Journal of Education
Rabindra Bharati University

ISSN: 0972-7175
A Peer Reviewed Journal

CERTIFICATE OF PUBLICATION

A REVIEW OF VARIOUS MACHINE LEARNING AND DATA MINING APPROACHES IN ANALYSING DATA

Authored By

Nayana K V
Asst. Professor, St Francis de Sales College, Electronic city, Bangalore-560100

Published in

Journal of Education: Rabindra Bharati University
ISSN : 0972-7175

Vol. : XXIV, No. :6, 2021-2022

UGC CARE, Peer Reviewed and Refereed Journal

Journal of Education

Editor
Journal of Education

JOURNAL OF EDUCATION: RABINDRABHARATI UNIVERSITY
ISSN : 0972-7175

A REVIEW OF VARIOUS MACHINE LEARNING AND DATA MINING APPROACHES IN ANALYSING DATA

Nayana K V
Asst. Professor, St Francis de Sales College, Electronic city, Bangalore-560100

Sailaja M
Asst. Professor, St Francis de Sales College, Electronic city, Bangalore-560100

Data mining DM and KDD has emerged as a problem-solving technique for analysing data for pre-existing databases, growing data industry issues and resulting consumer demands for different approaches to extract useful information from large data stores. This paper reviews the different machine learning algorithms used in the UCI repository for different training data sets. Machine learning is known as supervised and unsupervised learning, so supervised learning is acquired from different classification definitions, i.e. a new instance classifier. Unsupervised learning problems in separate unclassified classes. Predictive datamining is often referred to as supervised learning and, based on different association principles, descriptive datamining is unsupervised. The approach to machine learning and datamining focuses on categorical, on-numeric, and interpretable data processing. Cross-industry standard method for CRISP-DM datamining of mining techniques for KDD-based market solutions. Different research papers on datamining tools and algorithms and their effect on supervised learning are checked for fruitful data decisions.

Key words: UCI, Datamining (DM), knowledge discovery databases KDD, supervised, unsupervised

1. INTRODUCTION

Data is massive, so it's beyond human beings' understanding capacity to make a successful discovery of knowledge. The primary objective of datamining is to retrieve valuable knowledge from vast databases in a humanly understandable format. We may conclude that data mining is an intersection of different fields, such as machine learning, artificial intelligence, etc. In areas such as game engineering, biological, analytics and visualization, datamining applications are vast. Different datamining instruments such as R method, Rapid miner, keel, weka, orange etc. are available on the market. Datamining techniques such as grouping, clustering and regression methods are used to discover details and prepare for the future.

C lustering has three methods in which instances are grouped into groups that have been defined. The approach to clustering is focused on unsupervised learning, as there are no predefined groups. Data can be grouped together as a cluster in this method. Classification is a common activity in data mining, especially in the discovery of information and future plan, it offers smart decision-making, classification is not only used to research and analyze existing sample data, but also predicts the sample data's future actions. Two phases are included in the classification, first the step of the learning process in which the training data is evaluated, then the rules and patterns are formed. The second step checks the data and documents the consistency of the patterns of classification. Regression is used to map data items into a very useful estimation variable. Various algorithms such as decision tree, nearest neighbor, genetic algorithm support vector machine (SVM) etc. in the classification technique. We discuss the different classification algorithms and compare them in this paper. We first give Decision Tree Principles, Bayesian Network, and K-Nearest Neighbor Support Vector Machine in the rest of this paper.

2. LITERATURE REVIEW

A lot of study on the comparison of different machine learning algorithms and their effects on different datasets of training has been reviewed. In a study conducted in 2004, cross-validation of breast cancer data sets elicited in bin Othman[1] was conducted using the weka tool to perform multiple analyses using the K-fold cross-validation method, but there is no processing of thorough research. Likes writers. A new spatial decision tree algorithm was proposed by Imas SukaesihSitanggang et al [2] . Accordingly, in relational datamining, a data set is divided into two layers of spatial and referential relationship between different tuples, so its accuracy is called 74.72 percent . In order to construct multiple decision tree, Khatwani, S. Et al[13] suggested the Id3 algorithm and genetic algorithm to each predict the output based on the data sets function. Research on the comparison of different training datasets utilizing datamining methods such as weka, tanagra in 2018[4,5] compared the accuracy of algorithms such as KNN, SVM, Naïve bayee, C4.5, IR and OR algorithms, thus achieving accuracy using regression and fuzzy learning functionalities, however there are no connection laws. The new Bayesian classification technique proposed by Qin et al[7] is based on uncertain data by taking 20 data sets from the UCI repository and applying the uncertain Bayesian classification and prediction technique, the proposed Weka algorithm and showing that the outcome of the proposed method is better than that of the Bayesian classification. In spatial databases, David Tania[5] provides an absolute taxonomy of Nearest Neighbor Queries. There are four perspectives in the taxonomy: space, the outcome, query-point, and relationship. Research on the use of weka provided its advantages in 2013[5] relative to other instruments dealing with medical datasets. [6] The key subject of the Weka Tool Survey. In [7, 11, 12], different classification strategies were studied in a comparative way. Jain[8] concentrated on the decision tree of C4.5 and compared its work with the different mining instruments. A thorough description of the use of weka, tanagra and knime is given [14-16].

3. REVIEW COMPARISONS

3.1 Supervised algorithm: Predictive data mining Induction of Models

Analysis focuses primarily on predictive induction, arising from the classification process of datamining, i.e. induction of the decision tree and induction of the rule set.

3.1.1 Decision Tree Induction

Decision tree classification model, has nodes and arcs where each node is identified by the name of the attribute and arc is the valid value of the node-associated attribute. The topmost node is a root node and the originating sub nodes or arcs are leaf nodes where top-down traversal is performed. The Decision Tree is used to build predictive accuracy on the training data set for new instances. Training data collection is collected from the UCI repository according to the analysis carried out from different literature studies. By calculating the purity of the node, the main step in the decision tree is to define the attribute used for node selection.

3.1.2 C4.5 and ID3 Decision Tree Algorithm

It uses information-theoretical entropy as a purity measure in the C 4.5 decision tree algorithm, (quinlan, 1986),

Identify the largest utility attribute, i.e. the difference between the original purity value and the sum of in the analysis of exploratory data. Such decision tree and rule set induction resulting in classification models, association rule learning that is an unsupervised learning method without class label, similarly clustering is also an unsupervised learning method. Although subgroup discovery is a descriptive induction method for pattern learning aimed at finding descriptions of interesting population subgroup.

3.2.1 Association Rule Learning

In the data mining culture, the topic of inducing association rules (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1995) has gained a lot of priority. It is defined as follows: given the set of transactions (examples) where each transaction is a set of items, the rule of association is an expression where the sets of items are B and H, and B! H is interpreted as IF B THEN H, implying that the transactions containing B in a database also appear to contain H. An in-depth survey of the discovery of association law is beyond the scope.

3.2.2 Subgroup Discovery

The task of subgroup discovery is to identify sufficiently large population subgroups that have a class distribution that is substantially different from the whole dataset. Discovery by subgroup results in individual rules, where the conclusion of the law is a class (the property of interest). The key difference between the learning of classification rules and the discovery of subgroups is that it causes single interest rules (subgroups) aimed at revealing interesting characteristics of groups of instances, not necessarily at establishing a classification law.

4. TOOLS USED FOR DATAMINING The data mining tools that we have comparisons

The **WEKA** toolkit[14] was developed in New Zealand by the University of Waikato. For study and didactic purposes, it is commonly used. Weka is the app that is easiest to use and therefore has a wide user base.All the algorithms present in it for machine learning and data mining were written in Java. WEKA includes various functions, such as splitting, validation, regression, and mining.

Tanagra [15] is a free tool for analyzing data. It also supports mathematical learning algorithms, in addition to machine learning and data processing. An open source tool for mining is the primary feature of Tanagra. The secondary role is to allow researchers to establish and compare their own rules of classification with pre-existing methods. The tertiary aim is to provide their open source code to developers so that they can learn how the tool was made.

KNIME [16] is a robust instrument for data processing, discovery, and visualization. KNIME was developed using systematic methodology and is used in the science academy by a number of people.

5. MACHINE LEARNING ALGORITHM COMPARISONS

5.1 Machine learning algorithm

5.1.1 Bayesian Network

A Bayesian Network (BN) is a graphical model of relationships between a collection of features of different variable data sets. A Directed Acyclic Graph (DAG) is this graphical model structure S and all the nodes in S correspond to the characteristics of a data set in one-to-one correspondence. Among the characteristics of datasets, the arcs reflect forces, while the lack of possible arcs in S encodes conditional independence. When applied to large datasets, the Bayesian classifier showed high accuracy and speed[18][19] Bayesian networks are used to model knowledge of bioinformatics, engineering, medicine, biomonitoring, search image processing.

5.1.2. K-Nearest Neighbor

The K-Nearest Neighbor (NN) is the simplest machine learning approach and has some powerful results for accuracy. Based on the nearest training instance in the feature space, the object is classified by KNN. This specifically measures the decision boundary and the decision estimation explicitly. The function of boundary complexity[21] is therefore the computational complexity of NN. From a set of objects for which the correct classification is established, the neighbors are chosen. No special training phase is required, which can be considered as the training set for the algorithm.

The k-NN algorithm is adaptive to the data set's local structure. This is the unique case when the nearest neighbor algorithm is called k = 1. The best choice of k depends on the data set; higher values of k decrease the classification effect of noise[24] but make the boundaries less distinct between groups. The algorithm is guaranteed to yield an error rate lower than the Bayes error rate as the infinity approaches the results. [24]. [24] If the k value is small, then it leads to errors in misclassification, so the k value should be high. Different heuristic techniques are then used to pick the right K.

5.1.3 Support Vector Machine

A training algorithm is the support vector machine [SVM]. To predict the class of the new sample, it trains the classifier. SVM is based on the notion of decision planes defining the decision boundary and the point that forms the decision boundary as a parameter between the classes called support vector treatment. SVM is based on the machine learning algorithm that Vapnik developed in the 1960s. To avoid over fitting, it is also based on the structure risk minimization theory. Mathematical programming and the kernel function are two central implementations of the SVM technique.

6. COMPARATIVE STUDY

Some common datasets from the uci repository have been taken in common according to research from different publications and their accuracy is calculated accordingly on each dataset for each machine learning algorithm. Sound, Cancer Breast, Evaluation Vehicle, and Ablone, bcw, DNA Country Honour, Alphabets, Plant Culture, Broad Soybean, and Spamming and Animal data are the datasets used. The datasets have been selected as they are very distinct from each other, ranging from 100 to 1000 data objects. The number of attributes often varies in addition to this, and some of the datasets are multi-attribute ones as well. So choosing this way makes the research all the more informative and trustworthy. The average accuracy for decision tree is 76 percent for NB, KNN is 86 percent and SVM is 88 percent, according to the classification algorithm accuracy performed using MATLAB tool Ablone, Australian, bcw, bio, car & DNA. It is also found that the accuracy of Weka comes first as it was able to run all the algorithms followed by Tanagra and finally KNIME on other datasets using datamining methods such as weka, Tanagra & knime. Finally, after switching from a percentage split to a 10 fold cross-validation approach, Weka obtained the highest efficiency assessment. After that comes KNIME, followed by Tanagra. Thus, SVM and KNN range from 68% to 97% and 34% to 99%, respectively. Showing tests of precision for Tanagra. It is not possible to incorporate 1R and 0R in it. For certain datasets, SVM and KNN do not have readings, but for the one they provide, it ranges from 90 percent to 97 percent and 26 percent to 98 percent. Naive Bayes ranged from 60% to 96%. The range of C4.5 is 58 to 97 per cent.

in the analysis of exploratory data. Such decision tree and rule set induction resulting in classification models, association rule learning that is an unsupervised learning method without class label, similarly clustering is also an unsupervised learning method. Although subgroup discovery is a descriptive induction method for pattern learning aimed at finding descriptions of interesting population subgroup.

3.2.1 Association Rule Learning

In the data mining culture, the topic of inducing association rules (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1995) has gained a lot of priority. It is defined as follows: given the set of transactions (examples) where each transaction is a set of items, the rule of association is an expression where the sets of items are B and H, and B! H is interpreted as IF B THEN H, implying that the transactions containing B in a database also appear to contain H. An in-depth survey of the discovery of association law is beyond the scope.

3.2.2 Subgroup Discovery

The task of subgroup discovery is to identify sufficiently large population subgroups that have a class distribution that is substantially different from the whole dataset. Discovery by subgroup results in individual rules, where the conclusion of the law is a class (the property of interest). The key difference between the learning of classification rules and the discovery of subgroups is that it causes single interest rules (subgroups) aimed at revealing interesting characteristics of groups of instances, not necessarily at establishing a classification law.

4. TOOLS USED FOR DATAMINING The data mining tools that we have comparisons

The **WEKA** toolkit[14] was developed in New Zealand by the University of Waikato. For study and didactic purposes, it is commonly used. Weka is the app that is easiest to use and therefore has a wide user base.All the algorithms present in it for machine learning and data mining were written in Java. WEKA includes various functions, such as splitting, validation, regression, and mining.

Tanagra [15] is a free tool for analyzing data. It also supports mathematical learning algorithms, in addition to machine learning and data processing. An open source tool for mining is the primary feature of Tanagra. The secondary role is to allow researchers to establish and compare their own rules of classification with pre-existing methods. The tertiary aim is to provide their open source code to developers so that they can learn how the tool was made.

KNIME [16] is a robust instrument for data processing, discovery, and visualization. KNIME was developed using systematic methodology and is used in the science academy by a number of people.

5. MACHINE LEARNING ALGORITHM COMPARISONS

5.1 Machine learning algorithm

5.1.1 Bayesian Network

A Bayesian Network (BN) is a graphical model of relationships between a collection of features of different variable data sets. A Directed Acyclic Graph (DAG) is this graphical model structure S and all the nodes in S correspond to the characteristics of a data set in one-to-one correspondence. Among the characteristics of datasets, the arcs reflect forces, while the lack of possible arcs in S encodes conditional independence. When applied to large datasets, the Bayesian classifier showed high accuracy and speed[18][19] Bayesian networks are used to model knowledge of bioinformatics, engineering, medicine,