CHAPTER I

INTRODUCTION

1.1. Statement of the Problem

In the era of very large data available called Big Data, the massive amounts of data management is very interesting and challenging in almost research fields such as business, science, botany, computer vision, information retrieval, data mining, and so forth. The term "Big Data" is currently defined by five data characteristics including volume variety, veracity, valve and velocity [1]. The term Big Data Analytics is defined as the process of analyzing and understanding the characteristics of massive size datasets by extracting useful geometric and statistical patterns. Ideally these five characteristics of a dataset increase the complexity of the data and thus make the current techniques and technologies stop functioning as expected within a given processing time [2]. For dealing with very large data classification problem, batch incremental learning algorithms are not interested among researchers because of long taken learning time and large memory usage. Currently, incremental learning algorithms play role in dealing with Big Data classification issue. Various incremental learning algorithms have been proposed to deal with the classification problem with large data sets for learning time reduction.

Besides the very large data is concern, for real-world applications, the situation of complete training dataset is merely given in classification just when constructing a classification model. Actually in most applications in pattern recognition, data mining, time-series prediction, and so on, data are available in a stream of chunks. Since the data distribution does not know and cannot control in advance, both the size of the chunks and the datum in each chunk are given randomly. The data stream classification has attracted extensive attentions in pattern recognition, data mining and time series prediction and so on. However, the most important challenge in this field is how to timely and accurately classify the continuous stream of data chunks with fast computational time and less storage unit [3].

In this work, the Data-throwaway Learning for Streaming Chunk data classification (DLSC) is proposed. The DLSC is an expanded work of [4] to deal with the parameter update for multiple data points under one-pass-throwaway concept

in which the data are presented to the network only once and then, thrown away from the learning process. In [4], a hidden neuron is added and expanded incrementally to cover the data entirely according the incoming datum. Since the number of added neurons depends on the location of incoming datum in the feature space, the total number of neurons is sensitive to the incoming datum sequence. The performance in terms of classification accuracy and the number of hidden neurons also depends on the sequence of the incoming datum. Therefore, the VEBF with incremental learning algorithm for one datum is not efficient to manage the data stream scenario. The structure complexity and learning time complexity is improved by extending the constraint on one incoming datum to a chunk of multi-class data.

1.2. Related Works

Recognizing and classifying a pattern into an appropriate target is one of the most important problems prevailing in vast scientific and engineering researches e.g. face recognition [5-8], object recognition [9-12], pattern recognition [13-15], and pattern classification [16-19]. In the past, this problem concerned only a set of stationary and non-temporal data. Furthermore, the amount of data was rather small (less than 10,000 patterns). However, the previous approaches to this problem, as summarized in the reference list, cannot be deployed to the situation where tremendous amount of new data are generated in every second. An obvious example of this case is the data in bioinformatics and business areas. Many learning algorithms still learn a new data set with combining the previously learned data set with the old learned data set. [It is very costly to the taken learning time and memory usage such as standard MLP neural network and Support Vector Machine (SVM) so forth.] In recent years, the problem of developing fast learning algorithms with high classification accuracy has been a considerable interesting subject in machine learning. Various incremental learning algorithms have been proposed to deal with the classification problem with large data sets for learning time reduction.

The incremental learning algorithms can be categorized into two types based on the number of samples presented in the learning system [20], i.e. (i) online incremental learning of which only one sample is presented for each epoch, and (ii) batch incremental learning of which a suitable subset of samples is presented for each epoch. Polikar et al. [21] proposed an ensemble classifier for incremental learning called *Learn++* in which weak hypotheses were generated and combined by

weighted majority voting for class prediction. In their work, a relatively small MLP acted as a base classifier or weak classifier. Their experimental results showed that *Learn++* classifier outperformed fuzzy ARTMAP on four benchmarks and real-world data sets but the classifier is sensitive to parameters of the used network. Wilson and Martinez [22] proposed a general inefficiency of batch learning for gradient descent learning. Their results of recognition tasks demonstrated that the gradient descent incremental learning is significantly faster than batch learning, with no apparent difference in accuracy. An incremental learning method for the probabilistic RBF or PRBF network was proposed to handle the classification problems [23] by sequentially adding a new component for stationary environment. The procedure of sequential component addition started with one component and repeated until there was no component containing data points belonging to more than one class. Their experimental results showed that the incremental PRBF method outperformed the standard hierarchical PRBF and SVM methods in terms of accuracy and computational time. Furao and Hasegawa [24] proposed a fast prototype-based nearest neighbor classifier called Advanced SOINN Classifier (ASC). The ASC method automatically learned the number of prototypes needed to determine the decision boundary and learned new information without losing old information. The ASC was empirically compared with other prototype-based classifiers. The results showed that ASC provided the best performance. The limit of ASC was the difficulty of used parameters determination and it cannot be applied to real-time data. All learning data must be stored for deleting a prototype with no usage for classification. Hua Duan et al. [20] proposed an incremental learning algorithms for Lagrangian Support Vector Machines (LSVM) in both online and batch incremental manners. They introduced the matrix inverse computation based on previous information. The results showed that LSVM was a fast and efficient algorithm compared with other online and batch incremental learning based on LSVM. Jaiyen et al. [4] proposed a new study based on the condition of one-pass-throw-away learning for static environment. They also introduced the VEBF neural network using only new incoming datum presented to the network for learning process. This technique could be considered as prototype-based classifier. Their technique reached the lowest bound on time complexity and achieved the smallest network structure, but the situation of more than one datum for parameter update has not been considered.

Currently, the data stream classification has attracted extensive attentions in pattern recognition, data mining and time series prediction and so on. However, the most important challenge in this field is how to timely and accurately classify the

continuous stream of data chunks with low computational time and less storage unit [3]. Some incremental learning cannot cope with the data stream classification such as [24] for which the complete training data was assumed. Furthermore, many sequentially incremental learning algorithms, such as [4] were affected on the order of presented data sample or chunk of data sample. So, various incremental learning for stream data have been widely interested and proposed. Domingos and Hulten [25] proposed Hoeffding trees, a method for online learning from the high-volume data stream called VFDT. The experimental studies showed its effectiveness in taking advantage of massive numbers of samples but this method obtained tree with quite large size. Pang et al. [26] proposed an incremental linear discriminant analysis ILDA, considered as incremental feature extraction, in both types of incoming data: sequential and chunk. The proposed ILDA was tested on various data sets ranged from small numbers of classes and features to large numbers of classes and features. The results showed that the proposed ILDA could effectively evolve a discriminant eigenspace over a fast and large data stream, and extract features with superior discriminability in classification, when compared with traditional LDA. Wan and Banta et al. [27] proposed parameter incremental learning for Multilayer Perceptron (MLP) neural network. The proposed method was evaluated on both function approximation and classification. The experimental results on three benchmark data sets were shown that the proposed learning algorithm for MLP was measurably superior to the standard online backpropagation (BP) algorithm and the stochastic diagonal Levenberg–Marquardt (SDLM) algorithm in terms of the convergence speed and accuracy. Ozawa et al. [28] proposed chunk incremental principal component analysis called chunk IPCA in which feature extraction and classifier learning are simultaneously performed. The discussion of scalability of chunk IPCA under one-pass incremental learning environments was provided. The experimental results showed that chunk IPCA could reduce the training time comparing with sequential IPCA and obtain major eigenvectors with fairly good approximation as well. Ye et al. [29] proposed an Incremental Learning Vector Quantization (ILVQ) algorithm for pattern classification also viewed as a prototype-based classifier. The ILVQ was compared with other incremental learning in stationary and incremental environment. The experimental results showed that the ILVQ outperformed other incremental learning methods in terms of accuracy and compression ratio.

1.3. Objectives

1. To propose the incremental learning algorithm to deal with large data classification problem.
2. To propose the incremental learning algorithm to deal with the streaming chunk data classification problem.
3. To propose the learning algorithm under one-pass-throwaway concept.

1.4. Scope of Work

In this dissertation, the scope of work is constrained as follows:

1. The performance of a model is measured in terms of classification accuracy (%), the number of hidden neurons and the learning time.
2. The experimental data sets were mainly obtained from UCI machine learning repository [30] to generate the streaming chunk.

The dissertation is organized as follows. In Chapter II, an overview of Versatile Elliptic Basis Function Neural Network (VEBFNN) is given with the previous incremental learning algorithm for only one datum for parameter update [4]. Chapter III presents the studied problem and the Data-throwaway Learning for Streaming Chunk data classification (DLSC). Chapter IV presents discusses the model evaluation, experimental setting and experimental results. Finally, Chapter V concludes this work.