Logistic Regression Learning Model for Handling Concept Drift with Unbalanced Data in Credit Card Fraud Detection System

Pallavi Kulkarni and Roshani Ade

Abstract Credit card is the well-accepted manner of remission in financial field. With the rising number of users across the globe, risks on usage of credit card have also been increased, where there is danger of stealing credit card details and committing frauds. Traditionally, machine learning area has been developing algorithms that have certain assumptions on underlying distribution of data, such as data should have predetermined and fixed distribution. Real-word situations are different than this constrained model; rather applications often face problems such as unbalanced data distribution. Additionally, data picked from non-stationary environments are also frequent that results in the sudden drifts in the concepts. These issues have been separately addressed by the researchers. This paper aims to propose a universal framework using logistic regression model that intelligently tackles issues in the incremental learning for the assessment of credit risks.

Keywords Logistic regression learning · Concept drift · Class imbalance · Credit card fraud detection

1 Introduction

The credit card fraud detection has long been appraised as a basic issue in the educational and commercial society. It has become an important point for enterprise communities to assess threats in credit, improve cash flow, cut down frauds, and perform important decision management activities. To earn profit in case of economic organizations, correct results must be obtained and that is critical in

P. Kulkarni (✉) · R. Ade
Dr. D.Y. Patil School of Engineering and Technology,
Savitribai Phule Pune University, Pune, India
e-mail: kulkarnipallavi4@gmail.com

R. Ade
e-mail: rosh513@gmail.com

S.C. Satapathy et al. (eds.), *Proceedings of the Second International Conference on Computer and Communication Technologies*, Advances in Intelligent Systems and Computing 380, DOI 10.1007/978-81-322-2523-2_66

commercial credit card fraud detection. There are various data mining techniques that can be applied to real-world problems like this mentioned earlier to get accurate analysis and scientific solution of that problem [1].

If learning from enormous data which proceeds into iterations is required, then incremental learning algorithms are preferred. Stream data mining requires a learner that can be incrementally modified to achieve the full advantage of novelty of data, while simultaneously maintaining performance of a learner on older data. Multiple classifier systems, also known as ensemble techniques of machine learning area are pretty popular to learn from non-stationary environment and incremental learning. Non-stationary environment is kind of an environment where sudden concept drift/change can occur. Credit card frauds are considered as rare since their occurrence is less as compared to normal transactions. So the corresponding data are unbalanced because it consists of fraudulent events and regular transactions. Over the time period, attackers are inventing various ideas to commit frauds in financial systems. In case of credit card transactions intruders may steal credentials related to credit cards, they may hack the confidential web sites of banks, the plunder credit card itself. Hence the idea is to develop an algorithmic framework that tackles the problem of credit card risk assessment in non-stationary environment while considering the unbalanced nature of input data in incremental fashion while maintaining efficiency and improvement over existing techniques [2, 3].

2 Related Work

The following section illustrates existing work done in this research field.

2.1 *Credit Card Threat Assessment*

The technique to detect credit card frauds usually applies some classification mechanism on similar data of previous consumer to look for a relation between the characteristics and probable points of failure. One important ingredient needed to achieve this goal is to find an accurate classifier that is able to categorize new applicants or existing consumers as loyal or fraudulent. Many statistical techniques and optimization models, like linear discriminant analysis logit analysis, probit analysis, linear programming, integer programming, k-nearest neighbor (KNN) came into picture. These methods are capable to evaluate threats in credit card system, but they have some drawbacks. There is still room to improve the capability to distinguish between faithful and delinquent customers [1, 4].

Researchers noticed the fact that transpiring artificial intelligence approaches like artificial neural networks, support vector machines are beneficial over statistical models and optimization methods to assess credit card faults. Since ensemble techniques combine two or more distinct learners, have proven higher accuracy to predict risks than any specific methods. There is no doubt that credit card fault evaluation and modeling is one of the most critical themes in the area of economic risk assessment. In recent years, due to intense difficulties in commercial field, financial crises occurred, so more attention should be paid to credit threat assessment in banking and monetary service [5–10].

2.2 Concept Drift

In simpler terms, concept drift is change in class definition over the time. In such cases, it has been assumed that at time instance t, the algorithm A is stipulated with a group of labeled samples $\{X0,\ldots, X\mathrm{t}\}$, where Xi is considered as a v-dimensional vector and every sample has matching class label yj. If an unclassified sample arrives at time $t + 1$ as $Xt + 1$, then the algorithm is anticipated to supply a class label for $Xt + 1$. One of the class labels is predicted for newly arrived instance, the actual label $Yt + 1$ and a new testing sample $Xt + 2$ arrive so that one can continue with its testing. There is a data generating function which is responsible for creation of samples at a particular time instance and hence determining nature and distribution of data. In non-stationary environment, underlying and hidden distribution function (fh) changes over time. Change in distribution function can be described using several methods [11–16].

2.3 Combined Problem of Concept Drift with Unbalanced Data

In real world, tremendous applications in non-stationary environments suffer from class imbalance. This is challenging situation in many real-world applications which lead to inaccurate results. Examples include climate monitoring, network intrusion detection system, spam e-mail identification, and finally credit card fraud detection. However, it has been observed that the joint problem of concept drift and unbalanced data has drawn little attention from researchers worldwide. Ensemble techniques are intelligent methods that are able to handle concept drift along with class imbalance. Learn++ family of algorithms has been progressed over the time and has been addressing various issues associated with incremental learning in each new version of Learn++. Among them, Learn++.CDS (Concept Drift with SMOTE) and Learn++.NIE (Non-stationary and Imbalanced Environments) are truly incremental i.e., one pass approaches that do not access previous data and also handle class imbalance problem well. Learn++.CDS applies SMOTE algorithm to balance the classes and its samples. It is integration of two algorithms, Learn++.NSE (Non-stationary Environment) and Synthetic Minority Oversampling TEchnique (SMOTE), which are existing approaches for handling concept drift and class imbalance, respectively. Learn++.NIE is more efficient framework, which will be discussed in detail in next sections. Both the approaches are able to obtain new knowledge and preserve former knowledge about the environment, which is useful for recurring concepts (e.g., spiral dataset) [5, 6, 11, 15, 16]. A complete survey of concept drift, class imbalance and other incremental learning issues can be found in the review paper [13].

3 Methodology

This section describes a framework that addresses both the issues of stream data mining, i.e., concept drift and class imbalance for credit card threat assessment. For balancing the data, the proposed framework implements sub-ensembles with bagging along with alternate error measures. Bagging variation is a function that smartly tackles class imbalance, it neither generates synthetic samples nor faces under-sampling problem and hence is efficient in terms of performance.

Note that this framework is designed for handling unbalanced data. So obviously it works well when there is concept drift in majority and minority classes of unbalanced data. Figure 1 represents the basic architecture of the proposed framework.

This paper proposes an improved algorithmic system which smartly handles unbalanced data in non-stationary environment with enhanced accuracy in results. Instead of using f-measure as in earlier systems for weight modification, proposed framework uses mapping function which in turn is to be implemented by training Multi-Layer Perceptron neural network (MLP) with backpropagation algorithm. Figure 2 depicts the basic organization of mapping function. As proved in the experiments in next sections, use of MLP in a mapping function leads to knowledge transformation and improved accuracy in the results. This system can also be used for different applications which require handling of concept drift with unbalanced data like spam e-mail identification, student performance prediction, and weather forecast.

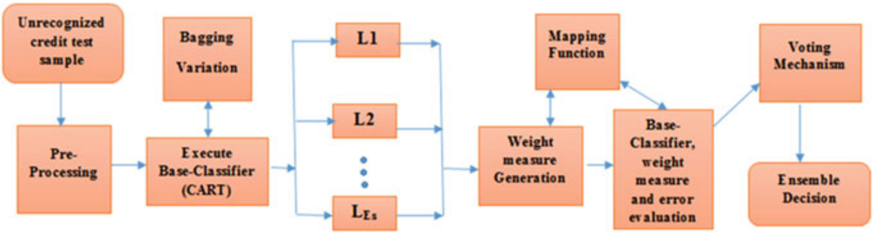

Unrecognized credit test sample
Pre-Processing
Bagging Variation
Execute Base-Classifier (CART)
L1
L2
L_{Es}
Weight measure Generation
Mapping Function
Base-Classifier, weight measure and error evaluation
Voting Mechanism
Ensemble Decision

Fig. 1 Architecture of the proposed framework using logistic regression model

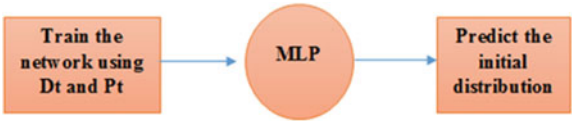
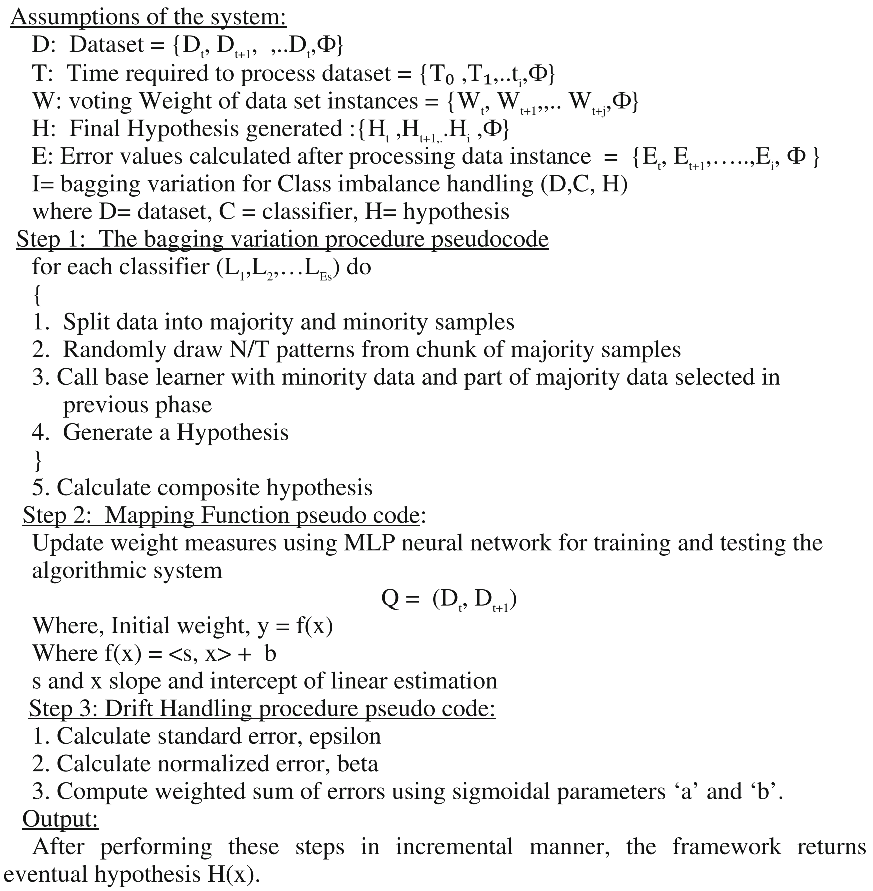

Train the network using Dt and Pt
MLP
Predict the initial distribution

Fig. 2 Organization of mapping function

3.1 Proposed Framework Pseudocode



```
Assumptions of the system:
  D:  Dataset = {$D_t$, $D_{t+1}$,  ,..$D_t$,Φ}
  T:  Time required to process dataset = {$T_0$ ,$T_1$,..$t_i$,Φ}
  W: voting Weight of data set instances = {$W_t$, $W_{t+1}$,,.. $W_{t+j}$,Φ}
  H:  Final Hypothesis generated :{$H_t$ ,$H_{t+1,..}$.$H_i$ ,Φ}
  E: Error values calculated after processing data instance  =  {$E_t$, $E_{t+1}$,…..,$E_i$, Φ }
  I= bagging variation for Class imbalance handling (D,C, H)
  where D= dataset, C = classifier, H= hypothesis
Step 1:  The bagging variation procedure pseudocode
  for each classifier ($L_1$,$L_2$,…$L_{Es}$) do
  {
  1.  Split data into majority and minority samples
  2.  Randomly draw N/T patterns from chunk of majority samples
  3. Call base learner with minority data and part of majority data selected in
      previous phase
  4.  Generate a Hypothesis
  }
  5. Calculate composite hypothesis
Step 2:  Mapping Function pseudo code:
  Update weight measures using MLP neural network for training and testing the
  algorithmic system
                                          $Q = (D_t, D_{t+1})$
  Where, Initial weight, y = f(x)
  Where f(x) = <s, x> +  b
  s and x slope and intercept of linear estimation
Step 3: Drift Handling procedure pseudo code:
  1. Calculate standard error, epsilon
  2. Calculate normalized error, beta
  3. Compute weighted sum of errors using sigmoidal parameters ‘a’ and ‘b’.
Output:
  After performing these steps in incremental manner, the framework returns
eventual hypothesis H(x).
```

4 Results and Experiments

This section illustrates the experiments held and corresponding output. In the first experiment, data is partitioned into 11 approximately equal slices. In each phase, one chunk is chosen for testing on ad hoc basis, and remaining 9 splits are stipulated sequentially to the proposed framework. Depending on the size of data piece, system produces 4–5 standard error values. Here, these error values are averaged across all executions for simplicity. Table 1 summarizes standard and normalized error values obtained across executions of 11 data bags. The bag size is calculated using standard formula. It is as follows: bag size = no. of instabce * m_bagsize-percent/100. Base classifier employed here is CART. PCA is another option. Ensemble size (Es) is another parameter which can change the performance of the system. Here, it is calculated at runtime depending on the data and its seed values. Table 2 summarizes overall performance of the algorithm in terms of accuracy and error. It has been observed that out of bag error for each iteration resulted as 0. This is absolutely good if we consider the nature of the data and system. While determining the performance prediction of any algorithm, the nature and size of the data plays an important role. This system uses German Credit dataset to solve credit risk evaluation problem.

This is the standard dataset downloaded from UCI repository having numerical instances suitable for incremental learning. It consists of 2 classes: good and bad which describe that the threat is present if the resulting sample is found to be bad. Statistical nature of this data has been tested and it is found that data is discrete.

Table 1 Averaged error values computed in executions of the framework

Data bag no.	Standard error (epsilon)	Normalized error (beta)
1	0.2594	0.9782
2	0.2452	1.0427
3	0.2224	1.1335
4	0.2657	0.9652
5	0.2837	0.9905
6	0.2169	1.1668
7	0.2744	0.9448
8	0.2743	0.9544
9	0.2020	1.0165
10	0.2436	1.0611
11	0.2031	1.0148

Table 2 Performance of the system

Mean absolute error	Relative absolute error	Accuracy	Kappa statistics
74.83 %	0.406	96.6243 %	0.2847

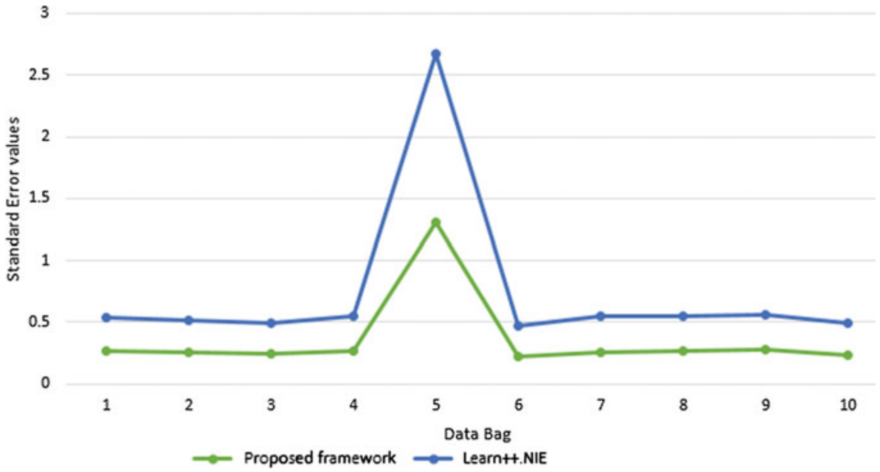


Fig. 3 Comparing performance of Learn++.NIE with proposed framework

Note that data bag is the part of entire dataset used for rebalancing the dataset. Databags get generated while processing of the algorithm with the provided dataset. The major comparison of this proposed framework with existing technique (i.e., Learn++.NIE) is depicted in Fig. 3 which shown below. It compares the averaged standard error values of Learn++.NIE and proposed framework with respect to credit card fraud detection problem.

As the graph illustrates, the error is less if one looks at performance of proposed credit risk assessment framework. In other words, it performs better and overcomes limitations of existing techniques. As said earlier, it has major benefit of knowledge transformation since MLP is applied in proposed framework.

Pseudocode can be mapped to results in the sense that the error values represents drift in the system and data bag is used for rebalancing the data after applying it to Bagging variation procedure.

Since this framework uses base classifier as CART, the results can be represented in the form of decision tree whereby leaf nodes will show the classes and intermediate nodes and path toward leaf node is modeled as complete conditions for the risk assessment. Consider an example, attribute values obtained as: checking status is <=0, duration is <=11, foreign_worker = yes, existing credits <=1, property magnitude = real estate, then class = good. It means that if the fields are having these particular conditions as true, then the corresponding customer is good and there is no fraud in credit. Drawbacks of the proposed system is it becomes complex and takes more time to train and test the system as it uses neural network approach. Since this system addresses wide range of issues, this complexity is affordable for large scale applications (Fig. 4).

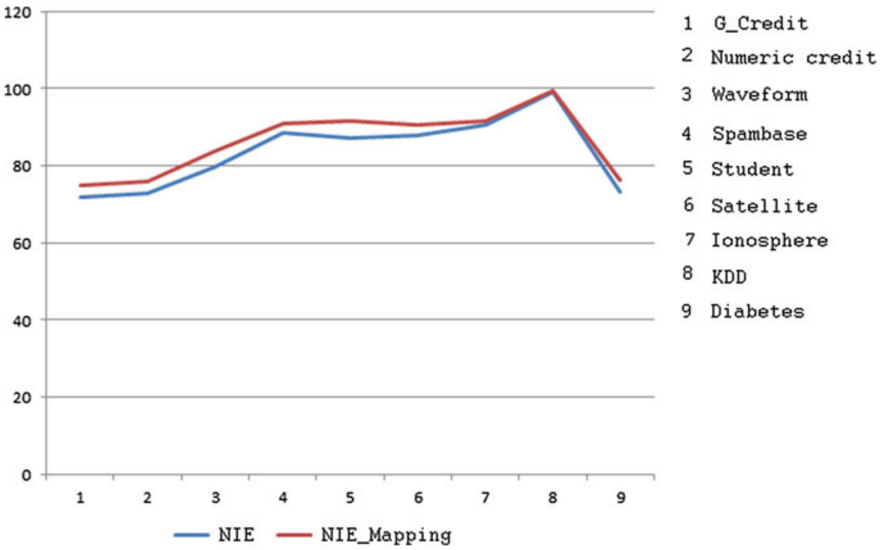


Fig. 4 Comparing performance of proposed framework on different real-time datasets

5 Conclusion and Future Enhancement

If the final objective of clever machine learning techniques is to be able to address a broad spectrum of real-world applications, then the necessity for a universal framework for learning from and adapting to environment where drift in concepts may occur and unbalanced data distribution is present can be exaggerated using the proposed algorithmic system. This system intelligently tackles challenges in incremental learning using logistic learning model of deep learning which in turn leads to knowledge transformation and preservation.

Scope of future work is leading toward a smart technique that will optimally discard drawbacks of the existing research and additionally assessment of these approaches on large scale, real-world applications consisting of formal statistical analyses of these systems, on certain non-stationary environments like Gaussian distribution drifts.

References

1. Wang, G., Ma, J.: A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. Expert Syst. Appl. **39**(5), 5325–5331 (2012)
2. Ditzler, G., Polikar, R.: Incremental learning of concept drift from streaming imbalanced data. IEEE Trans. Knowled. Data Eng. **25**(10), 2283–2301 (2013)
3. He, H., Chen, S., Li, K., Xin, X.: Incremental learning from stream data. IEEE Trans. Neural Networks **22**(12), 1901–1914 (2011)

4. Razavi-Far, R., Baraldi, P., Zio, E.: Dynamic weighting ensembles for incremental learning and diagnosing new concept class faults in nuclear power systems. IEEE Trans. Nucl. Sci. **59** (5), 2520–2530 (2012)
5. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Inf. Comput. **108**(2), 212–261 (1994)
6. Kulkarni, P., Ade, R.: Prediction of student's performance based on incremental learning. Int. J. Comput. Appl. **99**(14), 10–16 (2014)
7. Yu, L., Wang, S., Keung Lai, K.: Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst. Appl. **34**(2), 1434–1444 (2008)
8. Denison, D.G.T., Mallick, B.K., Smith, A.F.M.: A bayesian CART algorithm. Biometrika **85** (2), 363–377 (1998)
9. Ade, R., Prashant D.: Efficient knowledge transformation for incremental learning and detection of new concept class in students classification system. In: Information Systems Design and Intelligent Applications, pp. 757–766. Springer, India (2015)
10. Ade, R., Deshmukh, P.R.: Classification of students using psychometric tests with the help of incremental Naïve Bayes algorithm. Int. J. Comput. Appl. **89**(14), 26–31 (2014)
11. Elwell, R., Polikar, R.: Incremental learning of concept drift in nonstationary environments. IEEE Trans. Neural Networks **22**(10), 1517–1531 (2011)
12. Ade, M.R., Pune, G., Deshmukh, P.R., Amravati, S.T.: Methods for incremental learning: a survey. Int. J. Data Mining Knowled. Manage. Process **3**(4), 119–125 (2013)
13. Pallavi, K., Ade, R.: Incremental learning from unbalanced data with concept class, concept drift and missing features: a review. Int. J. Data Mining Knowled. Manage. Process **4**(6) (2014)
14. Polikar, R., Upda, L., Upda, S.S., Honavar, V.: Learn++: An incremental learning algorithm for supervised neural networks. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **31**(4), 497–508 (2001)
15. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational learning theory. Springer, Heidelberg, pp. 23–37 (1995)
16. Muhlbaier, M.D., Apostolos T., Polikar, R.: Learn. NC: combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes. IEEE Trans. Neural Networks **20**(1), 152–168 (2009)