Data Preprocessing for credit card dataset using synthetic minority oversampling technique

Mr Subhash Babu Bathala[1], Dr Muthuluru Nagendra[2]
[1]Research Scholar, Department of Computer Science and Technology, SK University, Ananthapur
[2]Professor, Department of Computer Science and Technology, SK University, Ananthapur

Abstract

Currently, the biggest challenge faced by the data integration process is data heterogeneity and data of low quality. It may lead to the results, which may not match with the situation. Data source pre-processing affects directly the standard of data mining. The strategies are totally different in step with explicit application fields and industries. Transformation includes data cleaning, which is used to clear the futile data, which is present in the original database. Before data cleaning, analysis of data quality is required.

Data loading is to load the data, that has been extracted and remodelled into the destination database through loading tools. This paper describes the data pre-processing of credit card fraud detection intimately. Firstly, some tables are hand-picked from the credit card data, that is bothered with the analysis topic. Next, the paper handles the difficulties in chosen initial data like strident data, missing values by data pre-processing that primarily includes data cleansing, integration, transformation, and loading. The reduction has been studied deeply and training sample data is obtained where needed. The model is applied to the bank credit card fraud detection and evaluation system, which proved to be more efficient in implementation, it can be used as the principle guiding of data processing in the bank system. This paves the way in reducing the running time of the mining algorithm, thus speeding the mining procedure.

Index Terms : consuming behaviour; data cleaning; integration; reduction

INTRODUCTION

Data mining algorithms are usually strict with the data, which has better data integrity, less data redundancy and a smaller correlation between attributes and so on. However, in the actual system, primary data is generally dirty, incomplete and inconsistent. It is hard to meet the requirements of data mining algorithms directly. Furthermore, there is plenty of insignificant data in the actual data, seriously affecting the efficiency of data mining algorithms. Therefore, primary data should be converted to formats through data pre-processing which are more suitable for data mining, which brings security to data mining algorithms[1][2].

Data pre-processing is an essential and important step in the data mining process and it has a huge influence on the success of a data-mining project. Data pre-processing is a step of the Knowledge discovery in databases (KDD) process that overcomes the complexity of the data and gives better conditions to subsequent analysis. Within this, the quality of the data is better known and the data analysis is performed more accurately and efficiently. Data pre-processing is difficult because it involves vast manual effort and time in developing the information operation scripts. There are a range of varied tools and techniques used for pre-processing, furthermore as sampling, that selects a representative set from an outsized population of data; transformation, that manipulates data to provide one input; denoising, that removes noise from data; normalisation, that organizes data for additional economical access; and have extraction, that pulls out outlined data that's vital in some specific context[3][4][5].

Data pre-processing is a complex process. Data pre-processing is a complex process. Artificial processing data isn't available due to its unskillfulness. At present, the strategies are infinitely varied as a result of totally different areas have different processing desires. This paper describes the data pre-processing of a credit card in the field of fraud detection.

1. First, we totally order the attributes of transaction records, and then classify the values of every attribute.

2. Then, the paper handles with initial data through data cleaning, transformation, integration, and reduction, and obtains training sample data needed by mining.

DATA PRE-PROCESSING TECHNIQUES

One of the biggest problems associated with researchers in fraud detection is the lack of real-life data because of the sensitivity of data and privacy issue. Many researchers have done research with real-life data of bank with agreements to deal with this downside, several tools are accessible to come up with artificial data[6].

The second problem is to deal with imbalance data or skewed distribution because a number of fraudulent transactions are very less comparing to legitimate transactions. To overcome this problem, synthetic minoring oversampling methods are used to increase the number of low incidence data in a dataset that generate synthetic transactions related to the original data set. Cost based mostly sampling is employed to get artificial fraudulent transactions to balance data set.

Overlapping of data is one more problem as some of the transactions look like fraudulent transactions when actually they are a legitimate transaction, it's additionally attainable that deceitful transactions seem to be traditional transactions.

(a) (b)

Figure: Example of Noisy data

For example, in Figure (a), black points indicate the fraudulent instances, and white points indicate the non-fraudulent instances. Point A is regarded as the outlier or noisy" instance in the data since it should belong to the area of black points. In addition, it may affect the generalization performance of a classifier, if the classifier over-learns the outlier, which is shown in Figure (b). When such classifier is applied and evaluated on a testing dataset, some white points (i.e. non-fraudulent instances) cannot be correctly classified because of the poor generalization performance. Secondly, when using some modelling approaches (i.e. Neural Networks), the results are not straightforward for general users to understand. Thirdly, except for noisy data, real-life financial data have other features, such as biased data distribution.

The task of data pre-processing is to arrange the initial business data with the new “business model", clear those attributes impertinent to the aim of data mining, provide clean, accurate, simplified data thus improving the standard and potency of excavation beneath the steering of domain knowledge.

The data pre-processing mainly concludes data cleaning, integration, transformation, and reduction. In this way, the dirty, incomplete and inconsistent data can be corrected in the real world.

DATA PRE-PROCESSING IN CREDIT CARD CONSUMING BEHAVIOR MINING

Data extraction, transform and loading architecture in the design of the bank credit evaluation system should have the following functions: management simple; using metadata method, centralized management; interface, data format, transmission are strict norms; external data source; data extraction system processes automation, and automatic scheduling; extracted data timely, accurate and complete; can provide the interface with various data systems, provide software framework for the system, system adaptability; functional changes, the application can adapt to new needs with very little to change; scalability.

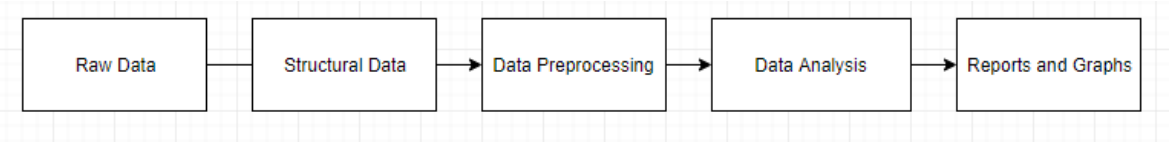
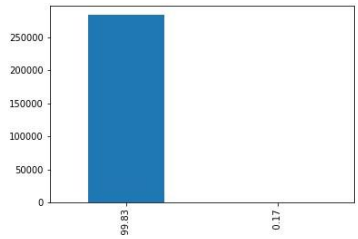


Figure: Data pre-processing flow

Practical Example: Credit Card Fraud Prediction

Github link—https://github.com/Kelechukwu1/Dealing-with-Imbalanced-Datasets, The ratio of non-fraudulent transactions to fraudulent transactions was a whopping—99.83% to 0.17%

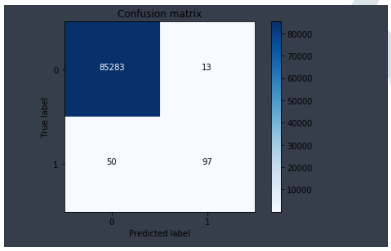


Imbalance in the dataset

We proceeded to model, selecting Logistic Regression as our algorithm. we trained the algorithm with the unbalanced dataset. Below was our result:

```
Accuracy score: 0.9992626663389628
Precision score: 0.8818181818181818
Recall score: 0.6598639455782312
F1 score: 0.754863813229572
Matthew Coefficient Score: 0.7624690650850618
```

Result for imbalanced data



Confusion matrix

Here the accuracy is very high, despite the presence of false positives and false negatives. We then proceeded to balance the dataset.

We used SMOTE (synthetic minority oversampling technique) to balance the dataset[7].

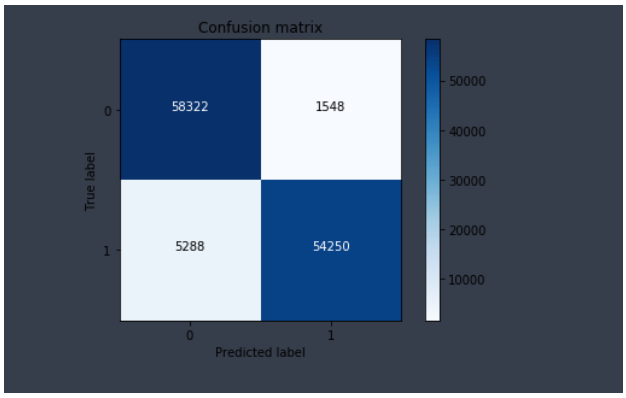
```
In [10]: #Oversample data
from imblearn.over_sampling import SMOTE
smote_algo = SMOTE(random_state=0)
smote_data_X,smote_data_Y = smote_algo.fit_sample(X_train, y_train)
smote_data_X = pd.DataFrame(data=smote_data_X,columns=X_train.columns )
smote_data_Y= pd.DataFrame(data=smote_data_Y,columns=["Class"])

#Join X and Y smote data into one
smote_data = smote_data_X
smote_data["Class"] = smote_data_Y["Class"]
```

Applying SMOTE algorithm

We then trained the algorithm with the balanced dataset and obtained these results:

```
Accuracy score: 0.9427509044620126
Precision score: 0.9722570701458834
Recall score: 0.911182774026672
F1 score: 0.9407296941111187
Matthew Coefficient Score: 0.887224086776177
```

Smote result



Confusion Matrix for smote As we can see, this model is much better than the others

CONCLUSION

This paper spreads out a discussion on data pre-processing in credit card data processing and did well for the procession. we have a tendency to remodelled the credit card data to the format of data mining, primarily by the subsequent means that, like adding the applied mathematics fields of the trade, data cleansing, conversion, data integration, and data discretization. This paper analyzed the performance of credit card dataset and it used SMOTE technique to overcome the imbalanced dataset difficulties.

REFERENCES:

1. Xiaohua Hu, DB-hreduction: A Data Preprocessing Algorithm for Data Mining Applications, 0893-9659/03- 2003 Elsevier Science Ltd
2. Fatin Zulkepli, Roliana Ibrahim- Data Preprocessing Techniques for Research Performance Analysis, from book Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2016 (pp.157-162)
3. R.M. Suresh ; R. Padmajavalli- An Overview of Data Preprocessing in Data and Web Usage Mining-1st International Conference on Digital Information Management- IEEE June2007
4. Mirela Danubianu, step by step data preprocessing for data mining. A case study, Proceedings of the International Conference on Information Technologies (infotech-2015) 17-18 September 2014, Bulgaria
5. Balaji Padmanabhan.Data Mining for Customer Segmentation: A Behavioral Pattern-Based Approach[A]. The Wharton School, University of Pennsylvania Jan,2004.
6. Pornwatthana Wongchinsri ; Werasak Kuratach -A survey - data mining frameworks in credit card processing, 2016 13th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)
7. Jia Li, Hui Li, Jun-Ling Yu, —”Application Of randomsmote On Imbalanced Data Mining”, 2011 Fourth International Conference On Business Intelligence And Financial Engineering