DATA ANALYZE AND PREDICTION USING DATA SCIENCE TECHNIQUE IN BANK LOAN APPROVAL

Bhava Dharani.P
Bhava Nandini.P
Dharani.M
Final Year, Dept. of CSE
Dr.M.G.R. Educational and Research Institute

Dr.V.RameshBabu
Professor
Department of CSE
Dr.M.G.R. Educational and Research Institute

Abstract
Anomaly detection relies on individuals' behavior profiling and works by detecting any deviation from the norm. When used for online banking fraud detection, however, it mainly suffers from three disadvantages. The historical behavior data are often too limited to profile his/her behavior pattern. Due to the heterogeneous nature of transaction data, there lacks a uniform treatment of different kinds of attribute values, which becomes a potential barrier for model development and further usage. The transaction data are highly skewed, and it becomes a challenge to utilize the label information effectively. The three disadvantages result in both poor generalization and high false positive rate of anomaly detection, and we propose a ranking metric embedding based multi-contextual behavior profiling (ReMEMBeR) model to battle them effectively. With the idea of collaborative filtering, for an individual, information from other similar individuals can be used to establish his/her behavior profile. The proposed model can, thus, integrate the multi-contextual behavior patterns and allow transactions to be examined under the different contexts. Extensive experiments on a real-world online banking transaction dataset demonstrate that our model not only outperforms benchmarks on all metrics but also can be combined with them to achieve even better performance.

Key words: SVM, NN2L, Filter, Machine Learning

I. INTRODUCTION

With the popularity of computer and Internet technology, online banking systems have flourished nowadays. They bring great facilities to people's daily life. As a coin has two sides, however, online banking is more inclined to fraudulent activities, and online banking fraud has become a serious financial crime that could cause massive losses. As a matter of fact, fraud detection is a permanent issue for online banking systems. Anomaly detection is the most popular technique used for fraud detection. It relies on an individual's behavior pattern and detects any deviation from the norm. The assumption is that the behavior pattern of an individual is generally stable, and any deviation from it might indicate anomaly or fraud. When used in the online banking scenario, traditional anomaly detection methods are faced with two major challenges.

First, anomaly detection requires elaborate behavior profiling for individuals, but this is impossible in most cases due to their limited historical behavior data. A straightforward idea is to utilize information from other similar individuals, but it brings another problem to find out similar individuals. Second, even if the behavior profile can be established, the highly skewed distribution of the legitimate class and the fraudulent one has always been troublesome. How to make full use of the label information (whenever it exists) is not an easy thing.

II. RELATED WORK

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

Review of Literature Survey

Title : Impact of Personal Loan Offered by Banks and Non Banking

Financial Companies in Coimbatore City

Author: C.Sankar Ph.D

Year : 2017

It is a unsecured loan as no securities are taken while availing the loan. Other form of loan like home loan, loan against shares/ securities demands collateral security while availing it. The personal loan is completely free from all the conditions and can be availed with ease.

Even though a few financial institutions may ask for a guarantor based on their comfort level of the applicant&#39;s profile, it is a hassle free form loan.

According to the Credit Information Bureau of India (CIBIL), your credit score should be between 100 and 999. The lower your credit score, the more chances of your application being rejected. The higher the credit score, the lower will be the interest rates charged. Another factor to be considered when applying for a personal loan is if you want to opt for reduced-balance interest rate or flat interest rate. With reduced-balance interest rate, the interest on the loan keeps on reducing as it is calculated on the reduced principle amount which reduces daily, monthly, quarterly or annually.

III. EXISTING SYSTEM

Anomaly detection relies on individuals' behavior profiling and works by detecting any deviation from the norm. When used for online banking fraud detection, however, it mainly suffers from three disadvantages. First, for an individual, the historical behavior data are often too limited to profile his/her behavior pattern. Second, due to the heterogeneous nature of transaction data, there lacks a uniform treatment of different kinds of attribute values, which becomes a potential barrier for model development and further usage. Third, the transaction data are highly skewed, and it becomes a challenge to utilize the label information effectively. Anomaly detection often suffers from poor generalization ability and a high false alarm rate. We argue that individuals' limited historical data for behavior profiling and the highly skewed nature of fraud data could account for this defect. Since it is straightforward to use information from other similar individuals, measuring similarity itself becomes a great challenge due to heterogeneous attribute values. We propose to transform the anomaly detection problem into a pseudore commender system problem and solve it with an embedding based method. By doing so, the idea of collaborative filtering is implicitly used to utilize information from similar users, and the learned preference matrices and attribute embedding provide a concise way for further usage.

IV. PROPOSED SYSTEM

Exploratory Data Analysis:

Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

Data Wrangling:

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis.

Make sure that the document steps carefully and justify for cleaning decisions.

Data collection:

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

Building the classification model:

The predicting the loan approval, decision tree algorithm prediction model is effective because of the following reasons: It provides better results in classification problem. It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.

It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

V. ARCHITECTURE DIAGRAM

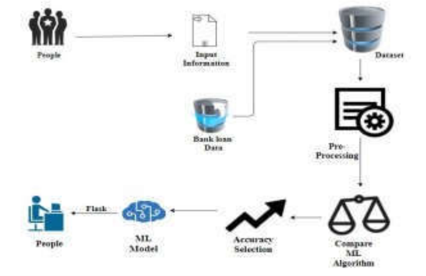
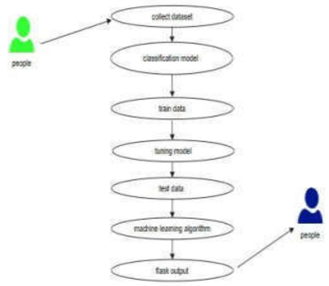
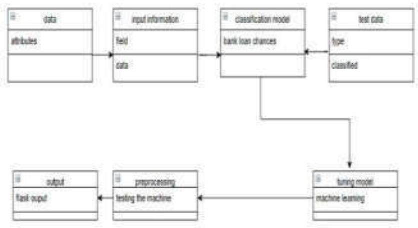


Fig 1: Block diagram

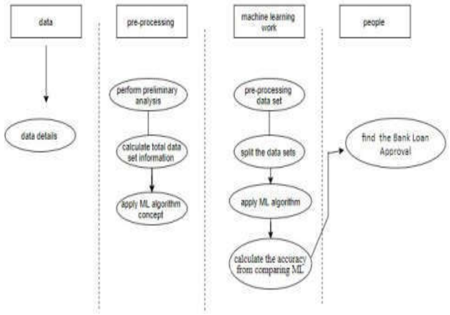
USE CASE DIAGRAM



CLASS DIAGRAM



ACTIVITY DIAGRAM

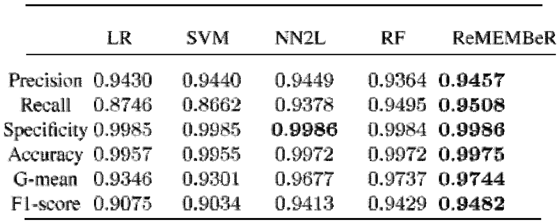

data
pre-processing
machine learning work
people
data details
perform preliminary analysis
calculate total data set information
apply ML algorithm concept
pre-processing data set
split the data sets
apply ML algorithm
calculate the accuracy from comparing ML
find the Bank Loan Approval

VI. RESULTS AND DISCUSSION

We mainly report two kinds of model performances: the point performance with the best F1-score and the overall performance. For the former, we choose the fraudulence threshold 0 such that our model gets the largest F1-score. The point performance is reported on multiple measures. For the latter, we incrementally change 0 to get a set of (Precision, Recall) pairs and a set of (TPR, FPR) pairs, with which we draw the so-called Precision-Recall curves and ROC curves. The overall performance indicates tradeoffs between different measures.

1) Significance Test of Differences Between Model Performances: As we have only one train/test split of our dataset, we turn to the cross-validation process to test the significance of differences between ReMEMBeR and benchmarks.

POINT PERFORMANCES OF DIFFERENT MODELS ON MULTIPLE MEASURES



	LR	SVM	NN2L	RF	ReMEMBeR
Precision	0.9430	0.9440	0.9449	0.9364	**0.9457**
Recall	0.8746	0.8662	0.9378	0.9495	**0.9508**
Specificity	0.9985	0.9985	**0.9986**	0.9984	**0.9986**
Accuracy	0.9957	0.9955	0.9972	0.9972	**0.9975**
G-mean	0.9346	0.9301	0.9677	0.9737	**0.9744**
F1-score	0.9075	0.9034	0.9413	0.9429	**0.9482**

We see that, when $p < 0.05$, the zero value is not included in any of those confidence intervals. That is to say, all performance differences between ReMEMBeR and the benchmarks are believed to be significant.

A. Performance Against Benchmarks: The point perfor- mances on multiple measures are shown in above diagram and the overall performance.

We can see that RF turns out to be slightly better than NN2L, and both of them outperform LR and SVM.Our ReMEMBeR model outperforms all of the benchmarks on all measures, which justifies its great effectiveness. In Fig. 2(b), ReMEMBeR has the fastest growth in TPR as FPR increases, which guarantees that we can detect much more fraudulent transactions at the expense of further disturbing a relatively smaller number of legitimate users. Note that we deliberately constrain the FPR to extremely low values (much smaller than 0.01, the FPR value usually required for online banking systems in practice) and still get very high TPR. From the Precision-Recall curves shown in Fig. 2(a), we also find that ReMEMBeR gives consistently better performance than benchmarks. This superiority reflects that our model can better distinguish fraudulent transactions from legitimate ones than benchmarks.

3) Performance Against Dynamically Skewed Data: So far, we have only tested our model on fixed-skewed datasets. However, in practice, this is often not the case. In order to simulate the dynamically skewed data environment, we filter out part of fraudulent transactions from the original training and testing data to create more imbalanced datasets and use the notation TR_x and TE_x to denote them. Here, the subscript x means that a proportion of 1 − x fraudulent transactions is removed from the original dataset, while all legitimated transactions are retained. As RF always shows the best performance among benchmarks, for simplicity, we only consider RF in this test. We train both ReMEMBeR and RF on TR_x and compare their point performances on TE_x, where x = 1, 0.5, 0.1, 0.05, 0.01. The (F1-score)–(1-x) curve is displayed in Fig. 3. Two important observations can be obtained: Not only that ReMEMBeR always shows superior performance to RF but also that, as x drops, the ReMEMBeR's curve decreases much more slowly than that of RF. In conclusion, ours can utilize the label information more effectively.

VI. CONCLUSION

Anomaly detection often suffers from poor generalization ability and a high false alarm rate. We argue that individuals' limited historical data for behavior profiling and the highly skewed nature of fraud data could account for this defect.

Since it is straightforward to use information from other similar individuals, measuring similarity itself becomes a great challenge due to heterogeneous attribute values. We propose to transform the anomaly detection problem into a pseudo-recommender system problem and solve it with an embedding based method. By doing so, the idea of collaborative filtering is implicitly used to utilize information from similar users, and the learned preference matrices and attribute embedding provide a concise way for further usage. Furthermore, by tackling the data sparsity problem and overcoming the implicit feedback difficulty, we depend on a ranking-based learning scheme to fully explore the label information as well. Finally, relying on the observation that individuals are determined by contexts, we extend the original behavior profiling model to be a multi-contextual one. Elaborate experiments on a real-world dataset demonstrate that our ReMEMBeR model outperforms benchmarks on all metrics, shows superior effectiveness and robustness in the face of a dynamically skewed data environment, and could be combined with benchmarks to achieve even better performance.

Future work falls into four aspects. First, we would like to find out more effective methods for model combination.

Second, we have so far considered all possible contextual attributes, without investigating whether there exists a (or a set of) optimized contextual attribute(s) for a specific transaction.

We believe it worth investigation. Third, as deep learning techniques have shown great advantages in almost all research areas, we also plan to extend our model to contain deep architectures, in the purpose of obtaining even more performance gain. Last but not least, sequential information is very important in modeling user behaviors. As a matter of fact, we look forward to exploiting this notable information directly in the following work.

REFERENCES

[1] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," World Wide Web, vol. 16, no. 4, pp. 449–475, Jul. 2013.

[2] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," J. Netw. Comput. Appl., vol. 68, pp. 90–113, Jun. 2016.

[3] D. J. Weston, D. J. Hand, N. M. Adams, C. Whitrow, and P. Juszczak, "Plastic card

fraud detection using peer group analysis," Adv. Data Anal. Classification, vol. 2, no. 1, pp. 45–62, Apr. 2008.

[4] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," 2013, arXiv:1305.1707. [Online]. Available: http://arxiv.org/abs/ 1305.1707

[5] R.-C. Chen, T.-S. Chen, and C.-C. Lin, "A new binary support vector system for increasing detection rate of credit card fraud," Int. J. Pattern Recognit. Artif. Intell., vol. 20, no. 2, pp. 227–239, Mar. 2006.

[6] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in Proc. ICML, Corvallis, OR, USA, Jun. 2007, pp. 935–942.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in Proc. NIPS, Vancouver, BC, Canada, Dec. 2007, pp. 1257–1264.

[8] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in Proc. Adapt. Web, 2004, pp. 291–324.

[9] S. Feng, X. Li, Y. Zeng, G. Cong, Y. M. Chee, and Q. Yuan, "Personalized ranking metric embedding for next new poi recommendation," in Proc. IJCAI, Buenos Aires, Argentina, Jul. 2015, pp. 2069–2075.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 15:1–15:58, 2009.

[11] S. Dhankhad, E. Mohammed, and B. Far, "Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study," in Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI), Jul. 2018, pp. 122–125.

[12] S. Xuan, G. Liu, Z. Li, L. Zheng, and C. Jiang, "Random forest for credit card fraud detection," in Proc. IEEE 15th Int. Conf. Netw., Sens. Control (ICNSC), Mar. 2018, pp. 1–6.

[13] I. Sohony, R. Pratap, and U. Nambiar, "Ensemble learning for credit card fraud detection," in Proc. ACM India Joint Int. Conf. Data Sci.Manage. Data, Jan. 2018, pp. 289–294.

[14] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," Expert Syst. Appl., vol. 51, pp. 134–142, Jun. 2016.

[15] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," Data Mining Knowl. Discovery, vol. 18, no. 1, pp. 30–55, Feb. 2009.