Chapter 54
Machine Learning for Health Data Analytics: A Few Case Studies of Application of Regression

Muralikrishna Iyyanki
https://orcid.org/0000-0002-4961-9010
Independent Researcher, India

Prisilla Jayanthi
Administrative Staff College of India, India

Valli Manickam
Administrative Staff College of India, India

ABSTRACT

At present, public health and population health are the key areas of major concern, and the current study highlights the significant challenges through a few case studies of application of machine learning for health data with focus on regression. Four types of machine learning methods found to be significant are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In light of the case studies reported as part of the literature survey and specific exercises carried out for this chapter, it is possible to say that machine learning provides new opportunities for automatic learning in expressive models. Regression models including multiple and multivariate regression are suitable for modeling air pollution and heart disease prediction. The applicability of STATA and R packages for multiple linear regression and predictive modelling for crude birth rate and crude mortality rate is well established in the study as carried out using the data from data.gov.in*. Decision tree as a class of very powerful machine learning models is applied for brain tumors. In simple terms, machine learning and data mining techniques go hand-in-hand for prediction, data modelling, and decision making. The health analytics and unpredictable growth of health databases require integration of the conventional data analysis to be paired with methods for efficient computer-assisted analysis. In the second case study, confidence interval is evaluated. Here, the statistical parameter CI is used to indicate the true range of the mean of the crude birth rate and crude mortality rate computed from the observed data.*

DOI: 10.4018/978-1-6684-6291-1.ch054

1. INTRODUCTION

Data in digital form is the new oil, as being considered globally. For any developmental activity data are essential and hence data science. According to Bernard (2015) by the end of the year 2025, Forbes estimated that the digital data is quite sure to increase automatically by an order of magnitude from 4.4 ZB. On this planet, every human generates new information of 1.7 MB in every second. Innovative data mining techniques and machine learning techniques are necessary for facilitating related information through data modelling, prediction and prescription. Data mining and Machine learning have a good amount of commonality as the two transect to enhance the collection and usability of large amounts of data for analytics purposes.

According to "*Bio IT World"* statement, the predictive analysis is the future of data mining as it can be seen in advanced analytics across industries like medical applications (Agosta 2004). Arthur Samuel one of the forerunners and developer of machine learning define that "machine learning relates to the study, design and development of the algorithms that give computer's the capability to learn without being explicitly programmed". The process of unstructured data that tries to extract knowledge and/or unknown interesting patterns is defined as data mining. During this process, Machine Learning (ML) algorithms are traditionally used. ML is further associated with the query, how machines can learn, i.e., to the algorithmic part. In ML, an agent learns from rewards (data) in the environment, but not from patterns or pattern-label pairs. In data mining, the question is how to learn from patterns or pattern-label pairs. "ML techniques are fairly generic and can be applied in various settings. Data mining has emphasis on utilizing data from a domain e.g., social media, sensor data, video streams, etc. to understand some questions in that domain". New questions arise that may not be answered in the algorithmically oriented ML perspective including preprocessing of data and the complete data mining process chain (Souhila 2013 and Xavier 2016). From large historical datasets, the data mining intent is to find out unseen patterns and relationships and derive a business value. Its interest is upon uncovering relationships between two or more variables in the dataset and extracting insights. These insights include mapping the data into information and predicting outcomes from incoming events and prescribing actions. Multiple data sorting techniques can be used to achieve this target such as clustering, classification, and sequence analysis. Typically, data mining uses batch-process information to reveal a new insight at any specific point. And DM is not automated process but DM requires human involvement and cannot be implemented without humans.

2. MACHINE LEARNING [ML] AND DATA MINING [DM]

ML uses human-based algorithm and works everything without the use of humans' interference; once implemented, the outcome is accurate because the process is automated. Also, ML is capable to take the own decision and resolve the issue. Ever growing ML techniques, overwhelms problems associated with DM techniques as ML techniques are more accurate and less error prone compared to DM. This self-learning technique is not available in DM whereas ML uses self-learning algorithms to improve its performance as an intelligent task with experience over time (Brooks & Dahlke 2017). To summarize the foregoing, it can be stated the following are the definitions and differences or commonalities, if any, between ML and DM: