Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 97–104

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

A New Data-Driven Deep Learning Model for Pattern Categorization using Fast Independent Component Analysis and Radial Basis Function Network. Taking Social Networks resources as a case

Choukri Djellali, Mehdi Adda

Mathematics, Computer Science and Engineering department.
University of Quebec At Rimouski
300, Allée des Ursulines, Rimouski , QC G5L 3A1, Canada

Abstract

This paper investigates the categorization problem using Data Mining techniques. We present a new conceptual model, which is named FICARBFN, for classifying patterns by using Fast Fixed-Point Algorithm for Independent Component Analysis and Radial Basis Function Network. It uses an artificial neural network model to find a single consolidated categorization, which is composed of tree process, variables selection, categorization, and finally models selection. Our categorization model used a hybrid technique that combines the advantages of factorial analysis and Neural Network approaches. Comparative study and experimental results showed that our scheme optimized the bias-variance on the selected model and achieved an enhanced generalization for Social Networks patterns recognition.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining; Pattern Recognition; Deep learning; Categorization; Variables Selection; Factor analysis.

1. Introduction

Categorization is a supervised machine learning technique that has been extensively and successfully used for Data Mining (or KD), which is defined as follows: *the placement of entities in groups whose members bear some similarity to each other*[3].
Categorization is a supervised learning task in which patterns are categorized into different categories. Each pattern

* Corresponding author. Tel.: +1 (514) 987-3000 ; fax: +1 (514) 987-8477.
E-mail address: Choukri.Djellali@uqar.ca

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.320

is represented by a tuple (P, L), where P is the variables set representing the input pattern and L is the category label (also known as predictor).
The problem of pattern categorization arises in many engineering systems, including, Pervasive computing[27], social Web categorization[4], text Mining[5], intrusion detection[6], medical diagnosis[7,8] and many other fields[10,15]. All these disciplines show the practical importance of categorization algorithms.
However, the categorization is a crucial challenge, especially for pattern recognition. Most categorization algorithms are sensitive to outliers, noise, presentation order, architecture configuration, Bellman's curse of dimensionality, and complex shapes.
On one hand, the curse of dimensionality associated with the exponential increase of the space size adds unnecessary noise within the decision boundary and learning generates significantly lower performance. Therefore, exhaustive search strategies are prohibitively time-consuming and computationally expensive even for problem instances of moderate size of the search space. Choosing the relevant variables can greatly reduce the variance.
On the other hand, models selection is the most common technique in machine leaning, which is a meta-model or an averaging scheme designed to assess the categorization stability and improve the recognition accuracy.
Based on these premises, we used a new categorization scheme based on an efficient Wrapper model, subset validity assessment criterion, Artificial Neural Network, and finally models selection.
This paper is divided into six sections, including the introduction. In the next section, we present the current state of the art, our research questions and the problematic of categorization. In the third section we describe the conceptual architecture of our approach. In section four we give a short evaluation with benchmarking models for our conceptual model. In the last section, we present the conclusions and outline some goals for future works.

2. State of the art, problem and research questions

Categorization is a machine learning task that assigns patterns in a training set to the target categories. It has become a widely used method for Data Mining as it automatically groups patterns according to category label from a set of high-dimensional data.
Numerous categorization models have been proposed during the last few decades. These models include: *Combinatorial models*[10,9], *Kernel Based models*[11], *Rule Based models*[12], *Decision Tree models*[13], *Neural Networks models*[16,18], *K-Nearest Neighbor models*[10,14], *Bayesian models*[17,19], etc. A more detailed study of different models can be found in the papers[15,26] .
This revolution, that the categorization witnessing, has led to the appearance of several approaches. Martin[23] used a monolithic fuzzy-based systems (TSK-FIS) to categorize real-world patterns into several categories. The conceptual model of the AwarePen is divided into hardware and software parts. The hardware part is an electronic PCB that contains microcontroller unit (DSP-MCU, dsPIC) and a collection of sensors. In order to map sensor data onto context categories, the authors used an Online Fuzzy Inference System as a software part. Ruixi Yuan[22] introduced a kernel learning model based on SVM (Support Vector Machine) and models selection for Internet traffic categorization. The data used in this study comes from a set of 8-hour traffic data on a Gbps Ethernet interfaces during a one week period. This biased scheme using 10 fold cross-validation showed an improvement of categorization accuracy 99.42% compared to unbiased learning 97.17%. Lakshmi[24] used a k-Nearest Neighbour classification model for Intrusion Detection. Variables selection is the first step of data cleaning, where patterns are prepared for categorization, which in turn allowed for categorization improving. The benchmarking data used in this study comes from the KDD Cup 1999 Data Set. Each pattern contains 41 variables representing the connection features. This categorization model based on vote scheme showed an improvement of recognition accuracy. Some recent works[25] proposed a Big Data Driven network optimization framework that includes Big Data collection, storage management, data analytics and network optimization. This framework used a categorization model based on machine learning techniques to find the correlation between the relevant factors and Quality of Service. The evaluation based on Big Data from the users and operators perspectives showed that the proposed approach yield good results.
In most previous models, there is no guarantee for convergence to the global optimum and the results depend on the representation space. Moreover, the recognition accuracy does not satisfy the user needs.
In order to overcome the obstacles mentioned above, we present in this paper a new conceptual model for pattern categorization, using variables selection, categorization, and finally models selection.