Advance analytics and Comparison study of Data & Data Mining

Review paper on Concepts and Practice with RapidMiner

Vandana Kaushik¹, Dr. Vikas Siwach²
M.Tech (Software Engineering)
Kaushikvandana22@gmail.com
University Institute of Engineering and Technology
MDU ROHTAK, HARYANA
Abstract

The aim of this paper is to study the algorithms and compare them before implementing them. The comparison is based on different parameters, type of data and its frequency and relationship of attributes in datasets. There are lots of algorithms which are used to learn patterns, analyse data and categorise the data. The implementation of algorithms is done with the help of RapidMiner. RapidMiner is a data mining tool which is used for data analytics, machine learning and for abstracting the data. RapidMiner extract meaningful data from the raw data. The problem is to find out the most accurate algorithm according to the given dataset. The result must be with higher accuracy. These algorithms are used in data mining and machine learning and in the field of knowledge discovery.

Keywords

RapidMiner, Data mining, Machine learning, Data analysis.

Introduction

Data Mining

It is a computing process to extract data patterns from the datasets. In this process various intelligent methods are applying to extract knowledgeable data and represent it in the structured form so that it is easy to use and understand the data in small period of time with less efforts. There are various tools and technologies used for data mining process. Here we work with a data mining
tool i.e RapidMiner, which is most useful and recent tool of data mining in market now a days.

RapidMiner

It is a data mining tool which was developed by the company of the same name i.e RapidMiner in 2006. It is a software platform which provides integrated environment for data preparation, machine learning, deep learning and predictive data analysis. It is developed on an open core model. The first paper on RapidMiner was YALE i.e yet another learning environment in 2001. RapidMiner is used client and server model either in public or private infrastructure.

RapidMiner provides data mining and machine learning procedures to ETL the data i.e Extract, Transform, and Load the data.

There are lots of algorithms which are applying on raw data to analyse the useful data. It is little bit complex to make sure that which algorithm is best suited on the datasets to get the accurate and fast result of data. So here RapidMiner is used to select the particular algorithm with the help of which it is easy to extract the knowledgeable data from a raw dataset.

Uses of RapidMiner for Data Mining

- It is used for many purposes like business, education, research, training, data preparation, and data optimization and to represent data in structured form.
- It uses client and server model on premise either on public and private infrastructure.

Merits of RapidMiner

- It provides approx. 100% solution of data with high speed and accuracy.
• It eliminates the need of coding.
• It provides ETL procedure i.e extract, transform and load the data.
• It is written in java programming language.
• It provides GUI interface to design and execute analytical data.
• Its functionality can be extended with additional plug-ins.

Why RapidMiner

• It makes data science fast and simple.
• It is easy to extend for java programmers.
• It is stable.
• It is powerful due to its learning operators and framework.
• It is free of cost.
• It is an open source tool so that we can easily understand its functionality.
• You can easily integrate it into your own software.
• It used less memory hence it is scalable.
• It supports visualization i.e white box testing.

Developers and Ranking of RapidMiner

RapidMiner is developed by the company named RapidMiner in 2006. The company that developed RapidMiner is partnership with venture capital firms like Nokia, Ascent, Longworth, Earlybird and Open Ocean.

In 2017 RapidMiner is ranked as one of the most popular data analytics software.
How to install and use RapidMiner

It is available on internet in the form of RapidMiner Studio. The process of data analysing is done by RapidMiner by drag and drop of operators. RapidMiner contains almost 1500 operators which are used on datasets. To work with RapidMiner studio you must need to download and install it first.

Installation:-

Step 1:-

You can find this on RapidMiner website i.e

http://www.rapidminer.com

Download the supported version of RapidMiner from here according to your operating system. It works on windows as well on Linus and Unix etc.

Step 2:-

If you are a new user of RapidMiner then first you have to create a new repository. For this you have to specify a name and then define any directory on your hard disk. This directory store your data for future use.

Step 3:-

After creating directory you are welcomed on home page. In right you can see latest news about RapidMiner if you are connected with internet. In centre there is working section where actions are performed.

Now you can use RapidMiner for data mining. It contains many typical actions like:-

- New process
- Open
- Application wizard
- Tutorials
Literature Review

Thirunavukkarasu Kand Dr. Manoj Wadhawa

Comparison of algorithms is much needed before implementing them for the needs of organisation. The comparison is based on different parameters, type of data and its frequency and relationship of attributes in datasets. There are lots of algorithms which are used to learn patterns, analyse data and categorise the data. The problem is to choose the best algorithm which fulfil all desires of organisation on datasets. The taken algorithm must provide most accurate and high speed result for datasets.

Ingo Mierswa and Michael Wurst

This paper describes Yale, a free open-source environment for KDD and machine learning. Yale provides a rich variety of methods which allows rapid prototyping for new applications and makes costly re-implementations unnecessary. Additionally, Yale offers extensive functionality for process evaluation and optimization which is a crucial property for any KDD rapid prototyping tool. Following the paradigm of visual programming eases the design of processing schemes. While the graphical user interface supports interactive design, the underlying XML representation enables automated applications after the prototyping phase.

Venkatesh Umaashankar Sangkyun Lee

Optimization in general means selecting a best choice out of various alternatives, which reduces the cost or disadvantage of an objective. Optimization problems are very popular in the fields such as economics, finance, logistics, etc. Optimization is a science of its own and machine learning or data mining is a diverse growing field which applies techniques from various other areas to find useful insights from data. Many of the machine learning problems can be modelled and solved as optimization problems, which means
optimization already provides a set of well-established methods and algorithms to solve machine learning problems.

Anna Lisa Gentile, Sabrina Kirstein, and Christian Bizer

The data mining tools that are currently available on the market offer a wide range of powerful data mining methods but hardly support analysts in searching for suitable data as well as in integrating data from multiple sources. This demo shows an extension to Rapid Miner, a popular data mining framework, which enables analysts to search for relevant datasets and integrate discovered data with data that they already know. In particular, we support the iterative extension of data tables with additional attributes.

Alfredo Bolt, Massimiliano de Leoni, Wil M. P. van

This paper structures the basic building blocks needed for process mining and describes various analysis scenarios. Based on these requirements we implemented RapidProM, a tool supporting scientific workflows for process mining. Examples illustrating the different scenarios are provided to show the feasibility of the approach.

Tim ruhe

Data mining, however, goes beyond the simple training and application of a learning algorithm. It also incorporates finding a good representation of data in fewer dimensions without losing relevant information, as well as a thorough validation of the results throughout the entire analysis. A data mining-based event selection chain has been developed for the measurement of the atmospheric spectrum with IceCube.

Conclusion

We compared learning algorithms including SVM and Neural Net. The results states that both Neural Net and SVM are capable to model both linear and non-linear datasets. The Neural Net is more effective when it comes with complex datasets whereas SVM is successfully classifying and modelling non-linear...
relationships. The mapping of features into a higher dimensional space is efficient and useful while recognizing patterns in SVM. The performance of SVM was found to be better than Neural Net while comparing.

References