Chapter XVIII
Virtual Sampling with Data Construction Method

Chun-Jung Huang
National Tsing Hua University, Taiwan

Hsiao-Fan Wang
National Tsing Hua University, Taiwan

ABSTRACT

One of the key problems in supervised learning is due to the insufficient size of the training data set. The natural way for an intelligent learning process to counter this problem and successfully generalize is to exploit prior information that may be available about the domain or that can be learned from prototypical examples. According to the concept of creating virtual samples, the intervalized kernel method of density estimation (IKDE) was proposed to improve the learning ability from a small data set. To demonstrate its theoretical validity, we provided a theorem based on Decomposition Theory. In addition, we proposed an alternative approach to achieving the better learning performance of IKDE.

INTRODUCTION

Lack of referenced data is very often responsible for poor performances of learning. In many cases, difficulty, if not impossible, in collecting additional data often cause unsatisfactory solutions. For example, due to the shorter life cycle of a product, prior management is very important, because it acquires management knowledge in the early stages, yet the accumulated information is often not sufficient for supporting the decisions. It indeed has been a challenge for managers. Because of insufficient data in the initial period of a manufacturing system, the derived management model may not be reliable and stable.
Virtual Sampling with Data Construction Method

One major contribution to the above issue has been given by Abu-Mostafa (1993) who developed a methodology for integrating different kinds of "hints" (prior knowledge) into usual learning-from-example procedure. By this way, the "hints" can be represented by new examples, generated from the existing data set by applying transformations that are known to leave the function to be learned invariant. Then, Niyogi, Girosi, and Tomaso (1998) modified "hints" into "virtual samples" and applied it to improve the learning performances of artificial neural networks such as Back-Propagation and Radial Basis Function Networks. In fact, it is evident that generating more resembling samples from the small training set can make the learning tools perform well.

Recently, Li and Lin (2006) combined the concept of virtual sample generation and the method of intervalized kernel density estimation (IKDE) to overcome the difficulty of learning with insufficient data at the early manufacturing stages. From the Li and Lin’s results, it can be noted that, when the size of virtual data increases, the average learning accuracy would decrease. This is because that when using IKDE for virtual sample generation, an actual probability density function should firstly be estimated from the original data set. Then, due to the unbounded universal discourse of the density estimation function, the more are the virtual samples generated, the larger is the Type Two error in Statistics as shown in Figure 1 (Wackerly, Mendenhall, & Scheaffer, 1996). Therefore, even though creating virtual samples by an estimated function is a way to overcome the difficulty in learning with insufficient data, the prediction capability remains a critical issue.

In this chapter, after introducing the main concept and procedure of IKDE, we used Decomposition Theory to provide a theoretical support for using IKDE to improve the learning ability from a small data set. In addition, to overcome the possible Type Two error in prediction, an alternative method named Data Construction Method (DCM) was proposed. To compare their performance, we demonstrated a numerical case adopted from Li and Lin’s and finally, discussion and conclusions were drawn.

The Procedure of Virtual Sample Generation

Because insufficient data is very often responsible for poor performances of learning, how to extract the significant information for inferences is a critical issue. It is well known that one of the basic theories in Statistics is the Central Limit Theorem (Ross, 1987). This theorem asserts that when a sample size is large (≥25 or 30), the x-bar distribution is approximately normal without considering the population distribution. Therefore, when a given sample is less than 25 (or 30), it is regarded as a small sample. Although "small" is not well defined, it is in general related to the concept of accuracy. Huang (2002) has clarified this concept which will be adopted in our study and thus presented by the following definition:

Definition 2.1 Small Samples (Huang, 2002)

Let \( X = \{x_i\ i = 1, 2, \ldots, n\} \) be a given sample drawn from a population. We assume that \( X \) will
Related Content

Fuzzy Neural Network Models for Knowledge Discovery
[www.irma-international.org/chapter/fuzzy-neural-network-models-knowledge/24214/](www.irma-international.org/chapter/fuzzy-neural-network-models-knowledge/24214/)

Conceptual Data Warehouse Design Methodology for Business Process Intelligence
Svetlana Mansmann, Thomas Neumuth, Oliver Burgert and Matthias Röger (2010). *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications* (pp. 129-173).
[www.irma-international.org/chapter/conceptual-data-warehouse-design-methodology/39591/](www.irma-international.org/chapter/conceptual-data-warehouse-design-methodology/39591/)

An Overview of Shallow and Deep Natural Language Processing for Ontology Learning
Amal Zouaq (2011). *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (pp. 16-37).
[www.irma-international.org/chapter/overview-shallow-deep-natural-language/53879/](www.irma-international.org/chapter/overview-shallow-deep-natural-language/53879/)

A Successive Decision Tree Approach to Mining Remotely Sensed Image Data
[www.irma-international.org/chapter/successive-decision-tree-approach-mining/24903/](www.irma-international.org/chapter/successive-decision-tree-approach-mining/24903/)

A Case Study: Closing the Assessment Loop with Program and Institutional Data
Robert Elliott (2012). *Teaching Cases Collection* (pp. 326-338).
[www.irma-international.org/chapter/case-study-closing-assessment-loop/60858/](www.irma-international.org/chapter/case-study-closing-assessment-loop/60858/)