Chapter 2
Synthesis of Classification Models and Review in the Field of Machine Learning

Venkatram Kari
VIT University, India

Geetha Mary Amalanathan
VIT University, India

ABSTRACT

Classification method is an important technique used in machine learning for predictive analytics. Classification enables business to predict future trends and behaviors of an enterprise with the help of their past data. Classification is a supervised learning model, which is built in twostep process, first building the classification model and second predicting the outcome for unknown data. This chapter describes various classification models by learning mechanisms and categorizes them into different statistical, probabilistic, and heuristic methods, and explains them with example dataset. It also compares these models and their efficiencies with model evaluation techniques and briefs some blended classification models. The goal of this chapter is to provide a comprehensive review of different classification techniques and give a quick refresher on classification models in big data analytics. The comparison of various classification models helps the readers to quickly decide which classification model to choose for the given business scenario.

INTRODUCTION

The easy way to understand the data is to distribute the data into different classes on the basis of their data attribute and characteristics. This process of organizing the data is known as classification. It is very difficult to interpret the data which is unorganized, such data can neither be easily compared nor interpreted.

Machine learning models are primarily built in collaboration of two important techniques - Supervised learning (classification & regression), unsupervised learning (clustering). A target function (f) is

estimated by classification model which maps input variables (A) to an output variable (B) i.e., B=f(A). Target function f(A) is built from labeled training data and will predict the outcome of unknown data into a categorical value called target class (Jothi, Rashid, & Husain, 2015). Classification is a twostep process (Wahbeh, Al-Radaideh, Al-Kabi, & Al-Shawakfa, 2011), 1) Model construction 2) Predicting the unknown data. Prediction is performed after rigorous testing for acceptance.

Classification objectives - simple to understand, easy to compare and intuitive to interpret by enabling data to represent in brief and simple way. It will summarize the large data into condensed form which is easy to understand. For instance 250 patients' in a hospital classified them on medical problems, status into 100, 70, 50, 30 as self-care, minimum-care, intermediate-care, ICU patients. Now it easy to interpret and compare data. This is motivation for classification & interpret the similarities and dissimilarities among the data visualize and goal to build a function to make distinction data into classes.

At times some important features will be missing in data, eg., patient data without age or tests etc. which are critical for predicting treatment. When attribute values are missing, there are several actions mentioned in the literature. These are 1) Discard the missing value's instances. 2) Acquire missing values 3) Imputation.

LITERATURE REVIEW

Researchers and health care industry practitioners are widely using machine learning and classification models for building decision support system. Tarik A and all has used Artificial Neural Network to classify of diabetic patients to predict the rate of fasting blood sugar and diabetes mellitus and figure out the relations between the FBS and symptoms with prediction rate and identify the impact and symptoms of the patient's health(Tarik A. Rashid, Abdullah, & Abdullah, 2016). Artificial bee colony ABC and genetic algorithm (GA), and back propagation neural network (BPNN) for better prediction of diagnosing diabetes mellitus (DM) (Tarik A Rashid & Abdullah, 2018). It enables decision support system and help physicians to provide patients good health treatments and recommendations for further medication(Tarik A. Rashid, 2016).

Premature Ventricular Contraction Detection on ECG Trace is predicted with classification by Manuel M. Casas and All. Bayesian classification algorithms trained from the patient ECG traces and concluded PVC complexes with 95% accurately. This automated mechanisms enable for the detection of PVC complexes(Casas, Avitia, Gonzalez-Navarro, Cardenas-Haro, & Reyna, 2018). This is a very good achievement to predict the PVC detection based on ECG trace.

Abraham Yosipof & All, in their recent paper build ensemble model -AL Boost for the visualization and identifying of interesting trends, which opens up new opportunities for designing new drugs(Yosipof, Guedes, & García-Sosa, 2018).

Singh and all tried to diagnose the disease by analyzing with pattern matching and neural networks (Singh, Prasad, Das, Poddar, & Choudhury, 2018).

BACKGROUND

A typical classification model is shown in the following diagram Figure 1, with two processes 1) Model construction for known dataset with predefined class labels. Build the model with a training dataset. Validating the model on the validation dataset. 2) Classify the unknown data.

Related Content

Application of Geographical Information System and Interactive Data Visualization in Healthcare Decision Making

Zhecheng Zhu (2016). *International Journal of Big Data and Analytics in Healthcare (pp. 49-58).*

www.irma-international.org/article/application-of-geographical-information-system-and-interactive-data-visualization-in-healthcare-decision-making/171404

Hadoop Tools

(2019). *Big Data Processing With Hadoop (pp. 169-215).*

www.irma-international.org/chapter/hadoop-tools/216604

Unstructured Healthcare Data Archiving and Retrieval Using Hadoop and Drill

Hang Yue (2018). *International Journal of Big Data and Analytics in Healthcare (pp. 28-44).*

www.irma-international.org/article/unstructured-healthcare-data-archiving-and-retrieval-using-hadoop-and-drill/223165

Big Data Thinking of Economy, Investment, and Business in COVID-19

(2022). *Applied Big Data Analytics and Its Role in COVID-19 Research (pp. 316-345).*

www.irma-international.org/chapter/big-data-thinking-of-economy-investment-and-business-in-covid-19/300389

Real Time Analysis Based on Intelligent Applications of Big Data and IoT in Smart Health Care Systems

Mamata Rath (2018). *International Journal of Big Data and Analytics in Healthcare (pp. 45-61).*

www.irma-international.org/article/real-time-analysis-based-on-intelligent-applications-of-big-data-and-iot-in-smart-health-care-systems/223166