 9. Predictive curves of the test set for ridge regression.

the three methods are not obvious. Moreover, the predictive error of granular regression is smaller than those of linear regression and ridge regression, the predictive curve of granular regression is slightly better than those of the other two.

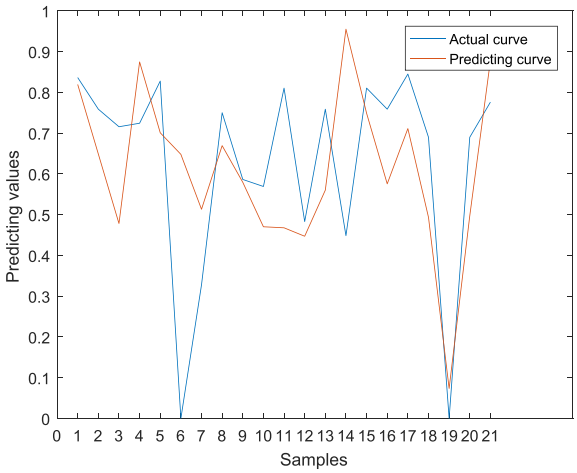

Fig. 10. Predictive curves of the test set for granular regression.

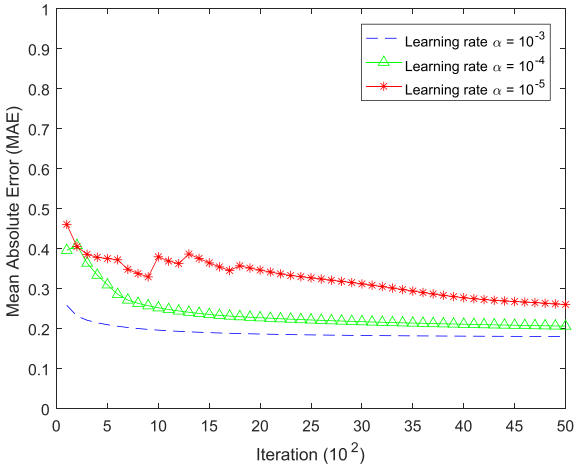

Fig. 11. Learning rate influence of granular regression for Slump data set.

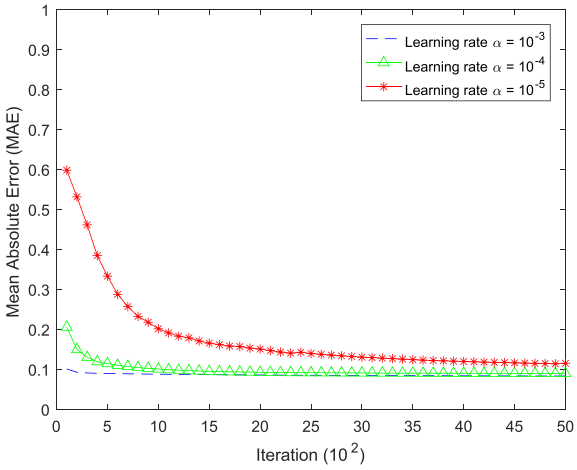

Fig. 12. Learning rate influence of granular regression for Concrete data set.

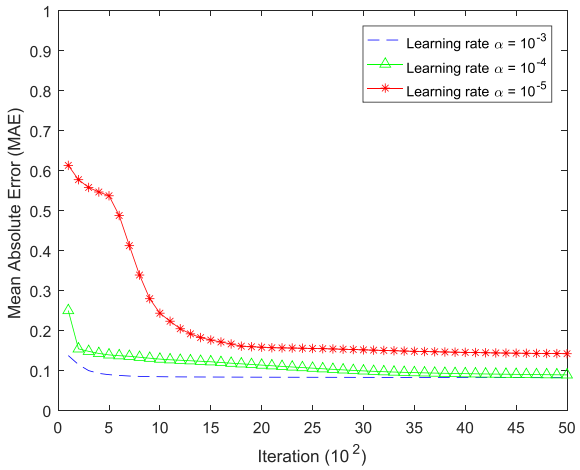

Fig. 13. Learning rate influence of granular regression for QSAR data set.

5.3. Influence of learning rate of granular regression

The three data sets Slump, Concrete and QSAR are used for testing the learning rate of granular regression. The 80% samples are randomly selected for training while the rest for testing. The learning rate sets three values with $\alpha = 10^{-3}, 10^{-4}$ and 10^{-5}. The experimental results are shown in Figs. 11–13. The horizontal axis represents the number of iterations and the vertical axis shows the mean absolute error. The maximum number of iterations is five thousand.

From Figs. 11–13, we can see that the mean absolute error decreases with the number of iterations increasing for the granular regression. As for different data sets, the mean absolute error decreases faster while the learning rate is bigger. In the experimental process, we also find that the mean absolute error will be not convergent when the learning rate is too big. While the learning rate is too small, it will be convergent slowly.

6. Conclusion

In the traditional regression models, the values involved in operations are real numbers. Starting from the study of sample granulation, a new set-form regression model is proposed by defining the concepts of granular vector and granular matrix. Firstly, a single feature granulation method is introduced to construct information granules and granular vectors in information systems, and the size measurement and operation rules of granular vectors are defined. Furthermore, the granular matrix and its related operations are proposed. The granular vector and granular matrix are applied in a classification field and the granular regression model is put forward. Secondly, the granular regression model is optimized to obtain an approximate solution, and the gradient descent algorithm of the granular regression is proposed. Finally, the feasibility and validity of the granular regression model are demonstrated by an experimental analysis. In the future work, the local granulation method will be studied for regression and prediction in big data systems. By defining the norm of a granular vector, the granular sparse regression model will be explored for feature selection and classification in machine learning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work is supported by the National Natural Science Foundation of China (Nos. 61976183, 61976158, 61871464) and the Natural Science Foundation of Fujian Province (No. 2019J01850).

References

[1] R. Al-Hmouz, W. Pedrycz, A. Balamash, et al, Description and classification of granular time series, Soft Comput. 19 (4) (2015) 1003–1017.
[2] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, et al, Rough Set Algorithms in Classification Problem, Springer-Verlag, 2000, pp. 49–88.
[3] S.B. Chen, Y.L. Xu, et al, A nonnegative locally linear KNN model for image recognition, Pattern Recogn. 83 (2018) 78–90.
[4] Y.M. Chen, N. Qin, W. Li, et al, Granule structures, distances and measures in neighborhood systems, Knowl.-Based Syst. 165 (2019) 268–281.

[5] Y.R. Chen, A. Rezapour, W.G. Tzeng, Privacy-preserving ridge regression on distributed data, Inf. Sci. 451 (2018) 34–49.
[6] J.H. Dai, H.F. Han, X.H. Zhang, et al, Catoptrical rough set model on two universes using granule-based definition and its variable precision extensions, Inf. Sci. 390 (2016) 70–81.
[7] U.M. Fayyad, K.B. Irani, On the handling of continuous-valued attributes in decision tree generation, Mach. Learn. 8 (1) (1992) 87–102.
[8] H. Fujita, A. Gaeta, V. Loia, et al, Resilience analysis of critical infrastructures: a cognitive approach based on granular computing, IEEE Trans. Cybern. 49 (5) (2019) 1835–1848.
[9] Z. Geng, C.C. Zhang, H.Y. Zhang, Improved K-means algorithm based on density canopy, Knowl.-Based Syst. 145 (2018) 289–297.
[10] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.
[11] Q.H. Hu, D.R. Yu, Z.X. Xie, Neighborhood classifiers, Expert Syst. Appl. 34 (2) (2008) 866–876.
[12] X.P. Kang, D.Q. Miao, A study on information granularity in formal concept analysis based on concept-bases, Knowl.-Based Syst. 105 (2016) 147–159.
[13] S.S. Keerthi, E.G. Gilbert, Convergence of a generalized SMO algorithm for SVM classifier design, Mach. Learn. 46 (2002) 351–360.
[14] J. Lee, Y. Choe, Robust PCA based on incoherence with geometrical interpretation, IEEE Trans. Image Process. 27 (4) (2018) 1939–1950.
[15] G.P. Lin, J.Y. Liang, J.J. Li, A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems, Knowl.-Based Syst. 91 (2016) 102–113.
[16] T.Y. Lin, Data mining: granular computing approach, Lect. Notes Comput. Sci. 1574 (1) (1999) 24–33.
[17] T.Y. Lin, L.A. Zadeh, Special issue on granular computing and data mining, Int. Intell. Syst. 19 (7) (2004) 565–566.
[18] Q. Liu, S.L. Jiang, Reasoning about information granules based on rough logic, Lect. Notes Comput. Sci. 2475 (2002) 139–143.
[19] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2106–2112.
[20] S.H. Nguyen, A. Skowron, J. Stepaniuk, Granular computing: a rough set approach, Comput. Intell. 17 (3) (2001) 514–544.
[21] S. Pang, A. Du, M.A. Orgun, et al, A novel fused convolutional neural network for biomedical image classification, Med. Biol. Eng. Comput. 57 (1) (2019) 107–121.
[22] A. Paul, D.P. Mukherjee, P. Das, et al, Improved random forest for classification, IEEE Trans. Image Process. 27 (8) (2018) 4012–4024.
[23] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[24] W. Pedrycz, Information granules and their use in schemes of knowledge management, Sci. Iran. 18 (3) (2011) 602–610.
[25] W. Pedrycz, H. Izakian, Cluster-centric fuzzy modeling, IEEE Trans. Fuzzy Syst. 22 (6) (2014) 1585–1597.
[26] W. Pedrycz, B.J. Park, S.K. Oh, The design of granular classifiers: a study in the synergy of interval calculus and fuzzy sets in pattern recognition, Pattern Recogn. 41 (12) (2008) 3720–3735.
[27] L. Polkowski, A. Skowron, Towards adaptive calculus of granules, in: Proceedings of 1998 IEEE International Conference on Fuzzy Systems, 1998, pp. 111–116.
[28] Y.H. Qian, H. Zhang, Y.L. Sang, et al, Multigranulation decision-theoretic rough sets, Int. J. Approx. Reason. 55 (1) (2014) 225–237.
[29] S.B. Roh, S.K. Oh, W. Pedrycz, A fuzzy ensemble of parallel polynomial neural networks with information granules formed by fuzzy clustering, Knowl.-Based Syst. 23 (3) (2010) 202–219.
[30] S.B. Roh, W. Pedrycz, T.C. Ahn, A design of granular fuzzy classifier, Expert Syst. Appl. 41 (15) (2014) 6786–6795.
[31] A. Skowron, A. Jankowski, S. Dutta, Interactive granular computing, Granul. Comput. 1 (2) (2016) 95–113.
[32] A. Skowron, J. Stepaniuk, Information granules: towards foundations of granular computing, Int. J. Intell. Syst. 16 (1) (2001) 57–85.
[33] Z.G. Su, T. Denoeux, BPEC: belief-peaks evidential clustering, IEEE Trans. Fuzzy Syst. 27 (1) (2019) 111–123.
[34] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. 58 (1) (1996) 267–288.
[35] G.Y. Wang, Q.H. Zhang, X.A. Ma, Q.S. Yang, Granular computing models for knowledge uncertainty, J. Software 22 (4) (2011) 676–694.
[36] W. Wang, W. Pedrycz, X. Liu, Time series long-term forecasting model based on information granules and fuzzy clustering, Eng. Appl. Artif. Intell. 41 (2015) 17–24.
[37] W.Z. Wu, Y. Leung, Theory and applications of granular labelled partitions in multi-scale decision tables, Inf. Sci. 181 (18) (2011) 3878–3897.
[38] J. Xu, G.Y. Wang, H. Yu, Review of big data processing based on granular computing, Chin. J. Comput. 38 (8) (2015) 1497–1517.
[39] W.H. Xu, J.H. Yu, A novel approach to information fusion in multi-source datasets: a granular computing viewpoint, Inf. Sci. 378 (2017) 410–423.
[40] R.R. Yager, An extension of the naive Bayesian classifier, Inf. Sci. 176 (5) (2006) 577–588.
[41] Y.Y. Yao, Three perspectives of granular computing, J. Nanchang Inst. Technol. 25 (2) (2006) 16–21.
[42] Y.Y. Yao, Information granulation and rough set approximation, Int. J. Intell. Syst. 16 (1) (2001) 87–104.
[43] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inf. Sci. 111 (1) (1998) 239–259.
[44] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (1997) 111–127.
[45] M.L. Zhang, Z.H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recogn. 40 (7) (2007) 2038–2048.
[46] Z. Zhang, Z.H. Lai, Y. Xu, et al, Discriminative elastic-net regularized linear regression, IEEE Trans. Image Process. 26 (3) (2017) 1466–1481.
[47] P.F. Zhu, Q.H. Hu, Y.H. Han, et al, Combining neighborhood separable subspaces for classification via sparsity regularized optimization, Inf. Sci. 370 (2016) 270–287.
[48] X. Zhu, W. Pedrycz, Z. Li, Granular data description: designing ellipsoidal information granules, IEEE Trans. Cybern. 47 (12) (2017) 4475–4484.
[49] X. Zhu, W. Pedrycz, Z. Li, Granular encoders and decoders: a study in processing information granules, IEEE Trans. Fuzzy Syst. 25 (5) (2017) 1115–1126.