Knowledge-Rich Data Mining in Financial Risk Detection

Yi Peng^{1,2}, Gang Kou^{1,2,*}, and Yong Shi^{2,3}

¹ School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, P.R. China, 610054

² CAS Research Center on Fictitious Economy and Data Sciences, Beijing 100080, China

³ College of Information Science & Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA
kougang@uestc.edu.cn

Abstract. Financial risks refer to risks associated with financing, such as credit risk, business risk, debt risk and insurance risk, and these risks may put firms in distress. Early detection of financial risks can help credit grantors to reduce risk and losses, establish appropriate policies for different credit products and increase revenue. As the size of financial databases increases, large-scale data mining techniques that can process and analyze massive amounts of electronic data in a timely manner become a key component of many financial risk detection strategies and continue to be a subject of active research. However, the knowledge gap between the results data mining methods can provide and actions can be taken based on them remains large in financial risk detection. The goal of this research is to bring the concept of chance discovery into financial risk detection to build the knowledge-rich data mining process and therefore increase the usefulness of data mining results in financial risk detection. Using six financial risk related datasets, this research illustrates that the combination of data mining techniques and chance discovery can provide knowledge-rich data mining results to decision makers; promote the awareness of previously unnoticed chances; and increase the actionability of data mining results.

Keywords: knowledge-rich data mining, chance discovery, financial risk detection, classification, risk analysis.

1 Introduction

Financial risks refer to risks associated with financing, such as credit risk, business risk, debt risk and insurance risk, and these risks may put firms in distress. Take health insurance risk as an example. According to the National Health Care Anti-Fraud Association's (NHCAA) estimation, at least 3% of the United States' annual health care expenditure, which in calendar-year 2003 alone amounted to \$1.7 trillion, is lost to outright fraud or erroneous payment [20]. Another example is credit card debt risk. The total credit card debt at the end of the first quarter of 2002 in the U.S. is about \$660 billion [21] and the total credit card holders declared bankruptcy in 2003 are more than 1.6 million [14].

* Corresponding author.

As the size of financial databases increases, large-scale data mining techniques that can process and analyze massive amounts of electronic data in a timely manner become a key component of many financial risk detection strategies and continue to be a subject of active research [4, 11, 16, 17]. While most data mining research focus on developing algorithms or applying data mining methods to detect financial risk, how to better utilize data mining results to help decision makers identify financial risks and integrate data mining results into decision-making routines have not been widely studied in the data mining community. The lack of interaction between industry practitioners and academic researchers makes it hard to discover financial risks or opportunities in data mining projects and hence weaken the value that data mining methods may bring to financial risk detection. The knowledge gap between the results data mining methods can provide and taking actions based on them [26] remains large in financial risk detection.

The goal of this paper is to bring the concept of chance discovery into financial risk detection to build the knowledge-rich data mining process and therefore increase the usefulness and actionability of data mining results in financial risk detection. The rest of this paper is organized as follows: Section 2 gives an overview of the datasets employed for this study; discusses the data mining techniques used in financial risk detection; and proposes a knowledge-rich data mining process for financial risk detection. Section 3 presents an empirical study that examines the proposed process using six financial risk datasets and Section 4 summarizes the paper.

2 Financial Risk Detection Problem

There is no one universally accepted definition for financial risk. Any risk that relates to financing can be considered as financial risk. According to its sources, financial risk can be broadly categorized as investment risk, credit risk, and business risk [24]. Investment risk is the probability that an investment may produce an undesirable outcome. Credit risk is the probability that debtors will not pay their debts. Business risk denotes the possibility that income is less than expected and/or expenditure is larger than expected. The datasets used in this work are examples of credit and business risk: consumer credit card application and existing credit cardholders' behavior belong to credit related risk; corporation bankruptcy and disability income insurance datasets are business related risks. Financial risk detection is the practice of identifying or predicting these financial risks in an attempt to control risk and minimize losses. The following subsections describe the data sources and examine the major data mining techniques that have been applied in this work.

2.1 Financial Risk Datasets

The datasets used in this study come from five countries and represent four aspects of financial risk: credit approval (credit card application and loan approval), credit behavior, bankruptcy risk, and fraud risk.

German credit card application dataset [22]

The German credit card application dataset comes from UCI Machine Learning databases. It contains 1000 instances with 24 predictor variables and 1 class variable.

The 24 variables describe the status of existing checking account, credit history, education level, employment status, personal status, age, and so on. The class variable indicates whether an application is Accepted or Declined. 70% of the instances are accepted applications and 30% are declined instances.

Australian credit card application dataset [19]

This dataset was provided by a large bank and concerns consumer credit card applications. It has 690 instances with 15 predictor variables plus 1 class variable. The class variable indicates whether an application is Accepted or Declined. 55.5% of the instances are accepted applications and 44.5% are declined instances.

Credit cardholders' behavior dataset [1, 11, 16]

The dataset was from a major US bank and contains 6000 credit card data with 64 predictor variables plus 1 class variable. Each instance has a class label indicating its credit status: either Good or Bad. 84% of the data are Good accounts and 16% are Bad accounts. Good indicates good status accounts and Bad indicates accounts with late payments, delinquency, or bankruptcy. The predictor variables describe account balance, purchase, payment, cash advance, interest charges, date of last payment, times of cash advance, and account open date.

Japanese bankruptcy dataset [12]

This set collects 37 bankrupt Japanese firms and 111 non-bankrupt Japanese firms from various sources during the post-deregulation period of 1989 to 1999. Final sample firms are ones traded in the First Section of Tokyo Stock Exchange, and their financial data are available from 2000 PACAP database for Japan compiled by the Pacific-Basin Capital Market (PACAP) Research Center at the University of Rhode Island. Each case has 13 predictor variables and 1 class variable (Bankrupt or Non-bankrupt). The predictor variables describe financial state and performance of firms.

Korean bankruptcy dataset [13]

This dataset collects bankrupt firms in Korea from 1997 to 2003 from public sources. It consists of 65 bankrupt and 130 non-bankrupt firms whose data are available and publicly trading firms in the Korean Stock Exchange. Each case has 13 predictor variables with one class variable (Bankrupt or Non-bankrupt).

Disability income insurance dataset [17]

Disability insurance provides financial support when a disability leaves the policyholder unable to work. The data was provided by an anonymous U.S. corporation. Each instance concerns a disability income insurance claim. The set has 18,875 instances with 103 variables. A binary class attribute indicates whether an instance is a Normal claim or Abnormal claim. There are 353 abnormal claims and 18,522 normal claims. The abnormal instances represent fraudulent or erroneous claims and were manually collected and verified.

2.2 Data Mining Techniques

Current data mining research in financial risk detection concentrates on fraud detection and credit risk analysis. Major data mining functions used in these areas include classification, prediction, cluster analysis and outlier analysis. The selection of data mining functions depends on data mining tasks. Since the datasets employed in this

study are examples of classification applications, the data mining function used in this paper is classification. Eight classification methods: Bayesian network [23], naïve Bayes [8], Support Vector Machine (SVM) [18], linear logistic regression [3], K-nearest-neighbor [7], C4.5 [19], Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule induction [6] and radial basis function (RBF) network [2], are used in our empirical study and all of them are implemented in WEKA [25].

Bayesian network and naïve Bayes classifier both model probabilistic relationships between predictor variables and the class variable. While naïve Bayes classifier estimates the class-conditional probability based on Bayes theorem and can only represent simple distributions, Bayesian network is a probabilistic graphic model and can represent conditional independencies between variables. SVM classifier uses a nonlinear mapping to transform the training data into a higher dimension and search for the linear optimal separating hyperplane, which is then used to separate data from different classes [10]. Linear logistic regression models the probability of occurrence of an event as a linear function of a set of predictor variables. K-nearest-neighbor classifies a given data instance based on learning by analogy, that is, assigns it to the closest training examples in the feature space. C4.5 is a decision tree algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. RIPPER is a sequential covering algorithm that extracts classification rules directly from the training data without generating a decision tree first [10]. RBF network is an artificial neural network that uses radial basis functions as activation functions. In addition to the eight classification techniques, we also employ ensemble method, which aggregates the predictions of the above mentioned eight classifiers.

2.3 Knowledge-Rich Data Mining in Financial Risk Detection

Even though data mining has become a crucial tool in financial risk detection, most data mining research focus on developing learning algorithms or improving existing algorithms that can identify suspicious patterns or predict future behaviors efficiently from financial databases and have not paid enough attention to the involvement of end users and the actionability of the final data mining results. This is mainly due to two reasons: (1) the difficulty in accessing real-life financial risk data; (2) limited access to domain experts and background information. The lack of interaction between industry practitioners and academic researchers makes it hard to discover financial risks or opportunities in data mining projects and hence weaken the value that data mining methods may bring to financial risk detection.

In an attempt to improve the usefulness of data mining results and increase the probability of identifying unusual chances in financial risk analysis, this paper proposes a chance discovery and data mining process for financial risk detection (Figure 1). Chance discovery (CD) is defined as “the awareness of a chance and the explanation of its significance” [15]. Ohsawa and Fukuda [15] suggested three keys to chance discovery: communicating the significance of an event; enhancing user’s awareness of an event’s utility using mental imagery; revealing the causalities of rare events using data mining methods. Users, communication and data mining are the main parts of chance discovery. Figure 1 combines the KDD (Knowledge Discovery in Database) process model [9], the chance discovery process [15] and the CRISP-DM process model [5]. It emphasizes three keys to chance discovery and knowledge-rich data

mining: users, communication and data mining techniques. Users refer to domain experts and decision makers. Domain experts are knowledgeable of the field information, data collection procedures and meaning of variables. With the assistance of data miners, domain experts can gain insights of financial risk data from different aspects and potentially observe new chances. To turn the identified knowledge into financial or strategic advantages, decision makers, who understand the operational and strategic goals of a company, are required to provide feedbacks on the importance of the potential new chances and determine what actions should be taken. Moving back and forth between steps is always required. The cyclical nature of chance discovery is illustrated by the outer circle of the chance discovery process in Figure 1.

Fig. 1. The process of chance discovery and data mining for financial risk detection

3 Empirical Study

In this section we apply the proposed process in Section 2 to the six financial risk datasets. The datasets were collected from five countries and represent four aspects of financial risk: credit approval, credit behavior, bankruptcy risk, and insurance fraud. The experiment was carried out according to the following process:

Chance discovery process for financial risk detection

Input: a financial risk related dataset

Output: Decision function; Results of performance metrics; new chance(s)

Step 1. Understand business requirements, dataset structure and data mining task

Step 2. Prepare target datasets: select and transform relevant features; data cleaning; data integration. Communicate any findings during data preparation to domain experts.

Step 3. Train and test multiple data mining models in randomly sampled partitions (e.g. k-fold cross-validation) using WEKA 3.5.7 [25].

Step 4. Evaluate data mining models using a set of performance metrics. The best model/s is the decision function.

Step 5. Discuss the data mining results with domain experts. Explore potential chance(s) from data mining results. If identify new chance(s), communicate the chance(s) with decision makers and determine the appropriate actions. Go back to Step 1 if new business questions are raised during the process.

END

Because different performance metrics are appropriate in different settings, this paper utilizes six performance metrics: True Positive rate, True Negative rate, False Positive rate, False Negative rate, Overall Accuracy and Area Under ROC. TP (True Positive) is the number of correctly classified Abnormal (Bankrupt, Bad, or Declined) instances. FP (False Positive) is the number of Normal (Non-bankrupt, Good, or Accepted) instances that is misclassified as Abnormal class. TN (True Negative) is the number of correctly classified Normal instances. FN (False Negative) is the number of Abnormal instances that is misclassified as Normal class. Accuracy is one the most widely used classification performance metrics. It is the ratio of correctly predicted instances to the entire instances or instances in a particular class.

$$\text{Overall Accuracy} = \frac{TN + TP}{TP + FP + FN + TN}$$

$$\text{True Positive rate} = \frac{TP}{TP + FN}, \text{True Negative rate} = \frac{TN}{TN + FP}$$

$$\text{False Positive rate} = \frac{FP}{FP + TN}, \text{False Negative rate} = \frac{FN}{FN + TP}$$

ROC curves stands for Receiver Operating Characteristic which detects and characterizes the tradeoff between TP rate and FP rate. The area under the curve represents the probability of the ranking by the classifier of a randomly sampled positive instance over a randomly sampled negative one. It indicates a classifier's performance; the larger the area, the better the classifier.

The classification results of eight data mining methods plus ensemble method for the six datasets are summarized in Table 1. In the dataset column, Australian indicates Australian credit card application data; Chase indicates credit cardholders' behavior data; DI indicates the disability income insurance data; German indicates German credit card application data; Japan indicates Japanese bankruptcy data and Korea indicates Korean bankruptcy data. The nine classification methods described in Section 2 were applied to each dataset using 10-fold cross-validation.

Table 1. Classification results

Dataset	Algorithm	Overall Accuracy	Area Under ROC	True Positive rate	True Negative rate	False Positive rate	False Negative rate
Australian	Bayesian Network	0.8522	0.9143	0.7980	0.8956	0.1044	0.2020
Australian	Naive Bayes	0.7725	0.8978	0.5863	0.9217	0.0783	0.4137
Australian	SVM	0.8551	0.8622	0.9251	0.7990	0.2010	0.0749
Australian	Linear Logistic	0.8623	0.9312	0.8664	0.8590	0.1410	0.1336
Australian	K Nearest Neighbor	0.7942	0.7922	0.7752	0.8094	0.1906	0.2248
Australian	C4.5	0.8348	0.8346	0.7948	0.8668	0.1332	0.2052
Australian	RBFNetwork	0.8304	0.8995	0.7524	0.8930	0.1070	0.2476
Australian	RIPPER Rule Induction	0.8522	0.8714	0.8534	0.8512	0.1488	0.1466
Australian	Ensemble	0.8551	0.9289	0.8274	0.8773	0.1227	0.1726
Chase	Bayesian Network	0.7055	0.8424	0.8656	0.6750	0.3250	0.1344
Chase	Naive Bayes	0.6933	0.8395	0.8740	0.6589	0.3411	0.1260
Chase	SVM	0.8372	0.5632	0.1604	0.9661	0.0339	0.8396
Chase	Linear Logistic	0.8532	0.8539	0.3031	0.9579	0.0421	0.6969
Chase	K Nearest Neighbor	0.8028	0.6327	0.3802	0.8833	0.1167	0.6198
Chase	C4.5	0.8170	0.6245	0.3542	0.9052	0.0948	0.6458
Chase	RBFNetwork	0.8400	0.8256	0.0000	1.0000	0.0000	1.0000
Chase	RIPPER Rule Induction	0.8443	0.6380	0.3333	0.9417	0.0583	0.6667
Chase	Ensemble	0.8382	0.8432	0.3990	0.9218	0.0782	0.6010
DI	Bayesian Network	0.8261	0.8361	0.6686	0.8291	0.1709	0.3314
DI	Naive Bayes	0.9695	0.9707	0.9943	0.9691	0.0309	0.0057
DI	SVM	0.9813	0.5000	0.0000	1.0000	0.0000	1.0000
DI	Linear Logistic	0.9809	0.7546	0.0000	0.9996	0.0004	1.0000
DI	K Nearest Neighbor	0.9723	0.5961	0.2040	0.9870	0.0130	0.7960
DI	C4.5	0.9786	0.6656	0.1898	0.9937	0.0063	0.8102
DI	RBFNetwork	0.9813	0.7097	0.0000	1.0000	0.0000	1.0000
DI	RIPPER Rule Induction	0.9806	0.5774	0.1586	0.9962	0.0038	0.8414
DI	Ensemble	0.9817	0.8443	0.0765	0.9989	0.0011	0.9235
German	Bayesian Network	0.7250	0.7410	0.3600	0.8814	0.1186	0.6400
German	Naive Bayes	0.7550	0.7888	0.5067	0.8614	0.1386	0.4933
German	SVM	0.7740	0.6938	0.4933	0.8943	0.1057	0.5067
German	Linear Logistic	0.7710	0.7919	0.4933	0.8900	0.1100	0.5067
German	K Nearest Neighbor	0.6690	0.6064	0.4500	0.7629	0.2371	0.5500
German	C4.5	0.7190	0.6607	0.4400	0.8386	0.1614	0.5600
German	RBFNetwork	0.7400	0.7520	0.4633	0.8586	0.1414	0.5367
German	RIPPER Rule Induction	0.7340	0.6557	0.4500	0.8557	0.1443	0.5500
German	Ensemble	0.7620	0.7980	0.4533	0.8943	0.1057	0.5467
Japan	Bayesian Network	0.7568	0.7292	0.5135	0.8378	0.1622	0.4865
Japan	Naive Bayes	0.7432	0.7197	0.4595	0.8378	0.1622	0.5405
Japan	SVM	0.7500	0.5000	0.0000	1.0000	0.0000	1.0000
Japan	Linear Logistic	0.7770	0.7290	0.4595	0.8829	0.1171	0.5405
Japan	K Nearest Neighbor	0.7770	0.6595	0.4324	0.8919	0.1081	0.5676
Japan	C4.5	0.7162	0.5270	0.3784	0.8288	0.1712	0.6216
Japan	RBFNetwork	0.7162	0.6533	0.2162	0.8829	0.1171	0.7838
Japan	RIPPER Rule Induction	0.7365	0.6193	0.4324	0.8378	0.1622	0.5676
Japan	Ensemble	0.7905	0.7424	0.3243	0.9459	0.0541	0.6757
Korea	Bayesian Network	0.8667	0.8773	0.7846	0.9077	0.0923	0.2154
Korea	Naive Bayes	0.7744	0.8168	0.5538	0.8846	0.1154	0.4462
Korea	SVM	0.8718	0.8682	0.8615	0.8769	0.1231	0.1385
Korea	Linear Logistic	0.8462	0.8749	0.7692	0.8846	0.1154	0.2308
Korea	K Nearest Neighbor	0.8154	0.7993	0.7538	0.8462	0.1538	0.2462
Korea	C4.5	0.8359	0.7948	0.7077	0.9000	0.1000	0.2923
Korea	RBFNetwork	0.8256	0.8033	0.7231	0.8769	0.1231	0.2769
Korea	RIPPER Rule Induction	0.8667	0.8577	0.8308	0.8846	0.1154	0.1692
Korea	Ensemble	0.8564	0.9026	0.8154	0.8769	0.1231	0.1846

To summarize, the empirical study demonstrated that introducing chance discovery into the KDD process can promote the awareness of previously unnoticed chances and increase the usefulness of data mining results.

4 Conclusion Remarks

In financial risk detection, fail to recognize significant chances may cause credit grantors and companies serious financial losses. In order to capture significant chances in financial risk detection, this paper proposed a knowledge-rich data mining process model and tested the model using real-life financial risk related datasets. The empirical study indicated that the combination of data mining techniques and chance discovery can provide human inspectors more information than data mining alone; promotes the awareness of identified financial risks among decision makers; and increases the usefulness of data mining results in the decision-making process. Due to the space limitation, this paper only reported the application of the knowledge-rich data mining process model to fraud detection in insurance data. In the future, we will continue the study about the usability of the process in other areas of financial risk detection, such as bankruptcy prediction and credit cardholders' behavior analysis.

Acknowledgements

This work was supported by the Youth Fund of University of Electronic Science and Technology of China (UESTC) and the National Natural Science Foundation of China (NSFC) under the Grand No. 70621001, No. 70531040, and No. 70472074 and 973 Project #2004CB720103, Ministry of Science and Technology, China.

References

1. Peng, Y., Kou, G., Shi, Y., Chen, Z.: Improving Clustering Analysis for Credit Card Accounts Classification. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005. LNCS, vol. 3516, pp. 548–553. Springer, Heidelberg (2005)
2. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)
3. le Cessie, S., Houwelingen, J.C.: Ridge estimators in logistic regression. *Applied Statistics* 41(1), 191–201 (1992)
4. Chan, P.K., Fan, W., Prodromidis, A.L., Stolfo, S.J.: Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems* 14(6), 67–74 (1999)
5. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0: Step-by-Step data mining guide (2000), <http://www.crisp-dm.org>
6. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann, San Francisco (1995)
7. Dasarthy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos (1991)

8. Domingos, P., Pazzani, M.: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 29(203), 103–130 (1997)
9. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM* 39(11), 27–34 (1996)
10. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2006)
11. Kou, G., Peng, Y., Shi, Y., Wise, M., Xu, W.: Discovering Credit Cardholders' Behavior by Multiple Criteria Linear Programming. *Annals of Operations Research* 135(1), 261–274 (2005)
12. Kwak, W., Shi, Y., Eldridge, S., Kou, G.: Bankruptcy Prediction for Japanese Firms: Using Multiple Criteria Linear Programming Data Mining Approach. *International Journal of Business Intelligence and Data Mining* 1(4), 401–416 (2006)
13. Kwak, W., Shi, Y., Kou, G.: Bankruptcy Prediction for Korean Firms after, Financial Crisis: Using Multiple Criteria Linear Programming Data Mining Approach (working paper) (1997)
14. New Generation Research, Inc.,
<http://www.bankruptcydata.com/default.asp> (as of April 27, 2008)
15. Ohsawa, Y., Fukuda, H.: Chance discovery by stimulated groups of people-application to understanding consumption of rare food. *Journal of Contingencies and Crisis Management* 10(3) (September 2002)
16. Peng, Y., Kou, G., Shi, Y., Chen, Z.: A Multi-Criteria Convex Quadratic Programming Model for Credit Data Analysis. *Decision Support Systems* 44(4), 1016–1030 (2008)
17. Peng, Y., Kou, G., Sabatka, A., Matza, J., Chen, Z., Khazanchi, D., Shi, Y.: Application of Classification Methods to Individual Disability Income Insurance Fraud Detection. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) *ICCS 2007, Part III. LNCS*, vol. 4489, pp. 852–858. Springer, Heidelberg (2007)
18. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schotolkopf, B., Burges, C.J.C., Smola, A. (eds.) *Advances in Kernel Methods-Support Vector Learning*, pp. 185–208. MIT press, Cambridge (1998)
19. Quinlan, J.R.: *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco (1993)
20. The National Health Care Anti-Fraud Association,
http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_cent&wpscode=TheProblemOfHCFraud (as of April 27, 2008)
21. The U.S. Payment Card Information Network,
<http://www.cardweb.com/cardlearn/stat.html> (as of April 27, 2008)
22. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA,
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
23. Weiss, S.M., Kulikowski, C.A.: *Computer Systems that Learn: Classification and Predication Methods from Statistics*. In: *Neural Nets Machine Learning and Expert Systems*. Morgan Kaufmann, San Francisco (1991)
24. Wikipedia.org, http://en.wikipedia.org/wiki/Financial_risk (as of April 28, 2008)
25. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2005)
26. Domingos, P.: Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery* 15, 21–28 (2007)