Science & Technology Development Journal – Engineering and Technology 2024, ():1-7

Open Access Full Text Article

Research Article

Clustering fuzzy data by hedge algebra and regression approach

Phu Phuoc Huy[1], Doan Van Thang[2,*], Hoang Tuan[1], Nguyen Xuan Nhut[3]

ABSTRACT

Fuzzy clustering has been extensively explored across various methodologies, yielding diverse results within the realm of data mining. The plethora of research outcomes underscores the complexity inherent in fuzzy data mining, particularly when confronted with diverse data types aiming to delineate objects' affiliation with specific clusters. This intricacy is further compounded by the ubiquity of incomplete data, commonly referred to as missing data, posing a formidable challenge in this domain. Addressing the missing value predicament becomes imperative for a more nuanced and accurate enhancement of fuzzy clustering.

In response to these challenges, a novel approach has emerged, leveraging the synergies between hedging algebra and the linear regression model. This innovative methodology seeks to overcome the intricacies associated with diverse data types and missing values. By integrating algebraic principles with linear regression techniques, the proposed method introduces a robust framework for classifying objects within a cluster. The fusion of these mathematical tools offers a unique solution that not only navigates the complexities of fuzzy data mining but also addresses the pervasive issue of missing data.

The paper delves into the advantages of adopting hedging algebra and the linear regression model in tandem, presenting a comprehensive methodology that significantly contributes to the refinement of fuzzy clustering. The collaborative interplay of algebraic principles and regression models not only enhances the accuracy of object classification within clusters but also provides a robust strategy for handling missing values in the dataset. This integrated approach represents a noteworthy advancement in the field of fuzzy clustering, offering a more comprehensive and effective solution to the intricate challenges posed by diverse data types and the prevalent issue of missing data.

Key words: linear regression, statistical theory, missing data, hedge algebra, data mining

[1]*Institute of Information Technology, AMST, Viet Nam.*

[2]*Ho Chi Minh city Industrial University, Viet Nam*

[3]*Ho Chi Minh City College of Industry and Trade, Viet Nam.*

Correspondence

Doan Van Thang, Ho Chi Minh city Industrial University, Viet Nam

Email: vanthangdn@gmail.com

History

- Received: 15-9-2023
- Accepted: 26-4-2024
- Published Online:

DOI :

INTRODUCTION

Data clustering stands as a pivotal technique within the realm of data mining, falling under the category of unsupervised learning methods in machine learning. While various definitions exist, the core concept of clustering revolves around the identification of methods to categorize a set of objects into clusters. These clusters are formed with the objective of grouping similar objects together, ensuring that objects within the same cluster exhibit similarity, while those in different clusters are dissimilar.

The aim of clustering is to discern the inherent characteristics within data groups. Clustering algorithms are capable of creating clusters, yet there is no universally accepted criterion to judge the effectiveness of clustering analysis. The choice of evaluation criteria is contingent upon the specific purpose of clustering, whether it be data reduction, identification of "natural clusters," extraction of "useful" clusters, or the detection of outliers

According to researches, there is currently no general clustering method that can fully handle all types of data cluster structures. Furthermore, clustering methods need a way to represent the structure of data clusters, for each different representation method there will be a corresponding appropriate clustering algorithm. Therefore, data clustering is still a difficult and open problem, because it must solve many basic problems in a complete and appropriate way for many different types of data, especially for mixed data, that is increasing in data management systems. This problem is also one of the major challenges in machine learning.

Data cleaning is an important step in the discovery process because if the data is not of good, the mining results are also poor quality, for example, duplicate or missing data can be the cause of wrong statistics. Clearly, quality decisions must be based on quality data. Currently, there are many approaches to solve the problem such as: models based on similarity relationships [1], statistics & probability [2–4], similarity reasoning [5,6] using random forest.... All of the above approaches aim to adequately capture and handle incomplete, inaccurate or uncertain information.

Cite this article : Huy P P, Thang D V, Tuan H, Nhut N X. **Clustering fuzzy data by hedge algebra and regression approach**. *Sci. Tech. Dev. J. – Engineering and Technology* 2024; ():1-7.

Doan et al. employ fuzzy dependencies for managing missing attribute values[7]. Introducing fuzzy attribute dependency and fuzzy method dependency expands upon the concept of fuzzy functional dependency within the context of fuzzy relational databases. Building upon this foundation, the paper utilizes these fuzzy dependencies to approximate the correct response to Null queries

In research[8–10] using the theory of similarity inference, if S is called the source object set, T is the target object set, the set of source and target objects with similar properties is P. Then , if S has property P′ then it follows that T may also have P′ based on property P present in both S and T. Similar inference can be applied to handle missing values and find approximate answer to Null query quite efficiently.

In Tang *et al.* (2017)[6], the researchers introduced a theoretical model employing analogous reasoning to address Null queries within a fuzzy relational database model reliant on ability distribution. Nonetheless, Dutta's model in this context does not take into account data characterized by discrete similarity domains, which involve modeling data through similarity relationships.

In this article, we study the regression model and the hedge algebra for handling the missing values in data preprocessing and conducting clustering more accurately on the data with the following information: Information is incomplete, inaccurate or uncertain. The theoretical basis will be presented in the next section.

SOME RELATED CONCEPTS

Hedge Algebra (HA)

Within this section, we encapsulate key notions pertaining to quantitative mapping as presented in[7], and elucidate the process of recognizing systems associated with quantitative semantic neighborhoods.

Given a HA number X = (X, G, H, ≤), in there X = LDom(X), G = {1, c-, W, c+, 0} is the set of generating elements, H represents the collection of hedge elements, regarded as unary operations and is a semantic ordering relationship on X. The set X is generated from the set G by the operations in H. Thus, each element of X will have a representation x = $h_n h_{n-1}$.......h_{1x}, x ∈ G. The set of all elements generated from an element x is denoted by H(x). Given set of hedges $H = H^- \cup H^+$, in there H^+ = {h_1,..., h_p} and H^- = {h_{-1}, ..., h_{-q}}, are all linear with the following order: h_1 < .. .< h_p và h_{-1} < .. .< h_{-q}, where both p and q are greater than 1. Subsequently, the subsequent definitions are interrelated:

Definition 2.1 Functions $fm : X \to [0,1]$ is called a measure of fuzziness on X if it satisfies the following conditions:

(1) fm is the full fuzzy measure on X, i.e $\sum_{-q \le i \le p. i \ne 0} fm(h_i u) = fm(u)$.

(2) If X is a clear concept, that is $H(x) = \{x\}$, $fm(x) = 0$, $so\ fm(0) = fm(W) = fm(1) = 0$.

(3) With $\forall x, y \in X, \forall h \in H$, we have $\frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)}$

It means, this ratio does not depend on x and y, is denoted by μ(h) and is called the fuzziness measure of the hedge h.

Definition 2.2 (Semantic quantifier function ν)

Let fm be the fuzziness measure on X, the semantic quantitative function v on X is defined as follows:

(1) $v(W) = \theta = fm(c^-)$, $v(c^-) = \theta - \alpha fm(c^-)$ and $v(c^+) = \theta + \alpha fm(c^+)$

(2) If 1 ≤ j ≤ p then:

$v(h_j x) = v(x) + Sign(h_j x) \times \left[\sum_{i=1}^{j} fm(h_j x) - \omega(h_j x) fm(h_j x)\right]$ if -q ≤ j ≤ -1 then: $v(h_j x) = v(x) + Sign(h_j x) \times \left[\sum_{i=1}^{j} fm(h_j x) - \omega(h_j x) fm(h_j x)\right]$ in there:

$\omega(h_j x) = \frac{1}{2}\left[1 + Sign(h_j x) Sign(h_q h_j x)(\beta - \alpha)\right] \in \{\alpha, \beta\}$ Partitioning based on fuzziness measure of linguistic values in hedge algebra

Since the measure of fuzziness of words is an interval of the interval [0, 1] and a family of such intervals of words of the same length will form the partition of [0, 1]. Partitions corresponding to larger word lengths will be finer, and when the length is infinitely large, the length of the partition intervals gradually decreases to 0.

Example 1: Consider hedge algebra AX = (X, C, H, ≤), in there H^+ = {More, Very} with More < Very, H^- = {Little, Possibly} with Little > Possibly, và C = {Small, Large} with Small is a negative element, Large is a positive element.

Given W=0.5, fm(Little) = 0.4, fm(Possibly) = 0.1, fm(More) = 0.1, fm(Very) = 0.4

Then, we have the following quantitative value, and results are shown in table 1.

Definition 2.3 Given $P^k = \{I(x) : x \in X_k\}$ with $X_k = \{x \in X : x = k\}$ is a partition [0, 1]. We say that u is equal to v by level k in P^k, is denoted u v if and only if I(u) and I(v) belong to the same inner range P^k. That means, ∀u, v ∈ X, $u \approx_k v \Leftrightarrow \exists \triangle^k \in P^k : I(u) \sqsubseteq \triangle^k\ and\ I(v) \sqsubseteq \triangle^k$ và I(v) ⊠ Δk.

Single Linear Regression

Given two random variables X and Y, observed experimentally by two samples of size n: X: X1, X2,…, Xn; Y: Y1, Y2,…, Yn.

Table 1: Quantitative value v

Linguistic values	function
Very Very Small	0.04
Very Small	0.10
Possibly Very Small	0.11
Little Very Small	0.16
Small	0.25
Very Possibly Small	0.26
Little Small	0.40
More Little Small	0.41
Very Little Small	0.46
Very Very Small	0.04
Very Small	0.10

Y has a linear relationship with X, if $Y_i = \alpha + \beta X_i + \varepsilon_i$, i = 1,2,..,n

With: ε_i is a random variable according to the law of normal distribution N(0;σ2)

α: is called intercept, β: is called slop hay gradient.

These coefficients are estimated from the data. The estimation method is the least squares method. This method finds α, β để $\sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2$ reaching the smallest value.

When $\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$,

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} \tag{1}$$

Attention: $\widehat{\alpha}$, $\widehat{\beta}$ are approximate estimates of α,β. With $\widehat{\alpha}$, $\widehat{\beta}$ we have $\widehat{y}_l = \widehat{\alpha} + \widehat{\beta}\widehat{x}_l$, then the quantity $(y_i - \widehat{y}_l)$ is called residual. The variance of the residuals can be estimated by (2).

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)}{n-2} \tag{2}$$

Example 2: Consider research data on blood cholesterol levels of 18 male subjects as follows (BMI: ratio of weight (kg) to height squared (cm2)). Estimated correlation coefficient between age and Cholesterol (results are shown in table 2).

To analyze simple linear regression for the two quantities *age* and *chol*, We need to calculate alpha-beta. Figure 1 is the code to calculate alpha-beta in R language

In this result, chol is described as a function of age, with $\widehat{\alpha}$ = 1.0892; $\widehat{\beta}$ = 0.05779, that means we have a linear equation $\widehat{y}_i$ = 1.08922+0.05779*$\widehat{x}_i$

```
> lm(chol~age)

Call:
lm(formula = chol ~ age)

Coefficients:
(Intercept)          age
    1.08922      0.05779
```

Figure 1: The code calculates alpha-beta.

Correlation Analysis

Correlation Analysis serves the purpose of quantifying the degree of a linear association between two random variables, with the intensity of this association conveyed by the correlation coefficient

Correlation Coefficient Pearson r

Let two random variables X and Y follow the law of normal distribution, observed experimentally by two samples of size n: [X: $X_1, X_2, .., X_n$; Y:$Y_1, Y_2, .., Y_n$]

Correlation coefficient Pearson r is determined:

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\left(Y_i - \bar{Y}\right)^2}} \tag{3}$$

Correlation coefficient $x_{XY} \in [-1, +1]$, in fact, is convented:

$|r_{XY}|$ > 0.8: xtremely robust linear correlation

$|r_{XY}| \in$ (0.6,0.8)Robust linear correlation

$||r_{XY}| \in$ (0.4,0.6): There exists a linear correlation

$|r_{XY}| \in$ (0.2,0.4):Weak linear correlation

$|r_{XY}|$ < 0.2: Extremely faint linear correlation or absence of a linear correlation.

Based on the correlation coefficient, we will know the relationship between two variables. Through this, we can know the strength and weakness of the relationship between the two variables under consideration. The closer the absolute value of the correlation coefficient is to 1, shows that the relationship between two variables is stronger.

Correlation matrix

The correlation matrix is a tabular representation revealing the correlation coefficients among variables when dealing with more than two variables in a dataset. Each cell within the matrix denotes the correlation between two specific variables. Typically, the correlationmatrix finds utility both before and after conducting exploratory factor analysis, serving

Table 2: Cholesterol data

id	age	bmi	chol
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8
9	22	19.8	2.1

to scrutinize correlations between factors and identify multicollinearity in multivariate linear regression models. It's worth noting that the assessment of multicollinearity is inherently relative, as variables may exhibit multicollinearity even in the absence of high correlation. Table 3 presents the correlation values numerically and figure 2 shows the correlation graphically

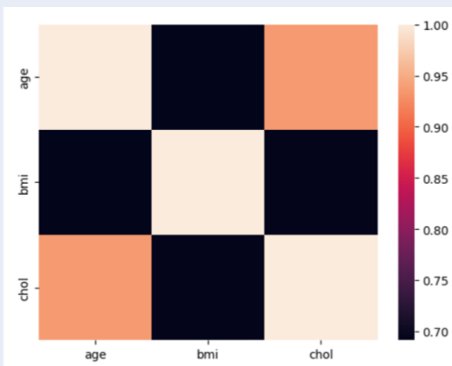

Figure 2: The correlation matrix is shown in the heat chart.

FUZZY DATA CLUSTERING

To apply the research method, in this article we use the Employee database (with 9 attributes, 93 records) and the structure is as follows (table 4)

We notice that the Salary attribute currently has three employees E006, E028, E037 whose value is NULL, and two employees E043, E52 whose Salary value are the *Language* values and is shown in table 5

So the question is how to perform data clustering for the Salary attribute with missing and incomplete values? As we stated in section 2 of the article, many research directions have been proposed to solve this problem. Within this section, the paper introduces an innovative resolution founded on the methodology of the simple linear regression model in addressing NULL values, coupled with hedge algebra incorporating linguistic values.

Handling Missing Values

Linguistic Values

We consider the attribute's value domain as an hedge algebra and transform the quantity values to the corresponding values in [0, 1], defined as follows:

Let X_{salary}= (X_{salary}, G_{salary}, H_{salary}, £) is hedge algebra, with G_{salary}= *{high, low}*, $H^{+}{}_{salary}$ = *{very, more}*, $H^{-}{}_{salary}$ = *{ability, less}*, *very > more và less > ability*.

Choose W_{salary} = 0.5, *fm(low)* = 0.5, *fm(high)* = 0.5, *fm(very)* = 0.2, *fm(more)* = 0.3, *fm(ability)* = 0.3, *fm(less)* = 0.2, và Dom(*salary*) = [760, 1500].

We have *fm(very low)* = 0.1, *fm(more than low)* = 0.15, *fm(less low)* = 0.1, *fm(likely low)* = 0.15. Since very low < more low < low < low possibility < less low, we have *I(very low)* = [0, 0.1], *I(more than low)* = [0.1, 0.25], *I(low possibility)* = [0.25, 0.4], I*(less low)*=[0.4, 0.50]. *I(less high)* = [0.50, 0.60], *I(high possibility)* =[0.60, 0.75], *I(more likely)* = [0.75, 0.90], *I(very high)* = [0.90, 1].

From definition 1.6, we can calculate the semantic value of the words as follows: v(very low)=0.05; v(higher low)=0.175; v(low)=0.25; v(low possibility)=0.325; v(low screw)=0.45; v(high screw)=0.55; v(high possibility)=0.675; v(high)=0.75; v(higher)=0.825; v(very high)=0.95.

After converting the salary attribute values to the range [0,1], and then determining which language those values belong to, we find that the average of the

Table 3: Correlation matrix between attributes

id	age	bmi	chol
age	1.000000	0.691420	0.936726
bmi	0.691420	1.000000	0.693392
chol	0.936726	0.693392	1.000000

Table 4: Employee database structure

ENO	Age	Dept	Gender	Skill	WorkinYear	Salary	TrainedYear	OfficeCity
E001	29	HR	Female	SQL	3	833.238061	3	Danang
E002	39	IT	Male	Java	7	1459.62983	7	Danang

Table 5: Representing the NULL value and Language of the Salary attribute

ENO	Age	Skill	WorkinYear	Salary	TrainedYear
E006	36	C#	5	Null	3
E028	28	C#	3	Null	5
E037	31	C#	5	Null	4
E043	23	Python	3	less high	3
E052	36	C#	5	less low	3

15 tables belongs to the language *'more than low'* is **888.740** and the average of the 19 tables belonging to the *'higher'* language is **1364.300.**

So the salary values of the two employees whose corresponding language values are filled in are E043 = 1364.300 and E052 = 888.740

NULL Value

We build the regression equation as follows:

Step 1: Determine the correlation between the attributes in Employee and Salary

Step 2: From figure 3, we see that the TrainedYear attribute is the strongest correlation with the Salary attribute.

Step 3: Build a linear regression equation Salary=$\alpha+\beta$TrainedYear

Linear regression equation: Salary = 166.949*Trained Year + 486.139.

Step 4: Fill in the missing Salary value for three employees E006, E028 and E037. The results are shown in table 6

Data Clustering

After the data has been preprocessed in step 3.1. We conduct clustering using weka software and the results are as follows

=== Run information ===

Relation: Employee_data - missing values

Instances: 92

Attributes: 9

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 9

Within cluster sum of squared errors: 0.27968034577371803

Initial starting points (random):

Cluster 0: 1222.260515

Cluster 1: 1425.293587

Cluster 2: 1412.100438

Cluster 3: 1263.193346

Cluster 4: 1191.312046

Final cluster centroids: Cluster#

Attribute

Full Data 0 1 2 3 4

(92.0) (10.0) (16.0) (18.0) (15.0) (33.0)

==============================

Salary 1153.9342 1162.845 1449.2437 1346.8959 1243.6651 862.015

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 10 (11%)

1 16 (17%)

2 18 (20%)

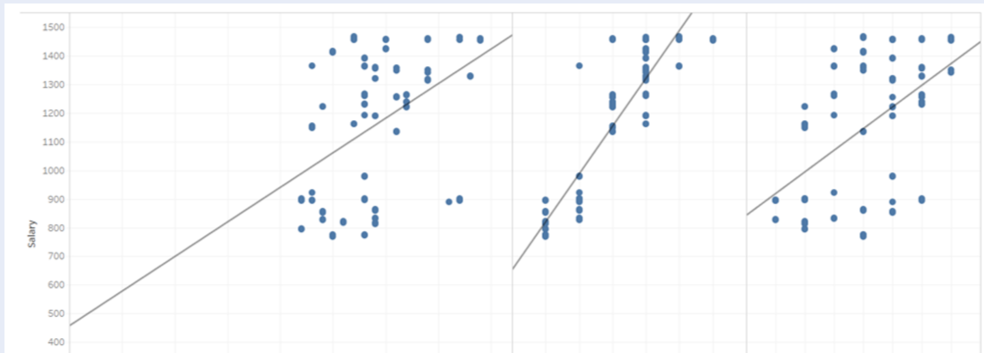

Figure 3: The correlation is shown on the trendline chart.

Table 6: Salary data of 2 employees is complete

ENO	Age	Skill	WorkinYear	Salary	TrainedYear
E006	36	C#	5	986.986	3
E028	28	C#	3	1320.884	5
E037	31	C#	5	1153.935	4
E043	23	Python	3	1364.3	3
E052	36	C#	5	888.74	3

3 15 (16%)

4 33 (36%)

CONCLUSION

The data mining process is a complex process that includes data as well as computing technologies. In particular, data preprocessing is the most important step because the collected data can be considered unclean, missing or incomplete. The article proposed a new method combining the hedge algebra and the linear regression for data preprocessing. This combination ensures the most complete handling of attribute values with incomplete, inaccurate or uncertain information. With the hedge algebra approach, based on the semantic quantitative values, viewing the attribute as a hedge algebra structure makes the processing of linguistic attribute values simple and effective. With the linear regression approach in statistical theory, determine the correlation between attributes and thereby build a regression equation for handling Null values. Finally, with applying the clustering method in data mining after using two approaches of of hedge algebra and linear regression, the data can be cleaned.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest in publishing the article.

AUTHOR'S CONTRIBUTION

Phu Phuoc Huy: Ideas for articles.

Doan Van Thang: Research and write drafts, present at conferences.

Hoang Tuan, Nguyen Xuan Nhut: Edit formatting and check for errors.

REFERENCES

1. Yu L, Liu L, Peace KE. Regression multiple imputation for missing data analysis. Stat Methods Med Res. 2020;29(9):2647-64;PMID: 32131673. Available from: https://doi.org/10.1177/0962280220908613.
2. Crambes C, Henchiri Y. Regression imputation in the functional linear model with missing values in the response. J Stat Plan Inference. 2019;201:103-19;Available from: https://doi.org/10.1016/j.jspi.2018.12.004.
3. Dominic Edelmann, Tamás F. Móri, Gábor J. Székely, 'On relationships between the Pearson and the distance correlation coefficients', Statistics & Probability Letters. Vol 169, February 2021, 108960;Available from: https://doi.org/10.1016/j.spl.2020.108960.
4. G.Jay Kens. Introduction to Probability and Statistics Using R, First Edition. 2010;.
5. Stekhoven DJ. missForest: nonparametric missing value imputation using random forest. Astrophysics Source Code Library. 2015;1505;.
6. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min ASA Data Sci J. 2017;10(6):363-77;PMID: 29403567. Available from: https://doi.org/10.1002/sam.11348.
7. Doan Van Thang, Doan Van Ban, 'Defining membership function based on approach to hedge algebras'. Journal of computer science and cybernetics, Vol 31, No 4 (2015), pp. 277-289;Available from: https://doi.org/10.15625/1813-9663/31/4/6189.

8. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20(1):1-12;PMID: 32711455. Available from: https://doi.org/10.1186/s12874-020-01080-1.
9. Dutta S. Approximate reasoning by analogy to answer null queries. Int. J. of Appr. Reasoning, (5) (1991), 373-398;Available from: https://doi.org/10.1016/0888-613X(91)90018-H.
10. Wang, S.L, and Huang, T.J., 'Analogical reasoning to answer null queries in fuzzy object-oriented data model', Proc. of the 6th IEEE Inter. Conf. on Fuzzy Systems, Barcelona, Spain, July 1-5 (1997), pp. 31-36;.

Tạp chí Phát triển Khoa học và Công nghệ – Engineering and Technology 2024, ():1-1

Open Access Full Text Article

Bài nghiên cứu

Phân cụm dữ liệu mờ theo tiếp cận đại số gia tử và mô hình hồi quy

Phù Phước Huy[1], Đoàn Văn Thắng[2,*], Hoàng Tuấn[1], Nguyễn Xuân Nhựt[3]

TÓM TẮT

Phân cụm mờ đã được khám phá một cách sâu rộng qua nhiều phương pháp khác nhau, mang lại các kết quả đa dạng trong lĩnh vực khai thác dữ liệu. Sự đa dạng trong kết quả nghiên cứu cho thấy sự phức tạp có sẵn trong việc khai thác dữ liệu mờ, đặc biệt khi đối mặt với các loại dữ liệu đa dạng nhằm phân định sự liên kết của các đối tượng với các cụm cụ thể. Sự phức tạp này càng được gia tăng khi dữ liệu không đầy đủ, thường được gọi là dữ liệu thiếu, trở thành một thách thức đáng kể trong lĩnh vực này. Việc giải quyết vấn đề giá trị thiếu trở nên quan trọng để cải thiện một cách tinh tế và chính xác hơn cho việc phát triển phân cụm mờ.

Để đối mặt với những thách thức này, một phương pháp mới đã xuất hiện, tận dụng sự tương hợp giữa đại số hedging và mô hình hồi quy tuyến tính. Phương pháp đổi mới này cố gắng vượt qua những sự phức tạp liên quan đến các loại dữ liệu đa dạng và giá trị thiếu. Bằng cách tích hợp các nguyên tắc đại số với các kỹ thuật hồi quy tuyến tính, phương pháp được đề xuất giới thiệu một khung nhìn mạnh mẽ để phân loại các đối tượng trong một cụm. Sự kết hợp của những công cụ toán học này cung cấp một giải pháp duy nhất không chỉ điều hướng qua các phức tạp của việc khai thác dữ liệu mờ mà còn giải quyết vấn đề phổ biến về dữ liệu thiếu.

Bài báo đi sâu vào ưu điểm của việc áp dụng đại số hedging và mô hình hồi quy tuyến tính song song, trình bày một phương pháp toàn diện đóng góp đáng kể vào việc làm rõ sự tinh tế của phân cụm mờ. Sự tương tác hợp tác giữa nguyên tắc đại số và mô hình hồi quy không chỉ nâng cao độ chính xác của việc phân loại đối tượng trong các cụm mà còn cung cấp một chiến lược mạnh mẽ để xử lý các giá trị thiếu trong tập dữ liệu. Phương pháp tích hợp này đại diện cho một bước tiến đáng chú ý trong lĩnh vực phân cụm mờ, mang lại một giải pháp toàn diện và hiệu quả hơn đối với những thách thức phức tạp do các loại dữ liệu đa dạng và vấn đề phổ biến về dữ liệu thiếu.

Từ khoá: hồi quy tuyến tính, lý thuyết thống kê, missing data, đại số gia tử, khai phá dữ liệu

[1]*Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ quân sự, Việt Nam.*

[2]*Trường Đại học Công nghiệp Thành Phố Hồ Chí Minh, Việt Nam*

[3]*Trường Cao đẳng Công thương Thành Phố Hồ Chí Minh, Việt Nam,*

Liên hệ

Đoàn Văn Thắng, Trường Đại học Công nghiệp Thành Phố Hồ Chí Minh, Việt Nam

Email: vanthangdn@gmail.com

Lịch sử

- Ngày nhận: 15-9-2023
- Ngày chấp nhận: 26-4-2024
- Ngày đăng:

DOI :

Trích dẫn bài báo này: Huy P P, Thắng D V, Tuấn H, Nhựt N X. **Phân cụm dữ liệu mờ theo tiếp cận đại số gia tử và mô hình hồi quy** . *Sci. Tech. Dev. J. - Eng. Tech.* 2024; ():1-1.