Unsupervised and Semi-Supervised Learning
Series Editor: M. Emre Celebi

Michael W. Berry
Azlinah Mohamed
Bee Wah Yap *Editors*

Supervised and Unsupervised Learning for Data Science

Springer

Unsupervised and Semi-Supervised Learning

Series Editor

M. Emre Celebi, Computer Science Department, Conway, Arkansas, USA

Springer's Unsupervised and Semi-Supervised Learning book series covers the latest theoretical and practical developments in unsupervised and semi-supervised learning. Titles – including monographs, contributed works, professional books, and textbooks – tackle various issues surrounding the proliferation of massive amounts of unlabeled data in many application domains and how unsupervised learning algorithms can automatically discover interesting and useful patterns in such data. The books discuss how these algorithms have found numerous applications including pattern recognition, market basket analysis, web mining, social network analysis, information retrieval, recommender systems, market research, intrusion detection, and fraud detection. Books also discuss semi-supervised algorithms, which can make use of both labeled and unlabeled data and can be useful in application domains where unlabeled data is abundant, yet it is possible to obtain a small amount of labeled data.

Topics of interest in include:

- Unsupervised/Semi-Supervised Discretization
- Unsupervised/Semi-Supervised Feature Extraction
- Unsupervised/Semi-Supervised Feature Selection
- Association Rule Learning
- Semi-Supervised Classification
- Semi-Supervised Regression
- Unsupervised/Semi-Supervised Clustering
- Unsupervised/Semi-Supervised Anomaly/Novelty/Outlier Detection
- Evaluation of Unsupervised/Semi-Supervised Learning Algorithms
- Applications of Unsupervised/Semi-Supervised Learning

While the series focuses on unsupervised and semi-supervised learning, outstanding contributions in the field of supervised learning will also be considered. The intended audience includes students, researchers, and practitioners.

More information about this series at http://www.springer.com/series/15892

Michael W. Berry • Azlinah Mohamed
Bee Wah Yap
Editors

Supervised and Unsupervised Learning for Data Science

Springer

Editors
Michael W. Berry
Department of Electrical Engineering
and Computer Science
University of Tennessee at Knoxville
Knoxville, TN, USA

Azlinah Mohamed
Faculty of Computer & Mathematical
Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

Bee Wah Yap
Advanced Analytics Engineering Centre,
Faculty of Computer
and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

ISSN 2522-848X ISSN 2522-8498 (electronic)
Unsupervised and Semi-Supervised Learning
ISBN 978-3-030-22474-5 ISBN 978-3-030-22475-2 (eBook)
https://doi.org/10.1007/978-3-030-22475-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Supervised and unsupervised learning algorithms have shown a great potential in knowledge acquisition from large data sets. Supervised learning reflects the ability of an algorithm to generalize knowledge from available data with target or labeled cases so that the algorithm can be used to predict new (unlabeled) cases. Unsupervised learning refers to the process of grouping data into clusters using automated methods or algorithms on data that has not been classified or categorized. In this situation, algorithms must "learn" the underlying relationships or features from the available data and group cases with similar features or characteristics. When small amounts of labeled data are available, the learning is specified as "semi-supervised." This volume provides both foundational knowledge for novice or beginning researchers in machine learning and new techniques for improving both the accuracy and computational complexity of supervised and unsupervised learning in the context of relevant and practical applications.

Part I of this volume is dedicated to the discussion of state-of-the-art algorithms used in supervised and unsupervised learning paradigms. In Chap. 1, Alloghani et al. provide a systematic literature review of scholarly articles published between 2015 and 2018 that address or implement supervised and unsupervised machine learning techniques in different problem-solving paradigms. In Chap. 2, C. Lursinsap addresses recent approaches to overcome commonly observed problems in big data analytics, such as data overflow, uncontrollable learning epochs, arbitrary class drift, and dynamic imbalanced class ratios. In Chap. 3, T. Panitanarak discusses recent improvements in the performance of graph-based shortest path algorithms that are commonly used in machine learning. Finally, in Chap. 4, R. Lowe and M. Berry illustrate the use of tensor-based algorithms for the unsupervised learning of influence in text-based media.

Part II of this volume highlights the various applications of learning algorithms including cancer diagnosis, social media and text mining, and prediction of stress-strain parameters in civil engineering. In Chap. 5, Prasetyo et al. demonstrate the use of support vector machines (SVMs) in cancer survival data analysis. In Chap. 6, D. Martin et al. discuss the use of latent semantic analysis (LSA) for unsupervised word sense disambiguation in textual documents. In Chap. 7, Pornwattanavichai

et al. explain how hybrid recommendation systems with latent Dirichlet analysis (LDA) can improve the unsupervised topic modeling of tweets. Finally, in Chap. 8, Jebura et al. demonstrate the use of artificial neural networks (ANNs) to predict nonlinear hyperbolic soil stress-strain relationship parameters in civil engineering models.

Some of the research described in this volume was supported by the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Robinson Pino, program manager, under contract number DE-AC05-00OR22725.

Knoxville, TN — Michael W. Berry
Shah Alam, Selangor, Malaysia — Bee Wah Yap
Shah Alam, Selangor, Malaysia — Azlinah Mohamed

Contents

Part I
Algorithms

Chapter 1
A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science

Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J. Aljaaf

1.1 Introduction

The demand for advanced data analytics leading to the use of machine learning and other emerging techniques can be attributed to the advent and subsequent development of technologies such as Big Data, Business Intelligence, and the applications that require automation. As Sandhu [1] explains, machine learning is a subset of artificial intelligence, which uses computerized techniques to solve problems based on historical data and information without unnecessarily requiring modification in the core process. Essentially, artificial intelligence involves creation of algorithms and other computation techniques that promote smartness of machines. It encompasses algorithms that think, act, and implement tasks using protocols that are otherwise beyond human's reach.

M. Alloghani (✉)
Applied Computing Research Group, Liverpool John Moores University, Liverpool, UK

Abu Dhabi Health Services Company (SEHA), Abu Dhabi, UAE
e-mail: M.AlLawghani@2014.ljmu.ac.uk; mloghani@seha.ae

D. Al-Jumeily · A. Hussain
Applied Computing Research Group, Liverpool John Moores University, Liverpool, UK

J. Mustafina
Kazan Federal University, Kazan, Russia
e-mail: dnmustafina@kpfu.ru

A. J. Aljaaf
Applied Computing Research Group, Liverpool John Moores University, Liverpool, UK

Centre of Computer, University of Anbar, Anbar, Iraq
e-mail: A.J.Kaky@ljmu.ac.uk; a.j.aljaaf@uoanbar.edu.iq

M. W. Berry et al. (eds.), *Supervised and Unsupervised Learning for Data Science*,
Unsupervised and Semi-Supervised Learning,
https://doi.org/10.1007/978-3-030-22475-2_1

Machine learning is a component of artificial intelligence although it endeavors to solve problems based on historical or previous examples [2]. Unlike artificial intelligence applications, machine learning involves learning of hidden patterns within the data (data mining) and subsequently using the patterns to classify or predict an event related to the problem [3]. Simply, intelligent machines depend on knowledge to sustain their functionalities and machine learning offers such a knowledge. In essence, machine learning algorithms are embedded into machines and data streams provided so that knowledge and information are extracted and fed into the system for faster and efficient management of processes. It suffices to mention that all machine learning algorithms are also artificial intelligence techniques although not all artificial intelligence methods qualify as machine learning algorithms.

Machine learning algorithms can either be supervised or unsupervised although some authors also classify other algorithms as reinforcement, because such techniques learn data and identify pattern for the purposes of reacting to an environment. However, most articles recognize supervised and unsupervised machine learning algorithms. The difference between these two main classes is the existence of labels in the training data subset. According to Kotsiantis [4], supervised machine learning involves predetermined output attribute besides the use of input attributes. The algorithms attempt to predict and classify the predetermined attribute, and their accuracies and misclassification alongside other performance measures is dependent on the counts of the predetermined attribute correctly predicted or classified or otherwise. It is also important to note the learning process stops when the algorithm achieves an acceptable level of performance [5]. According to Libbrecht and Noble [2], technically, supervised algorithms perform analytical tasks first using the training data and subsequently construct contingent functions for mapping new instance of the attribute. As stated previously, the algorithms require prespecifications of maximum settings for the desired outcome and performance levels [2, 5]. Given the approach used in machine learning, it has been observed that training subset of about 66% is rationale and helps in achieving the desired result without demanding for more computational time [6]. The supervised learning algorithms are further classified into classification and regression algorithms [3, 4].

Conversely, unsupervised data learning involves pattern recognition without the involvement of a target attribute. That is, all the variables used in the analysis are used as inputs and because of the approach, the techniques are suitable for clustering and association mining techniques. According to Hofmann [7], unsupervised learning algorithms are suitable for creating the labels in the data that are subsequently used to implement supervised learning tasks. That is, unsupervised clustering algorithms identify inherent groupings within the unlabeled data and subsequently assign label to each data value [8, 9]. On the other hand, unsupervised association mining algorithms tend to identify rules that accurately represent relationships between attributes.

1.1.1 Motivation and Scope

Even though both supervised and unsupervised algorithms are widely used to accomplish different data mining tasks, the discussion of the algorithms has been mostly done singly or grouped depending on the need of learning tasks. More importantly, literature reviews that have been conducted to account for supervised and unsupervised algorithms either handle supervised techniques or unsupervised ones with limited focus on both approaches in the same. For instance, Sandhu [1] wrote a review article on machine learning and natural language processing but focused on supervised machine learning. The author did not conduct a systematic review and, as such, the article does not focus on any specific period or target any given database. Baharudin et al. [10] also conducted a literature review on machine learning techniques though in the context of text data mining and did not implement any known systematic review methodology. Praveena [11] also conducted a review of papers that had implemented supervised learning algorithms and, as such, did implement any of the known systematic review approaches. However, Qazi et al. [12] conducted a systematic review although with a focus on the challenges that different authors encountered while implementing different classification techniques in sentimental analysis. The authors reviewed 24 papers that were published between 2002 and 2014 and concluded that most review articles published during the period focused on eight standard machine learning classification techniques for sentimental analysis along with other concept learning algorithms. Unlike these reviews, the systematic review here conducted focused on all major stand-alone machine learning algorithms, both supervised and unsupervised published during the 2015–2018 period.

1.1.2 Novelty and Review Approach

The systematic review relied on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) tool to review studies that have used different supervised and unsupervised learning algorithms to address different issues [13]. The approach used in the search was such that different papers published between 2013 and 2018 dealing with the use of machine learning algorithms as methods of data analysis were included. The identification and subsequent inclusion and exclusion of the articles reviewed was based on whether the paper is peer-reviewed, scholarly, full-text, and year of publication that ranges between 2015 and 2018 [13–15]. The search was conducted on EBSCO and ProQuest Central Databases. The search queries used are as follows, and they were implemented in the two databases. In conventional PRISMA review, it is a requirement to check and identify the search criteria in the title and the structure of the abstract alongside introduction (rationale and objectives) and methods including information sources, data items, summary measures, and synthesis results [16]. However, such an approach was adopted, and

Table 1.1 Summary of the queries used to search ProQuest Central and EBSCO databases

Query
(“Machine learning”) AND (Supervised Learning AND Unsupervised Learning)
(“Data mining”) AND (Supervised machine learning algorithms)
(“Supervised Machine Learning”) AND (“Unsupervised Machine Learning”)

applied to published articles instead of being implemented on review articles. Table 1.1 summarizes the search queries that were run in the two databases.

The inclusion criteria deferred for both databases with EBSCO relying on date of publication and full-text to narrow the search, while ProQuest Central search filters included Abstract (AB), Document Text (FT), Document Title (TI), and Publication Title (PUB). An instance of search implemented in ProQuest Central with some of the above criteria is as shown below.

```
ft(Supervised machine learning) AND ft(Unsupervised machine
learning) OR ti(Supervised machine learning) AND ti(Unsupervised
machine learning) OR pub(Supervised machine learning) AND
pub(Unsupervised machine learning)
```

1.2 Search Results

The search and screening results based on PRISMA and elements of meta-analysis are presented in the following section. The major steps used to arrive at the final articles and subsequent analysis included screening (rapid title screening), full test screening, data extraction including extraction of the characteristics of the study, and meta-analysis based on specific check lists and aspects of the machine learning algorithm used.

1.2.1 EBSCO and ProQuest Central Database Results

The search results obtained from the two databases before the commencement of the review process were as follows. The EBSCO search identified 144 articles that were published between 2015 and 2018. Of the 144 documents, 74 had complete information including name of authors, date of publication, name of journal, and structured abstracts. However, only 9 of the 74 articles had full-text and, as such, selected for inclusion in the review process. As for the search results from ProQuest Central, the initial search yielded over 19,898 results, but application of the filters reduced 3301 articles, of which 42 were reviews and 682 covered classification techniques, while 643 covered or had information related to algorithms in general. However, the subject alignment of the research papers was not considered because of the wide spectrum of application of the algorithms such that both supervised

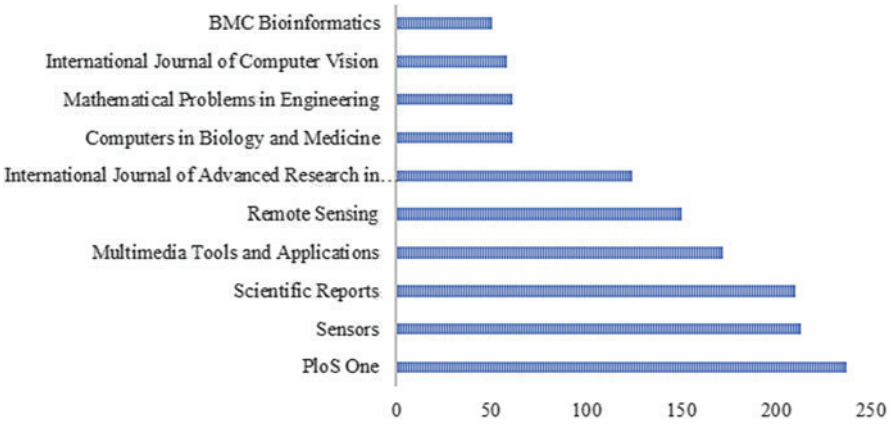

Fig. 1.1 The distribution of ProQuest Central Search Results as per the top ten publication titles (journals)

and unsupervised methods were also applied in other subjects. The distribution the search result based on top ten journals is as shown in Fig. 1.1.

Figure 1.1 shows that PloS One had the highest number of articles published on supervised and unsupervised machine learning. Sensors and Scientific Reports (Nature Publisher Group) had 213 and 210 articles. Multimedia Tools and Applications (172), Remote Sensing (150), and International Journal of Computer Vision (124) had over 100 articles. Even though Mathematics Problems in Engineering and Internal Computer Vision had 61 and 58 articles, the two publications were better placed at exploring the mathematical and algorithmic aspects of supervised and unsupervised machine learning algorithms. The inclusion and exclusion criteria focused on the algorithms as well as their mathematical discourse and application in different fields.

Based on the PRISMA checklist, a total of 84 articles were included in the study and their content analyzed for the implementation of supervised and unsupervised machine learning techniques.

The final number of articles used in the review is 84, although 20 of them underwent meta-analysis when each study was vetted for clarity of the objectives and study questions. Regarding study questions and the effectiveness of the approached used to implement the chosen machine learning algorithms resulted in exclusion of 1290 articles (Fig. 1.2). The rest (1985) met the required study question criteria but also screened for the comprehensiveness of the literature search, data abstraction, evaluation of the results, and the applicability of results [17–19]. It is imperative to note that publication bias and disclosure of funding sources were not considered as part of the screen process. The 84 articles met these meta-analysis requirements and were subsequently included in the analysis (Fig. 1.2).

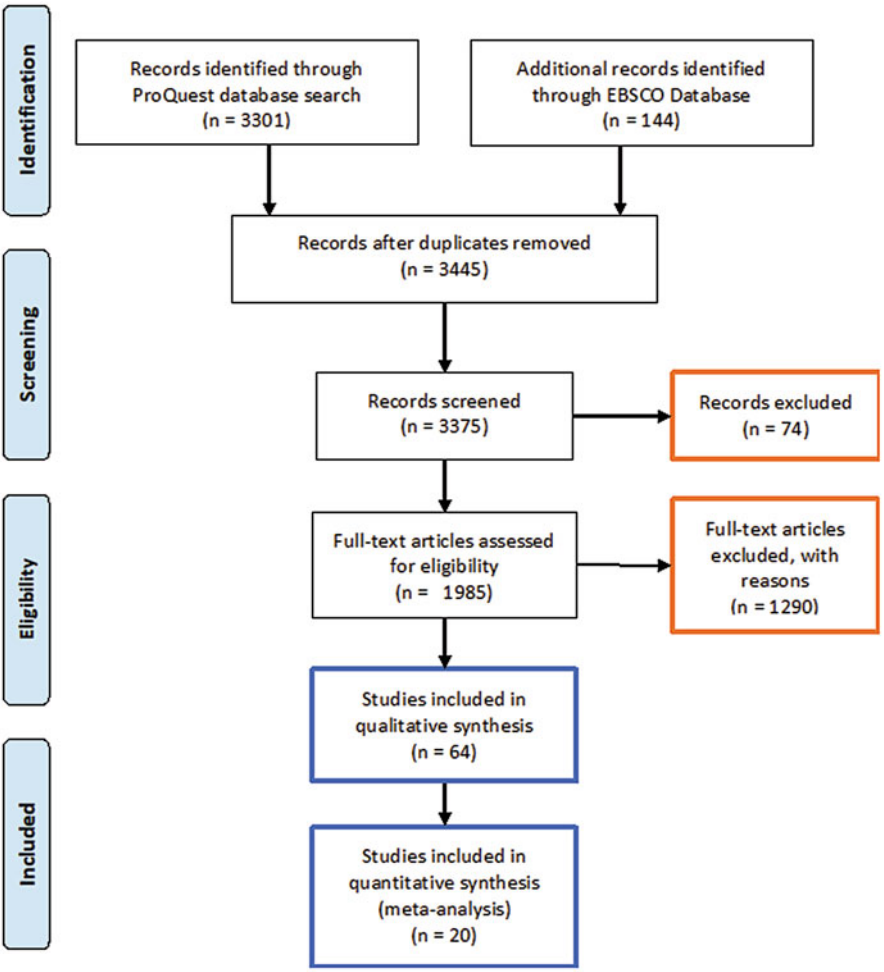

Fig. 1.2 The PRISMA flow diagram for the search conducted on ProQuest Central and EBSCO and the final number of studies included the analysis

It is crucial to note that of the 84 articles that were included in the study, 3 were published in 2013 and 3 were published in 2014 but were not filtered out by the data of publication restriction.

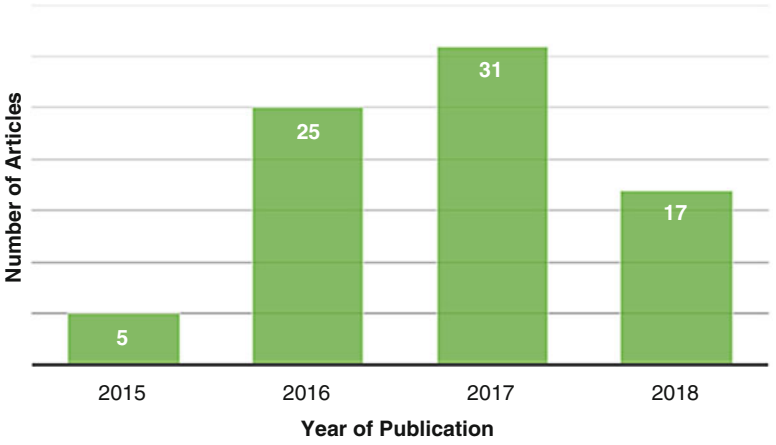

Fig. 1.3 Distribution of articles based on year of publication

1.2.2 Distribution of Included Articles

The articles used in the study consisted of Feature, Journal Articles, General Information, Periodical, and Review types with a distribution represented in the following chart.

From Fig. 1.3, 78 articles were published between 2015 and 2018, while the missing articles were published in 2013 [20–22] and 2014 [23–25] and their inclusion can be associated to publication biasness, which is also observed in the type of documents or study. According to the search, inclusion, and inclusion criteria, the final results ought to have only journal articles, but others were features, general information, periodicals, and reviews. The six papers that were published between 2013 and 2014 were included, because they met all the criteria required for meta-analysis and the indexed meta-data showed that the papers were published in 2015. Regarding the misinformation, we can deduce that the publications had an inaccuracy of about 7.2%.

1.3 Discussion

The 84 articles discussed different supervised and unsupervised machine learning techniques without necessarily making the distinction. According to Praveena [11], supervised learning requires an assistance born out of experience or acquired patterns within the data and, in most cases, involves a defined output variable [26–30]. The input dataset is segregated into train and test subsets, and several papers address the concept of training datasets based on the desired outcome [31–34]. All the algorithms that use supervised learning approach acquire patterns within the

training dataset and subsequently apply them to the test subset with the object of either predicting or classifying an attribute [35–37]. Most of the authors described the workflow of a supervised machine learning and, as it also emerged from the review, decision tree, Naïve Bayes, and Support Vector Machines are the most commonly used algorithms [8, 38–42].

1.3.1 Decision Tree

It is important to recall that supervised learning can either be based on a classification or regression algorithm, and decision tree algorithm can be used as both although it is mainly used for classification as noted in these articles [20, 43–45]. The algorithm emulates a tree, and it sorts attributes through groupings based on data values [46]. Just like a conventional tree, the algorithm has branches and nodes with nodes representing variable group for classification and branches, assuming the values that the attribute can take as part of the class [47, 48]. The pseudocode illustrating the decision tree algorithm is as shown below. In the algorithm, D is the dataset, while x and y are the input and target variables, respectively [49, 50].

Algorithm 1.1: Decision Tree

Protocol *DT Inducer* (D, x, y)

1. $\boldsymbol{T = Tree\ Growing\ (D, x, y)}$
2. Return Tree Pruning (D, T)

Method Tree Growing (D, x, y)

1. Create a tree T
2. **if** at least one of the Stopping Criteria is satisfied **then**;
3. label the root node as a leaf with the most frequent value of y in D as the correct class.
4. **else**;
5. Establish a discrete function f(x) of the input variable so that splitting D according to the functions outcomes produces the best splitting metric
6. **if** the best metric is greater or equal to the threshold **then**;
7. Mark the root node in T as f(x)
8. **for** each outcome of f(x) at the node **do**;
9. $\boldsymbol{Subtree = Tree\ Growing\ \left(\delta_{f(x)=t_1}, D, x, y\right)}$
10. Connect the root of T to Subtree and label the edge $\boldsymbol{t_1}$
11. **end for**
12. **else**
13. Label the root node T for a leaf with the frequent value of y in D as the assigned class
14. **end if**

15. **end if**
16. Return T

Protocol *Tree Pruning* (D, T, y)

1. **repeat**
2. Select a node t in T to maximally improve pruning evaluation procedure
3. **if** $t \neq 0$ **then**;
4. $\boldsymbol{T} = \boldsymbol{pruned}\ (\boldsymbol{T}, \boldsymbol{t})$
5. **end if**
6. **until** $t = 0$
7. Return T

As illustrated in the pseudocode, Decision Tree achieves classification in three distinct steps. Firstly, the algorithm induces both tree growing and tree pruning functionalities [51]. Secondly, it grows the tree by assigning each data value to a class based on the value of the target variable that is the most common one at the instance of iteration [52, 53]. The final step deals with pruning the grown tree to optimize the performance of the resultant model [19, 53, 54]. Most of the reviewed studies involved application of decision trees for different applications, although most involved classification cancer and lung cancer studies, clinical medicine especially diagnosis of conditions based on historical data as well as some rare forms of artificial intelligence applications [40, 52, 55–57]. Most of the studies have also recognized decision tree algorithms to be more accurate when dealing with data generated using the same collection procedures [43, 44, 52].

1.3.2 Naïve Bayes

The Naïve Bayes algorithm has gained its fame because of its background on Bayesian probability theorem. In most texts, it is considered a semisupervised method, because it can be used either in clustering or classification tasks [58, 59]. When implemented as a technique for creating clusters, Naïve Bayes does not require specification of an outcome and it uses conditional probability to assign data values to classes and, as such, is a form of unsupervised learning [47, 60–62]. However, when used to classify data, Naïve Bayes requires both input and target variables and, as such, is a supervised learning technique [55, 63, 64]. As a classifier, the algorithm creates Bayesian networks, which are tree generated based on the condition probability of an occurrence of an outcome based on probabilities imposed on it by the input variables [65, 66]. The pseudocode for the Naïve Bayes algorithm is presented below [49, 67, 68].

Algorithm 1.2: Naïve Bayes Learner

Input: training set $\boldsymbol{T_s}$, Hold-out set $\boldsymbol{H_s}$, initial components, $\boldsymbol{I_c}$, and convergence thresholds $\boldsymbol{\rho_{EM}}$ and $\boldsymbol{\rho_{add}}$

Initial M using one component
$\boldsymbol{I} \leftarrow \boldsymbol{I_c}$.
repeat
Add I components to M thereby initializing M using random components drawn from the training set $\boldsymbol{T_s}$
Remove the I initialization instances from $\boldsymbol{T_s}$
repeat
E-step: Proportionally assign examples in $\boldsymbol{T_s}$ to resultant mixture component using M
M-Step: Calculate maximum likelihood parameters using the input data.
if ***log P*** $(\boldsymbol{H_s/M})$ is the best maximum probability, then save M in $\boldsymbol{M_{best}}$
every 5 cycles of the two steps, prune low-weight components of M
until P $(\boldsymbol{H_s/M})$ fails to increase by the ratio $\boldsymbol{\rho_{EM}}$
M$\leftarrow\boldsymbol{M_{best}}$
Prune low weight components of M
$\boldsymbol{I} \leftarrow \boldsymbol{2I}$.
until P $(\boldsymbol{H_s/M})$ fails to increase by the ratio $\boldsymbol{\rho_{add}}$
Execute both E: step and M: step twice on $\boldsymbol{M_{best}}$ using examples from $\boldsymbol{H_s}$ and $\boldsymbol{T_s}$
Return M$\leftarrow\boldsymbol{M_{best}}$

As the pseudocode illustrates, Naïve Bayes algorithm relies on Bayes' theorem represented mathematical below to assign independent variables to classes based on probability [31, 58].

$$P\,(H|D) = \frac{P(H)P\,(D|H)}{P(D)} \tag{1.1}$$

In Eq. (1.1), the probability of H when the probability of D is known is defined in terms of the product probability of H, probability of D given the probability of H divided by the probability of D. The H and D are events with defined outcome and they can represent Heads and Tails in coil tossing experiments [12, 45, 69, 70]. The extension of the theorem in supervised learning is of the form represented in Eq. (1.2).

$$P\,(H|D) = P\,(x_i, \ldots, x_n|H) = \prod_i P\,(x_i|H) \tag{1.2}$$