Complex Big Data Analysis Based on Multi-granularity Generalized Functions

https://doi.org/10.3991/ijoe.v14i04.8368

Xueya Zhang(✉), Jianwei Zhang
Baoji University of Arts and Sciences, Baoji, China
534700786@qq.com

Abstract—A new method for the big data analysis - multi-granularity generalized functions data model (referred to as MGGF for short) is put forward. This method adopts the dynamic adaptive multi-granularity clustering technique, transforms the grid like "Hard partitioning" to the input data space by the generalized functions data model (referred to as GFDM for short) into the multi-granularity partitioning, and identifies the multi-granularity pattern class in the input data space. By defining the type of the mapping relationship between the multi-granularity model class and the decision-making category ftype:Ci→y, and the concept of the Degree of Fulfillment (referred to as DoF (x)) of the input data to the classification rules of the various pattern classes, the corresponding MGGF model is established. Experimental test results of different data sets show that, compared with the GFDM method, the method proposed in this paper has better data summarization ability, stronger noise data processing ability and higher searching efficiency.

Keywords—Generalized Function Set; Generalized Functions Data Model; Multi-granularity Clustering; Data Mining; Multi-granularity Generalized Functions Data Model

1 Introduction

In the knowledge discovery in database (referred to as KDD for short) and data mining (referred to as DM for short), there are usually two basic problems existing, that is, the selection of the appropriate data expression and the formalization of model indexes. Surrounding these two problems, a variety of big data analysis model methods, such as the decision tree model, neural network model, belief networks model, generalized functions data set model and so on, have emerged. Among them, in the generalized functions data set model put forward in the literature [1], a pair of clear lower and upper approximation sets is used to describe from the inner and outer sides any subset of the concepts defined on the universe of discourse. And the approximate accuracy [2] is taken as the quality indicator of the model, with the characteristics of relatively simple learning process, no need for any priori knowledge points and other characteristics [3], hence making it a powerful tool for the classification analysis in the KDD and DM.

However, since the generalized functions data set model is classified strictly in terms of equivalence relation, the classes that it processes must be accurate or completely contained [4]. In practice, however, the data set often contains noise, which leads to the result that the classification effect of the Pawlak model is not good or even fails. In order to solve this problem, in the literature [5], a variable precision function data set model (VPM) is introduced, which can "Soften" the completely exact inclusion relation into the inclusion relation to a certain degree, that is, if at least (1-β)% of the elements in the equivalent class C belong to the subset X, the class C $^{\beta}$ is referred to as being included in X, and $0 \leq \beta < 0.5$. And by changing β, the approximate sets of different precisions of X can be obtained. Recently, the concept of the generalized functions data model (referred to as GFDM for short) [6] has been put forward, which attempts to describe a certain subset concept X in the universe of discourse through the establishment of the mapping relation type:C→X between the equivalence class in the universe of discourse and the decision class. Due to the difference in the defined mapping relation type, the approximation set of X thus obtained is different. And both of the methods have achieved highly successful applications [7].

However, the above two methods still have limitations. First of all, in the data expression, the concept and knowledge that are involved in the two models are both clear. However, in the practical problems, it is more often that a number of multi-granularity concepts and multi-granularity knowledge is involved, which is reflected in the generalized functions data set model as the following: Either the knowledge of the knowledge base (equivalence relation) is clear and the approximated concept has multi-granularity, or the knowledge of the knowledge base and the approximate concept has multi-granularity. In the two cases, it is difficult for both the VPM and the GFDM to deal with the situation effectively. Secondly, in the perspective of the model index, the VPM still adopts the model accuracy as the index. Although GFDM can adopt different performance index functions as needed, it is an NP difficult problem to look for the optimal VPM or the GFDM model [8]. At present, most of the solving methods make use of a certain kind of complex heuristic algorithm, or adopt a time-consuming exhaustive search strategy, and there is still lack of a unified and effective method. On the other hand, however, we notice that in the multi-granularity data set theory [9], a membership degree is specified for each object in the universe of discourse in the interval [0,1] through the multi-granularity membership function, so as to make sophisticated description on any inaccurate concept in the domain of discourse. The generalized functions data set is combined with the multi-granularity data set to look for a stronger uncertainty model, which has been developed into an important research direction [10].

In this paper, on the basis of the VPM and GFDM, a new big data analysis method - multi-granularity generalized functions data model (referred to as MGGF for short) is put forward, which transforms the equivalent partition of the VPM and GFDM to the data space into the multi-granularity clustering classification, so as to transform the solving process of the optimal GFDM model into the optimization process of the classification objective function, which effectively avoids the NP difficult problem. At the same time, by defining the degree of fulfillment DoF, the transformation from

the MGGF model to the GFDM model is implemented, which can better describe the inaccurate concepts in the domain of discourse. In particular, due to the application of the adaptive multi-granularity clustering technology, MGGF model can reflect the feature pattern class with the super-ellipsoid, hyper plane or super linear type in the data set by a relatively small number of multi-granularity model classes. And the results of the partitioning can achieve the optimum in the whole domain of discourse space of the data set according to the different classes of the various patterns, and thus has very strong data generalization ability, which is not sensitive to the noise, thus overcoming the defects that the generalization ability of the GFDM model depends heavily on its prior partitioning of the domain of discourse space.

2 Multi-granularity Functions Data Model

2.1 Generalized Functions Data Model (GFDM)

In the generalized functions data model (GFDM), the elements in the same equivalent class qi(i=1,2…,n) in the input space may belong to the different decision class yj(j=1,2,…,p). However, as the various elements in the same equivalence class in the input space are indistinguishable from each other, they have to be grouped into the same decision class. This kind of artificial correspondence relationship between the input space and the decision space is defined as the type function in the GFDM, type:q→y. GFDM characterizes each of the decision classes in the decision space through the type function and some of the corresponding statistics. The values of these statistics are used to sort the classes $q_i = \{i = 1,2,\mathrm{L}\ ,n\}$ in accordance with the standards determined by some of the users. Therefore, the generalized functions data model (GFDM) of the information table $S = (U, A \cup d)$ can be defined as the triples as the following:

$$T = \langle q, type, \underline{\mathrm{p}} \rangle \tag{1}$$

In which $\underline{\mathrm{p}}$ stands for a linear ordering relation which is defined on the $q = \{q^1, q^2, \mathrm{L}\ q^n\}$. A basic GFDM model has the form as the following:

U/R	*type*(q_i)	*acc*
q_1	y_1	acc_1
q_2	y_2	acc_2
M	M	M
q_n	y_p	acc_n

In which acc1≥acc2≥…≥accn

For a given training data set, the grid like partition of the data space by the GFDM has the "Granularity". And the classification accuracy of the model is seriously dependent on the degree of sophistication of the pre-division of the data space, and the number of models is increased exponentially with the increase in the number of at-

tributes as well as the attribute discrete intervals. The inspiration from this to us is that, in fact, we do not need to make complete equivalence partitioning to the data space, but only need to identify the various pattern classes existing in the data space. These pattern classes can have multi-granularity, so that fine adjustment on the partitioning of the data space can be carried out to avoid the excessive hard partitioning and reduce the burden of computation.

2.2 Adaptive Dynamic Multi-granularity Partitioning of the Input Data Space

Multi-granularity clustering is a kind of important tool for the identification of the data patterns. Gustafson-Kessel (GK) algorithm is the multi-granularity generalization of the adaptive distance dynamic clustering algorithm, which can effectively search for the hyper-ellipsoid, planar or liner data class.

Assuming that $x \in^{n}$ is the input data vector, and $y \in$ is the categorical data, that is, the output variable, which value is usually taken as the integer. Let $y = \{y_1, y_2, \mathrm{L}\ , y_p\}$. Denote $Z_k = \left(x_k^T, y_k\right)^T$, in which, k stands for the k-th data point. In the GK algorithm, the distance from the midpoint x_k of n-dimensional data space to the clustering center v_i is a square inner product distance norm as the following:

$$D_{ikM_i}^2 = \|x_k - v_i\|_{M_i}^2 = \left(x_k - v_i\right)^T M_i \left(x_k - v_i\right) \tag{2}$$

In which, $M_i = det\left(F_i\right)^{1/n} F_i^{-1}$, where F_i is the covariance matrix of the i-th cluster center, which is a positive definite symmetric matrix. The estimate by using the data covariance F_i to make adaptive adjustment on M_i, so as to identify the different topologies on n (super-ellipsoid, planar or linear) pattern class. The data set $\{x_1, x_2, \mathrm{L}\ , y_n\}$ is divided into c multi-granularity classes by minimizing the objective function as the following

$$J\left(X;U,V\right) = \sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}\right)^m \|x_k - v_i\|_{M_i}^2 \tag{3}$$

to complete, in which, $U = u_{ik}$ stands for the multi-granularity partition matrix of the data set, and it meets the following equation

$$\sum_{k=1}^{c} u_{ik} = 1,\ \ 1 \le k \le N, u_{ik} \in [0,1] \tag{4}$$

$m \in [1, \infty)$ is a weighted index, which determines the degree of the multi-granularity of the classes which have been obtained (for a clear model, m = 1; for the multi-granularity model m > 1, it is usual that m = 2). The Lagrange multiplier λ_k can

be used to transform the objective function (3) and its constraint (4) into a new objective function as the following

$$\overline{J}\left(X;U,V,\lambda\right)=\sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}\right)^{m}D_{ikM_i}^{2}+\sum_{k=1}^{N}\lambda_k\left(\sum_{i=1}^{c}u_{ik}-1\right) \tag{5}$$

Assuming that $D_{ikM_i}^{2}>0,\forall_i$,k, and m>1. Let the gradient of $\overline{J}$ onU,V and λ be zero, then two necessary conditions for taking the minimum value for the equation (3) can be obtained as the following:

$$u_{ik}=\frac{1}{\sum_{j=1}^{c}\left(D_{ikM_i}/D_{jkM_j}\right)^{2/(m-1)}},1\le i\le c,1\le k\le N \tag{6}$$

$$v_i=\frac{\sum_{k=1}^{N}\left(uik\right)^{m}x_k}{\sum_{k=1}^{N}\left(uik\right)^{m}},1\le i\le c \tag{7}$$

Alternate optimization calculation is performed on the equation (6) and equation (7), and the sequence $\left\{\left(U^{(1)},V^{(1)}\right),\left(U^{(2)},V^{(2)}\right),\mathrm{K}\right\}$ will converge to the local optimal point or saddle point of the objective function equation (3).

The conduction of GK multi-granularity clustering on the training data set is actually to carry out the "Soft" partitioning to the input data space that is suitable for the prototype (such as ellipsoid, plane or linear type) in the data set, which changes with the data set and can be adjusted in the sophistication. The original data set is divided into a set of multi-granularity classes $C_i\left(1\le i\le c\right)$, its clustering center is $v_i\left(1\le i\le c\right)$, and the multi-granularity partitioning matrix is $U=\left[u_{ik}\right]\left(1\le i\le c,1\le k\le N\right)$. If the equivalent class q_i in the GFDM is replaced by the multi-granularity class C_i, how to define the pattern function $type\left(\cdot\right)$? We further study the division of the multi-granularity clustering on the classification data $y_k\left(k=1,2,\mathrm{L}\ ,N\right)$.

All the class data are weighted by using the multi-granularity partitioning matrix $U=\left(u_{ik}\right),\left(1\le i\le c,1\le k\le N\right)$ of the input space obtained as described above, and the weighted mean (that is, the class center point) that is corresponding to each input multi-granularity class can be obtained as the following

$$v_{iy}=\frac{\sum_{k=1}^{N}\left(uik\right)^{m}y_k}{\sum_{k=1}^{N}\left(uik\right)^{m}},1\le i\le c \tag{8}$$

Its corresponding variance is as the following

$$\sigma_{iy}^2 = \frac{\sum_{k=1}^{N} (uik)^m \left(y_k - v_{iy}\right)^2}{\sum_{k=1}^{N} (uik)^m}, 1 \le i \le c \tag{9}$$

2.3 Definition of Multi-granularity Functions Data Model

In order to be able to evaluate the effect of the multi-granularity clustering on the data set in the input space effectively, we assume that there is a one-to-one or many-to-one relationship between the pattern class in the input space and the class value of the decision variable, instead of the many-to-many relationship, that is, there is no such case where the same input pattern class belongs to multiple decision categories at the same time (Note: It is considered that such situation can be solved by adopting the supervised multi-granularity clustering method in the input-output product space, which has been elaborated in a separate paper), which is consistent with most of the actual situations.

Under the above assumptions, we have the following proposition.

Proposition 1. A good multi-granularity clustering on the data set in the input space corresponds to a relatively small class center variance σ_{iy}^2.

This conclusion is obvious, and the reason why the proposition does not take $\sigma_{iy}^2 = 0$ is that the impact of the data noise is taken into consideration. On the contrary, if the value of σ_{iy}^2 is relatively large, it indicates that there are at least two different classes of the patterns which are assigned to the same class by error.

To this end, we set an allowable error vector tolSig2=(s1,s2,…,sc), in which si>0,i=1,2,…,c,c is the number of the classes of the multi-granularity. Only when the class center variance σ_{iy}^2 of all the classes C_i in the clustering results meet the following equation

$$\sigma_{iy}^2 < tolSig2(i), i = 1,2,\mathrm{L}\ ,c \tag{10}$$

can the corresponding multi-granularity partitioning be accepted to be used for the construction of the MGGF model, so as to ensure the performance of the MGGF model. In this way, we can take the value of the nearest class from the distance class center as the type of the model class C_i.

Definition 1. In the case of good multi-granularity clustering, that is, $\sigma_{iy}^2 < tolSig2(i), i = 1,2,\mathrm{L}\ ,c$, the type functions of all the multi-granularity class C_i are given by the equation (11) as the following

$$ftype(C_i) = \left\{ y_i \middle| min\left(\left| v_{iy} - y_j \right|, j = 1,2,\mathrm{L}\ ,p \right) \right\}, i = 1,2,\mathrm{L}\ ,c \tag{11}$$

In this way, each multi-granularity class C_i, along with its type, corresponds to a decision-making rule as the following:

$$R_i : If\ x \in C_i, then\ d(x) = ftype(C_i) \tag{12}$$

Herein, $x \in {}^n$ stands for the input data, and C_i stands for the multi-granularity set.

Definition 2. Denote the degree of fulfillment of any input data $x \in {}^n$ for the decision rule R_i corresponding to the multi-granularity pattern class C_i as $DoF_i(x)$, and its definition is as the following

$$DoF_i(x) = \frac{\left(D_{iM_i}^2\right)^{-1/(m-1)}}{\sum_{j=1}^{c}\left(D_{jM_j}^2\right)^{-1/(m-1)}} \tag{13}$$

In which, $D_{jM_j}^2$ is the inner product distance between the data point x in the input space and the center $v_j \left(\in {}^n\right)$ of the pattern class C_j,

$$D_{jM_j}^2 = \left(x - v_j\right)^T M_j \left(x - v_j\right) \tag{14}$$

$$M_j = det\left(F_j\right)^{1/n} F_j^{-1} \tag{15}$$

It is apparent that $DoF_i(x) \in [0,1]$.

In order to make the classification rule (12) have better interpretability, the multi dimensional multi-granularity set C_i can be projected to each of the components corresponding to the conditional attribute, then the rule R_i can be expressed as the form in the following:

$$R_i^{'} : If\ x_1 \in C_{i1}\ and\ x_2 \in C_{i2}\ and\ \mathrm{L}\ \ and\ x_n \in C_{in}, then\ d(x) = ftype(C_i) \tag{16}$$

In this case, the matching degree $DoF_i(x)$ of x to the rule $R_i^{'}$ can be defined as the product of the membership degree $\mu_{C_{ij}}(x_j)(j = 1,2,\mathrm{L}\ ,n)$ of each component as the following,

$$DoF_i(x) = \prod_{j=1}^{n} \mu_{C_{ij}}(x_j) \tag{17}$$

According to the degree of fulfillment of the input data R_i to each classification rule $R_i(i = 1,2,\mathrm{L}\ ,c)$, the class value of x can be determined comprehensively. Firstly, a threshold constant is set for the degree of fulfillment of each classification rule R_i,

which is denoted as TC, and TC□c.. Only when $DoF_i(x) > TC(i)$ will $ftype(C_i)$be considered as a candidate for the class value of x, so as to inhibit the noise and improve the efficiency of the algorithm. And then, from all the selected multi-granularity pattern classes, the class with the degree of fulfillment of the rule greater than the threshold value and with the maximum class is selected as the final class value of x, that is, the following can be obtained,

$$d(x) = ftype(C_i) \tag{18}$$

In which, C_j meets the following equation

$$DoF_i(x) = max\left\{DoF_i(x) \middle| DoF_i(x) > TC(i), i = 1,2,\mathrm{L}\ ,c\right\} \tag{19}$$

In this way, the index statistics of the multi-granularity model class $C_i(i = 1,2,\mathrm{L}\ ,c)$in the type function $ftype(\cdot)$ such as the class classification accuracy $acc(C_i)$, the cumulative model accuracy cma of the other index statistics can be calculated.

In summary, we can provide a new method for the data classification analysis on the basis of the GFDM model: Multi-granularity generalized functions data model (referred to as MGGF for short).

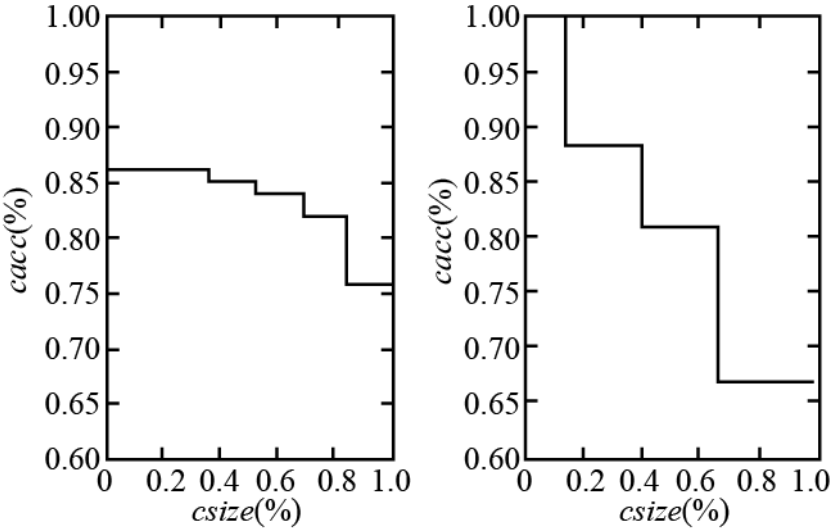
Definition 3. The multi-granularity functions data model (MGGF) of the information table $S = (U, A\,\mathrm{U}\,d)$ is defined as a triad as the following

$$FT = \langle C, ftype, \underline{\mathrm{p}} \rangle \tag{20}$$

In which, C={C1,C2,…,Cc} stands for the set of multi-granularity pattern classes on the input data space, and $ftype: C \to y$ stands for the mapping relation given by the equation (11), which is an ordering relation defined on C according to the corresponding statistics computed.

2.4 Characteristics of the Multi-granularity Functions Data Model

Compared with the GFDM model, the multi-granularity generalized functions data set model (MGGF) has a number of good features, which are mainly reflected in the following aspects: (1) The dynamic adjustability of the model MGGF is the same as that of the GFDM, with the characteristics of simplicity, intuitiveness and clear internal character at a glance. In addition, by using the relevant statistics, the corresponding model performance curve can be plotted. As shown in Figure 1(a) and Figure 1(b), the threshold values of the degree of fulfillment of the rules of 5 multi-granularity pattern classes obtained by the MGGF model in the Pima Indians Diabetes Data Set are given. And the threshold values are taken as TC=[0.10.10.10.10.1] and TC=[0.90.020.80.10.02], the cumulative model accumulation model *cmα* cumulative model element coverage csize curve.



(a) *TC*=[0.1 0.1 0.1 0.1 0.1] (b)*TC*=[0.9 0.02 0.8 0.1 0.02]

Fig. 1. Cumulative model element coverage csize curve of the cumulative model accuracy cma

It can be seen from Figure 1 that, in the MGGF, only by adjusting the tolerance threshold value tolSig2 and TC of different pattern class C_i according to the needs, the MGGF model with different performance can be obtained, so as to adapt to different actual situation; while in the GFDM method, it is only possible to modify the performance index function of the model and re-establish the new GFDM model, which is time-consuming and lack of flexibility.

(2) Adaptation of the model class prototype

Another superiority of the MGGF model compared with the GFDM model is that MGGF can identify the multi-granularity pattern class $C_i\left(1 \le i \le c\right)$with different shapes and directions in the same data set, such as super-ellipsoid, plane or linear type and so on, and has certain degree of self-adaptability to the pattern class prototype in the data set.

The self-adaptability of this kind of pattern class prototype is derived from the GK multi-granularity clustering algorithm adopted by MGGF. This algorithm makes use of the estimation information of the multi-granularity covariance matrix F_i of each pattern class C_i ($1 \le i \le c$, similar below), while the feature structure of F_i can provide the information on the shape and direction of its corresponding multi-granularity class C_i .

Proposition 2. The ratio of the length of each axis of the hyper ellipsoid prototype of the multi-granularity model class $C_i\left(1 \le i \le c\right)$in the data set $\left\{x_k \in {}^n, k = 1,2,\mathrm{L}\ ,N\right\}$ is equal to the ratio of the square root of the corresponding eigenvalue of its covariance matrix F_i .

As shown in Figure 2, the equation $(x-v)^T F^{-1}(x-v)=1$ has defined a super-ellipsoid. And the length of the j-th axis of the super-ellipsoid is equal to $\sqrt{\lambda_j}$, the direction is given by ϕ_j, λ_j and ϕ_j are the j-th eigenvalue of F_i respectively and their eigenvectors.

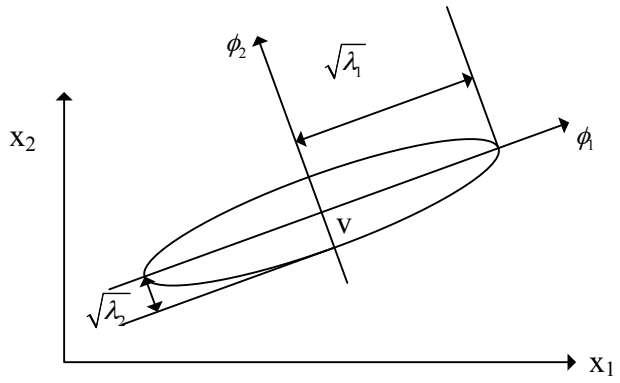


Fig. 2. Relationship between eigenvalues and eigenvectors of the hyperella and covariance matrix Fi

The linear subspace of the data space can be expressed as the flat super-ellipsoid, or also as a hyper plane, which means that when there is one or there are several eigenvalues λ close to zero in the covariance matrix F_i, the super-ellipsoid prototype is degraded to the corresponding hyper plane, super linear prototype. GK multi-granularity of the multi-granularity C_i will be degraded into the corresponding hyper plane and super linear prototype. The GK multi-granularity clustering algorithm makes use of the estimation information of F_i to correct the calculation of the inner product distance $D^2_{ikM_i}$, so that C_i can fit with the pattern class prototype in the data set.

(3) Global character of the pattern class recognition

MGGF can reflect the prototype of the linear feature pattern in the data set with a relatively small number of multi-granularity pattern classes, which has the integrity and globality in the recognition of the data prototype. Therefore, it has relatively strong noise data processing ability and data generalization ability. However, GFDM can only rely on the grid like partition of the domain of discourse space of the data set in advance to approximate a certain data class from the local fine grid point. The smaller the block shapes of the grid point, the higher the data partition accuracy, hence the more the required number of points of division. However, the increase in the number of the points of division will not only lead to the sensitivity of the model to the data noise, but also result in the exponential increase in the number of the GFDM models to be calculated, hence causing great computational burden.

(4) Effective processing of the concept of multi-granularity

By comparing the proximity between two multi-granularity sets, MGGF can effectively deal with the multi-granularity concept as defined in the data space.

Definition 4. Assuming that B is any multi-granularity set in the data space $x \in {}^{n}$, and $C_i \left(1 \le i \le c\right)$ is the multi-granularity pattern class of MGGF. The degree of fulfillment $DoF_i \left(B\right)$ of the classification rule R_i of B to C_i is calculated according to the following equation:

$$DoF_i \left(B\right) = \frac{\int_{\Omega_x} \left[C_i \left(x\right) \wedge B\left(x\right)\right] dx}{\int_{\Omega_x} \left[C_i \left(x\right) \vee B\left(x\right)\right] dx}, i = 1,2,\mathrm{L}\ ,c \tag{21}$$

In which, Ω_x stands for the universe of discourse of the data set $x_k \in {}^{n}$. It shall be pointed out that, both $B\left(x\right)$ and $C_i \left(x\right)$ here are multi dimensional multi-granularity set (n), in the equation, multiple integral calculus is conducted. And the equation (21) can also be interpreted as the form of "Volume" as the following

$$DoF_i \left(B\right) = \frac{\text{Volume under the hypersonic surface}\left(C_i \text{ I } B\right)\left(x\right)}{\text{Volume under the hypersonic surface}\left(C_i \text{ U } B\right)\left(x\right)}, i = 1,2,\mathrm{L}\ ,c \tag{22}$$

Similarly, we only need to calculate the multi-granularity concept B and the degree of fulfillment $DoF_i \left(B\right)$ of the classification rule of each multi-granularity pattern class $C_i \left(1 \le i \le c\right)$, and then take the class of the pattern with the $DoF_i \left(B\right)$ value exceeding the threshold value $TC\left(i\right)$ and the maximum pattern class as the class of the multi-granularity concept B.

3 Studies of the Experimental Data

Example 1. It is considered that the three pattern classes, that is, C1, C2 and C3 data set, have circular, linear and elliptical prototypes as shown in Figure 3. The applied data noise is ε~N(0,σ2),σ=0.1. 200(C1), 150(C2) and 200(C3) sample data points are first generated as the training data set, and the same number of data points are generated separately as the test set according to the same mechanism. The method put forward in this paper and the GFDM method are operated for 10 times, respectively, and the average performance of the methods is calculated as shown in Table 1.

It can be seen from Table 1 that the method put forward in this paper can be used to better identify the prototypes of the three pattern classes. The cumulative classification accuracy of the established MGGF model is 99.46%, and the time used is far less than that used by the GFDM method. The classification accuracy on the test set is up to 99.46%; 10 small partitions are divided first for the x,y input variables, and then four points of division, that is (1,1), (2,2), (3,3) and (4,4) are taken from the corre-

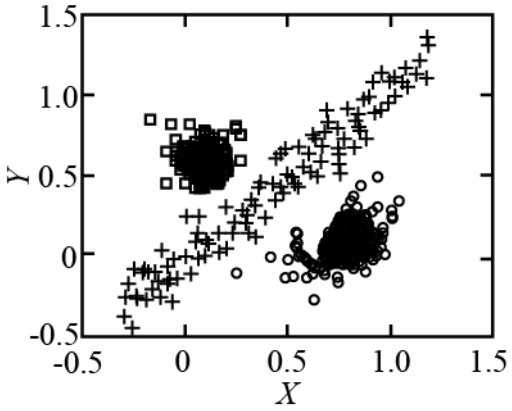


Fig. 3. Data scatter plot of Figure 1

Table 1. Main performance of the GFDM obtained by various methods on the data set in the Example 1

	Class number c or point of division	cma(%)	Time(s)	valid(%)
MGGF	c=3	99.46	0.410	99.46
GFDM	c=4(1,1)	97.46	3.174	97.46
	c=9(2,2)	97.82	56.902	97.82
	c=16(3,3)	98.00	348.902	98.00
	c=25(4,4)	98.37	924.539	98.37

sponding (9,9) points of division to establish the corresponding GFDM model. The results are shown in Table 1.

And it can be seen from the partition of the MGGF and GFDM on the data space in Example 1 that: (1) The partition direction of the input space by MGGF is determined by the data prototype of the clustering, which is not necessarily parallel to the axis; while the partition of the input space by GFDM is gradient, and the directions of each section must be parallel to the axis. (2) The partition of the input space by MGGF is smooth transition, which is in line with the characteristics and laws of the classification of the data affected by the noise, and thus more objectively reflects the practical problem; while the partition of the input space by GFDM has clear boundaries, which is divided completely, and is a kind of "Hard" division, and does not meet the characteristics of the noise data and the distinguishing law.

Example 2. The UCI database is a well-known database in the machine learning field, from which we have selected seven typical data sets (please refer to Table 2) to conduct testing and comparison of the method put forward in this paper. For each data set, about 2 / 3 of the case samples are randomly selected as the training set, and the remaining case samples are taken as the test set.

The percentage of the correct classification of the samples on the test set is recorded as the operational accuracy valid.

Table 2. The Main performance of the results obtained by various methods on the data set in the Example 2

Data set	Number of cases (training set / test set)	C4.5		GFDM			MGGF		
		cma(%)	*valid(%)*	*cma(%)*	*valid(%)*	*Class number*	*cma(%)*	*valid(%)*	*Class number*
Bupa-liver	345(245/100)	79.60	64.00	67.347*	61.00	7	75.918	71.00	8
Iris	150(100/50)	98.00	98.00	98.00	96.00	15	98.00	100.00	3
Monks-1	432(300/132)	100.00	100.00	96.67	96.97	61	96.33	99.242	4
Monks-2	432(300/132)	83.33	77.273	64.00	53.03	8	75.00	77.273	4
Monks-3	432(300/132)	100.00	96.10	96.67*	98.485	32	100.00	100.00	11
Pima	768(568/200)	79.60	63.00	76.585	71.50	11	76.056	71.00	5
Wine	178(128/50)	100.00	94.00	100.00*	98.00	54	92.188	100.00	5

Note: * stands for the results obtained using GFDM + GA method.

The method put forward in this paper, GFDM method and C4.5 classification tree method are applied to the aforementioned data set respectively, and are operated for 10 times respectively. The average results are taken as shown in Table 2.

C4.5 is a well-known classification tree algorithm. It adopts the recursive partitioning method to continually divide the training sample set into similar subsets until the leaf nodes contain only a single class of samples. In the process of learning, the classification tree method also has the problem of over-fitting to the training sample. Its solution is to adopt the "Pruning" strategy, that is, to give up one or a few "Sub-trees" and replace them with the "Leaves". In C4.5, the classification tree can also be expressed as a set of rules.

From the above examples, we find that: (1) For the MGGF model established by the method put forward in this paper, its cumulative model classification accuracy cma does not have significant difference when compared with that of the GFDM method and the C4.5 method. However, the method put forward in this paper has higher solving efficiency and consumes less time; (2) When the classification model obtained by the three methods is applied to the test data set, it is found that the MGGF model has higher classification accuracy, that is, the effective accuracy of the model, which means that the MGGF method has stronger data generalization capability. (3) The method put forward in this paper has very good "Flexibility", and a suitable MGGF model can be obtained by adjusting the tolerance threshold value tolSig2 of different pattern classes and the degree of fulfillment threshold value TC of all the classification rules according to the characteristics of the data set, so as to better deal with the impact of the noise data. At this point, it is similar to the "Pruning" strategy in C4.5.

4 Conclusion

The multi-granularity generalized functions data model (referred to as MGGF for short) combines the adaptive dynamic multi-granularity clustering technique with the generalized functions data model to provide a new method for the analysis of the big data. This method has not only overcome the defect that the classification accuracy of

the GFDM method of has serious dependence on the grid like partition performed on the input data space in advance, but also can identify the optimal feature pattern class with the super-ellipsoid, hyper planar or linear type in the data with highly efficient search strategy. Compared with the GFDM method, the MGGF method is more flexible, and has stronger noise data processing capabilities. Different types of experimental data set test results have further demonstrated the effectiveness of the MGGF method.

5 Acknowledgment

Project No.16JK1045 of the scientific research project of the Shaanxi Provincial Education Department; Project No.14GYGG-4 of Baoji science and technology program ; Project No.ZK14083 of key project of scientific research of Baoji University of Arts and Sciences

6 References

[1] Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205. https://doi.org/10.1126/science.1248506

[2] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352. https://doi.org/10.1001/jama.2013.393

[3] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86-94. https://doi.org/10.1145/2611567

[4] George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. Academy of Management Journal, 57(2), 321-326. https://doi.org/10.5465/amj.2014.4002

[5] Barrett, M. A., Humblet, O., Hiatt, R. A., & Adler, N. E. (2013). Big data and disease prevention: From quantified self to quantified communities. Big data, 1(3), 168-175. https://doi.org/10.1089/big.2013.0027

[6] Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 79, 3-15. https://doi.org/10.1016/j.jpdc.2014.08.003

[7] Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. Communications of the ACM, 57(3), 78-85. https://doi.org/10.1145/2500873

[8] Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234-246. https://doi.org/10.1016/j.ijpe.2014.12.031

[9] Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J. C., Hueske, F., Heise, A., ... & Naumann, F. (2014). The stratosphere platform for big data analytics. The VLDB Journal, 23(6), 939-964. https://doi.org/10.1007/s00778-014-0357-y

[10] Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing, 74(7), 2561-2573. https://doi.org/10.1016/j.jpdc.2014.01.003

7 Authors

Xueya Zhang is with the College of Computer Science,Baoji University of Arts and Sciences, Baoji, China.

Jianwei Zhang is with the College of physics and optoelectronic technology, Baoji University of Arts and Sciences, Baoji, China.

Article submitted 07 February 2018. Final acceptance 31 March 2018. Final version published as submitted by the authors.