REVIEW OF CURRENT DATA MINING TECHNIQUES USED IN THE SOFTWARE EFFORT ESTIMATION

Julius Olufemi Ogunleye[1][0000-0002-1085-9615]

[1] Tomas Bata University in Zlin, Nad stranemi 4511, 760 05 Zlín, Czech Republic
juliusolufemi@yahoo.com

ABSTRACT. Data Mining is a method of finding patterns from vast quantities of data and information. The data sources include databases, data centers, the internet, and other data storage forms; or data that is dynamically streaming into the network. Estimation of effort is very important in the cost estimation of a software development project, and very critical in the software life development cycle planning process. This paper offers a description of the latest data mining techniques used in estimating software effort, and these techniques are divided into two, namely: Classical and Modern, based on when they were developed and when they started to be used in business administration. *The Classical techniques are the ones that have been in use for decades and are still relevant until today, while the Modern ones are the ones that have been introduced recently and have gained wide acceptance in the system.* The Classical techniques are Statistical methods, Nearest Neighbours, Clustering and Regression Analysis, while Neural Networks, Rule Induction Systems and Decision Trees are included in the Modern techniques. This paper offers an overview of these strategies in terms of their features, benefits, drawbacks and use areas.

KEYWORDS: Software effort estimation · Data mining techniques · Regression Analysis · Classification techniques · Clustering techniques · Neural networks · Nearest Neighbours · Decision trees · Rule induction systems.

1 INTRODUCTION

Today's advances paved the way for an automated extraction of hidden predictive information from databases, along with many other fields of knowledge such as analytics, artificial intelligence, machine learning, database management, data visualization and recognition patterns. Through data mining, a person can use different analytical methods, data analysis and machine learning to explore and analyze large data sets, thereby extracting new and useful knowledge that will improve decision-making processes [1].

Companies in information technology are currently using data mining techniques in different fields with the goal of increasing decision-making efficiency and enhancing business results. The amount of data produced and processed is rising exponentially, owing in large part to the continuing developments in computer technology. This provides immense opportunities for those who are able to access the knowledge hidden in this data but also poses new challenges. This paper is a review on the latest methods /

techniques for data mining used in estimating the program effort. The methods/techniques use simple cognitive metrics which include all of the software's important parameters. This will serve as a reference for the professionals involved in the software development process in cost estimation, time schedule, and manpower requirement. It would be prudent, therefore, to use these estimates as additional feedback in the decision-making process. Since this paper is a review of current data mining techniques, the research methodology is based on secondary data.

Techniques of data extraction include data research, data reformatting and data restructuring. The structure of the necessary information is dependent on the methodology and the research to be carried out. Finally, all the techniques, approaches and frameworks of data mining help to explore new innovative technologies.

1.1 DATA MINING

The process of finding trends in large data sets involving approaches at the intersection of machine learning, statistics, and database systems [2] is data mining, also referred to as data or information discovery. It could also be defined as a computer science and statistics interdisciplinary subfield with an overall objective of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use [3]. The functionalities of data mining are used to define the patterns to be found in data mining tasks. Data mining activities can usually be divided into two categories: descriptive, and predictive. Descriptive mining activities are characteristic of the general data resources in the database. Predictive mining tasks conclude on the current data to make predictions [4]. Data mining benefits cover nearly every facet of life that includes; gaming, police, industry, research, engineering, human rights organizations, and surveillance. The best way to gain an understanding of data mining is to consider the types of tasks, or issues, it can solve. The advantages of data mining techniques emphasize their importance in software effort estimation.

Table 1. Advantages and disadvantages of Data Mining Techniques

Advantages	Disadvantages
• Helps to predict future trends • Signifies customer habits • Helps in decision making • Increases company revenue • Depends on market-based analysis • Quickly detects fraud	• Violates user privacy *(Information collected through data mining and intended for ethical purposes can be misused).* • Uses additional irrelevant information • Information can be misused • Accuracy of data is within its own limits

1.2 SOFTWARE DEVELOPMENT EFFORT ESTIMATION

Cost estimation of software development is very important for IT professionals and it's an important task that affects an organization's software investment decisions. It is a job that must be completed before any contract is entered into, or the dedication of the resources committed to any project. Both developers and customers need precise estimates of the cost of software to make far-reaching project decisions. Precise estimates of the cost of software can be used to make proposals and schedule, monitor and control requests. Although several software projects have been developed for accurate cost estimation purposes, it is still difficult to claim that there is a specific model that can offer estimation close to the actual cost. In reality, over-estimated or under-estimated costs can result in the production delay of the final software product, inefficient resource use, poor software project quality or unexpected budget increase. Therefore it is difficult to make correct decisions [5].

2 RELATED WORKS

2.1 ANALYSIS OF DATA MINING TECHNIQUES FOR SOFTWARE EFFORT ESTIMATION

Under Sehra, S. K. et al.(2014), Software effort estimation requires a high degree of precision but, sadly, accurate estimates cannot be easily obtained. The use of data mining to enhance software process efficiency for an enterprise is on the rise. There are several different method combinations available when conducting software effort estimation, but it had become difficult to pick the most appropriate combination. The analysis offered opportunities for how data pre-processing was applied and effort estimation using the COCOMO Model. OLS Regression and K-Means Clustering data mining techniques were subsequently applied on preprocessed data and were correlated with results obtained. Implementing the data mining techniques on pre-processed data was more effective than OLS Regression Technique [6].

2.2 DATA MINING TECHNIQUES FOR SOFTWARE EFFORT ESTIMATION: A COMPARATIVE STUDY

Dejaeger K. et al.(2012) considered that a predictive model had to be reliable and easily understandable in order to inspire trust in a business environment. Although both aspects (accuracy and understanding) were evaluated by previous studies in a software effort estimate environment, no definitive conclusion was drawn as to which technique was the most suitable. This issue was dealt with as a benchmark through reports of the results of a large-scale study. There were various types of techniques under consideration, including tree or rule-based models such as M5 and CART, linear models (linear regression), nonlinear models (MARS, multilayered perceptron neural networks, radial base function networks, and least squares support vector machines), and several other inference techniques that do not directly trigger a model (e.g. case-based reasoning).

Additionally, the function subset selection aspect was investigated using a generic backward input selection wrapper. The results were subjected to rigorous statistical testing which suggested that the best results were actually obtained by ordinary minus square regression in combination with a logarithmic transformation. In addition, another important finding was that a substantial increase in estimation accuracy can be achieved when selecting a subset of highly predictive attributes such as project size, growth, and environment-related attributes [7].

2.3 SOFTWARE TEST EFFORT ESTIMATION

D. S. Kushwaha & Misra A.K. (2008) demonstrated that software testing is an important software development process and is carried out to help and enhance the reliability and consistency of the program. The method involves estimating the test effort, choosing the correct test team, planning test cases, executing the program with the test cases and reviewing the results provided by those executions. Statistics indicate that more than fifty per cent of the software development cost is spent on testing, with the figure being higher for essential software testing. Unless we can predict the testing effort and find effective ways to perform effective testing, there will be a significant increase in the percentage of development costs spent on testing, coupled with the project costing and development schedule mismatch. This paper seeks to establish the Cognitive Information Complexity Measurement (CICM) as an appropriate estimation tool for estimating the test effort [8].

3 METHODS

Data scientists currently use multiple data mining techniques, and these techniques vary from each other based on their accuracy, performance, and the type and/or volume of data available for analysis [9]. These techniques can be classified into the Classical and Modern techniques of data mining.

3.1 CLASSICAL TECHNIQUES

STATISTICAL TECHNIQUES

Statistics is a branch of mathematics related to data collection and explanation. Software mining is not statistics or mathematical techniques. They were being used to apply to business applications long before the term data mining was coined. Nevertheless, the data drive statistical techniques and are used to discover trends and to construct predictive models. And from the user viewpoint, when solving a data mining problem, you'll face a conscious decision as to whether you want to solve it using statistical methods or other data mining techniques.

Statistics can be of great help in answering various important data questions, namely:

1. What trends to my database are there?

2. What is the probability there will be an event?
3. Which patterns are meaningful?
4. What is the high-level overview of the data that gives me some insight into what my database contains?

There are several different aspects of statistics but sometimes the principle of gathering and recording data is at the core of all these more complex techniques. The first step then in understanding statistics is to consider how the data is obtained in a higher-level context-with the histogram one of the most prominent ways to do so. A histogram provides an alternative way to show a quantitative variable distribution. Histograms are of particular interest to vast sets of data. A histogram divides the values of the variable into intervals of equal dimensions. Within each interval we can see the number of individuals [10].

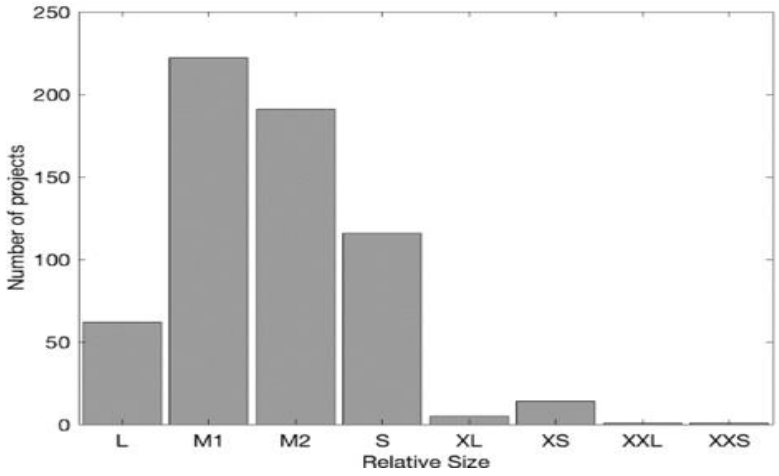

Fig. 1. A histogram showing the relative size of software projects and their frequencies [11].

This histogram shows clearly that the majority of the software projects in selected dataset are projects known as M1 i.e. their sizes are from interval <100, 300> [11]. Some of the overview statistics most widely used include:

- *Max* - maximum value for a predictor given.
- *Min* - the minimum of a given predictor value.
- *Mean* - average of a given predictor value.
- *Median* - the value for a given statistic that splits the database into two sets with equivalent numbers with records as close as possible.
- *Mode* - most common indicator value.
- *Variance* - a calculation of how the average value is spread out.

Numbers can also be used for study of Predictions and Linear Regression.

NEAREST NEIGHBOURS

This is one of the oldest data-processing techniques used. This is a supervised learning technique that is commonly used for predictions, and during a Euclidean space, instances are typically represented as points. The Nearest neighbor is defined in terms of Euclidean distance, e.g. distance(x1, x2), and the target function may be evaluated in a discrete or actual way. Distance-weighted nearest neighbour algorithm weight the contribution of every of the k neighbours consistent with their distance to the query and give greater weight to closer neighbours. Nearest neighboring technique by averaging k-nearest neighbors is robust to noisy results, and could be a predictive technique that is somewhat like clustering. Objects close to each other at least would have similar predictive values, so one object's predictive value can be used to predict that of its nearest neighbor. The nearest neighbor was commonly used in text retrieval for prediction. Space is described by the problem to be solved (supervised learning), and it uses distance metrics generally to evaluate closeness.

CLUSTERING

Clustering is one of the oldest methods used in data mining. It is an example of unsupervised learning, that is, the class labels are not present in the training because it is not understood to start with. It is the grouping of similar/related data points or records from raw, unlabeled data, based on the concept of maximizing object homogeneity in the same group or class and minimizing object heterogeneity in the different groups or classes. Often clustering is used instead of segmentation which gives a general overview of the data set. The cluster analysis output is a collection of groups (clusters) that form a partition or partition structure of the data set, and it is a simplified definition of each cluster that is particularly important for a deeper analysis of the data set's characteristics.

Cluster hierarchy is typically seen as a tree where the smallest clusters merge to create the next higher level of clusters, and those at that level merge to create the next higher level of clusters. There are two major types of strategies for clustering, namely: hierarchical and nonhierarchical.

- The hierarchical clustering techniques establish cluster hierarchy from the smallest to the largest. Hierarchical clusters are specified purely by data (not by the users predetermining the number of clusters), and by simply going up and down the hierarchy, the number of clusters may be increased or decreased.
- There are two major non-hierarchical clustering methods, both of which can be measured very easily on the database but have some disadvantages. The first is the single pass method which derives its name from the fact that in order to construct the clusters, the database must only be passed through once (i.e. each record is only read once from the database). The other class of techniques is called methods of relocation that derive their name from the movement or "reallocation" of records from one cluster to another to establish better clusters. The reallocation technique is faster than the hierarchical technique and uses multiple passes through the database [12].

- In clustering, space is either defined as the default n-dimensional space, or is defined by the user, or a predefined space driven by previous experience (unsupervised learning) and it can use other metrics besides distance to evaluate the closeness of two records, e.g. to link two points together.

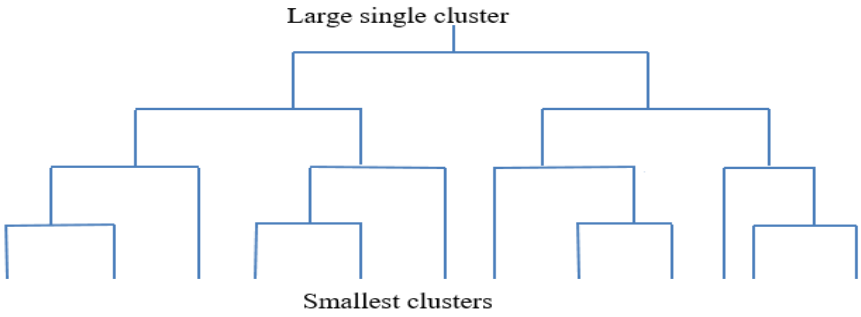

Fig. 2. Diagrammatic representation of the hierarchy

REGRESSION ANALYSIS

This form of analysis is supervised and determines which item sets are linked to or separate from each other among the different relationships. It can predict human actions, sales, income, temperature, etc. It has an already known data set value. When an input is given, the input and expected value will be compared with the regression algorithm, and the error will be determined to get to the exact result. Regression is generally synonymous with regression of some sort in the statistics. The purpose of the regression analysis is to find the best model which can relate the output variable to different input variables. This analysis is the method of evaluating the relationship between a variable Y and one or more other variables: X1, X2, to Xn. The dependent variable (or response output) is Y, while X1 to Xn are the independent variables or inputs. This technique is used to establish the dependence between the two variables so that causal relationship can be used to predict the outcome, and it helps to know the dependent variable's characteristic value. Regression is commonly used for forecasting and prediction. A model is generated in regression which maps values from predictors in such a way that the lowest error occurs when making a prediction. Simple linear regression is the simplest type of regression which contains just one predictor and one prediction [12].

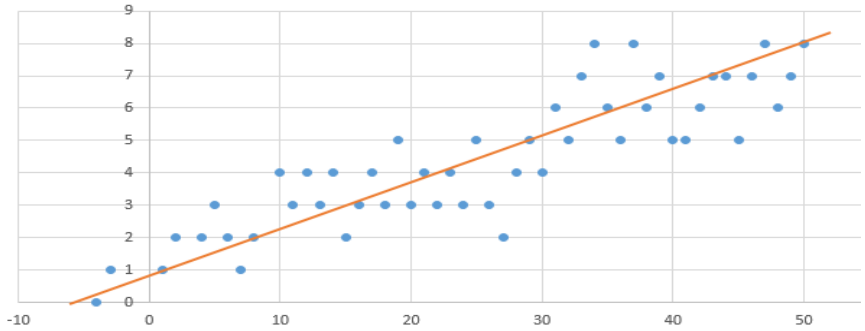

Fig. 3. Illustration example of linear Regression on a set of data

3.2 MODERN TECHNIQUES

NEURAL NETWORK

This is an effective predictive modeling technique but some of the power comes at the cost of user-friendliness and ease of use. It is an unsupervised learning method and during the formative stages of data mining technology has possibly been of greater interest than Decision trees. It produces very complex structures that are almost always difficult for even experts to completely understand. In a complex calculation, the model itself is defined by numeric values that allow all of the predictor values to be in the form of a number. Neural network performance is always numerical and must be translated if the real predictive value is categorical [13].

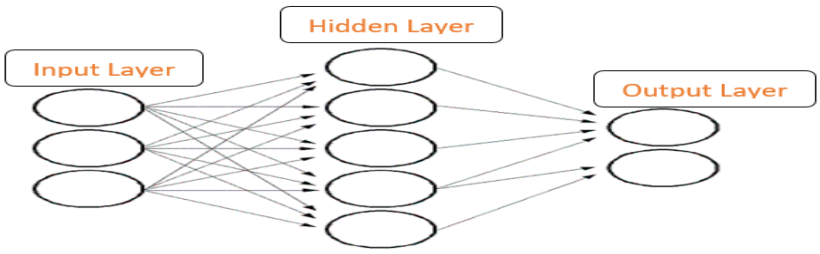

Fig. 4. Example of a neural network

RULE INDUCTION

Rule induction is a field of machine learning, in which a series of observations extract formal laws. The rules extracted can represent a complete data science model, or merely reflect local trends in the data. Rule induction generates rules that are not mutually exclusive and may be necessarily exhaustive, and produces a model based on If – Then – Else style rules. It can function with numerical values as well as categorical values

and the models have a number of input variables and one or more output variables, but are different from the neural networks in that we can actually see within the model and how it generates the output or outcome. The models and laws are typically built from decision trees in the rule induction data models. Rule induction is the most common form of discovery of knowledge in highly automated unsupervised learning systems, and is possibly the best form of data mining techniques for discovering all possible predictive patterns in a database. This can be modified for use in problems of prediction but the algorithms for integrating evidence from a variety of rules come more from thumb rules and practical experience.

DECISION TREES

This technique is used for categorizing or predicting data, and it produces rules that are mutually exclusive and collectively exhaustive with respect to the training database. The root of a decision tree is a simple question but have multiple answers. A decision tree is a predictive model that, as its name implies, can be viewed as a tree and each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision tree algorithms tend to automate the entire process of hypothesis generation and validation much more completely, and in a much more integrated way than any other data mining technique. They are also particularly adept at handling raw data with little or no pre-processing. Perhaps also because they were originally developed to mimic the way an analyst interactively performs data mining, they provide a simple to understand predictive model based on rules. They can be used in a wide variety of business problems for both exploration and prediction. Due to their tree structure and ability to easily generate rules, decision trees are the favored technique for building understandable models. As a result of this clarity Decision trees also allow for more complex profit and ROI (Return-On-Investment) models to be added easily on top of the predictive model. Because of their high level of automation and the ease of translating decision tree models into SQL for deployment in relational databases, the technology has also proven to be easy to integrate with existing IT processes, requiring little preprocessing and cleansing of the data, or extraction of a special purpose file specifically for data mining [13].

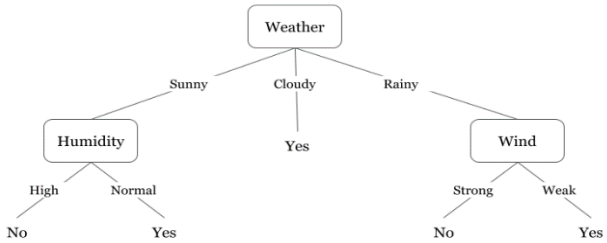

Fig. 5. Example of a decision tree

4 DISCUSSION

With an immense amount of data being collected every day, the businesses are now involved in figuring out the patterns from them. The methods for extracting data help turn the raw data into usable information. Computer software is needed to mine massive volumes of data, because it is difficult for a person to go through the vast volume of data manually. Below are illustrations on the areas of strengths and weaknesses of the traditional and modern data mining techniques:

4.1 Statistical Techniques

Advantages

- Because secondary data is usually cheap and it takes less time because someone else has compiled it.
- The trends and similarities are evident and consistent.
- Taken from large samples, to ensure high generalizability.
- Can be used many times to check different variables.
- Changes which improve reliability and representativeness can be imitated to test.

Disadvantages

- The researcher cannot test the validity and cannot consider a causation theory mechanism only to draw trends and associations from the data
- Statistical data are often secondary data, meaning that misinterpretation is simple.
- Statistical data is subject to manipulation and can be distorted and phrased to make the argument the researcher wishes to prove (effects objectivity).
- Since these are often secondary data, it is difficult to access and verify.

4.2 Nearest Neighbours

Advantages

- Pretty easy and intuitive
- Does not have any assumptions
- No Education Move
- It grows continuously
- Multi class problem very easy to implement.
- Can be used for Regression and Classification.
- It might take some time to select the first hyper parameter but the rest of the parameters are aligned with it after that.
- Has different distance parameters (Euclidean distance, Hamming distance, Manhattan distance, Minkowski distance) to choose from.

Disadvantages

- Irrelevant attributes could dominate the distance between neighbors.
- Implementation might be very easy but efficiency (or speed of algorithm) declines very fast with the growth of the dataset.
- Can accommodate small number of input variables but as the number of variables grows, the algorithm finds it difficult to predict the output of new data point.
- Features need to be homogeneous
- In the classification of new data entry, there is usually the problem of choosing an optimal number of neighbours to be considered.
- Problems with the use of imbalanced data.
- Because neighbours are simply chosen based on distance criteria, it is sensitive to outliers.
- Cannot deal with missing value problem because it inherently has no capability to do so.

4.3 Clustering

Advantages

- Hierarchical approaches allow the end user to choose between multiple clusters or only a few.
- Suitable for arbitrarily shaped data set and arbitrarily type attribute.
- The hierarchical relationship between clusters is easy to identify, with typically fairly high scalability.
- Various and well-developed models that provide a means to adequately describe the data and each model has its own special characters that may bring significant advantages in certain specific areas.

Disadvantages

- Usually fairly high in length of time.
- Cluster numbers must be preset.
- The assumption is not entirely right and the clustering outcome is sensitive to the parameters of the chosen models.

4.4 Regression Analysis *(MARS- Multivariate Analysis for Regression Splines, OLS - Ordinary Least Square regression, SVR-Support Vector Regression, Radial Basis Function Networks)*

Advantages

- Some very easy problems can be solved much faster and simpler by linear regression, where prediction is just an easy multiple of the predictors.
- Linear regression: the speed of modeling is high, does not require very complicated calculations and runs quickly when the data is big.

- Linear regression: the understanding and interpretation of each variable may be given by the factor.
- Linear Regression works well in linearly separable datasets.
- Linear regression is simpler to implement, easier to analyze and more efficient to practice.
- Dimensionality reduction, regularization (L1 and L2) and cross-validation techniques can easily be used to avoid over-fitting in linear regression.
- Multiple regression is capable of evaluating the relative contribution of one or more predictor variables to the value of the criterion.
- Multiple regression is capable of finding outliers, or anomalies.

Disadvantages

- Linear regression: Linear relationship is minimal.
- Linear regression: The outliers are quickly influenced.
- Regression solution is likely to be dense (because there is no regularization)
- Linear regression is noise- and overfitting-prone.
- Regression solutions achieved using various methods (e.g. optimization, less-square, decomposition of QR, etc.) are not inherently special."
- Vulnerable to multicollinearity: Multicollinearity (using dimensionality reduction techniques) should be eliminated before implementing linear regression since it implies there is no relationship between independent variables.
- Any drawback of using a multiple regression model is typically due to the data being used, either by using insufficient data or by misleadingly assuming that a correlation is the cause.

4.5 Neural Networks

Advantages

- Artificial Neural Networks(ANN) are capable of studying and modeling non-linear, complex relationships
- Has predictive models that are highly accurate and can be applied across a large number of different types of problems.
- Stores information on the network as a whole, not on a database, and the absence of a few pieces of information at one location does not prevent the network from functioning.
- Ability to work with inadequate information.
- Has fault tolerance i.e. contamination of one or more ANN cells does not prevent production generation.
- Has a given memory.
- Gradual corruption: A network slows over time and becomes increasingly compromised. The problem with the network doesn't corrode right away.

- Ability to train machines: Artificial neural networks learn events by commenting on similar events and making decisions.
- Parallel processing capability: Artificial neural networks have computational power capable of doing more than one job simultaneously.

Disadvantages

- Limiting usability and ease of deployment.
- Extraction of features-question of deciding which predictors are the most appropriate and the most important in building models that are predictably accurate. The predictors may be used on their own, or they may be used to shape function in conjunction with other predictors.
- Dependence on hardware: Artificial neural networks need parallel processing processors, by their nature. It is for this reason that the equipment realization is based.
- Assurance of proper network structure: There is no clear law for artificial neural network construction. The proper configuration of the network is accomplished by practice and trial and error.
- Unexplained network functioning: If ANN offers a sampling solution, it does not provide any hint as to why and how.
- Difficulty showing the network the problem: ANNs can work with the numerical details. Problems need to be converted into numerical values before integration into ANN.
- The duration of the network is unknown: Reducing the network to a certain value of the sample error means the training was completed. This interest does not produce optimal performance for us.

4.6 Rule Induction

Advantages

- IF-THEN rules are simple to understand and are supposed to be the most interpretable model, especially when dealing with a small amount of rules.
- The decision rules are as descriptive as the decision trees, while being more compact.
- IF-THEN rules are easy to predict, as only certain conditional statements have to be tested to decide which rules apply.
- Decision rules are resilient to monotonous input function transformations, as conditions change only at the threshold.
- IF-THEN rules create models that are typically sparse, meaning there are not many features. They select only the features pertinent to the model.
- Simple rules such as OneR can be used as benchmarks for more complicated algorithms.

Disadvantages

- IF-THEN rules focus on grouping, and ignore regression almost entirely.
- Functions must be categorical, too. This means this if you want to use them, numerical features must be categorized.
- Most of the older algorithms for rule-learning are susceptible to overfitting.
- Decision rules are poor in the analysis of linear feature-output relations.

4.7 Decision Trees *(CART – Classification and Regression Trees)*

Advantages

- Ideal to catch interactions in the data between apps.
- Data ends up in distinct groups which are often easier to understand than points on a multidimensional hyperplane as in linear regression.
- The tree structure, with its nodes and edges, also has a natural visualization.
- Usually the models to be constructed and the interactions to be detected are much more complex in real-world problems.
- CART immediately validates the Tree, i.e. the algorithm has the validation of the model and the discovery of the optimally general model (the algorithm) developed deep inside it.
- The CART algorithm is relatively sturdy in relation to missing data.
- Decision trees label so strongly on so many important data mining features.

Disadvantages

- Is not going to do well with some very simple problems where prediction is just a simple multiple of predictors.
- Trees do not handle the linear relationships. Any linear input-outcome relationship must be approximated by splits, creating a step function. This is not successful.
- Felt smooth. Slight changes in the input feature can have a major impact on the forecast result, which is usually not desirable.
- The trees are pretty unstable too. A few changes in the training dataset can build a different tree altogether. This is because any split is based on splitting the parent.

These techniques have specific tasks where they can be best applied in order to produce optimal results. The following table shows the data mining tasks with the appropriate data mining techniques to accomplish them:

Table 2. Data mining tasks with the appropriate data mining techniques

No	Data Mining Task	Data Mining Techniques
1	Classification	Decision trees, Neural networks, K-nearest neighbors, Rule induction methods, SVM-Support vector machine, CBR-Case based reasoning
2	Prediction	Neural networks, K-nearest neighbors, Regression Analysis
3	Dependency Analysis	Correlation analysis, Regression Analysis, Association rules, Bayesian networks, Rule Induction
4	Data description and summarization	Statistical techniques, OLAP (Online Analytical Processing)
5	Segmentation or clustering	Clustering techniques, Neural Networks
6	Consolidation	Nearest neighbours, Clustering

The table below shows the data mining techniques and their areas of use:

Table 3. Data mining techniques and their areas use.

Data Mining Techniques	Areas of use
Association Analysis	Designing store shelves, marketing, cross-selling of products.
Classification (K-nearest neighbor, etc.)	Banks, marketing campaign designs by organizations.
Decision Trees	Medicine discovery and prediction, engineering, manufacturing, astronomy etc. They were used for problems ranging from estimation of credit card depletion to estimation of time series of the exchange rate of various international currencies.
Clustering Analysis	Image recognition, web search, and security.
Outlier Detection	Detection of credit card fraud risks, novelty detection, etc.
Regression Analysis (K-nearest neighbor, …)	Marketing and Product Development Efforts comparison.

Artificial Neural networks	Data compression, feature extraction, clustering, prototype formation, function approximation or regression analysis (including prediction time series, fitness approximation and modeling), classification (including pattern and sequence recognition, novelty detection and sequential decision making), data processing (including filtering, clustering, blind source separation and compression), and robotic compression.
Support vector machines regression	Oil and gas industry, classification of images and text and hypertext categorization.
Multivariate Regression algorithm	Retail sector
Linear Regression	Financial portfolio prediction, salary forecasting, real estate predictions and in traffic estimated time of arrivals (ETAs).

5 CONCLUSION

Defining which methodology to use, and when, is obviously one of the hardest things to do when deciding to implement a data mining method. How are neural networks acceptable and when are the decision trees suitable? How is data mining appropriate to work with relational databases and reporting? How would it be acceptable to use only OLAP and a multidimensional database? Trial and error determine some of the criteria which are important in determining the technique to be used. There are clear variations in the types of problems that are most conducive to each approach, but the nature of data from the real world and the complex way in which markets, consumers and thus the data representing them are created means that the data is constantly evolving. Therefore, there is no clear law which recommends a particular technique over another. Sometimes, decisions are made based on the availability of experienced data mining analysts in one technique or the other. The preference of some classical techniques over the newer techniques depends more on getting good resources and good analysts.

ACKNOWLEDGMENT
This work was supported by the Faculty of Applied Informatics, Tomas Bata University in Zlín, under Projects IGA/CebiaTech/2020/001 and RVO / FAI / 2020/002.

REFERENCES

1. *Kamber H. Et al. (2011): Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.*
2. *ACM SIGKDD (2006-04-30),* Retrieved (2014-01-27): Data Mining Curriculum

3. *Clifton C. (2010): Encyclopædia Britannica: Definition of Data Mining"*. Retrieved 2010-12-09.
4. *Trevor H. Et al.(2009):* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Archived from the original on 2009-11-10. Retrieved 2012-08-07.*
5. Jiawei H. and Micheline K. (2000): *Data Mining: Concepts and Techniques*
6. *Sehra S.K. et al. (2014):* Analysis of Data Mining techniques for software effort estimation.
7. *Dejaeger K., et al.(2012):* Data Mining Techniques for Software Effort Estimation: A Comparative Study.
8. *Weiss G. M. and Davison B. D. (2010):* Data Mining (Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010).
9. *Berson A. et.al (2005):* An Overview of Data Mining Techniques (Excerpts from the book by Alex Berson, Stephen Smith, and Kurt Thearling).
10. *Mehmed K. (2011):* Data Mining Concepts, Models, Methods, and Algorithms (Second Edition).
11. *Silhavy, P., Silhavy, R., & Prokopova, Z. (2019): Categorical variable segmentation model for software development effort estimation. IEEE Access, 7, 9618-9626.*
12. *Software Testing Help (April 16, 2020):* Data Mining Techniques: Algorithm, Methods & Top Data Mining Tools
13. *Kushwaha D. S. and Misra A. K. (2008):* Sofware Test Effort Estimation.