CHAPTER 1

Introduction

1.1 Background and motivation

In recent years, scientific knowledge and discovery become more advanced. One factor is due to the rise of computers developed to have larger memory and more computational efficiency enabling scientists and research communities to develop sophisticated mathematical methods to solve several practical problems.

Supervised learning [24], an approach in science and engineering to solve data prediction problems, which includes classification and regression essentially involves a computing procedure called “training” on a collection of data. In general, these data are different in types, formats, sampling rates, and length, and in some cases, the number of data can be large. Some problems, e.g. load forecasting [8], electrocardiograph reconstruction [1-7], power system stabilizer [10], and flooding prediction, simply cannot avoid dealing with a large collection of data. Their training process is likely to be computationally expensive requiring large memory usages and lengthy computing time.

As a result, numerous research has tried to address this issue concerning large training data set. One of the solutions is to develop a reduction method to reduce the size of data. Generally, there are three types of data reduction, i.e. feature reduction [24], prototype selection [21], and prototype generation [34]. Feature reduction reduces data sizes by removing some features from the data. Although the number of features can be reduced, the number of samples does not change. The decrease in number of features possibly leads to poor prediction or regression accuracy.

Another type of data reduction is prototype selection. It reduces the number of samples by selecting and maintaining only some parts or prototypes of the data that are important for later uses in supervised learning tasks. Meanwhile, insignificant parts of data are removed. For prototype generation, data reduction is performed rather differently from prototype selection. It tries to construct a new smaller training data set based on the original data. The new data set may include samples which do not belong to the original data but are the best representatives of the overall data. All types of data reduction techniques pursue the same goal to decrease the number of data so that existing supervised learning algorithms are able to faster perform a calculation. In this thesis, we particularly concern about developing data reduction techniques that fall into two categories: prototype selection for classification problems and prototype generation for regression.

Considering the nature of the data in regression and classification, it is known that the output of the training data set is continuous for regression. That is the magnitudes of the output is real numbers. In contrast, the magnitudes of the output training data for classification are discrete and usually finite. Therefore, these properties lead to different reduction method for regression and classification problems.

Data reduction is an indispensable preprocessing step to allow supervised learning algorithms such as Support Vector Machine (SVM) [9] to process big data. In order to illustrate the possibility of data reduction, we provide a data reduction example for regression in an electrocardiograph (ECG) problem. An example for classification is also given based on the artificial banana dataset.

In an ECG reconstruction problem, an ECG signal consisted of four inputs and one output is shown in Figure 1.1. Each input signal contains 4 intervals where the sample points A to D, E to H, I to L, and M to P lay across the first, second, third, and fourth interval, respectively. For the first interval, the points A to D have different input values though the output values are the same. This characteristic also happens in the other intervals which implies unimportant details or redundancies in the data. The duplication of data not only occurs ECG signals but in other data types such as electrical load series shown in Figure 1.2. Therefore, a data deduplication process should be applied beforehand in order to reduce computing time and memory consumption.

Since the same output value can be produced from different inputs, and also there is a chance that the inputs can be redundant, we propose an idea to exploit the relation between the input and output of the training data set to reduce the number of training data. This process starts by grouping the output data at the same level, and then groups the similar input data and selects appropriate prototypes which represent the data in each group. By selecting the prototype, we are able to reduce the neighboring and redundant data. This approach is one solution to reduce the data for regression problem.

For classification, data reduction can also be performed and we provide an illustration for the two-class artificial banana data set in Figure 1.3. According to Figure 1.3(a), the number of the original training samples is relatively large. Therefore, the computational cost can be high in training. We applied the relation between the input and output to extract the class boundaries whose result is in Figure 1.3(b). The number of the extracted data is about 10 percent of the original data. When the reduced data is trained with a supervised learning algorithm, the resulted classification accuracy is nearly close to the one that is from the original data.

From the possibility in performing data reduction, this research intends to develop data reduction methods to reduce the number of samples for classification and regression problems. The basis is to remove unnecessary or duplicated data from the training set by considering the input-output relationship of the data in order to make a decision on which data points to keep or discard.

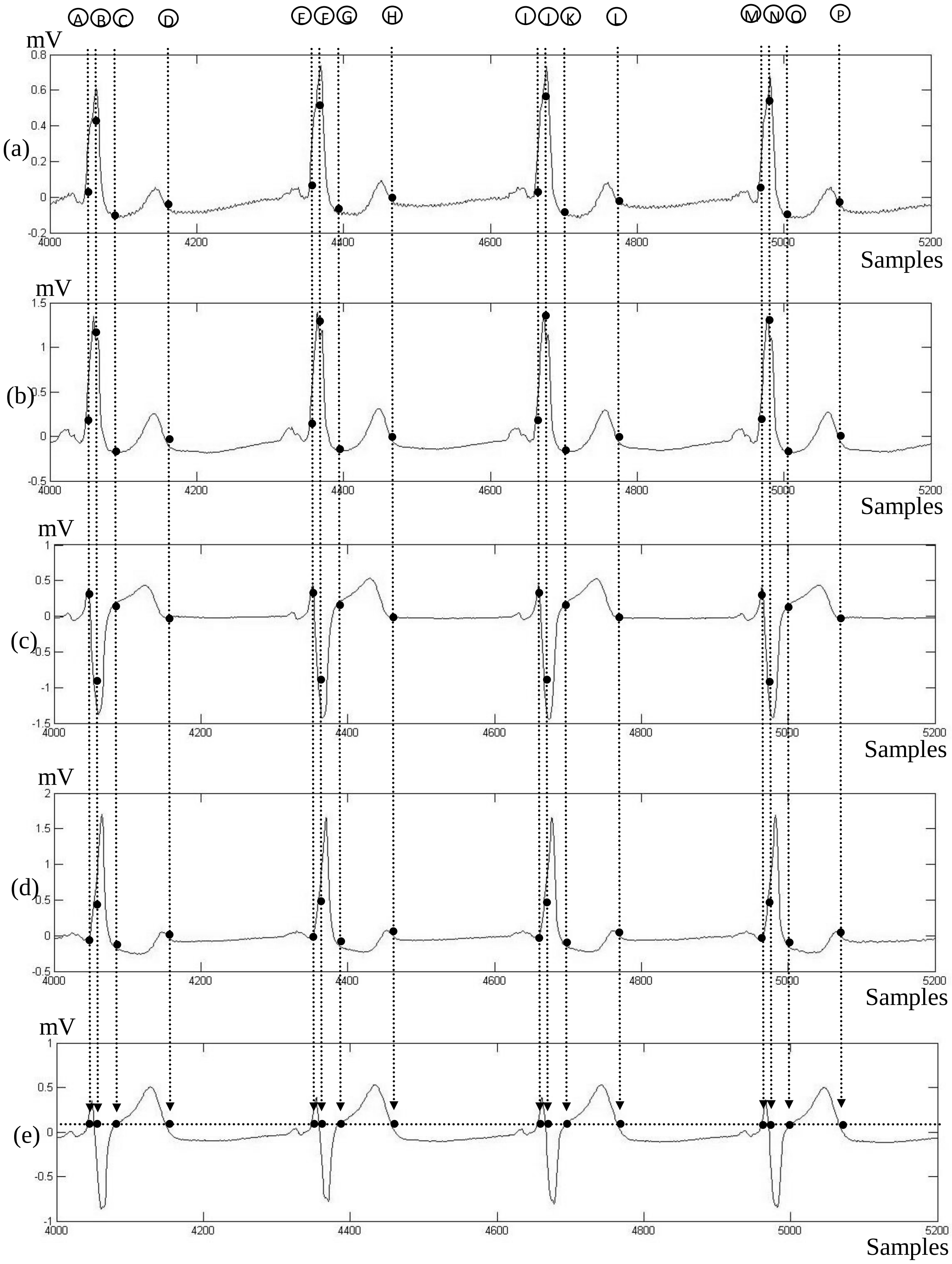

Figure 1.1 (a)-(d) ECG input signals, (e) ECG output signal.

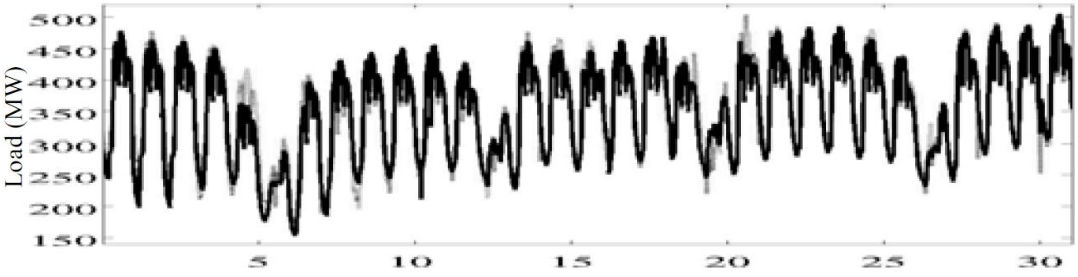

Figure 1.2 Electrical load series.

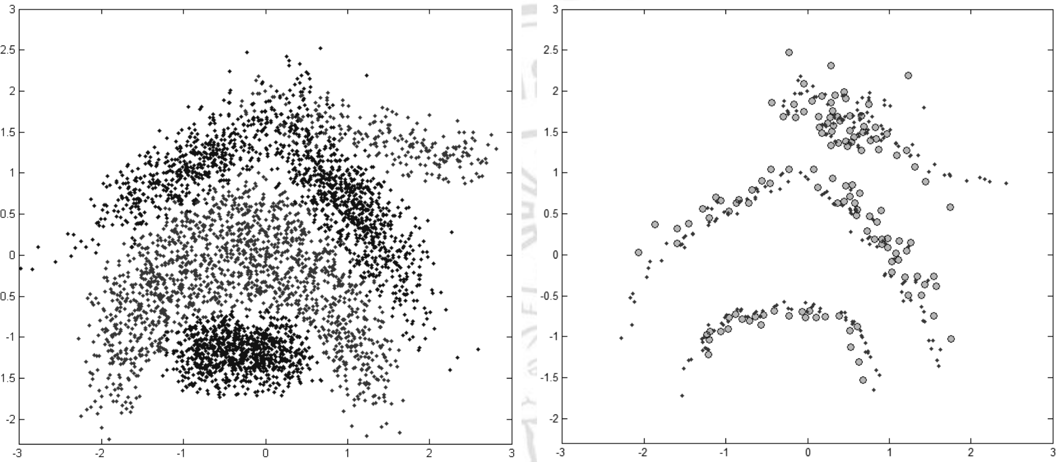

Figure 1.3 Banana data set (a) Original data set, (b) Class boundary data.

1.2 Literature review

Due to the different property of output data from regression and classification as mentioned in the previous section, this research separates data reduction techniques into 2 types, i.e., data reduction for regression and for classification problems. The related works concerning each type of reduction are explored and described as follows.

1.2.1 Data reduction for regression problems

Very few research works related to instance reduction for regression were done. Some of the studies include the works by S. Eschrich *et al*. [11] and S. Lee and X. Zeng [12] where they applied the fuzzy theory to solve the instance reduction problem.

S. Eschrich *et al*. [11] proposed brFCM which is the algorithm to reduce the number of distinct patterns that clustered without adversely affecting partition quality. The reduction processes started with aggregating similar examples and then used a weighted exemplar in the clustering process. The reduction of clustering data allowed a partition of the data to be produced faster. Their algorithm was applied to segment 32 magnetic resonance images into different tissue types and segment 172 infrared images into trees, grass, and target. The average speed up as much as 59–290 times of the traditional implementation of fuzzy c-means was obtained using brFCM while producing partitions that were equivalent to those produced by fuzzy c-means.

Next, S. Lee and X. Zeng [12] presented a clustering-based approach to fuzzy system identification. In order to construct an effective initial fuzzy model, they presented a modular method to identify fuzzy systems based on a hybrid clustering-based technique. The determination of the proper number of clusters and the appropriate location of clusters was one of the primary considerations on constructing an effective initial fuzzy model. A hybrid clustering algorithm concerning input, output, generalization, and specialization was introduced. The proposed clustering technique, three-part input-output clustering algorithm, integrates a variety of clustering features simultaneously, including the advantages of input clustering, output clustering, flat clustering, and hierarchical clustering, to effectively perform the identification of clustering problem. From the idea of this kind of input-output clustering utilized in this paper [12], this thesis also makes use of the relation between input and output data to develop the data reduction process for regression problems.

1.2.2 Data reduction for classification problems

Unlike regression, numerous studies have been conducted to solve instance reduction for classification problems. Some proposed methods in recent years usually concern solving integer programming or genetic algorithms such as in the following works.

Hamidzadeh *et al*. [18] developed a large margin instance reduction algorithm called LMIRA. LMIRA removed non-border instances and kept border ones based on keeping the hyperplane that separated two-class data and provided large margin separation. This instance reduction process was formulated as a constrained binary optimization problem and solved by employing a filled function algorithm. The performance was examined on 20 data sets and the results were compared to six states of the art instances reduction algorithms. LMIRA yielded the lowest classification error rate among the other algorithms and had the highest instance reduction percentage, while this algorithm maintained competitive execution times.

Nikolaidis *et al*. [19] proposed direct weighted pruning (DWP) algorithm that uses a set of weights to directly control which prototypes should be discarded or survived. It uses a simple genetic algorithm to optimize a bi-objective index which incorporated the condensation rate and a measure of the classification inaccuracy as reflected by the nearest neighbor rule. The experiments over 18 data sets with the comparisons on 8 condensation algorithms showed that DWP was effective and achieved the highest classification accuracy along with competitive condensation rates.

Although the above-mentioned techniques for data reduction could provide good accuracy and reduction results, the processes are overly complex and hard to apply with large data. Some methods that tried to solve data reduction problem with larger number of instances include the following works.

D.R. Wilson *et al*. [21] presented the paper that has two main purposes. First, his report provides a survey of existing algorithms used to reduce storage requirements in instance-based learning algorithms and other exemplar-based algorithms. Second, they presented decremental reduction optimization procedure 1-5 (DROP1–DROP5) and the decremental encoding length (DEL) that could be used to remove instances from the concept description. These algorithms and 10 previous algorithms were compared using 31 classification data sets. DROP3 gave the best result of generalization accuracy and storage requirements of the DROP methods and had significantly higher accuracy and lower storage than CNN, SNN, and IB2 algorithms. However, somewhat lower accuracy but significantly lower storage than ENN, RENN and ALL K-NN algorithms and significantly worse storage but significantly better accuracy than ELGROW and EXPLORE algorithms.

W. Lam *et al*. [13] proposed a new framework for discovering good prototypes, called ICPL (Integrated Concept Prototype Learner). Under this framework, two kinds of concept prototypes were separately learned by abstraction and filtering techniques. The concept prototypes were then integrated by maintaining a balance of different kinds of prototypes based on locality using specially designed integration methods. As for the abstraction component in ICPL, existing methods such as clustering required a high computation time so they are incapable of handling large data sets. To mitigate this problem, they proposed a novel abstraction method based on typicality, with relatively low-computational cost.

Fayed *et al*. [14] proposed the idea which based on chain. Chain was a sequence of nearest neighbors from alternating classes. They made the instance that patterns further down the chain close to the classification boundary and based on a cutoff value for the patterns which they keep in the training set. This algorithm is named template reduction for KNN (TRKNN).

Olvera-Lopez *et al*. [15] proposed a new fast prototype selection method for large data sets based on clustering algorithm which selected border and some interior prototypes. This method called Prototype Selection by Clustering (PSC).

Nikolaidis *et al*. [16] introduced a multi-stage method for instance selection called Class Boundary Preserving (CBP) algorithm. CBP is a hybrid method which select and abstracts instances from training set that are close to the class boundaries. The class boundaries were smoothed by using ENN algorithm in the first stage of CBP. Next, it tried to distinguish between border and non-border instances by using the geometric characteristics of the instance distribution. The border instances were pruned by using the concept of mutual neighborhood in the third stage. In the last stage, the non-border instances were clustered.

Verbiest *et al*. [17] proposed fuzzy rough prototype selection (FRPS) which uses the fuzzy rough set theory to express the quality of the instances and use a wrapper approach to determine which instances to prune. The experimental study comparing the FRPS with 22 state-of-the-art prototype selection algorithms on 58 data sets showed its good performance.

Leyva *et al*. [20] proposed 3 instance selection methods, namely the local set-based smoother (LSSm), the local set-based centroids selector (LSCo), and the local set border selector (LSBo). This method was develop based on local sets which used different and complementary strategies to reduce the number of instances in the training set without affecting the classification accuracy. They compared the experimental study with 11 most prominent state-of-the-art methods. The performance were evaluated using several publicly available databases grouped in two categories; standard classification problems and classification problems with induced noise. They applied the pareto dominance relation and the TOPSIS ranking to evaluate their methods. The three methods were in the Pareto front for the standard problems. From the TOPSIS perspective, LSBo and LSCo were the first and second best methods. For the noisy problems with attribute noise, the three methods are in the Pareto fronts for three of the four noise levels studied. LSCo has the best rankings for all the noise levels in the TOPSIS perspective.

A survey paper on prototype selection methods was also presented by G. Salvador *et al*. [22] which included the state-of-the-art techniques around before the year 2010. In the survey, they summarized the properties of prototype selection methods in different groups such as a direction of search, a type of selection, an evaluation of search, the other properties and the criteria for comparing prototype selection methods. In the last section, they evaluated the performance of prototype selection methods using testing accuracy and reduction ratio from three groups of data sets (small, medium, and large data sets).

1.3 Research objectives

1.3.1 To create a new data reduction method for supervised training processes.

1.3.2 To apply the proposed method to classification and regression problems.

1.4 Research scopes

1.4.1 Data reduction method is based on input-output clustering.

1.4.2 The number of input samples determined based on cluster validity approaches.

1.4.3 Only data sets for classification and regression problems are considered.

1.4.4 The performance is evaluated using standard data sets from cml.ics.uci.edu, physionet.org and sci2s.ugr.es.

1.5 Educational advantages

1.5.1 Two new data reduction methods are developed based on input-output clustering. For regression problems, linear quantization is applied to output data, and the input data in each quantization level are then clustered. For classification problems, we apply condensation and edition methods.

1.5.2 The new methods can be used to produce a new smaller data set containing data sufficiently from the original data set.

1.5.3 The reduced data set resulted from the proposed methods can be used instead of the original data set to reduce computational costs in the training process.

1.6 Research location

The research is conducted at the Computational Intelligence Research Laboratory (CIRL), Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Thailand.

1.7 Thesis organization

The thesis consists of 5 chapters. Chapter 2 gives the important information of data reduction for classification and regression problems such as theories and methods of data reduction. Chapter 3 explains about the research designs and the proposed methodology of data reduction. Chapter 4 describes the experimental results and discussion of the proposed method on the synthetic and real world data sets. Finally, conclusions are stated in chapter 5.