Data Mining and Machine Learning: Definitions, Basics and Applications

İbrahim Erdem KALKAN [*]

January 14, 2021

Introduction

With the increased usage of many types of database applications, it has been recorded an exponential growth at data during the last decades. Without hesitation, the institutions, that have effective tools to extract useful knowledge from that amount of data, could attain significant advantages. To stress this issue, it is mentioned in this paper about some aspects of artificial intelligence, data mining, machine machine learning, even statistical learning as concepts in terms of those effective tools.

One can be easily observed there is a common misconception or confusion between the terms. To underline this case first section focuses mostly on brief definitions and differences.

Definitions and Differences

Artificial intelligence is a field of research, which aims at developing software that can do some tasks that require intelligence. However, what tasks require

[*] Cukurova University, Institute of Natural and Applied Sciences, Industrial Engineering

intelligence is an arguable question. Expert systems or knowledge-based systems is one of the many type of artificial intelligence systems. Machine learning is also one of that kind of systems which is designed to learn by themselves from data and not preprogrammed[8]. With this definition, machine learning can be considered as an subbranch of artificial intelligence. In other words, artificial intelligence can be found more inclusive and broader term than machine learning.

Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples [10]. An equally basic scientific objective of machine learning is the exploration of possible learning mechanisms, including the discovery of different induction algorithms, the scope and theoretical limitations of certain methods, the information that must be available to the learner, the issue of coping with imperfect training data, and the creation of general techniques applicable in many task domains [6]. Machine learning is considered in literature as a set of techniques of training and testing.

Data mining, also known as knowledge discovery from data (KDD), is as name implies a process of extracting useful knowledge from data [10]. According to another definition [14], data mining is a step in the process KDD that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data. In addition, KDD is evolving from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. One can easily notice that both definitions use the knowledge discovery process as a key concept.

There is also one more part, that can be included in same domain and sometimes used instead of machine learning, called “statistical learning”. The most important feature that distinguishes machine learning from statistics is the purpose of their usage. While machine learning methods are trained to obtain predictions as accurate as possible, statistical methods are used

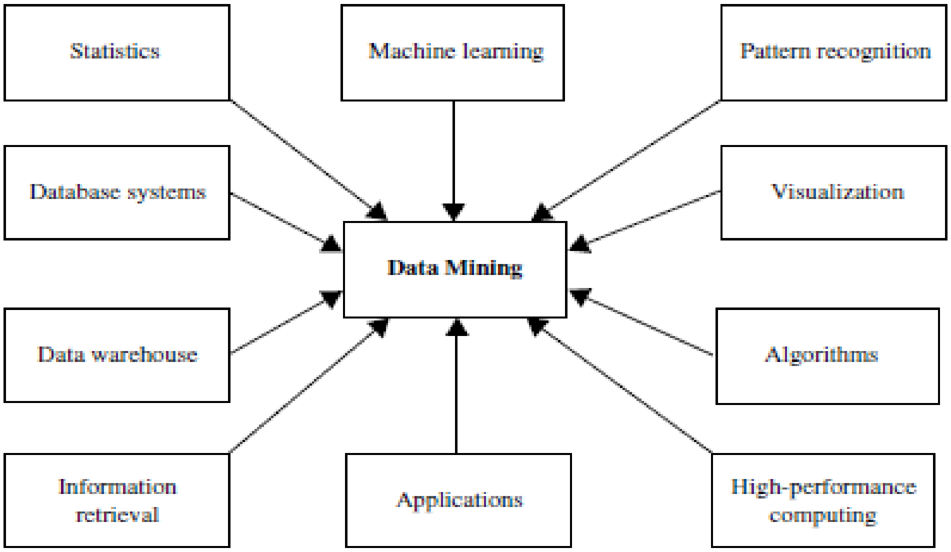

Figure 1: Data Mining Machine Learning Interaction [10]

for revealing the relationships between the variables [9]. In addition to this definition, According to James et al. [18] statistical learning refers to a set of tools for modeling and understanding complex datasets. It is an area in statistics and blends with parallel developments in computer science and, in particular, machine learning. Through that definitions the difference of statistical learning can be considered as having "inference" function instead of just "prediction" in terms of being a sub-field of statistics.

Data mining traditionally focus more on providing knowledge or models that are explainable or interpretable by humans, while machine learning studies are often more focused on what a model does [8]. Even though, all of them are interactive concepts and engaged in data, one can say that each reflects different aspects of data. Consequently, while artificial intelligence is a term which refers to generally intelligence systems, machine learning is a kind of method in artificial intelligence based on training and testing models using data. Data mining is an approach of knowledge discovery with using machine learning techniques.

Basic Concepts

As it is seen above, in terms of techniques there is no exact discrimination between machine learning and data mining. That is, there may be a different aspect of a method in machine learning corresponding to a data mining approach. For example, learning types of a model are related to some data mining concepts as detailed below.

Basic Data Mining Methods

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions [10]. The process of data mining typically consists of 3 steps, carried out in succession: Data Preprocessing, Data Analysis, and Result Interpretation (see Figure 2) [11]. Basic data mining framework starts with data preprocessing, and includes some machine learning methods.

Data preprocessing is considered as a crucial step in order to apply machine learning techniques to data. Data cleaning can be applied to remove noise and correct inconsistencies in data. Data integration merges data from multiple sources into a coherent data store such as a data warehouse. Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering. Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements [10].

Exploratory data analysis (EDA) is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a dataset [17].

Feature engineering involves the careful selection and manipulation of dataset features. The target is to use the most necessary part of input to

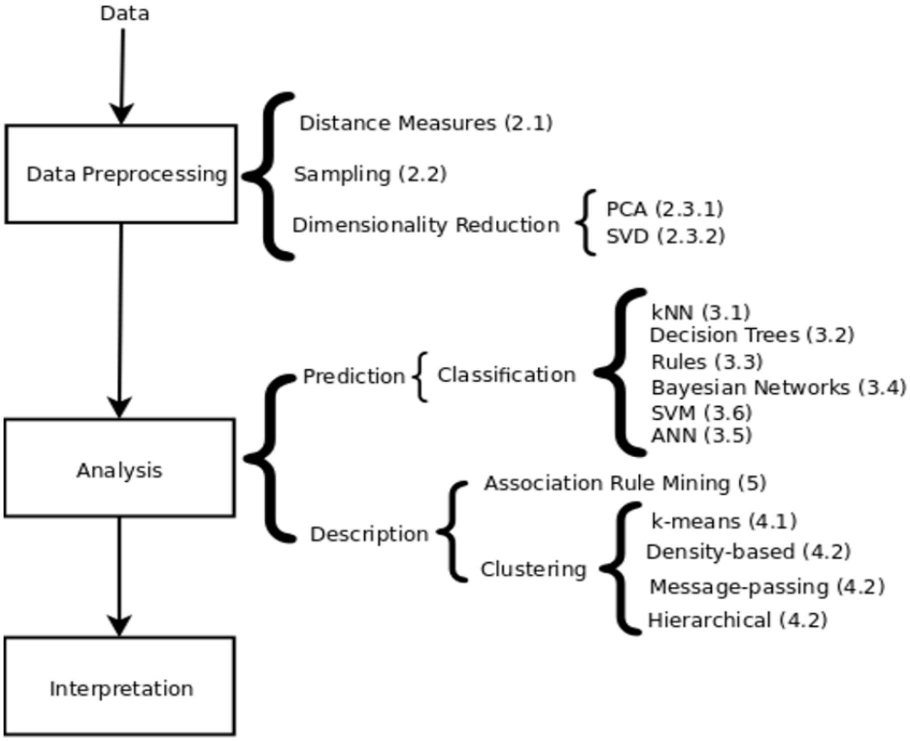

Figure 2: Basic Data Mining Method [11]

make accurate predictions and does not have to deal with any extra noise that comes from the rest of the data [19]. Whereas feature engineering is a preprocessing task, **feature selection** is different concept that refers to reducing irrelevant features during model assessment process. As machine learning aims to address larger, more complex tasks, the problem of focusing on the most relevant information in a potentially overwhelming quantity of data has become increasingly important. Thus, it is required to handle data sets containing large amounts of irrelevant information focusing on the problem of selecting relevant features, and the problem of selecting relevant examples [4].

Principal Component Analysis (PCA) [2] is a classical statistical method to find patterns in high dimensionality data sets. PCA allows to obtain an ordered list of components that account for the largest amount of the variance from the data in terms of least square errors: The amount of variance captured by the first component is larger than the amount of variance on the second component and so on. We can reduce the dimensionality of the data by neglecting those components with a small contribution to the variance [11]. PCA is a popular data processing and dimension reduction technique, and also considered as an unsupervised machine learning method [21].

Frequent pattern mining and association rule mining are both data mining techniques that aimed at detecting frequent items and extracting interesting association rules between items from especially transaction data. These methods, in particular, are explained regarding an analysis of an itemset. Association rule mining focuses on finding rules that will predict the occurrence of an item based on the occurrences of other items in a transaction. The fact that two items are found to be related means co-occurrence but not causality [11]. In general, association rule mining can be viewed as a two-step process: 1) Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count. 2) Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence [10]. “Apriori” [15], “FP-Growth” [20], “EClaT” [3] can be considered as some of basic algorithms for mining frequent patterns and association rules.

Common Concepts

Supervised learning is basically a synonym for **classification**. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model[10]. In supervised classification, a set of labels or categories is known in advance and we have a set of labeled examples which constitute a training set [11]. Classification algorithms can be grouped into three simple groups [10]: 1) Decision Trees with algorithms like “ID.3” [12], “C4.5” [13] or “CART” [16]; 2) Bayesian Methods; 3) Rule - Based classifications.

Unsupervised learning is essentially a synonym for **clustering**. The learning process is unsupervised since the input examples are not class labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits [10]. Clustering task is also known as unsupervised classification. Thus, the labels or categories are unknown in advance and the task is to suitably (according to some criteria) organize the elements at hand [11]. According to Han et al. [10], clustering methods can be grouped basically as partitioning methods with "k-Means" approach, hierarchical methods, density-based methods, and grid-based methods.

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes [10].

Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program [10].

Model Selection and Assessment

As machine learning is built on generating a model with a basic train and test approach, model selection and assessment model selection and assessment become an integral part. Tibshirani et. al [1] define this process: "in model selection we estimate the performance of various competing models with the hope of choosing the best one. Having chosen the final model, we assess the model by estimating the prediction error on new data." **Cross-validation** and **bootstrap** are mainly related concepts with model assessment and model selection. For example, cross-validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility. The process of evaluating a model's performance is known as "model assessment", whereas the process of selecting the proper level of flexibility for a model is known as “model selection”. On the other hand bootstrap is used in several contexts, most commonly to provide a measure of accuracy of a parameter estimate [18].

Bias-variance trade-off: Variance refers to the amount by which the model would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different model. On the other hand bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. As a general rule, as one use more flexible methods that fit the data given, the variance will increase and the bias will decrease [18]. Therefore, there can be considered that there is a trade-off between bias and variance.

Applications

Data mining applications are commonly used in banks and other financial institutions: Design and construction of data warehouses for multidimensional data analysis and data mining, loan payment prediction and customer credit policy analysis for example customer retention analysis [7], classification and clustering of customers for targeted marketing with plenty of recommendation analysis, detection of money laundering and other financial crimes (one can be found a complete review of those kind applications in Albashrawi [5]). There are also much more industries or engineering areas (see Figure 3) in which data mining applications are very common, such as retail and telecommunication [10].

Domains using applications of machine learning techniques also can be listed. The list, in alphabetical order, below specifies application areas to which various existing learning systems have been applied [6]: Agriculture, chemistry, cognitive modeling (simulating human learning processes), computer programming, education, expert systems (high-performance, domain-specific ai programs), game playing (chess, checkers, poker, and so on), general methods (no specific domain), image recognition, mathematics, medical diagnosis, music, natural language processing, physical object characterizations, physics, planning and problem-solving, robotics, sequence extrapola-

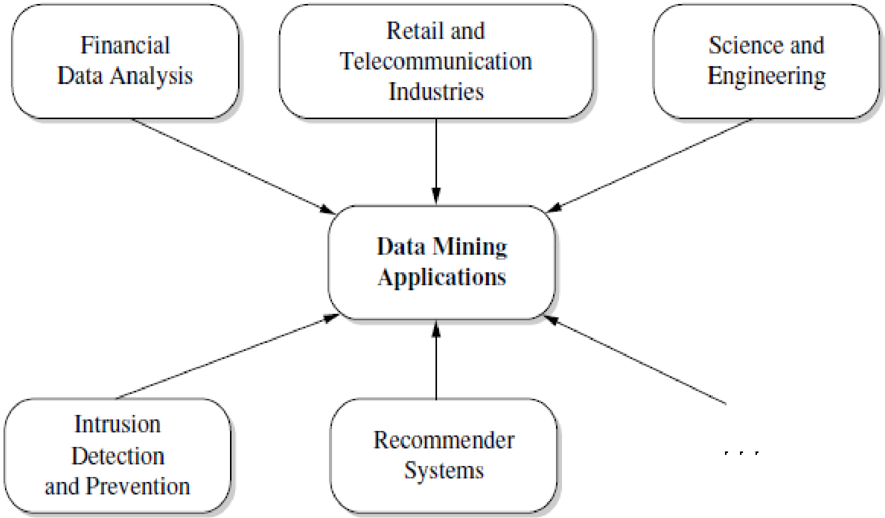

Figure 3: Data Mining Application Domains [10]

tion, speech recognition.

Conclusion

Whereas data mining is a systematic approach of extracting hidden and interesting knowledge from the data, machine learning is, on the other hand a model generating and evaluating technique itself in order to make predictions. Even though, sometimes statistical learning term is used interchangeably with machine learning, it is a little different field which inherits statistical inference. Artificial intelligence is such a term that embraces all fields that requires intelligence.

It is quite reasonable that seeming blended each concept when one considers the applications, because a ordinary data miner can use machine learning approaches in order to achieve his or her task.

References

[1] Tibshirani R. Friedman J., Hastie T. Model assessment and selection. *The Elements of Statistical Learning*, pages 128–155, 2001.

[2] Jolliffe I. *Principal Component Analysis*. Springer, second edition, 2002.

[3] Zaki M. J. Scalable algorithms for association mining. *IEEE Trans Knowl Data Eng*, 12(3):372–390, 2000.

[4] Blum A. L. Langley P. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, 1997.

[5] Albashrawi M. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, 14:553–570, 2016.

[6] Carbonel G. J. Mitchell M. T., Michalski R. S. Machine learning: A historical and methodological analysis. *AI Magazine*, 4(3):68–79, 1983.

[7] Liu H. Ng K. Customer retention via data mining. *Artificial Intelligence Review*, 14:569–590, 2000.

[8] Fournier-Viger P. What is the difference between data mining and machine learning? *The Data Mining Blog*, 2019. Available at: https://data-mining.philippe-fournier-viger.com/what-is-the-difference-between-data-mining-and-machine-learning/.

[9] Thomas S. Paluszek S. An overview of machine learning. *MATLAB Machine Learning*, 1:3–15, 2016.

[10] Han J. Pei J., Kamber M. *Data Minig Concepts and Techniques*. Elsevier, third edition, 2012.

[11] Jaimes A. Amatriain X. Pujol J. M., Oliver N. Data mining methods for recommender systems. *Recommender Systems Handbook*, pages 39–71, 2010.

[12] Quinlan J. R. Induction of decision trees. *Machine Learning,*, 1(1):81–106, 1986.

[13] Quinlan J. R. *C4.5: Programs for Machine Learning*. 1993.

[14] Fayyad U. Smyth P., Piatetsky-Shapiro G. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):68–79, 1996.

[15] Agrawal R. Srikant R. Fast algorithms for mining association rules. *Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94)*, pages 487–499, 1994.

[16] Friedman J. Breiman L. Stone C. J., Olshen R. A. *Classification And Regression Trees*. Chapman Hall/CRC, 1983.

[17] Behrens J. T. Principles and procedures of exploratory data analysis. *Psychological Methods*, 2(2):131–160, 1997.

[18] Witten D. James G. Thibshirani R., Hastie T. *An Introduction to Statistical Learning*. 1983.

[19] Zhang M. Xue B. Yao X., Browne W. N. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions On Evolutionary Computation*, 20(4):131–160, 2016.

[20] Han J. Yin Y., Pei J. Mining frequent patterns without candidate generation. *ACM SIGMOD Rec.*, 29(2):1–12, 2000.

[21] Tibshirani R. Zou H., Hastie T. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.