Chapter 23

Adaptive Network Structures for Data/Text Pattern Recognition (Application)

Emmanuel Buabin
Methodist University College Ghana, Ghana

ABSTRACT

The objective of this chapter is implementation of neural based solutions in real world context. In particular, a step-wise approach to constructing, training, validating, and testing of selected feed-forward (Multi-Layer Perceptron, Radial Basis function) and recurrent (Recurrent Neural Networks) neural based classification systems are demonstrated. The pre-processing techniques adopted in extracting information from selected datasets are also discussed. In terms of future practical directions, a catalogue of intelligent systems across selected disciplines, are outlined. The main contribution of this book chapter is to provide basic introductory text with less mathematical rigor for the benefit of students, tutors, lecturers, researchers, and/or professionals who wish to delve into foundational (practical) representations of bio-intelligent intelligent systems.

INTRODUCTION

Human and computer systems generate large amounts of data each day. From airline booking web portals through credit card systems to space monitoring systems, data (structured and unstructured) has grown into overwhelming quantities. The inception of the Internet has increased the production of unstructured data formats. Capitalizing on the data explosion situation, scientists have developed various kinds of data mining systems to investigate hidden trends and heuristics. One area that has in the past, received less attention, is the use of bio-inspired models. Scientists, in recent times, have developed interest in this area due to the fact that, (1) biological knowledge bases and datasets have and are still been developed to support data exploration and information mining process (2) salient clues to contemporary scientific research and experiment could be unearthed through interestingness (3) governments are under pressure to investigate efficient ways of unraveling solutions for newly detected medical problems. Although data mining research and its applications within

DOI: 10.4018/978-1-4666-2661-4.ch023
biology is infantile, different kinds of approaches have been developed. Neural based methods, a section of the approaches, have proven better than other counterparts—e.g. statistical, mathematical. Their (i.e. neural based methods) improvement in performance has largely been attributed to the robust features they possess. Their fault tolerance levels, ability to adapt under varying conditions and resemblance to human information processing, make them better candidates for rigorous and complex tasks.

In this book chapter, a practical approach to implementing intelligent solutions with artificial neural networks is adopted. In particular, supervised learning based neural methods are implemented within the context of directed graphs. The neural network formation, training, validation and testing regimes of the Radial Basis Function (RBF), Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN) are explained in practical terms. The pre-processing steps of selected datasets are also explained. In basic English language terms, underpinning mathematical rigor is lessened for easy reading and better digestion. The level of comprehension of this book chapter is that of undergraduate computer science. However post-graduate students/tutors/lecturers/professionals and/or researchers who wish to study or research into the foundations, representations, concepts, theory and applications of these supervised neural solutions will also find this book chapter beneficial. A description of the datasets, how they are pre-processed and partitioned into training, validation and test sets are given. A step-wise approach to constructing neural based intelligent agents is done. Lastly, real life applications of engaged techniques and future directions of neural solutions are outlined.

DATASETS

A data set (or dataset) is a collection of data, processed in tabular or non-tabular formats for scientific research purposes. In tabular datasets, each row is an instance whereas each column is an attribute. A Cell value is known as datum. Cell values may be labels, numeric or a combination of labels and numeric values. For supervised learning purposes, each row (i.e. instance) has a target. The target could be single or multiple. Unlike tabular forms, non-tabular datasets exist more in real life situations as text based datasets. Non tabular datasets are usually expressed in XML files for easy retrieval and processing. The University of California Irvine (UCI) Data repository makes available various datasets for research purposes, some of which have been explained in this book chapter.

Iris Dataset

The Iris dataset is a commonly used dataset within the Machine Learning community. The dataset was introduced by R.A Fisher (but collected by Edgar Anderson) for standardized statistical research analysis. Often called Anderson’s Iris dataset, the data was meant to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula from the same pasture, picked on the same day and measured at the same time by the same person with the same apparatus”. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features (sepal length, sepal width, petal length and petal width) were measured from each sample they are the length and the width of sepal and petal, in centimeters (see Table 1).

Splice Junction Dataset

The task posed in this splice junction dataset is to recognize the demarcations between exons and introns, given a DNA sequence. Exons in biological terms are EI sites (donors), whereas Introns are IE sites (acceptors). The dataset has been used in Knowledge Based Artificial Neural Networks