Chapter 18
An Evolving System in the Text Classification Problem

Elias Oliveira
Universidade Federal do Espírito Santo, Brazil

Patrick Marques Ciarelli
Universidade Federal do Espírito Santo, Brazil

Evandro Ottoni Teatini Salles
Universidade Federal do Espírito Santo, Brazil

ABSTRACT

Traditional machine learning techniques have been successful in yielding good results when the data are stable along the time horizon. However, in many cases, these techniques may be inefficient for data that are constantly expanding and changing over time. To address this problem, new learning techniques have been proposed in the literature. In this chapter, the authors discuss some improvements on their technique, called Evolving Probabilistic Neural Network (ePNN), and present the aspects of this recent learning paradigm. This technique is based on the Probabilistic Neural Networks. In this chapter the authors compare their technique against two other competitive techniques that can be found in the literature: Incremental Probabilistic Neural Network (IPNN) and Evolving Fuzzy Neural Network (EFuNN). To show the better performance of their technique, the authors present and discuss a series of experiments that demonstrate the efficiency of ePNN over both the IPNN and EFuNN approaches.

INTRODUCTION

The volume of registered information in both the public and private domains has increased steadily over the past decade. The information may be in textual, graphical image or audio format. Information in any of these formats can be considered as data to be processed. In many contexts, the data analysis is performed manually (Mitra et al., 2002). A few examples in which information is registered mainly as text include the multitude of medical registers that physicians access each day, the numerous documents that governmental authorities must analyze within a month to make important decisions; and the many academic essays that are analyzed and graded during a semester. In such cases, specialists analyze the data and make decisions based on the information acquired from the analysis.

DOI: 10.4018/978-1-4666-2518-1.ch018

As the amount of data grows, the time required for analysis, the demand for specialized labor and the processing costs all increase. The use of computational technologies by specialists to automate some processes and make them less time-consuming and costly is therefore becoming increasingly frequent. Machine learning has been applied in this context. The machines usually learn from data sets prepared by humans and are subsequently able to infer knowledge from new data that come in. As time goes by and the machines' capacity for extracting knowledge from new data decreases, a new training procedure is needed. Therefore, from time to time, these algorithms are retrained from scratch using a new human prepared data set. However, such techniques tend to become inadequate or inefficient for data that are constantly expanding and/or changing over time (Salles et al., 2010).

An off-line model has the advantage of reaching an optimized structure from a previously obtained training data set, but its performance may decline suddenly when certain features of the environment change. Although there are many cases in which this approach produces good results in many contexts, an off-line model usually needs to be completely re-designed when new circumstances occur; for example, a new class or substantial changes in one or more features of the environment. Another drawback is that the performance of an off-line model is directly related to the quality of the available data set. The process of learning in these models is based on the assumption that the whole data set is already available during the training phase and therefore does not consider the possibility of accommodating new knowledge when new data are acquired. However, the acquisition of adequate data representative of the problem is often onerous and time-consuming; furthermore, the data are usually available in small quantities over a period of time (Bhattacharyya et al., 2008).

New intelligent algorithms have therefore been developed to overcome these difficulties of the previous learning approaches. A new paradigm in the field of computational intelligence, based on building models from incremental algorithms and data flows, emerged around a decade ago. The new models, called evolutionary models, not only minimize the problem of storing large amounts of data by processing them altogether at once, but also offer important characteristics for the modeling of non-linear adaptive processes. The main characteristics of the evolutionary models are continuous learning, self-organization and the ability to adapt to unknown environments (Watts, 2009). To attain these goals, the training processes of these models must be able to obtain new knowledge (plasticity) without forgetting the previously acquired knowledge (stability). Thus, a trade-off between the properties of stability and plasticity is necessary (Polikar et al., 2001).

In this chapter, we discuss a new technique of evolutionary model, called Evolving Probabilistic Neural Network (ePNN), based on a technique presented by Vlassis et al. (1999). Experimental evaluations showed that the proposed technique is highly competitive with other techniques in the literature.

This chapter is divided into 8 sections. In The Problem Section, we present the formalism of the text classification problem. This problem is well known in the field, particularly because of the substantial increase in the volume of data published on the World Wide Web during the past decade. We therefore treat this problem within this chapter as an example. Next, an introduction to Expectation Maximization Gaussian Mixture Model, following by the basic concepts regarding Artificial Neural Networks are presented in The Background Techniques Section. In this section it is also described the Probabilistic Neural Network and discuss the principles of evolving systems. In Evolving Probabilistic Neural Network Section it is presented the neural network proposed. In