A Proposed Model of Machine Learning Algorithms for Business Intelligence

Sajid Hasan Sifat

Faculty of Information Iechnology, American International University-bangladesh, Dhaka, Bangladesh.
Email: sajidhasan054@gmail.com

ABSTRACT

Business intelligence (BI) has steadily become a popular information systems terminology. There are a variety of BI software packages in the industry today although it is essentially a combination of data mining, statistical analysis, and advanced reporting features. Data mining searches for hidden patterns from a huge data warehouse so it can help managers to make business decisions. To determine the hidden pattern from a huge data warehouse, BI software's use a variety of algorithms to find out the relationship between different data and variables. Although most of the BI the algorithms seem to be similar to traditional statistical techniques, data scientists are passionate to create BI software as a new decision tool. Thus, the purpose of this paper is to define most popular the algorithms and propose a model for BI software to use the algorithms according to the datasets.

Keywords Business Intelligence, Algorithms, Machine Learning, Predictive algorithms, Data mining.

1 INTRODUCTION

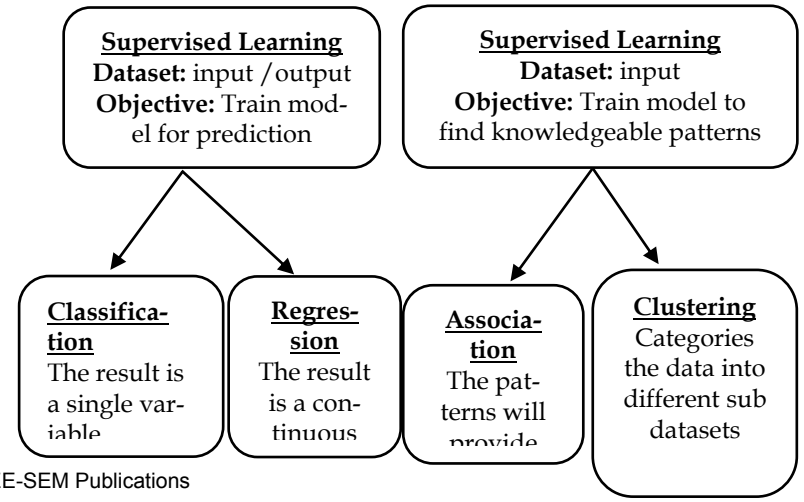
Data mining (DM) and business intelligence (BI) are among the IT applications that have commercial value. This paper will first outline what data mining algorithms are, then move on to practical usages in various business intelligence of those specific algorithm and propose a model.

Considering the needs for Business Intelligence the model can be divided into three main particular purpose.

1.1 Customer Segmentation (Clustering)
1.2 Recommended System (Classified Algorithms)
1.3 Predicting Outcome (Regression Algorithms)

2 SUPERVISED LEARNING AND UNSUPERVISED LEARNING

Data are mainly divided into two sections for performing machine learning algorithms to gain knowledge from the data sets.



2.1 Supervised Learning

It is a process of an algorithm learning from the training dataset can be assumed of as a teacher supervising the learning process. We know the correct results; the algorithm iteratively makes predictions on the training data and is modified. Learning stops when the algorithm accomplishes a satisfactory level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

Classification
A classification problem is when the output variable is a category, such as "*red*" or "*blue*" or "*true*" and "*false*".

Regression
A regression problem is when the output variable is a real value, such as "*Rupee*" or "*height*".

2.2 Unsupervised Learning

Unsupervised learning has no correct result and there is no perfect strategy. Algorithms are left to their own devises to determine and present the stimulating structure in the data.
Unsupervised learning problems can be further grouped into clustering and association problems.

Clustering
A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

Association
An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

3 Pros and Cons of Popular Algorithms

3.1 Clustering and Dimensionality Reduction

3.1.1 K-Means

Results are simple to understand by a human but Requires the number of clusters to be known in advance.

3.1.2 Expectation Maximization (EXP.MAX)

Can be naturally used for finding outliers but Slow linear convergence.

3.1.3 Latent Dirichlet Allocation (LDA)

Excellent for empirical results and mitigates overfitting well for unsupervised data. Can be used in only specific data sets

3.1.4 UMAP

Has very fast implementation in multiple programming languages, including Python. Has hyperparameters one has to tune to find a good visualization

3.1.5 APRIORI Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by recognizing the frequent specific items in the database.

3.1.6 Frequent Pattern-Tree Growth Algorithm (FP-tree)

It involves first of reducing the database into a compressed structure called FP-tree (Frequent Pattern tree), then separating it into sub-projections of the database called conditional databases.

3.1 Classification and Regression

3.2.1 Gradient Boosting Machine (GBM)

Can hold huge datasets, very accurate and can be used for both classification and regression tasks.

3.2.2 Random Forest (RF)

Deal well with uneven datasets that have missing variables and Rarely overfits. Random Forest is biased in favour of those with more levels. Works with irregular datasets that have missing variables.

3.2.3 Support Vector Machine (SVM)

Handles both classification and regression while supporting very high-dimensional data. Not an ideal choice for Big datasets due to wide range for hyperparameter.

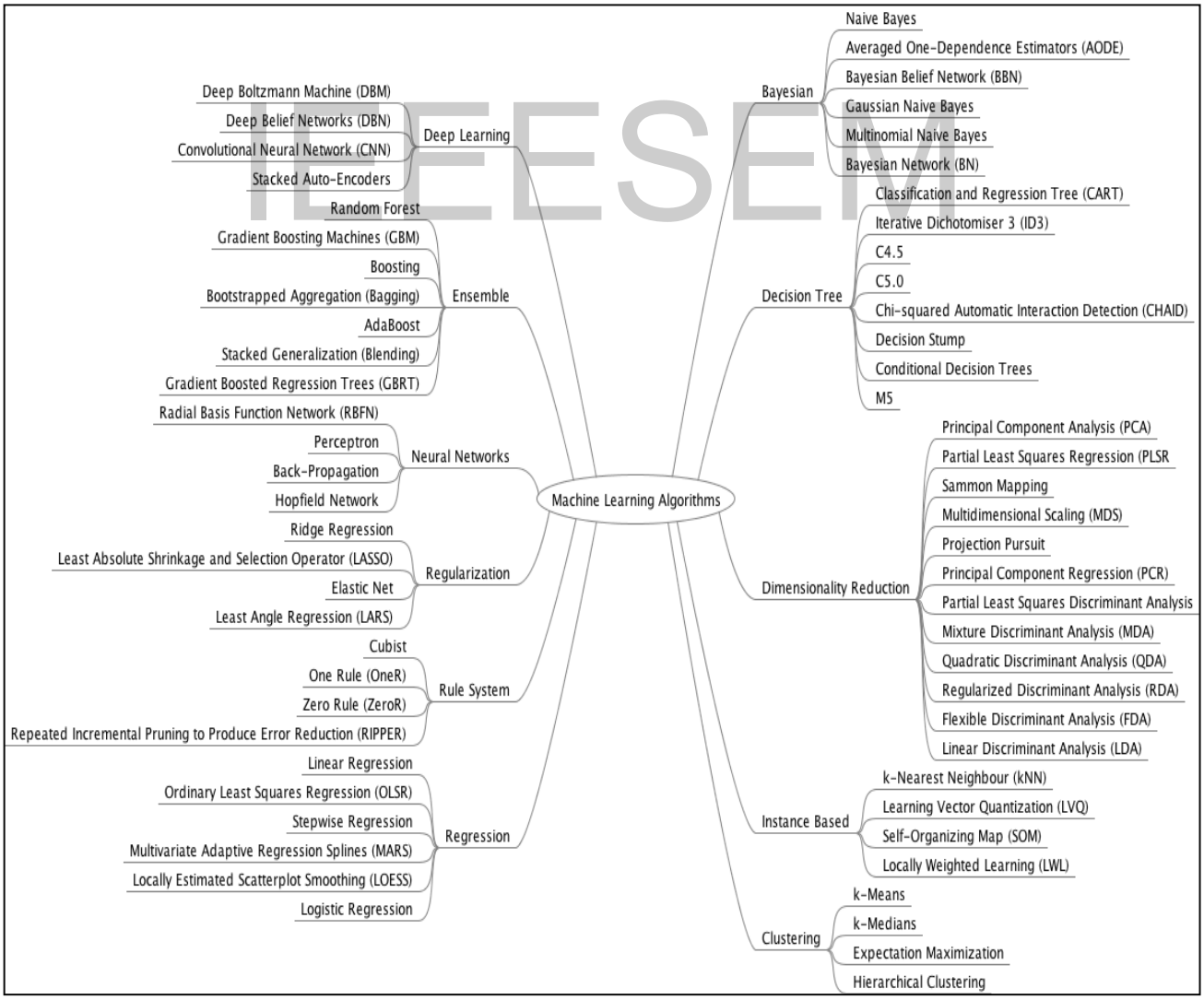
3.2.4 Convolutional Neural Network (CNN)

Very accurate at image classification and many other tasks, including language processing but requires a lot of training data and works slow on CPU.

3.2.5 K-Nearest Neighbours (K-NN)

Does well in practice with enough representative data and can be used for both classification and regression. Sensitive to irrelevant features and the scale of the data.

There are more algorithms and technique available and even more hybrid algorithms which are not suitable for generic purposes. This paper focuses on the general model. Figure 1 shows most of the machine learning algorithms used in Business Intelligence.



4 PROPOSED MODEL

Here is a proposed model for using the algorithms for business intellegence purpose.

		BUSINESS INTELLIGENCE PURPOSE		
DATA TYPE	SCALE	SEGMENTATION	RECOMMENDED SYSTEM	PREDICTION OUTCOME
SUPERVISED	LARGE		CNN, GBM, LSTM	CNN, LSTM, GBM, RF
	SMALL		K-NN, RF, SVM	K-NN, RF, SVM
	CATEGORY		K-NN, SVM, CNW	CNN, SVM
	VALUE		RF, CNN	RF, CNN
UNSUPERVISED	LARGE	EXP.MAX, LDA		FP-TREE
	SMALL	K-MEANS, EXP.MAX		APRIORI
	CATEGORY	K-MEANS, EXP.MAX, LDA		APRIORI
	VALUE	K-MEANS, EXP.MAX, UMAP		FP-TREE

The proposed model is shown by a table which indicates the BI systems purpose and the data scaling, if the data set is too large or small, categorical or a value. The table identifies the purpose and scaling and projects the best algorithms for that specific Business Integence.

4 CONCLUSION

This review paper's motivation was to relate the popular machine learning algorithms to the business intelligence or the data analysis platforms. Identifying the quality and the usefulness of the algorithms towards different query and needs for various froms of data.

REFERENCES

[1] https://www.infoworld.com/article/3259512/machine-learning-what-developers-and-business-analysts-need-to- know.html

[2] https://emerj.com/ai-sector-overviews/machine-learning-algorithms-for-business-applications-complete-guide.

[3] http://www.myacme.org/ACMEProceedings09/p6.pdf.

[4] https://tdwi.org/Articles/2018/07/02/ADV-ALL-5-Algorithms-for-Big-Data.aspx.

[5] http://www.hrpub.org/download/20181230/CSIT1-13512271.pdf

[6] https://pdfs.semanticscholar.org/d9a8/61a851f65249987a14b0d0a4ff227ac76ce2.pdf

[7] https://semanti.ca/blog/?the-most-important-machine-learning-algorithms

[8] http://datasciencetoday.net/index.php/en-us/machine-learning/179-association-rules

[9] https://dl.acm.org/citation.cfm?id=1143865

[10] https://academic.oup.com/imaman/article-abstract/4/1/43/656001