Chapter 23
Learning from Unbalanced Stream Data in Non-Stationary Environments Using Logistic Regression Model:
A Novel Approach Using Machine Learning for Assessment of Credit Card Frauds

Pallavi Digambarrao Kulkarni
Dr. D. Y. Patil School of Engineering and Technology, India

Roshani Ade
Dr. D. Y. Patil School of Engineering and Technology, India

ABSTRACT

There are several deep learning approaches that can be applied for analyzing situations in real world problems and inventing their solution in a scientific technique. Supervised data mining methods that predicts instance values, using previously obtained results from already collected data are pretty popular due to their intelligence in machine learning area. Stream data is continuous form of data which can be handled by using incremental learning approach. Stream data learning may face several challenges in real world like concept drift or class imbalance. Concept drift occurs in non-stationary environment where data distribution generation function is dynamic in nature and has no fixed formula to predict the future data distribution nature. Neural network techniques are intelligent enough to improve performance of algorithmic systems that work in such problem domains. This chapter briefly describes how MLP technique is integrated in system so that the system becomes a complete framework for handling unbalanced data with concept drift in the incremental learning strategies.

DOI: 10.4018/978-1-7998-0414-7.ch023

INTRODUCTION

The fundamental activity for both human and machine learning is the knowledge acquisition. It is the ultimate objective of any learning system which in turn is expected to lead towards the wisdom of the entire system. The crude definition of data mining system is to dig massive amounts of data and extracting information out of it. The term is quite misleading as it should actually be knowledge extraction as the system is extracting processed knowledge out of raw data. Supervised systems have knowledge since it has already classified labeled data to be used as of training data. It form instructions or postulates using this already classified data and uses these assumptions to calculate the classes of forthcoming data illustrations. These approaches are attractive due to their intellect in the machine learning area. However, real world problems are quite complex and need expertize to have accurate results. There are several classification algorithms to apply on such situations such as nearest neighbor classification technique, probabilistic algorithms, entropy based classifiers, support vector machines, artificial neural networks, decision tree classifiers etc. Additionally, there are various deep learning approaches that can be applied for analyzing circumstances in real world problems and inventing their solution in a scientific technique(Polikar et al., 2001, Kulkarni, & Ade, 2014).

Real world problem consists of situations where the generated data is not static, instead it is dynamically produced and mostly this happens in the computer networks. As this is the era of internet and widespread smart devices connected to each other across the globe, huge amount of data explosion takes place in the online environment. Real challenge is to handle time factor associated with this online stream data and preserve the knowledge extracted at each time instance accordingly. Learning from such data is also known as incremental learning technique that is always fed with input data in its arrival sequence and should primarily possess following properties:

1. The algorithm should gain knowledge which is additional in recent data.
2. To train already present classifier, the algorithm need not access to the initial data.
3. It should keep up the knowledge which it has previously learned (no catastrophic forgetting should occur).
4. New data may bring in concept class.

Stream data evolves over the time period and one can observe rapid increase in the amount of data stored. Stream data mining field handles various issues associated with incremental data such as class imbalance, missing features, concept drift, concept class etc. The data may be unbalanced if it consists of few samples of one class and more data samples of another class. So this majority-minority nature of data creates new challenge to learn from it. Generally, the data generating function has some predetermined form, and if that function has not predefined format and may change over time then that particular environment is known as non-stationary environment where sudden concept drift can occur. Firsthand classes may be announced over the time in case of incremental records, such classes should be renowned and data should be clustered consequently. This issue is called as concept class problem (Polikar et al., 2001).

Logistic regression learning model is quite popular statistical model which can be used to solve several real world problems. If merged into machine learning tactic, it is perceived that this system certainly works sound. In this chapter, the concept of logistic regression learning model is used in the application of multilayer perceptron neural network and that system is incorporated in the algorithm such that it