Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 33 (2012) 1689 – 1694

Physics Procedia

2012 International Conference on Medical Physics and Biomedical Engineering

Outlier Detection Data Mining of Tax Based on Cluster[1]

Bin Liu[1]，Guang Xu[2], Qian Xu[2] and Nan Zhang[2]

[1]*Electrical and Information Engineering College*
Shaanxi University of Science and Technology
Xi' an, P. R. China
[2]*Department of Intelligent Robotics*
University of Huaguoshan
Huaguoshan, Jileshijie Province, China
blue-print@126.com；formated@126.com

Abstract

In order to solve the tax problem of mining industry of outlier data, analysis of the tax industry, the demand for data mining and data features, a clustering based data mining algorithms to solve the issue of tax discovery of outlier data, and an example to prove effectiveness of the algorithm.

Keywords - Tax; Data Mining; Outlier detect; clustering analysis.

Introduction

A database generally has some data does not meet the classification obtained by cluster analysis to predict or model, most of the data objects that do not meet the law posed by the data object model is called abnormal. However, the data in the tax work has great practical value. In the revenue management system, a sign of the tax shift data for key sources found in addition to screening, but also can be used to detect abnormal operating behavior of taxpayers, or even determine the existence of the taxpayers suspected of evasion and tax fraud, which speedy and accurate tax audit Suspects.

Data mining technology data mining, from the stored in databases, data warehouses or other large repository of data to discover useful knowledge [1-2]. The Pawlak Z. 1982 proposed rough set theory (Rough Sets Theory RST) [3] is an effective way to knowledge discovery. There Bayesian rough set (BRS) model and other knowledge discovery methods can solve the incomplete data problem [4-5].

Currently, cluster analysis and outlier detection is an important exception in data mining method [6], but the exception is not the same isolated point [7-9]. On the one hand, some data may appear to be a normal isolated points (the similar boundary points), so that isolated point is meaningless to users, but for

[1] This work is partially supported by education department of shaanxi provincial government special research project

doi:10.1016/j.phpro.2012.05.272

policy makers may be an opportunity or a sign; the other hand, in poly Among the data type is also possible that outlier data, especially for some continuity of data, the static clustering may be difficult for some of the abnormal performance of the outlier data, but also combined with Domain Expert Knowledge and other data mining Methods for further excavation, in order to find these hidden anomalies.

Tax is an important source of state revenue, the national economy healthy and sustainable development plays an important regulatory role. With the social development and technological advances, tax information to ensure revenue is becoming an important means for the successful completion of the work. Tax information development, some new problems have appeared, mainly in two aspects: First, from a technical point of view, with in-depth tax information, tax authorities at all levels of accumulation of a large number of business data, but lack of effective technical means of decision-makers is difficult to obtain from the data in depth, valuable information, the data become tasteless gesture of much use, it is difficult for policy-makers to provide comprehensive and efficient decision support information. Second, the demand from the user point of view, work with the tax information of further development, Ruhe help decision makers Zaimian Lin Hua semi-structured or structured Bijiao different question to Juece, tax information is a major task.

So far no specific application for the tax customized data mining algorithms. If the tax system, in-depth analysis of industry characteristics and data mining needs, based on the characteristics of the full use of tax data, effective data mining from the unique tax needs of the industry starting, the original data mining algorithms targeted improvements, will be able to effectively improve the efficiency of the tax system, data mining and accuracy, while this algorithm will also provide improved data mining applications in other industries to provide a good reference.

Outlier Data Mining

A. divisions of outlier data and clustering

Tax collection and management practice, cluster analysis can often be found in the taxpayer's group behavior. The same class of taxpayers in the business model, tax management should be very similar by cluster analysis of data can be divided, each type of data have some common characteristics, we can compare on this basis, thus more effective to tax regulation. If a business division of categories suddenly appeared different from the type of anomaly, the anomaly of the enterprise data should be enough attention.

For example, in the tax area, if a large enterprise of a sudden sharp drop in sales in one month (compared with the previous average of sales), resulting in outlier data, the reason may be due to human error, data entry operator error, it may be As the data itself unusual, the computer implementation of the error.

For another example, a "general taxpayer," the invoice amount and the amount of data such as a sudden increase, and the "key taxpayers," the data fairly, then the "key taxpayers," Statistical classification, the data are considered to be abnormal, the need for focus control.

Exception in accordance with the definition of the data given Hawkins [10], exception is data collection in different data, some wonder whether these data are not random error, but from completely different mechanisms. Unusual clustering algorithm is embedded in the definition of abnormal clustering of background noise. Outlier data clustering is usually a solitary point, outlier clustering is neither points nor background noise, and their behavior and normal behavior in general is very different, obviously deviate from other data, the data are inconsistent with the general model, and the existence of inconsistent data, other data objects.

In fact, the classification of the cluster after the change from one category to another category, the changes in the data object should be smooth and not jump. Jump if data should be handled as an exception data object. Outlier detection can be led to the discovery of small data set the model relative to the cluster, that data set were significantly different from the other data between the target and outlier analysis of the data than the information contained in general more valuable[11-12].

B. The cluster partitioning algorithm

The n data points into k groups, each group contains at least one object, and each of the images must belong to and only belong to a group.

k -means algorithm: the k as a parameter, the data is divided into k - n -cluster to cluster with high similarity, but low similarity between the cluster and the cluster, the calculation of similarity of objects under the cluster mean to carry out.[13-14]

Any given k objects, each object represents a cluster of initial average or center of each of the remaining objects, in accordance with various cluster centers of the distance to its assignment to the nearest cluster, and then re-calculated each cluster, on average, cycle this process until the criterion function converges. Convergence function is defined as follows:

$$E = \sum_{i=1,2,\cdots,k} \sum_{p \in C_i} |p - m_i|^2$$

E is the database where all objects of the sum of squared errors.

p points for the space.

m is the average cluster C_i , but p and m_i are multidimensional.

Tax clustering of outlier data

A. Algorithm Description

Tax outlier data clustering algorithm is the basic idea is: given the relative stability of the object data set, delete data does not cause too much of the cluster change, and change a lot of data object is smooth, if there is a sudden jump clustering, classifies the data as outlier data[15].

Setp1:

Distance-based anomaly detection algorithm to detect the data collection will be detected outliers data separated into the original data collection point data set and outlier data sets correctly

Setp2:

On the normal data sets separated by k - means algorithm for clustering, clustering results are given

Setp3:

The detection of abnormal points set and the second step of the clustering results combined output.

Input:

Parameter p ， D and the expected number of clusters M ;

Output:

M items of data containing M items of data clusters and data clusters exception.

B. Cluster Detection and Optimization of the distance calculation

Clustering is used to inspect the object similarity to cluster dissimilarity, while the dissimilarity between objects is based on object detection distance calculated. The most common methods of EUCLID distance measure distance [16-17].

EUCLID distance is defined as type as follows:

$$d(i,j) = \sqrt{|x_{i1} - y_{i1}|^2 + |x_{i2} - y_{i2}|^2 + \cdots + |x_{im} - y_{im}|^2} \qquad (1)$$

Where i ， j is the data objects of m

h is the object i to j in the tracks of any other object.

$d(i,j)$ distance between objects can be (1)-type push:

$$d(i,j) = \frac{\sqrt{diff^2(x_{i1},y_{i1}) + diff^2(x_{i2},y_{i2}) + \cdots + diff^2(x_{im},y_{im})}}{n} \quad (2)$$

$x_{i1}, y_{i1}, x_{i2}, y_{i2}, \cdots, x_{im}, y_{im}$ is the Attribute of object i and j.

$diff(x,y)$ is the distance between the properties, the property follows a standardized calculation function $norm$：

$$X_{norm} = {(X_i - \min_i)}/{(\max_i - \min_i)} \quad (3)$$

Cluster continuous attribute heart formula is：

$$Mod(i) = {\sum_{i=1}^{n} weight_t * value_i}/{\sum_{i=1}^{n} weight_t} \quad (4)$$

Where $weight$ weight the data.

Properties of clusters of discrete formula for the heart：

$$Mod(i) = MaxIndex(counts(attribute(attindex)numValues())) \quad (5)$$

In order to cluster the data to meet the standards and requirements of the tax, within the class interval in the past may be small, but the distance between classes and the class of almost likely the largest, in order to satisfy a similarity within the large and small similarity between the requirements of class. Optimization cluster following the situation and the number of clusters.

Clustering distribution in each sample, the assessment of a particular class and all other kinds of similarity calculation, $i = 1,2,\cdots,M_n$, M_n number of the current class.

Where $i \neq j$,

Then $R_i = \max_{j, j \neq i} \{R_{ij}\}$,

Where various types of similarity, the average maximum is:

$$\overline{R} = \frac{1}{M_n}\sum_{i=1}^{M_n} R_i$$

Then the most hours that the optimal clustering.

Experiment

Use of a corporate tax department of two annual tax amounts of data tables, as shown in Table Ⅰ. Data preprocessing, the data set to join the file name, attribute names, data, signs and other parameters. According to the above testing and optimization of clustering algorithm, clustering results are shown in Figure 1 and Figure 2.

Can be seen from the Figure 1 and Figure 2, the cluster 1 and cluster 2, the 0.16 and 142.13 points respectively, the isolation should be extracted as the data submitted for abnormal focus; 3.54 and 82.51 respectively, cluster 1 and cluster 2 The most typical point, the core of the point data, although in the second clustering process may no longer be a core point data type, but generally can not be moved out of the cluster.

Experiments show that the algorithm, while outlier data can be simultaneously clustering, application of the tax have high practical value.

TABLE I
The amount of corporate tax year tables

Time	Test1	TEST2
jan	0.16	69.23
feb	2.81	80.15
mar	3.35	97.90
apr	4.80	77.21
may	2.62	78.91
jun	5.06	81.92
jul	4.09	75.06
aug	3.54	65.30
sep	3.81	82.51
oct	4.01	100.80
nov	3.89	142.13
dec	2.90	77.83

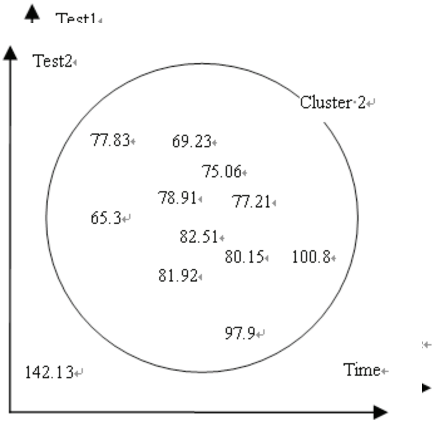

Fig. 1 Clustering results of the clustering 1.

Fig. 2 Clustering results of the clustering 2.

Conclusion

Outlier factor based clustering algorithm based on analysis of the tax industry, the demand for data mining and data features of the tax data on the clustering analysis, a tax data anomalies of data mining algorithms, on the one hand the tax can be found Data in the outlier data, improve the comprehensive evaluation of tax credit, it also enhances the accuracy of clustering for data preprocessing and neural network training laid the foundation. Using a tax example shows the effectiveness of the algorithm.

The algorithm used in tax administration system, massive tax data can be found in the abnormal data, key sources not only for the screening can also be used to detect abnormal operating behavior of taxpayers, or even determine whether the presence of the taxpayer evasion and tax fraud suspects, to rapidly obtain accurate tax check suspects. Data processing for the tax provides a broad application prospects.

References

[1] Fayyad U， Piatetsky-Shapiro G， Smyth P. The KDD Process for extracting useful knowledge from volumes of data。 Communications of the ACM， 1996， 39(11):27-34.

[2] Fayyad U，Piatetsky-ShaPiro G， Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad U， ed. Advances in knowledge discovery and data mining， AAAI/MIT Press， 1996， l-34.

[3] Pawlak Z. Rough sets. International Journal of Computer and Information Sciences, 1982(11): 341-456

[4] W. Z. Wu. Attribute reduction based on evidence theory in incomplete decision systems. Information Sciences, 2008, 178:1355-1371

[5] T. Q. Deng, Y. Chen, W. Xu, Q. Dai. A novel approach to fuzzy rough sets based on a fuzzy covering. Information Science, 2007, 177:2308-2326

[6] Yao Y.Y., Wang F.,Zeng D., Wang J. Rule+exception strategies for security information analysis. IEEE Intelligent Systems, 2005, 1095: 52-57.

[7] T. J. Li, W. X. Zhang. Rough fuzzy approximations on two universes of discourse. Information Sciences, 2008(178):892-906

[8] Kryszkiewicz M. Comparative studies of alternative type of knowledge reduction in inconsistent systems. International Journal of Intelligent Systems, 2001, 16:105-120.

[9] Liang J Y, Dang C Y, Chin K S, Yam Richard C M. A new method for measuring for rough sets and rough relational databases. Information Sciences, 2002, 31(4): 331-342

[10] D Hawkins. Dentification of out liers [M]. London: Chapman and Hall, 1980

[11] X. Z. Wang, E. C. C.Tsang, S. Y. Zhao, D. G. Chen, D. S. Yeung. Learning fuzzy rules from fuzzy samples based on rough set technique. Information Sciences, 2007(177):4493-4514

[12] Dominik, Wojciech Ziarko. The investigation of the Bayesian rough set model . International Journal of Approximate Reasoning, 2005, 40:81-91.

[13] M. Sarkar. Fuzzy-rough nearest neighbor algorithms in classification. Fuzzy Sets and Systems, 2007,158:2134-2152

[14] Shepard N.,Novland C.,Jenkins H.. Learning and memorization of classification. Psychological Monographs,1961,75(13):1-42.

[15] Nosofsky M.,Palmeri J.,Mckinley C.. Rule-plus-exception model of classification learning. Psychological Review,1994,101(1):53-79

[16] Zhou Y.,Wang J.. Rule+exception modeling based on rough set theory, RSCTC98, Warsaw,Poland,1998,529-536

[17] B. Tuttle, S. D. Vandervelde. An empirical examination of CobiT as an internal control framework for information technology. International Journal of Accounting Information Systems, 2007(8):240-263