EasyChair Preprint

№ 3103

Conceptual study-A review on various machine learning algorithms of datamining

C. B. Lakshmi

April 2, 2020

Conceptual study-A review on various machine learning algorithms of datamining

Lakshmi CB[1]

Asst. professor, Dept of Computer science,

St.francis desales college Bangalore

lakshmicbg@gmail.com

Abstract:

Rising concerns of database industry and resulting market needs for various methods to extract valuable knowledge from large data stores, thus data mining DM and KDD has emerged as a problem solving tool for analyzing data for the databases that are preexisting. This paper reviews on various machine learning algorithm used on various training data set used from UCI repository. Machine learning is categorized as supervised and unsupervised learning, accordingly supervised learning is obtained from various classified concepts i.e. classifier for new instance. Unsupervised learning concerns on various unclassified class. Predictive datamining is also called as supervised learning and descriptive datamining is unsupervised based on various association rules. Machine learning and Datamining approach focuses on analysis of categorical, on-numeric data and on the interpretable data. Cross industry standard process for datamining CRISP-DM a mining techniques involved for business solutions based on KDD. A review is done from various research papers on datamining tools and algorithms and its effect on supervised learning for fruitful decisions on data.

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

1. INTRODUCTION

Data is huge thus it's beyond the comprehending power of humans to make an efficient knowledge discovery. The main goal of datamining is to extract useful information in human understandable format from large databases. Datamining we can say as an intersection of various fields like machine learning, artificial intelligence etc. Datamining applications are vast in areas like gaming engineering, biological, analytics and visualization. There various datamining tools available in market like R tool, Rapid miner, keel, weka, orange etc. Datamining approach like classification, clustering and regression approaches are being used for knowledge discovery and future plan.

C lustering has three approaches in which instances are grouped into identified classes. Clustering approach is based on unsupervised learning because there are no predefined classes. In this approach data may be grouped together as a cluster. Classification is a popular task in data mining especially in knowledge discovery and future plan, it provides the intelligent decision making, classification is not only used to study and examine the existing sample data but also predicts the future behavior to that sample data. The classification includes two phases, first is learning process phase in which the training data is analyzed, then the rules and patterns are created. The second phase tests the data and archives the accuracy of classification patterns. Regression is used to map data item into a really valuable prediction

variable. In Classification technique various algorithms such as decision tree, nearest neighbor, genetic algorithm support vector machine (SVM) etc. In this paper, we examine the various classification algorithms and compare them. In the rest of this paper, we first give Decision Tree Concepts, Bayesian Network, and K-Nearest Neighbor Support Vector Machine.

2. LITERATURE REVIEW

A lot of research have been analyzed on comparison of various machine learning algorithm and its implications on various training datasets. In study conducted in 2004 say cross validation on breast cancer data sets elicited in bin Othman [1] conducted research using weka tool to perform various analysis using K-fold cross validation method but in depth research is not processed. Authors like. ImasSukaesihSitanggang et al [2] proposed a new spatial decision tree algorithm .Accordingly which a data set is classified in to two layers spatial and referential relationship among various tuples in relational datamining thus its accuracy is termed to be 74.72%. Khatwani, S. Et al [13] proposed Id3 algorithm and genetic algorithm to create multiple decision tree each of predicts the performance based on the feature of data sets. In 2018 [4,5] did research on comparison of various training datasets using datamining tools like weka ,tanagra have compared accuracy of algorithms like KNN, SVM ,Naïve bayee,C4.5,IR and OR algorithms thus accuracy is achieved using functionalities of regression and fuzzy learning but no association rules are implied. Qin et al [7] proposed novel Bayesian classification technique is based on the uncertain data by taking 20 data sets from UCI repository and apply uncertain Bayesian classification and prediction technique, proposed algorithm in Weka and show that the result of the proposed approach is better than the Bayesian Classification. David Tania [5] presents absolute taxonomy of Nearest Neighbor Queries in spatial databases,. The taxonomy comprises four perspectives: space, the result, query-point, and relationship. In 2013 [5] did research on using weka providing its advantages compared to other tools dealing with medical datasets. [6] Main focus of the survey of weka tool. Different classification techniques have been comparatively analyzed in [7, 11, 12]. Jain [8] has focused on C4.5 decision tree and compared its working in the various mining tools. [14–16] give a detailed explanation on the use of weka, tanagra and knime respectively.

3. REVIEW COMPARISONS

3.1 Supervised algorithm: Predictive data mining Induction of Models

Research mainly emphasis on predictive induction which the result of classification method of datamining i.e. decision tree induction and rule set induction.

3.1.1 Decision Tree Induction

*Decision tre*e classification model, has nodes and arcs where each node is labeled by attribute name and arc is a valid value of the attribute associated to the node. The top most node is a root node and sub nodes or arcs originated are leaf nodes where traversal is done in top-down manner. Decision tree is used to construct predictive accuracy on training data set on new instances. According to the review done from various literature study training data set is obtained from UCI repository. The crucial step of decision tree is to identify the attribute used for node selection by measuring the purity of the node.

3.1.2 C4.5 and ID3 Decision Tree Algorithm

In C 4.5 decision tree algorithm, (quinlan, 1986), it uses information-theoretic entropy as a purity measure,

Thus identifying attribute with largest utility i.e. the difference between original purity value and the sum of the purity value of the successor nodes weighed by its relative size of the nodes. ID3 Dichotomer 3. It is an older decision tree algorithm introduced by Quinlan Ross in 1986 [9]. The basic concept is to make a decision tree by using the top-down greedy approach.

To construct a decision tree by simply picking the next available attribute instead of most informative attribute. As a result recursive portioning can be done at every step of the data using top down tree construction process, thus reducing each node steadily. Thus, the reliability of the chosen attributes decreases with increasing depths of the tree. As a result, overly complex models are generated, which explain the training data but do not generalize well to unseen data. This is known as *overfitting*. As we *prune the branches* and nodes near the leaves, thus resulting in replacing some of the interior nodes of the tree with a new leaf, thereby removing the subtree that was rooted at this node. It is important to note that the leaf nodes of the new tree are no longer pure nodes, containing only data that belong to same class labeling the leaf; instead the leaf will bear the label of the most frequent class at the leaf. Many decision tree induction algorithms exist, the most popular being C4.5 and its variants: a commercial product SEE5, and J48, which is available in the WEKA workbench (Witten & Frank, 2005), as open source.

3.1.3 Rule Set Induction

The next important machine learning technique is the induction of rule sets. The Learning of rule-based models has been a main research goal in the field of machine learning since its beginning in the early 1960s. The rule-based induction techniques have also received increased attention in the statistical community (Friedman & Fisher,1999). Rule based classification model consists of set of if –then rules where every rule is a conjunction of attribute values or also called as features, where values are in the conditional part and consequent class. *probabilistic rules* are induced; in addition to the predicted class label, the consequent of these rules consists also of a list of probabilities or numbers of covered training instances for each possible class label (Clark & Boswell, 1991).Rule sets are typically simpler and more comprehensible than decision trees. Rules can also be implied using if-then rule where the first condition always uses the same attribute, namely, the one used at the root of the tree. Next to each rule, we show the proportion of covered examples for each class value. The main difference between the rules generated by a decision tree and the rules generated by a rule learning algorithm is that the former rule set consists of no overlapping rules, pruning is a good idea for rule learning, in which the rules only need to cover examples that are *mostly* from the same class. It turns out to be advantageous to prune rules after been learned, that is before successive rules are learned (F¨urnkranz, 1997).

3.1.4 Rule Sets versus Decision Trees

There are several aspects which make rule learning attractive. Decision trees are often quite complex and hard to understand. Quinlan (1993) has stated that even pruned decision trees may be too cumbersome, complex, and inscrutable to provide insight into the domain at hand and has consequently devised procedures for simplifying decision trees into pruned production rule sets (Quinlan, 1987a, 1993).Accordingly Rivest (1987) states, showing that decision lists (ordered rule sets) with at most k conditions per rule are strictly more expressive than decision trees of depth k. A similar result has been proved by Bostr¨om (1995), thus restriction of decision tree learning algorithms to no overlapping rules.

3.2 Descriptive Data Mining: Induction of Patterns

Decision tree and its set of rules are used for classification or prediction problem, Instead of model construction, the goal may be the discovery of individual patterns/rules of each data set thus describing regularities in the data. This form of data analysis is referred to as *descriptive induction* and is frequently used in exploratory data analysis. Those decision tree and rule set induction that result in classification models, association rule learning which is an unsupervised learning method without class label, similarly clustering is also an unsupervised learning method .while *subgroup discovery*—aimed at finding descriptions of interesting population subgroup it is a descriptive induction method for pattern learning.

3.2.1 Association Rule Learning

The problem of inducing *association rules* (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1995) has received much priority in the data mining community. It is defined as follows: given a set of transactions (examples), where each transaction is a set of items, an association rule is an expression, where B and H are sets of items, and B! H is interpreted as IF B THEN H, meaning that the transactions in a database which contain B tend to contain H as well. An in-depth survey of association rule discovery is beyond the scope.

3.2.2 Subgroup Discovery

In subgroup discovery the task is to find sufficiently large population subgroups that have a significantly different class distribution than the entire dataset. Subgroup discovery results in individual rules, where the rule conclusion is a class (the property of interest). The main difference between learning of classification rules and subgroup discovery is that it induces single rules (subgroups) of interest, which aim at revealing interesting properties of groups of instances, not necessarily at forming a rule set used for classification.

4. TOOLS USED FOR DATAMINING

The data mining tools that we have comparisons

• **WEKA toolkit** [14] is developed by the University of Waikato which is situated in New Zealand. It is widely used for research and didactic purposes. Weka is the easiest to use software and thus has a large user base.
All the machine learning and data mining algorithms present in it have been written
in Java. WEKA contains various functionalities like splitting, validation, regression and mining.

• **Tanagra** [15] is a free data analysis tool. In addition to machine learning and data analysis, it also supports statistical learning algorithms. The primary function of Tanagra is an open source tool for mining. The secondary function is to allow researchers to make their own classification rules and compare them with pre-existing methods. The tertiary purpose is to provide developers their open source code so that they could learn how the tool was made.

• **KNIME** [16] is a comprehensive tool for analysis, exploration, and visualization of data. KNIME has been made using rigorous techniques and is used by a lot of people in the research academia. It is a modular tool that could be used to make data flows by connecting various functions, selectively run a part of it and then

5. MACHINE LEARNING ALGORITHM COMPARISONS

5.1 Machine learning algorithm

5.1.1 Bayesian Network

A Bayesian Network (BN) is a graphical model for relationships among a set of various variable of data sets with features. This graphical model structure S is a directed acyclic graph (DAG) and all the nodes in S are in one-to-one correspondence with the features of a data set. The arcs represent influences among the features among datasets, while they lack of possible arcs in S encodes conditional independence. Bayesian classifier has exhibited high accuracy and speed when applied to large databases [18] [19] Bayesian networks are used for modeling knowledge Bioinformatics, engineering, medicine, Biomonitoring, Semantic search image processing.

5.1.2. K-Nearest Neighbor

The K-Nearest Neighbor (NN) is the simplest method of machine learning and it has some strong consistency results. KNN the object are classified based on the nearest training example in the feature space. Thus implicitly computes the decision boundary and computation of the decision explicitly. So the computational complexity of NN is the function of the boundary complexity [21]. The neighbors are selected from a set of objects for which correct classification is known. No explicit training step is required this can be thought of as the training set to the algorithm.

The k-NN algorithm is sensitive to the local structure of the data set. This is the special case when k = 1 is called the nearest neighbor algorithm. The best choice of k depends upon the data set; higher values of k diminish the effect of noise on the classification [24] but make boundaries between classes less distinct. As the infinity approaches to data, the algorithm is guaranteed to yield an error rate less than the Bayes error rate. [24]. If the value of k is small then it leads to misclassification errors therefore k value should be large. Thus various heuristic techniques are used to select the good K.

5.1.3 Support Vector Machine

The support vector machine [SVM] is a training algorithm. It trains the classifier to predict the class of the new sample. SVM is based on the concept of decision planes that defined decision boundary and point that form the decision boundary between the classes called support vector treat as parameter. SVM is based on the machine learning algorithm invented by vapnik in 1960‘s. It is also based on the structure risk minimization principle to prevent over fitting. There are 2 key implementations of SVM technique that are mathematical programming and kernel function.

6. COMPARITIVE STUDY

According to study from various journal few common datasets have been taken in common from uci repository and its accuracy is measured accordingly on each datasets for each machine learning algorithm. The datasets used are Sound, Cancer Breast, Evaluation Car, and Ablone, bcw, DNA Country Honour, Alphabets, Plant Culture, Soybean Large, and Spamming and Animal data. The datasets have been so chosen as they are very different from each other where the data objects range from 100 to 1000. In addition to this, the number of attributes vary as well and also some of the datasets are multi-attribute ones. So choosing like this make the analysis all the more comprehensive and credible.

According to classification algorithm accuracy performed using MATLAB tool Ablone, Australian, bcw, bio, car & DNA average accuracy for decision tree is 76% respectively for NB is 57%, KNN is 86% and SVM is 88%. Also other accuracy classification done on other datasets using datamining tools like weka, Tanagra & knime it is found that accuracy of Weka comes first as it was able to run all the algorithms followed by Tanagra and finally KNIME. At last, Weka has achieved the highest performance measure when moving from percentage split to 10 fold cross-validation approach. KNIME comes after it followed by Tanagra. Thus SVM and KNN range from 68%–97% and 34%–99% respectively. Showing accuracy measures for Tanagra. 1R and 0R cannot be implemented in it. SVM and KNN do not give readings for some datasets but for the one they give, it ranges from 90% to 97% and 26% to 98% respectively. Naive Bayes ranged from 60%–96%. 58% to 97% is the range of C4.5.

8. CONCLUSION & FUTURE RESEARCH

The above comparative study on datamining and machine learning algorithms using various mining tools and combine machine learning algorithm accuracy of various algorithms only classification or predictive learning was implemented. Accordingly KNN and SVM machine learning algorithm is stated as efficient method of measuring accuracy on various data sets also tool like weka shows the efficiency compared to other tools. As a future research, other methods like association rules, ensemble learning, regression or clustering could be done to get deeper insights on these tools and to be able to know in much detail about which of the following algorithms will be applicable than the other. In the future, intensive development and increased usage of datamining in specific domain areas, such as bioinformatics, multimedia, text and web data analysis. We also involve a shift of datamining towards automation in future in choosing relevant tool and algorithms to obtain performance accuracy. we can also improve the accuracy of performance in other applications of datamining

References

Balagatabi, Z. N. (n.d.). Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive Styles 2013 IEEE International Conference on Computational Intelligence and Computing Research in Academic Environment. Indian Journal of Automation and Artificial Int.

Han, J. &. (. (2001).). Data mining: Concepts and techniques. China Machine Press.

. Han, J. K. (n.d.). *(2006). Data mining: concepts and techniques. Morgan kaufmann.*

. Khatwani, S. &. (n.d.). a novel framework for envisaging a learner's. *IEEE,2012*.

. Quinlan, J. R. ((1986).). Induction of decisiontrees. Machine learning, .

Akarsh Goyal(✉), I. K. (2016). *A Comparative Analysis of the Different Data Mining Tools* (Vol. Proceedings of the Eighth International Conference on Soft computing). vellore, tamil nadu, india: Springer International Publishing AG 2018.

Bakar, A. A. (2009). *Building a new taxonomy for data discretization techniques.* IEEE.

J. F¨urnkranz et al. (2012). *Machine Learning and Data Mining* (Vols. DOI 10.1007/978-3-540-75197-7 1,). Foundations of Rule Learning, Cognitive Technologies,.

Merceron, A. &. ((2005, May)). Educational Data Mining: a Case Study.In AIED . 467-474).

Seema Sharma1, J. A. (2013). *Machine Learning Techniques for Data Mining: A Survey* . madhayapradesh.

Taniar, D. &. (2013). A taxonomy for nearest neighbour queries in spatial databases. Journal of Computer and System Sciences.

WEKA, the University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/. (n.d.).