International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1336-1343
© Research India Publications. http://www.ripublication.com

A Framework for Identifying and Analyzing the Latent Risk Based on the Knowledge Discovery Techniques

Huan Doan
Research Scholar, Department of Information Systems
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam.
Orcid Id: 0000-0002-6998-0962

Bao Quoc Ho
Associate Professor, Department of Information Systems
HCMC University of Science, VNU-HCM, Ho Chi Minh City, Vietnam.

Dinh Thuan Nguyen
Associate Professor, Department of Information Systems
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam.

Abstract

In the global trend of competition becomes higher with every passing day, the risk threats also more easy to arise. Now, almost all transactions of the enterprise are recorded in the database. In database may be a risk which is latent such as the financial risk, operational risk, etc. The discovering the latent risk in the database is an important task of the risk management. In this paper, we propose an approach to identify and analyze the latent risk in the database based on developing and applying the knowledge discovery techniques. First, the authors builds a definition of latent risk and gives an illustration of the building of the function which identifies the latent risk, which is underlying for studying afterward. Next, we propose a framework for identifying and analyzing the latent risk in the database based on applying the knowledge discovery techniques. Then, a method to illustrate for the framework is introduced. The authors also build a strength level measure of clusters for integrating into a clustering algorithm to support for building the method.

We choose objects are the customers in the enterprise and choose the type of risk is the receivable debt payment risk (a form of financial risk) and choose the strength level measure of clusters is the risk level measure to experiment. The experiment and evaluation have been performed to a real data set in an enterprise. The results obtained by the proposed method based on the framework will assist the managers in risk management usefully.

Keywords: Latent risk, framework for identifying and analyzing the latent risk, ranking measures of clusters, method for identifying and analyzing latent risk.

INTRODUCTION

Nowadays, applying the knowledge discovery techniques in the computer science for identifying and analyzing the risk of objects that especially is customers in enterprises is interested from not only companies but also researchers.

Cheng and Chen [1] proposed a procedure, joining quantitative value of RFM (Recency–Frequency–Monetary) attributes and K-means algorithm into rough set theory (the LEM2 algorithm), to extract meaning rules for classifying the segmentation of customer value. Hanafizadeh and Paydar [2] presented a two-phase model named ‘Auto Insurance Customers Segmentation Intelligent Tool’ to segment customers in insurance companies on the basis of risk by utilizing the concept of a self-organizing map. In the works of Liang [3], the paper systematically integrates numerous data mining technologies to analyze customer value and thus promote customer value. First, the K-means and SOM methods are selected to perform customer value analysis and segment customers based on customer value. Secondly, the decision tree is used to mine the characteristics of each customer segment. Third, different strategies are developed for differently valued customer segments and customer value is thus promoted. In [4], Cao and Do apply the clustering algorithm CLOPE for discovering the money laundering transactions in the banking industry. Their work supports the banks in risk management. On the other hand, Rajagopal [5] used the Demographic clustering process to find the high-profit, high-value and low-risk customers. Farajian and Mohammadi [6] proposed a two-stage framework of customer behavior analysis using K-means clustering algorithm and Apriori association rule inducer.

Almost the above works either only identify or only analyze the risk of objects without both. Now, applying knowledge discovery techniques specifically the data mining techniques for the risk management problem is still challenging and disunion.

In this paper, the first, we suggest a formal definition of latent risk of the objects which is inside of data and illustrate the building of the function which identifies the risk of the objects for making basic for studying of ours. According to the Rob Gerritsen [16], the data mining techniques separate models into two classes: predictive and descriptive. The next, based on the function of the data mining models: predictive and descriptive, we propose a framework which is two stages to identify and analyze the latent risk in the database based on applying the knowledge discovery techniques. Specifically, stage 1 applies classification techniques (crisp risk case) or fuzzy logic (fuzzy risk case) for identifying the risk and stage 2 applies clustering techniques for analyzing the risk. The concept of the crisp risk is defined in the next section of this paper and the concept of the fuzzy risk is defined in [17].

The traditional clustering algorithms don't integrate a measure to rank clusters after clustering. That causes difficulty in the use of clustering results. Thus, the authors propose the building of a strength level measure of clusters to rank clusters obtained after clustering. Then, by adding the measure of the strength level of clusters into algorithm FCM-E [7] for building an improved fuzzy clustering algorithm named FCM-R. FCM-E is improved from algorithms FCM [8,9] by integrating the computing of an indicator determined the appropriate number of clusters. Thus FCM-R is an algorithm not only creating clusters, finding the suitable number of clusters of the data set but also ranking clusters based on this measure. The FCM-R will be used to build a method for illustrating the framework in the next section.

Finally, for illustrating the framework, a two-stage method to identify and analyze the latent risk in the database is presented. In the first stage-identifying risk, the authors build a function which identifies the risk of objects and use it to classify the objects into two groups: risk and no risk. This function may be built based on a classifying technique for best appropriate with the data set. The classifying is to determine the risk objects which are input data for stage 2 because we only need to consider the group of risk objects in the risk analysis stage. In the second stage-analyzing risk, the authors use FCM-R to cluster the risk objects into some appropriate clusters and then to rank these clusters based on the risk level measure of clusters. In this stage, the value of the condition attributes of objects is normalized to avoid the effect of different measures of attributes to the clustering result and the computation of the risk level measure of clusters.

We experiment the proposed method on a real data of the Hong Ky company group in Ho Chi Minh City, Vietnam. The obtained results of the method are the clusters for appropriate with input data set and they are ranked according to the risk level measure from high to low. The managers use this result for evaluating the risk and making the suitable decision in the risk management for minimizing loss and maximizing advantage. This paper extends our previous work in [10].

The remainder of the paper is organized as follows. The next section presents some definitions, that comprises: the building of a definition of latent risk, the illustration of the building of the function which identifies the latent risk. In the main part, we propose a framework which is two stages to identify and analyze the latent risk in the database. After that, the authors build the method based on this framework for illustrating the framework. The successive section presents the results and discussion. The last section is the conclusion of the paper.

SOME DEFINITIONS

Build a definition of crisp latent risk of objects

At this time, according to our knowledge, it has not yet had a mathematical formal definition of latent risk in the database. To support for ourselves study, the authors use classification of risk in the works of Tsumoto and Hong [11]. They classify the types of risk into four categories as follows: risk for aversion, risk for benefit, latent risk, mathematical risk. Therein, the latent risk is the risk which may be inferred by mining databases. In this work, we choose the type of the latent risk to study. Hence, the authors build a definition of crisp latent risk as follows.

It gives a set O consisting n objects: o_1, o_2,…o_n, with each o_i having an ID-Identify Definition attribute o_{i0}, k condition attributes o_{ij}, j=1..k. A subset $H \subseteq O$ called crisp latent risk if every element in H, which is assigned the risk label which signs value 1 (opposite value 0). It may use the function χ to describe the concept "belong". If $x \in H$, $\chi(x) = 1$, opposite if $x \notin H$, $\chi(x)=0$. The function χ called the characterized function of the risk set H.

Definition 1. It gives a universe set O, a risk set $H \subseteq O$, an object $x \in O$ is called the crisp latent risk if $x \in H$ or in other word $\chi(x) = 1$ and contrary an object x is called no risk if $x \notin H$ or in other word $\chi(x) = 0$ with χ is the characterized function of the risk set H.

There is whether to exist another function which is other with the characterized function χ, which determines a subset of risk objects in O. The authors will find such function in the next part.

An illustration of the building of the function which identifies the latent risk

Apply a classifying algorithm (e.g. ID3, See5, and so on) on a training data set and a test data set extracted from O as in Table 1, it obtains a decision tree.

Table 1: Data set

ID (identification number of customers) (O_{i0})	Total debt (O_{i1})	Average of the overdue number of days of debt (O_{i2})	Overdue total debt (O_{i3})	Bad total debt (O_{i4})
C00001	17930000	14	17930000	0
C00003	21114480	21	15086480	14502480
C00004	29736784	0	0	0
C00006	660000	3	660000	0
C00007	13353400	0	0	0
C00008	69918000	4	38121000	0
C00009	162975000	12	88690000	4000000
C00010	75000	29	75000	0
C00011	18755852	75	18755852	18755852
C00012	332000	169	332000	332000
C00013	7812000	47	7812000	3690000
…..				

The authors build the function which identifies the latent risk based on this decision tree to determine the risk for new objects.

The function has the form as follows:

$$f(o_i) = f(o_{i1}, o_{i2}, \ldots o_{ik}) = \begin{cases} \mathbf{0},\ \boldsymbol{n\acute{\hat{e}}u}\ \mathbf{o_{i2} \leq 75000\ or\ (o_{i2} > 75000\ and\ o_{i4} \leq 1056000\ and\ o_{i2} \leq 6685000\ and\ o_{i1} > 2312780)} \\ \mathbf{1},\ \boldsymbol{n\acute{\hat{e}}u}\ \mathbf{(o_{i2} > 75000\ and\ o_{i4} > 1056000)}\ \boldsymbol{or}\ \mathbf{(o_{i4} \leq 1056000\ and\ o_{i2} > 6685000)} \\ \qquad \mathbf{or\ (o_{i2} > 75000\ and\ o_{i4} \leq 1056000\ and\ o_{i2} \leq 6685000\ and\ o_{i1} \leq 2312780)} \end{cases}$$

Here o_i is the new object and o_{i1}, o_{i2}, o_{ik} are the condition attributes of o_i.

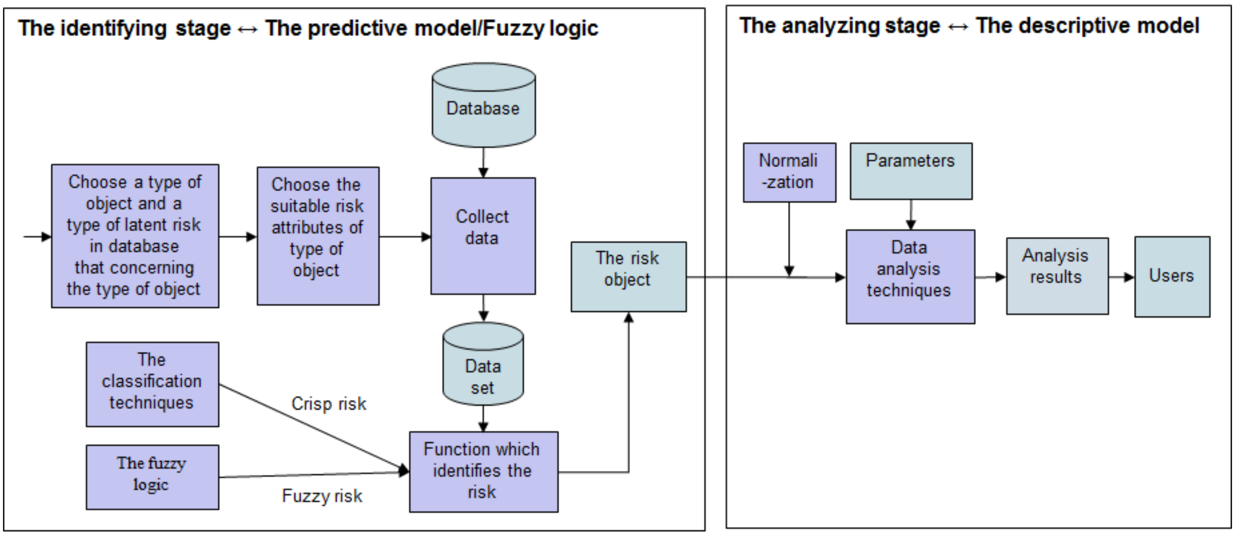
The identifying stage ↔ The predictive model/Fuzzy logic
Database
Choose a type of object and a type of latent risk in database that concerning the type of object
Choose the suitable risk attributes of type of object
Collect data
The risk object
Data set
The classification techniques
Crisp risk
The fuzzy logic
Fuzzy risk
Function which identifies the risk
The analyzing stage ↔ The descriptive model
Normali -zation
Parameters
Data analysis techniques
Analysis results
Users

Figure 1: Framework for identifying and analyzing the latent risk in the database

PROPOSE A FRAMEWORK FOR IDENTIFYING AND ANALYZING THE LATENT RISK

Based on the function of the data mining models: predictive and descriptive [16], we propose an framework which is two stages to identify and analyze the latent risk in the database.

The predictive model includes the classification and regression techniques. The descriptive model includes the clustering and association techniques. We find that the techniques of the predictive model, especially the classification techniques are suitable for identifying the risk and the techniques of the descriptive model, especially the clustering techniques are suitable for analyzing the risk. Terence [18] have also established a mapping between the clustering techniques and risk analyzing. Besides, fuzzy logic is also a good predictive technique. From this comment, we propose a framework to identify and analyze the latent risk by using the appropriate techniques of two models and fuzzy logic.

Stage 1 applies classification techniques for the type of the crisp risk and applies the fuzzy logic for the type of fuzzy risk to identify the latent risk. Stage 2 applies the clustering algorithms to analyze the latent risk. This framework orientates to build the methods for identifying and analyzing the latent risk in the database and depicts as in Figure 1. In [17], we have built a method for identifying and analyzing the fuzzy latent risk based on this framework. In the next section, for illustrating for the crisp risk case of the framework, we present the building a method for identifying and analyzing the crisp latent risk.

BUILD THE METHOD FOR ILLUSTRATING THE FRAMEWORK

Building of strength level measure of clusters to rank clusters

The traditional clustering algorithms normally only generate the clusters, but not rank the clusters. This is difficult for users to interpret and digest the clustering results. Up to this point, according to our knowledge, it has not yet had a work which comprehensively solves the ranking of clusters. In [12], Ziv Bar-Yossef et al only introduces a cluster strength measure for a special case, which is the integrated cohesion - which is applicable to arbitrary weighted networks. In order to address the above limit of the clustering algorithms, we propose the strength level measure of clusters to rank clusters after the clustering.

A cluster is built from objects so when making the strength level measure of clusters, we also are based on attributes of objects. Initially, the authors have to determine a ranking criteria concerned with objects in clusters. With objects is the customer in the enterprise, it can choose the ranking criteria as the profit, sales, or risk level. After determining the ranking criterion, we have to choose the attributes of objects that concern to this ranking criterion and assess these attributes interrelated positively or negatively on the ranking criterion. For example, with the ranking criterion as profit, the attribute: sales will positively interrelate, but the attribute: cost will negatively interrelate. Lastly, we build a formula of strength level measure of clusters according to the attributes which are chosen. In this paper, for easy illustration, we choose objects are customers in the enterprise and the ranking criterion is the risk level. With other objects and criteria, we can study to execute it similarly. The attributes of objects which are negatively interrelated and haven't the numeric data type also don't yet consider here. With the data set in Table 1, according to real knowledge and expert opinion, attributes of customer that positively interrelate with the risk level are Total debt, Average of the overdue number of days of debt, Overdue total debt, Bad total debt. After determining attributes which positively interrelate with the risk level, a risk level measure of an object can be built simply by adding the value of attributes of this object.

Suppose that G is a set of attributes of customer that positively interrelate to the risk level criteria and the clustering algorithm generates c clusters. We build the risk ranking measure of clusters as follows:

$$\boldsymbol{R_i} = \frac{\sum_{i=1}^{ni}\sum_{j\in G} A_j}{n_i}, \ i = 1..c \qquad (1)$$

Here R_i is the value of the risk level measure of the cluster i, A_j is the attribute that belongs the attribute group G, n_i is the number of objects in the cluster i.

Building of the improved fuzzy clustering algorithm FCM-R

With traditional clustering algorithm, the result obtained is the clusters which are not ranked. This will embarrass to the users in interpreting and using the result. To build an algorithm not only generating clusters, finding the suitable number of clusters of the data set but also ranking clusters, we additionally integrate the ranking measures of clusters into algorithm FCM-E [7]. With the risk level criterion, the ranking measures of clusters are built as in the expression (1). The improved algorithm is named FCM-R and it is presented as follows:

The algorithm FCM-R

Input: Data Set n objects x_i.

Output: c_{best} of data clusters ranked with c_{best} is the appropriate number of clusters

```
1) Input n objects $x_i$, fuzzy parameter m > 1, epsilon small enough.
2) Input weighted vector w.
3) Input $c_{min}$ and $c_{max}$ ($c_{min}$ >= 1, $c_{min} < c_{max} < n$).
4) Input the risk attribute group G
5) For c = $c_{min}$ to $c_{max}$
    a) Initialization matrix of members $U_{cxn}$
    b) Calculation of centre of cluster j $C_j$ (j = 1,..,c)
    c) Update the distance matrix D (c x n)
    d) Update matrix of members U
    e) If the change of the matrix U is small enough compared to the previous step, then go to step f) otherwise go to step b)
    f) Based on the matrix U, $x_i$ is arranged into clusters according to rules as follows: $x_i$ will belong to any cluster that it has the greatest degree
    g) Save the c of data clusters at step c
    h) Get all extremely marginal objects of all clusters
    i) Compute $\overline{\gamma}$ indicated the appropriate number of clusters
6) Select $c_{best}$ at neighborhood when $\overline{\gamma}$ get a local minimum value after $\overline{\gamma}$ fluctuates fast and may start to have a stable trend.
7) Take the $c_{best}$ of data clusters at step $c_{best}$
8) For i = 1 to $c_{best}$
        Compute $R_i$ based on (1)
9) Arrange clusters based on $R_i$, (i=1..$c_{best}$)
```

Evaluation of Computational Complexity of FCM-R

The computational complexity of the algorithm FCM is O(tcn). Where n is the number of objects in a data set, c is the number of clusters, t is the number of iterations. The computational complexity of the algorithm FCM-R is calculated as follows:

Set c_{max} is the number of iterations for selecting of a number of clusters, then based on J.C. Bezdek's suggestion [13], $c_{max} = \boldsymbol{\sqrt{n}}$, n is the number of objects.

$$O(FCM\text{-}R) = [O(FCM) + O(Step\ h) + O(\overline{\gamma})]*c_{max} + O(R_i)*c_{best}$$

By reason of $c_{best} <= c_{max}$, hence

$$O(FCM\text{-}R) = [O(FCM) + O(Step\ h) + O(\overline{\gamma})]*c_{max} + O(R_i)*c_{max}$$

$$O(FCM\text{-}R) = [O(FCM) + O(Step\ h) + O(\overline{\gamma}) + O(R_i)]*c_{max}$$

$$O(FCM\text{-}R) = [O(FCM) + O(Step\ h) + O(\overline{\alpha}) + O(\overline{\beta}) + O(R_i)]*c_{max}$$

Where O(FCM) is the computational complexity of FCM, O(Step h) is the computational complexity of step h.

$O(\overline{\alpha})$ is the computational complexity of $\overline{\alpha}$, $O(\overline{\beta})$ is the computational complexity of $\overline{\beta}$, $O(R_i)$ is the computational complexity of R_i in (1).

On an average, the number of objects in each cluster is $\frac{n}{c}$.

Set p is the number of extremely marginal objects in each cluster. On an average, it can assume $p \approx k$, with k is the number of dimensions.

So O(Step h) = c*(n/c)*k = n*k = O(nk)

$O(\overline{\alpha}) = c*p*(p-1) = c*k*(k - 1) = ck^2 - ck \ \le ck^2$.

Thus $O(\overline{\alpha}) \approx O(ck^2)$

$O(\overline{\beta}) = c*c*p = c*c*k = c^2k = O(c^2k)$

$O(R_i) = n_i*|G|$, on an average, the number of objects of cluster i is $\frac{n}{c}$, G is a set of attributes of the object that interrelate positively to risk level criteria so, $|G| <= k$.

Thus $O(R_i) = \frac{n}{c}*k$.

In fact, the number of clusters c normally is smaller n, the number of iterations t normally is smaller n, and the number of attributes k normally is smaller n. It can set: t = max (t, k, c).

O(FCM-R) =

$$[O(tcn) + O(nk) + O(ck^2) + O(c^2k) + O(\frac{n}{c}*k)]* c_{max}$$

$$= [O(tcn) + O(nk) + O(ck^2) + O(c^2k) + O(\frac{n}{c}*k)]*\sqrt{n}$$

$$\approx [O(nt^2) + O(nt) + O(t^3) + O(t^3) + O(\frac{n}{t}*t)]*\sqrt{n}$$

$$\approx [O(nt^2) + O(nt) + O(t^3) + O(t^3) + O(n)]*\sqrt{n}$$

$$\approx O(nt^2)*\sqrt{\boldsymbol{n}} \approx O(nt^2\sqrt{\boldsymbol{n}})$$

Here, it perceives that the computational complexity of FCM-R is higher than the computational complexity of FCM, but it is not much.

Proposes a method to identify and analyze the crisp latent risk in the database for illustrating the framework

In this section, we propose an method which identifies and analyzes the latent risk of objects includes two stages as follows.

Stage 1: The posterior steps are performed sequentially:

1. Select the object and the type of the latent risk which concerns object, which needs the risk management.
2. Choose the attributes of objects which are suitable for the goal of identification and analysis of risk.
3. Collect data and performs data preprocessing steps for building the data sets as the training data set, the test data set and such.
4. Apply the classifying techniques for building the function of identifying the risk.
5. Utilize this function to identify the latent risk of the objects by classifying the objects into two groups: risk and no risk.

After identifying the risk of the objects, we only need to analyze the risk level of the group of risk objects, hence it only needs to choose the group of the risk objects as an input of the second stage.

Stage 2: The posterior steps are performed sequentially:

1. To eliminate the effect of different measures of attributes to the clustering result and the computation of a risk level measure of clusters, the values of condition attributes of risk objects which get from stage 1 are normalized.
2. Apply the improved clustering algorithm FCM-R on the input data set. In building this method, the strength level measure of clusters in FCM-R is replaced by the risk level measure. The algorithm FCM-R performs some steps as follows:
   - Determine the appropriate number of clusters.
   - Cluster the objects into the appropriate number of clusters.
   - Calculate the risk level measure of clusters
   - Rank these clusters according to the risk level measure from high to low.

The results obtained are supplied to the managers to support them to make a decision in risk management. Diagram of the proposed method to identify and analyze the latent risk is depicted in Figure 2.

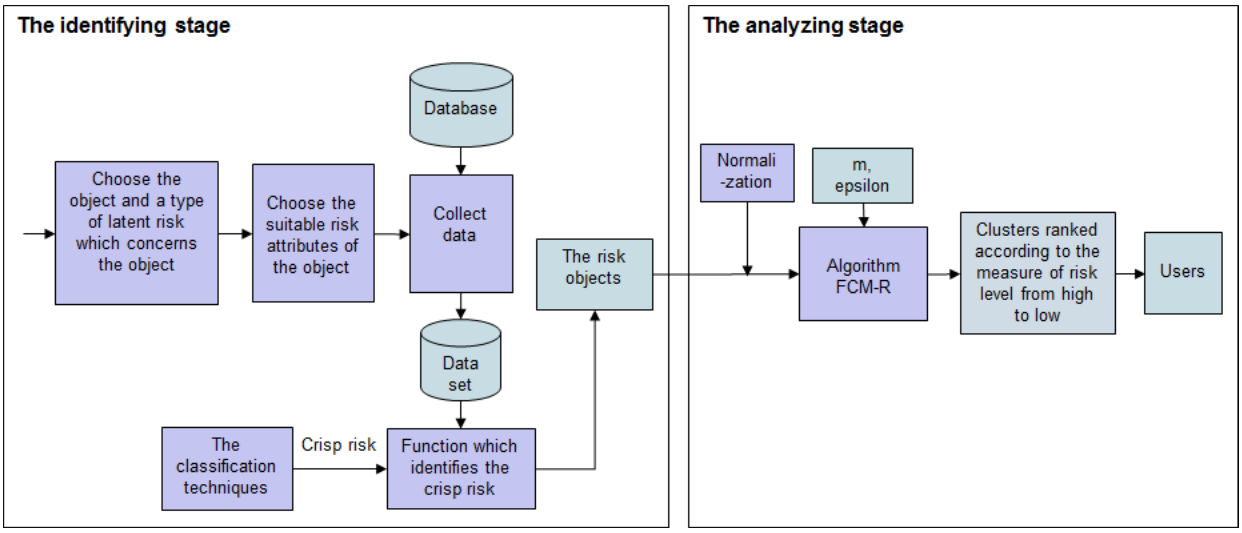
The identifying stage
Database
Choose the object and a type of latent risk which concerns the object
Choose the suitable risk attributes of the object
Collect data
The risk objects
Data set
The classification techniques
Crisp risk
Function which identifies the crisp risk
The analyzing stage
Normali -zation
m, epsilon
Algorithm FCM-R
Clusters ranked according to the measure of risk level from high to low
Users

Figure 2: Method for identifying and analyzing the crisp latent risk for illustrating the framework

RESULTS AND DISCUSSION

Experiment

The authors experiment the proposed method on a real data set extracted in 2014 from the database of software EnterERP of the Hong Ky company group in Ho Chi Minh City, Vietnam.

Stage 1:

The object which is selected to experiment is customers in the enterprise. The essential risk of enterprise that is concerning customers is the customer debt so attributes concerning the risk of customer also are the information about the debt of the customer. Based on the real knowledge and the domain expert opinion, the authors choose the attributes of customer such as: ID (identification number of customers), Total debt, Average of the overdue number of days of debt, Overdue total debt, Bad total debt (overdue total debt which is over 30 days) and a classification attribute: Risk.

The original data set has 2848 debt invoices of the customers. After preprocessing the steps, the data set has 443 records of the customer debt. From the above data set, it forms the training data set which has 250 records and the test data set which has 60 records. In both these data sets, based on the expert opinion and the real knowledge, set the Risk attribute which has the value: True or False.

Run the algorithm See5 (http://www.rulequest.com/see5-info.html, access: May 2014) with the input data is the training data set and the test data set, it obtains a decision tree. Based on the decision tree, it builds a function which identifies the risk. Utilize this function to identify the latent risk of 133 new customers by classifying them into 2 groups: risk or no risk. We have obtained 72 customers belonging the risk group and passed them to stage 2.

Stage 2:

After normalizing data received from stage 1, select the attributes of group G, which concerns to risk as follows: Total debt, Overdue total debt, Average of the overdue number of days of debt, Bad total debt.

Run algorithm FCM-R on the 72 customers data set belong to the risk group with the number of cluster c from 1 to 8 ($c_{\max} = 8 \approx \sqrt{72}$, based on J.C. Bezdek's suggestion [13]), parameters m = 2, epsilon = 0.0001, the weights of each condition attribute is 0.25, $\bar{\gamma}$ indicates the appropriate number of clusters that being 4.

After stage 2, the experiment result obtained by the proposed method is 4 clusters ranked according to the measure of risk level from high to low as in Table 2. The result will support the managers in the risk management as discussed in detail in next section.

Table 2: Clusters ranked according to the risk level measure from high to low

Risk Level Measure	Cluster	Customer ID	Total Debt	Average of the overdue number of days of debt	Overdue total debt	Bad total debt
42.98 (Very High Risk)	4	C00375	871,842,518	53	749,311,518	605,225,418
16.23 (High Risk)	3	C00402	100630198	757	100630198	63505000
		C00428	248053729	638	248053729	248053729
		C00429	291433992	638	291433992	291433992
		C00430	975000	1583	975000	975000
		C00433	92229600	496	92229600	92229600
		C00444	2500000	638	2500000	2500000
9.03 (Medium Risk)	2	C00440	454610465	7	149540506	0
		C00448	413814591	23	242944391	52635091
		C00443	384591982	17	131148182	49229182
		C00454	1067174224	14	527971524	0
		C00431	395218413	28	207261353	98518953
		C00395	227663183	39	111872183	111872183
		C00420	128112416	72	128112416	106017416
		C00386	437437000	3	437437000	0
		C00406	168867007	42	119322007	88902007
		C00322	267697720	57	195787720	137816840
		C00356	367735748	45	270853748	190078748

Risk Level Measure	Cluster	Customer ID	Total Debt	Average of the overdue number of days of debt	Overdue total debt	Bad total debt
-0.97 (Low Risk)	1	C00358	34846000	76	34846000	34630000
		C00361	195812060	11	46373020	0
		C00363	52316500	13	11300500	0
		C00364	33171800	9	20346800	0
		C00366	56389935	1	13009935	0
		C00367	58252670	57	58252670	4668920
		C00368	22207475	64	22207475	22207475
		……	………..	………..	…………	…………

Assessment of the clustering result

Based on the experiment, Wang et al [14] have assumed the indicator V_{PBMF} of Pakhira et al [15] is good in determining the number of clusters in fuzzy clustering. Therefore, the authors use indicator V_{PBMF} to assess the determining the number of clusters of algorithm FCM-R in the proposed method. The indicator V_{PBMF} is calculated as follows:

$$V_{PBMF} = \left(\frac{1}{c} \times \frac{E_1}{Jm} \times D_c\right)^2$$

where *c* is the number of clusters and here:

$$E_1 = \sum_{j=1}^{n} u_{ij} \left\|x_j - v\right\|$$

$$D_c = max_{i,j=1}^{c} \left\|v_i - v_j\right\|$$

$$J_m(\mathrm{U}, \mathrm{Z}) = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m \left\|x_i - x_j\right\|$$

n is the total number of patterns in the data set, $U(X) = [u_{ij}]c \times n$ is a partition matrix for the data and v_i is the centroid of the *i*th cluster.

Calculate indicator V_{PBMF} on the same data set as the above experiment with the same fuzzy parameter m = 2 and the number of clusters from 1 to 8, we obtain the result in Table 3.

With the result of the calculation in Table 3, indicator V_{PBMF} indicates the suitable number of clusters that is 4 as in Figure 3. This result is appropriate for the number of clusters of the indicator $\bar{\gamma}$ in algorithm FCM-R which is used in the proposed method.

Table 3: The result of calculation of index V_{PBMF}

Number of clusters	E1	Jm	Dc	V_{PBMF}
1	266	266	0	0
2	266	140	9	81
3	266	97	11	104
4	**266**	**83**	**25**	**417**
5	266	63	17	209
6	266	53	18	229
7	266	45	18	238
8	266	40	18	230

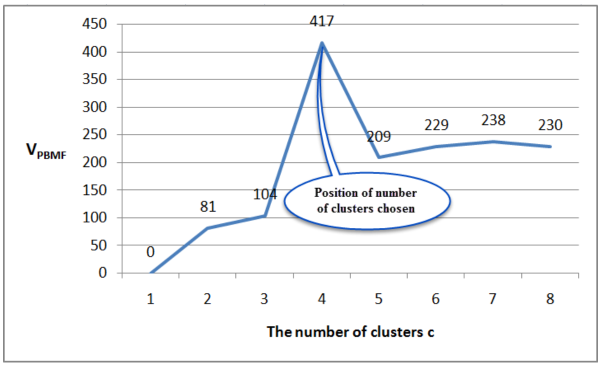

Figure 3: The diagram of index V_{PBMF} shows the appropriate number of clusters

Discussion

We realize that, in Table 2, the cluster 4 which has the risk level measure = 42.98 is the very high risk. The customers in the cluster 4 have values of attributes (Total debt, Overdue total debt, Bad total debt) which are much larger than the corresponding values of attributes of customers in the other clusters. For the customers in cluster 4, the manager should have the drastic solutions to minimize the damage that the firm can encounter.

The cluster 3 which has the risk level measure = 16.23 also is a high risk, although the risk level measure is the lower level. Especially, the customers in this cluster have the attribute "Average of the overdue number of days of debt" is extraordinary. Hence, the manager need consider the nature of this attribute to treat the customers in cluster 3 appropriately. In a nutshell, based on the results of the proposed method, the users can propound the appropriate solutions for each specific circumstance of each cluster to prevent loss skillfully.

CONCLUSION

The risk management is a survival of the enterprises. The identification and analysis of the risk are two important stages in the risk management. In the paper, we propose the two-stage framework to identify and analyze the latent risk by the knowledge discovery techniques. For illustrating the framework, a method to identify and analyze the crisp latent risk is introduced in this paper and a method to identify and analyze the fuzzy latent risk is also introduced in [17].

To support the building of method, the authors not only have suggested the definition of the latent risk and the function of identifying the latent risk but also built the strength level measure of clusters to rank clusters and integrated it into the improved fuzzy clustering algorithm, named FCM-R. The Algorithm FCM-R is used for analyzing the risk in stage 2 of the proposed method.

The authors have experimented with the proposed method based on the framework for the receivable debt payment risk (a form of financial risk) of customers of the enterprise. The results which obtain from experiment have shown the encouraging advantage of the method for the managers in the risk management.

REFERENCES

[1] Cheng C.H. & Chen Y.S. Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst Appl 2009; 36: 4176-4184. http://dx.doi.org/10.1016/j.eswa.2008.04.003

[2] Hanafizadeh P. & Paydar R.N. A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry. International Journal of Strategic Decision Sciences 2013; 4(1):52-78. http://dx.doi.org/10.4018/jsds.2013010104

[3] Liang Y.H. Integration of data mining technologies to analyze customer value for the automotive maintenance industry. Expert Syst Appl 2010; 37: 7489–7496. http://dx.doi.org/10.1016/j.eswa.2010.04.097

[4] D. K. Cao and P. Do. Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry. In: J.-S. Pan, S.-M. Chen, N.T. Nguyen, editors. Intelligent Information and Database Systems-LNCS vol 7197. Springer, Berlin, Heidelberg, 2012. pp 207-216.

[5] Rajagopal S. Customer Data Clustering Using Data Mining Technique, International Journal of Database Management Systems (IJDMS) 2011; Vol.3, No.4: pp.1-11, 2011. DOI: 10.5121/ijdms.2011.3401.

[6] Farajian A.M. & Mohammadi S. Mining the Banking Customer Behavior Using Clustering and Association Rules Methods. International Journal of Industrial Engineering & Production Research 2010; 21(4): 239-245.

[7] Doan H. & Nguyen D.T. A Method for Finding the Appropriate Number of Clusters, Int Arab J Inf Techn 2017 (Online Publication).

[8] Bezdek C.J., Ehrlich R. & Full W. FCM: The Fuzzy C-Means Clustering Algorithm. Comput Geosci 1984; 10(2-3): 191-203.

[9] Hathaway J.R. & Bezdek C.J. Recent Convergence Results for the Fuzzy c-Means Clustering Algorithms. Journal of Classification 1988; 5: 237-247.

[10] Doan H., Nguyen D.T. & Ho B.Q. Building a Measure to Integrate Into a Hybrid Data Mining Method to Analyze the Risk of Customer. In: H.A. Sulaiman, M.A. Othman, M.F.I. Othman, Y.A. Rahim & N.C. Pee, editors. Advanced Computer and Communication Engineering Technology-LNEE 362, Springer International Publishing Switzerland, 2015. pp. 843-851. http://dx.doi.org/ 10.1007/978-3-319-24584-3_71

[11] Tsumoto S. & Hong T.P. Special issue on data mining for decision making and risk management, J Intell Inf Syst 2011; 36:249–251.

[12] Ziv Bar-Yossef, Ido Guy, Ronny Lempel, Yoëlle S. Maarek, Vladimir Soroka. Cluster ranking with an application to mining mailbox networks. Knowledge and Information Systems, January 2008, Volume 14, Issue 1, pp 101-139.

[13] Bezdek C.J. Chapter F6: Pattern Recognition in Handbook of Fuzzy Computation. Boston, NY: IOP Publishing Ltd, 1998.

[14] Wang W & Zhang Y. On fuzzy cluster validity indices. Fuzzy Set Syst 2007; 158(14): 2095–2117.

[15] Pakhira K.M., Bandyopadhyay S. & Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition 2004; 37(3): 487-501.

[16] R. Gerritsen, "Assessing Loan Risks: A Data Mining Case Study," *IT Professional*, vol. 1, no. 6, pp. 16-21, 1999.

[17] Doan H. & Nguyen D.T. A Method for Identifying and Analyzing the Risk based on the Fuzzy Logic and the Algorithm FCM-R. Advanced Information Sciences and Service Sciences (AISS) 2017; Volume 9(2): 19-28.

[18] J. Terence, "Conceptual Mapping of Risk Management to Data Mining," in In Proceedings of the ICETET, PP. 636-641, © IEEE, 2010