Microsystem Technologies (2022) 28:2715–2721
https://doi.org/10.1007/s00542-022-05310-y

TECHNICAL PAPER

Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis

Pankaj Kumar Jadwal[1] · Sunil Pathak[2] · Sonal Jain[1]

Received: 11 November 2021 / Accepted: 27 April 2022 / Published online: 9 May 2022

Abstract

This research concentrates on segmenting credit card clients of Taiwan into optimal groups. Unsupervised Learning plays a significant role in dividing customers into similar groups based on several parameters. If customers are clustered in groups optimally, it leads towards the retrieval of better precision from machine learning models applied to customers associated with the clusters. Different machine learning algorithms (Linear discriminant analysis, logistic regression and random forest) were applied on the obtained clusters through K-means, hierarchical and HK Means clustering algorithm, and predictive accuracy is compared with the accuracy obtained via applying mentioned machine learning models on the whole dataset. In this paper, a novel approach of combining K Means and hierarchical clustering (HK Means) is used. In this approach, HK means clustering algorithms are applied on the factorial coordinates, obtained from multiple correspondence analyses for segmenting customers into optimal groups has been proposed. The accuracy of the clustering techniques is evaluated from the decomposition of inertia. The results demonstrate that the combination of K Means and hierarchical clustering proved to be optimal clustering techniques for customer segmentation which can be used further for applying Machine Learning techniques for credit risk analysis.

1 Introduction

Optimal assessment of Credibility of the borrowers is one of the most crucial tasks in the banking sector. Accurate classification and clustering of customers is a challenging and nontrivial task which contains multidimensional analysis of various factors of customer (Danenas and Garsva 2012). Banks might divide customers into different groups based on the financial and personal information. Credit scoring is a numerical expression to assess the credibility of the borrowers (West 2000). Credit scoring provides the foundation through which probability of providing bad loans can be eliminated. Credit risk scorecard consists of a group of characteristics that are used to give a credit score to a customer indicating their loan repayment risk level. Parameters which may affect credit score are illustrated in Fig. 1 and range of credit score which defines the creditability of the borrower is illustrated in Fig. 2 given below.

This credit scoring models discriminate borrowers through threshold value.Credit scoring is not a single stage model; It involves multiple stages (Wah et al. 2011):

- Development of the statistical model based on historical data of loan seekers
- Calculation of credit score from model
- Calculating and measuring the accuracy of the model
- Analysis and identify the indicators which may improve business performance (Siddiqi 2006).

A vast number of different classification algorithms for credit risk evaluation have been proposed in the literature (Jiang et al. 2018; Xiao et al. 2016). Classification technique can be categorized into two types: Statistical techniques and Artificial intelligence techniques (Wang et al. 2012). Predictive accuracy, resource efficiency and easy to use are the benefits of using statistical methods for credit risk evolution model (Hens and Tiwari 2012). A transformation in technology has been observed from statistical techniques to Artificial intelligence techniques in last three

✉ Pankaj Kumar Jadwal
pankajjadwal@gmail.com

Sunil Pathak
sunilpath@gmail.com

Sonal Jain
sonaljain@jklu.edu.in

[1] JK Lakshmipat University, Jaipur, India

[2] Department of Computer Science and Engineering, Amity University, Jaipur, India

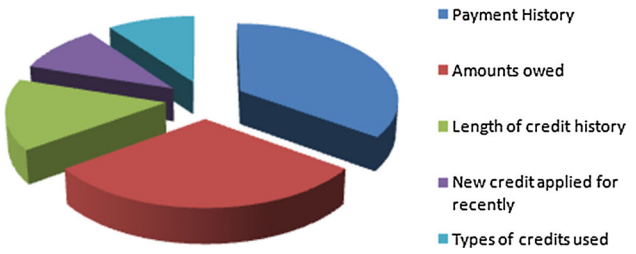

Fig. 1 Parameters affecting credit score

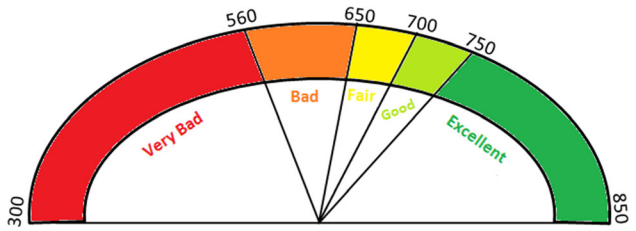

Fig. 2 Range of Credit Score

decades. Most popular methods in statistical techniques are logistic regression and linear discriminant analysis (Eisenbeis 2015). Advantages of using statistical techniques are easy to use and generate interpretable results, but these techniques do not provide promising results in the real-world dataset where data is not linearly separable.

Customer segmentation into clusters is another crucial and significant issue in the domain of credit risk analysis. Clustering of credit card holders indicates the tendency of customers towards their payment style which can help lenders in deciding risk associated lending amount. Clustering may be used for segmenting customers into groups in such a way that intra-class similarity index should have a higher value whereas interclass similarity should possess a smaller value. Most researchers have concentrated upon building a precise credit scoring model to choose whether or not to grant credit to new candidates. To strengthen consumer behavioural management for existing credit cards customers, Hsieh (2004) proposed a mining and behavioural scoring model to control the prevailing credit card customers. Performance assessment is the most significant parameter in credit risk evaluation. Performance of the evaluation parameters in supervised learning is accessed using accuracy and precision. Assessment of the Performance of the evaluation parameters in unsupervised learning is more complex than supervised learning due to the complex nature of cluster analysis (Witten and Eibe 1999). The objective of the proposed work is to segment the credit card holders into optimal clusters and serve it as an input for establishing better financial risk management. The proposed model has two processing phases. In the initial phase, the samples of preprocessed: removal of missing values, reformat categorical variables and recode continuous variables . then, credit cards holders are grouped into optimal clusters and machine learning models are applied on the dataset as well as on clusters.Further, predictive accuracy is obtained and compared in the both scenario.

The remaining of the research paper is structured as follows. Next section discusses the related works in credit risk evaluation using unsupervised and supervised methods. In Sect. 3, Proposed approach is elaborated. In Sect. 4, experimental results are discussed. Finally, Sect. 5 concludes the paper.

2 Related work

Credit risk evaluation is the area where clustering can be applied in the preprocessing phase to make groups of similar borrowers and such groups can assist machine learning models to deal in a more significant way. Machine learning algorithms are playing a significant role in credit risk evaluation. Several classification algorithms have been applied to predict the creditworthiness of the borrowers. Credit datasets contain qualitative and quantitative information of the borrowers. Instead of building machine learning models from the credit datasets, models can be

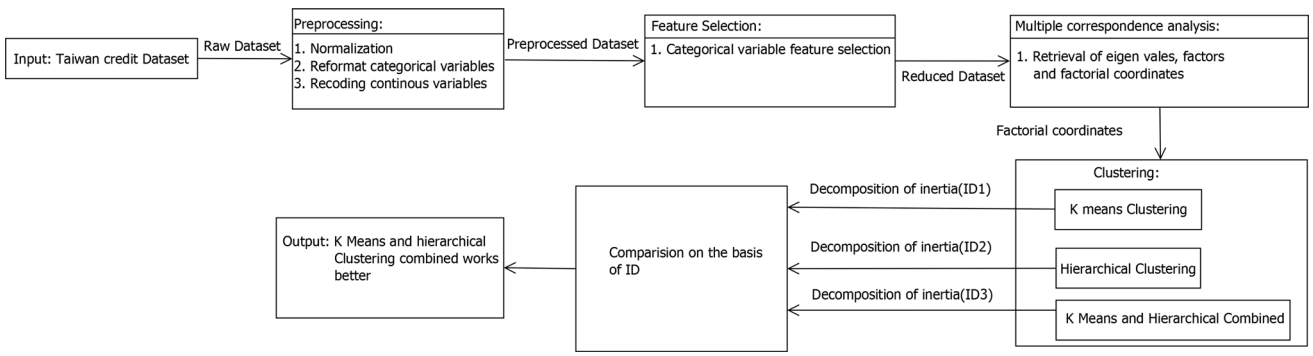

Fig. 3 Comparison of clustering algorithms w.r.t. credit card holders dataset

built from the segmented group of the borrowers. This will assist in generating more precise predictive outcomes, and generation of noise can be combated. Clustering algorithms have been extensively used in credit risk analysis domain (Wang 2010). Brockett et al. (1998) applied Kohonen's self-organizing feature map to detect frauds in automobile body injury claims. The validity of the proposed feature map approach is checked via backpropagation algorithm and feedforward neural networks.

Williams and Huang (1997) proposed the hot spot technology in which firstly the mining was performed on the dataset and later obtained models were explored to find essential nuggets. Two significant methods (data-driven approach based on hierarchical clustering and heuristic method) were used (Yeo et al. 2001) to predict the claim cost in the automobile industry and showcased benefits of employing the data-driven approach.

Cluster validation is one of the most prominent parameters that need to be considered in clustering analysis. Cluster analysis is a complex process because of lacking objective measures. Cluster validation is divided into three major types: internal examination, external assignment and relative test (Kou et al. 2014). In the external assessment, predicting labels are compared with actual data labels. This approach is used for choosing the optimal and right clustering algorithm for a related dataset but it requires knowledge of the internal structure of the data for evaluation of the clustering process which is normally missing in the dataset. It utilizes the internal information of the clustering procedure to look over the optimality of the cluster. Internal examination evaluates the quality of the clustering process via increasing the intra-class similarity and decreasing the interclass similarity for obtaining the optimal clusters (Fraley 1998). Relative cluster validation evaluates the optimality of the clusters by changing different parameters values for the same algorithm (Sarle et al. 1990). Other subjective criteria, such as computational complexity, interpretability (Berkhin 2006) and visualization (Bittmann and Gelbard 2007; Raveh 2000) can be used as an evaluation process for the clustering process. Apart from these three types of methods, evaluating clustering algorithms of a fuzzy partitioning of data or using fuzzy criteria have also been examined (Pedrycz and de Oliveira 2008; Pedrycz 2008).

Xiao et al. (2016) proposed a novel hybrid approach by integrating supervised and unsupervised models to achieve higher precision. Pankaj Jadwal et al. applied clustering algorithms on the credit datasets to obtain optimal clusters of credit card applicants and further state of art machine learning algorithms were applied on the clusters and obtained predictive accuracy is compared with the predictive accuracy of the models which applied on the complete dataset (Jadwal et al. 2019, 2017). Researchers (Bu 2017) came with the combination of a high-order k-means clustering algorithm with the dropout deep neural network model to segment the dissimilar records obtained from cyber-physical-social systems.

In order to preprocess the dataset, hybrid combination of clustering and classification is very popular. Researchers (Caruso et al. 2021; Niu et al. 2020) came with the different angles where clustering is applied on the credit dataset to refine the dataset. Most of the time, single clustering algorithm was applied over there. We introduced a novel approach where hybrid combination of two popular clustering algorithms(K-Means and hierarchical clustering) which provide assistance to classification algorithm.

3 Methodology used: comparison of clustering techniques applied on factorial coordinates obtained from multiple correspondence analysis

In the proposed approach, Taiwan credit dataset is taken for clustering. Firstly, the dataset is preprocessed. Categorical formats are reformatted, continuous variables are recoded and numerical variables are min-max normalized in the preprocessing stage. In the second stage, prominent categorical features are retrieved by applying boruta feature selection algorithm on the dataset.All the attributes except education are selected and further used for designing machine learning models. Further, Multiple correspondence analysis is applied on the final preprocessed dataset and eigen values, factors and factorial coordinates are obtained. Finally, K Means clustering algorithm, hierarchical clustering algorithm and proposed HK Means is applied on the factorial coordinates and quality of the clustering output is evaluated by decomposition of inertia.

Input: Taiwan credit dataset Data was taken from a reputed bank of Taiwan (a cash and credit card issuer) in Taiwan and main objective was to concentrate of credit card applicants. Total number of credit card applicants were 25,000 observations, where 5529 observations (22.12%) are the defaulted credit card applicants. Dataset consists of a binary variable—default payment (Yes = 1, No = 0), as the response variable and the following 23 variables are used as explanatory variables. Description of the attributes is provided in Table 1 given below:

Step 1: Pre-processing of dataset: Preprocessing of the dataset was done by applying three operations.

(a) Normalization of the dataset: Min–Max normalization was used for normalization of the dataset. In Min–Max normalization process, the minimum value of the attribute was used as centre and subtraction of maximum value with minimum value was used as the range of dataset. A1, A5,

Table 1 Description of input features

Input features	Description
A1	Amount of the given credit (NT dollar)
A2	Gender (1 = male; 2 = female)
A3	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
A4	Marital status (1 = married; 2 = single; 3 = others)
A5	Age (year)
A6–A11	History of past payment
A12–A17	Amount of bill statement (NT dollar)
A18–A23	Amount of previous payment

Table 2 Factor levels of categorical variables

Categorical variable	Factor levels								
Sex	Male	Female							
Marriage	Married	Single	Others						
PAY_0	Paid duly	Paid ontime	Delay 1	Delay2	Delay3	Delay4	Delay5	Delay6	Delay7
Education	High School	Graduate	University	Others					

A6–A11, A12–A17, A18–A23 are the attributes where normalization is applied.

(b) Reformat categorical variables: Categorical variables Sex, Marriage, PAY_0 and Education were reformatted. Categorical variables were specified as factors and factor levels are changed to meaningful values. Factor levels of categorical variables were shown in Table 2.

(c) Recoding of continuous variables: Continuous variables were recoded on the basis of 10 cut levels. Cut levels are [0,.1], [.1,.2], [.2,.3], [.3,.4], [.4,.5], [.5,.6], [.6,.7], [.7,.8], [.8,.9] and [.9,.10]. Continuous variables are categorized and recoded, now multiple correspondence analysis can be applied to the preprocessed dataset.

Step 2: Feature selection: The dataset contains 15 categorical variables (SEX, MARRIAGE, Age, PAY_0, PAY_2,PAY_3,PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6). Extraction of prominent features which impact on the response variable (default), is the essential task in feature selection. For getting significant features, the chi-square test was applied to the categorical variables of the data set and concern p values were generated. Variables are sorted in ascending order of p values and The significance level is assumed to be .05 and the NULL hypothesis is rejected if p values are smaller than or equal to significant level. Ranking of Associations with payment default is obtained by ordering the continuous variables according to P values. P values of all categorical variables are less than significant level so that all variables were accepted.

Table 3 K means clustering having K = 5, 6, 7, 8

Number of clusters	BSS	WSS	Id
5	49,083.48	39,808.42	55.21
6	50,530.19	38,361.71	59.84
7	71.447.84	17,444.07	80.37
8	71449.32	17,441.32	80.32

Step 3: Multiple correspondence analysis: The dataset contains 9 categorical variables and 21 continuous variables are converted into categorical variables. The aim is to apply clustering on the dataset to segment credit card holders. To identify feature attributes from categorical variables, Multiple correspondence analysis is applied. Multiple correspondence analysis (MCA) is applied to categorical and categorized continuous variables of the dataset and eigen values were obtained. Eigen values are represented by eig. After obtaining eigen values, significant dimensions in MCA were calculated. Dimensions associated with eigen values greater than (1/length(eig)), considered are significant dimensions. Here significant dimensions are 51.

Step 4: Clustering on factorial coordinates obtained from Multiple Correspondence Analysis:

(a) K Means Clustering: K means clustering was applied on factorial coordinates showcased in Table 3.7. Value of k was chosen as k = 5, k = 6, k = 7, k = 8. On the basis of elbow method, the value of k was chosen as 8 for further comparison with other clustering schemes. Decomposition

of inertia was taken as an evaluation parameter to check the quality of clustering algorithm.

$$Decomposition\ of\ inertia\ (Id) = \frac{(100 \times BSS)}{(BSS + WSS)} \quad (1)$$

where BSS = Between sum of squared error, WSS =Within sum of squared error.

As decomposition of inertia increases, quality of clustering also improves. Results obtained from K means clustering algorithm is shown in Table 3 which contains between sum of squared error (BSS), within sum of squared error(WSS) and decomposition of inertia. As per Table 3, Id increases as the number of clusters increases but after K = 8, no significant differences were observed in Id. So results obtained from elbow method (K = 8) were verified. In K Means clustering, most significant value of Id is obtained at K = 8.

(b) Hierarchical clustering: In hierarchical clustering, the dataset elements are not partitioned right into a particular cluster in a single step. Instead, some partitions require place, which might run from only one cluster containing all items belongs to n clusters that each contain only a single object. Hierarchical Clustering is divided into agglomerative methods, which proceed by some fusions of the n items into categories, and divisive methods, which separate n items successively into finer groupings. A step by step procedure was followed to perform hierarchical clustering.

- Distance matrix between individuals was calculated.
- Hierarchical clustering was applied to the distance matrix.
- The number of clusters was taken as 8. Table 4 shows the number of objects in each cluster in hierarchical clustering.
- Center of gravity of the clusters was calculated. Table 5 depicts centre of gravity of each cluster.
- Again, Decomposition of inertia was taken as an evaluation parameter for hierarchical clustering.

(c) Combined K Means and hierarchical clustering:

Table 4 Cluster wise objects

Cluster	Number of objects
1	14,896
2	3531
3	2825
4	102
5	2789
6	3153
7	1495
8	864

Table 5 Center of gravity of 8 clusters

Cluster	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1	− 0.28	− 0.24	0.44	− 0.26	0.05
2	− 0.08	− 0.04	− 0.13	0.21	− 0.07
3	− 0.22	− 0.18	0.23	− 0.06	− 0.01
4	7.79	3.03	7.41	4.79	0.38
5	0.06	0.10	− 0.62	0.71	− 0.22
6	0.24	0.27	− 1.04	0.97	− 0.21
7	0.81	0.76	− 1.41	− 0.12	0.61
8	2.46	1.97	− 0.38	− 2.33	− 0.16

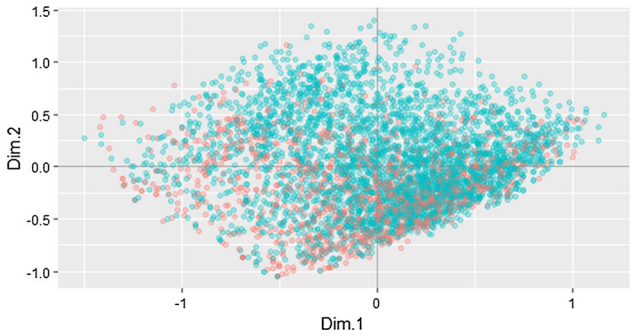

Fig. 4 MCA plot of individuals

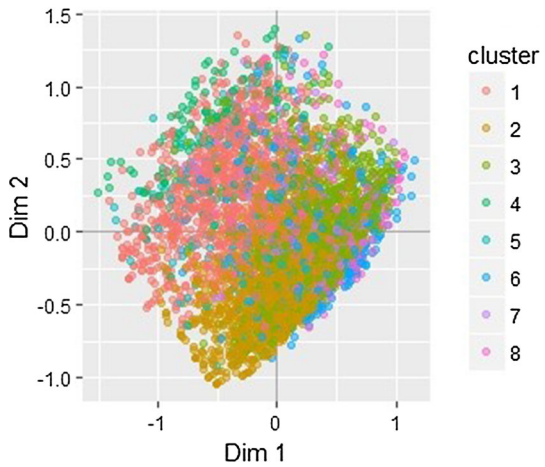

Fig. 5 MCA plot with cluster of individuals

- Hierarchical clustering is applied to the distance matrix using the centre of gravity from K-Means clustering. MCA plot of individuals and cluster of individuals were also shown in Figs. 4 and 5 respectively given below.
- Decomposition of inertia was taken as an evaluation parameter to check the quality of clustering algorithm.

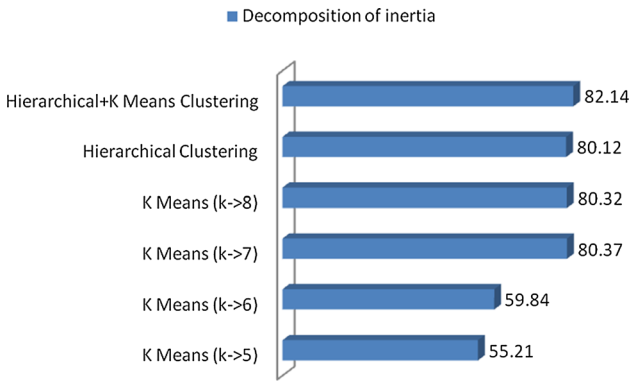

Fig. 6 Comparison of different clustering algorithms on the basis of decomposition of inertia

Output Clusters obtained from different clustering algorithms.

4 Results

Performance of different clustering algorithms is evaluated by results obtained from multiple correspondence analysis. Decomposition of inertia (Id) was taken as the evaluation parameter. Figure 6 compares K Means clustering, hierarchical clustering and combined clustering on factorial coordinates obtained from MCA results.

5 Conclusion and future scope

This paper proposes an approach in which different clustering algorithms were applied on the factorial coordinates, obtained from multiple correspondence analysis for segmenting credit card holders into optimal groups. The approach firstly preprocesses the dataset, and then multiple correspondence analysis was applied to the preprocessed dataset later on. K Means clustering (K = 5, 6, 7, 8), hierarchical clustering and combination of K Means and hierarchical clustering were applied on the preprocessed dataset and quality of the clustering algorithms was measured using the decomposition of inertia. The results show that the combination of K Means and hierarchical clustering with MCA results (factorial coordinates) outperforms other clustering algorithms. Credit card holders were segmented in an optimal way using the combination of K means and hierarchical clustering. Taking advantage of customer segmentation for improving the accuracy of machine learning algorithms is a future direction.

References

Berkhin PP (2006) A survey of clustering data mining techniques. Group Multidimens Data:25–71

Bao W, Lianju N, Yue K (2019) Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Syst Appl 128:301–315

Bittmann RM, Gelbard RM (2007) Decision-making method using a visual approach for cluster analysis problems; indicative classification algorithms and grouping scope. Expert Syst 24(3):171–187

Brockett PL, Xia X, Derrig RA (1998) Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. J Risk Insur 65(2):245

Bu F (2017) A High-order clustering algorithm based on dropout deep learning for heterogeneous data in cyber-physical-social systems. IEEE Access

Caruso G, Gattone SA, Fortuna F, Di Battista T (2021) Cluster analysis for mixed data: an application to credit risk evaluation. Socio-Econ Plan Sci 73:100850

Danenas P, Garsva G (2012) Credit risk evaluation modelling using evolutionary linear SVM classifiers and sliding window approach. Procedia Procedia Comput Sci 9:1324–1333

Eisenbeis RA (2015) Problems in applying discriminant analysis in credit scoring models, vol 4266

Fraley C (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 41(8):578–588

Hens AB, Tiwari MK (2012) Computational time reduction for credit scoring: an integrated approach based on support vector machine and stratified sampling method. Expert Syst Appl 39(8):6774–6781

Hsieh NC (2004) An integrated data mining and behavioural scoring model for analyzing bank customers. Expert Syst Appl 27(4):623–633

Huang X, Liu X, Ren Y (2018) Enterprise credit risk evaluation based on neural network algorithm. Cogn Syst Res 52:317–324

Jadwal PK, Jain S, Gupta U, Khanna P (2017) March. K-Means clustering with neural networks for ATM cash repository prediction. In: International conference on information and communication technology for intelligent systems. Springer, Cham, pp 588–596

Jadwal PK, Jain S, Gupta U, Khanna P (2019) Clustered support vector machine for ATM cash repository prediction. Progress in advanced computing and intelligent engineering. Springer, Singapore, pp 189–201

Jiang H, Ching W-K, Yiu KFC, Qiu Y (2018) Stationary Mahalanobis kernel SVM for credit risk evaluation. Appl Soft Comput 71:407–417

Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci (NY)

Liu C, Xie J, Zhao Q, Xie Q, Liu C (2019) Novel evolutionary multi-objective soft subspace clustering algorithm for credit risk assessment. Expert Syst Appl 138:112827

Niu K, Zhang Z, Liu Y, Li R (2020) Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. Inf Sci 536:120–134

Pedrycz W (2008) Decoding through fuzzy clustering. IEEE Trans Instrum Meas 57(4):829–837

Pedrycz W, Izakian H (2014) Cluster-centric fuzzy modeling. IEEE Trans Fuzzy Syst 22(6):1585–1597

Pedrycz W, de Oliveira JV (2008) A development of fuzzy encoding and decoding through fuzzy clustering. IEEE Trans Instrum Meas 57(4):829–837

Raveh A (2000) Co-plot: a graphics display method for geometrical representations of MCDM. Eur J Oper Res 125(3):670–678
Sarle WS, Jain AK, Dubes RC (1990) Algorithms for clustering data. Technometrics 32(2):227
Shen F, Ma X, Li Z, Xu Z, Cai D (2018) An extended intuitionistic fuzzy TOPSIS method based on a new distance measure with an application to credit risk evaluation. Inf Sci (NY) 428:105–119
Siddiqi N (2006) Credit risk scorecards: developing and implementing intelligent credit scoring, vol 1
Song Y, Wang Y, Ye X, Wang D, Yin Y, Wang Y (2020) Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. Inf Sci 525:182–204
Valente de Oliveira J, Pedrycz W (2007) Advances in Fuzzy clustering and its applications
Wah B, Huat S, Huselina N, Husain M (2011) Expert systems with applications using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst Appl 38(10):13274–13283
Wang S (2010) A comprehensive survey of data mining-based accounting-fraud detection research. In: 2010 Int. conf. intell. comput. technol. autom. ICICTA 2010, vol 1, pp 50–53
Wang G, Ma J, Huang L, Xu K (2012) Two credit scoring models based on dual strategy ensemble trees. Knowl Based Syst 26:61–68
West D (2000) Neural network credit scoring models, vol 27
Williams GJ, Huang Z (1997) Mining the knowledge mine. Adv Top Artif Intell 1342(December):340–348
Witten I, Eibe F (1999) Data mining—practical machine learning tools and techniques with java implementations
Xiao H, Xiao Z, Wang Y (2016) Ensemble classification based on supervised clustering for credit scoring. Appl Soft Comput J 43:73–86
Xiao H, Xiao Z, Wang Y (2016) Ensemble classification based on supervised clustering for credit scoring. Appl Soft Comput J
Yang Y, Gu J, Zhou Z (2016) Credit risk evaluation based on social media. Environ Res 148:582–585
Yeo AC, Smith K, Willis RJ, Brooks M (2001) Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. Int J Intell Syst Acc Financ Manag 10(1):39–50