

A Literature Survey on Classification Algorithms of Machine Learning

Priyanka Verma

Department of Computer Science & Engineering
Sagar Institute of science & Technology Bhopal
(M.P), India

Rajeev Kumar Gupta

Department of Computer Science & Engineering
Sagar Institute of science & Technology Bhopal
(M.P), India

ABSTRACT

The spreading amount of data usually generates interesting demand for the data analysis tools that spot regularities in these data. Data mining has turned up as great domain that contributes mechanism for data analysis, to find out the hidden knowledge, and self-ruling decision making in many operation domains. Supervised machine learning is using to find out the search for algorithms that reason from clearly supplied instances to produce general interpretation, which then makes predictions about future scenario or events. In other words, the goal of supervised learning is to make a small model of the distribution of class labels (distribution or classification) in terms of finding (predictor) features. The resulting classifier is then used to assign class labels (attributes) to the testing instances where the values of the predictor (attributes or properties) features are known, but the value of the class label is unknown. This paper explains various supervised machine learning classification techniques.

In this paper, we have discussed the about the classification algorithm which are available today, how they works, and what are their advantages and disadvantages. The algorithms which we will discuss are Naïve Bayes, SVM, random forest, decision tree and logistic regression.

General Terms

Survey about various classification algorithm used in machine learning algorithm

Keywords

Classification, Naïve Bayes, Random forest, Multiple regression dependent variable, independent variables, predictor variable, response variable

1. INTRODUCTION

Social Predictive modeling can be explained in terms of mathematics to find out the goal there is relationship between a target, response, or “dependent” variable and various predictor (attributes) or “independent” variables with the goal in mind of measuring future (attributes) values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable, it is mandatory to give some measure of mistrust for the predictions, typically a prediction interval that gives some assigned level of confidence like some percentage value i.e (95%). Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors. Regression, as a data mining technique, is supervised learning. Supervised learning partitions the database into training and validation data. The techniques used in this research were simple linear regression and multiple linear regressions. Some divergence between the uses of regression in statistics verses data mining are: in statistics the data is a sample from a (Data Storage) population, but in Data Mining

the data is taken from a large database (e.g. 1 million records). Also in stats the regression model is created from a given sample, but in Data Mining the regression model is created from a part of the data (training data). Predictive (To guess) analytics enclose a number of mechanism from stats, data mining and game theory that find out current and historical facts to make guesses about future events. The variety of techniques is sometimes divided in three ways: predictive models, descriptive models and decision models.

Predictive models explain for sure relationships and some patterns that usually edge to a certain behavior, point to fraud, predict system failures, and so many. By explaining the explanatory variables, we can find out or predict results in the dependent variables.

Descriptive models explain for creating partition or segment; generally it is used to classify (find out) customers based on for instance (behavior of customers in different locations) socio-demographic characteristics, life cycle, profits, required product and many more. Where predictive models focus on a specific (individual) event or behavior, descriptive models identify as many different (general) relationships as possible.

Decision models explain to find out optimization ways to predict (find out) results of decisions. This branch of predictive analytics apply in operations research, including areas such as resource optimization in networking & many places, route planning in many industries.

1.1 Data Mining Techniques

It is used for decision making in a business is very poor even though data storage grows exponentially. Data mining also known as knowledge discovery or finding some important information in some scenario. The Knowledge extracted allows predicting the behavior and future behavior. This allows the business owners to take positive, knowledge driven decisions. Data mining is applied on various industries like retail, finance, health care, aerospace, education etc. Knowledge is extracted from the historical data by applying pattern recognition, statistical and mathematical techniques those results in the knowledge the form of facts, trends, association, patterns, anomalies and exceptions. There are some areas where data mining will be applied.

Data Pre-processing: Data pre-processing make ready the real world data for mining process.

Data Mining: data mining is the process of extracting some important patterns from a large amount of data.

Pattern Evaluation: This process evaluates the pattern that is generated by the data mining. The patterns are evaluated according to the interestingness measure given by user or system.

Knowledge presentation: To show your Knowledge Presentation of your data uses visualization mechanism that can explore the important patterns (behaviour) and help the user to understand and interpret the resultant patterns.

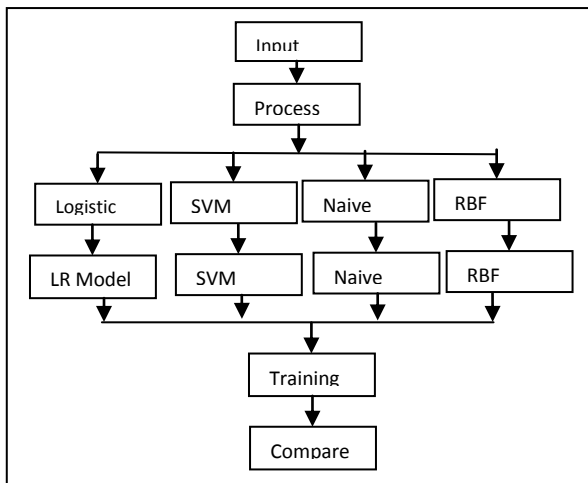


Figure 1: Different Machine Learning Methods [1]

1.2 Overview of data mining

Data mining is really attracted many industries to explore information or required data i.e. it provides a great deal of attention in the information industry and in society in upcoming years, due to availability of large amount of data and required need for turning such data into useful information and knowledge. Data mining is the process of data and looking for meaningful and important trending patterns. The information extracted from previous knowledge can be used for future applications for analysis of market, detecting much fraud in industries, and how to retain our customers. It follows (Data mining) iterative process.

Data cleaning: It is a process of removing noise and inconsistent data.

Data integration: In this step data from multiple sources are combined.

Data selection: In this step data relevant for mining task is selected.

Data transformation: In this step data will be transformed into form that is appropriate for mining.

Data mining: In this step some intelligent methods are applied for extracting data patterns.

Pattern evaluation: In this step we concentrate upon important patterns representing knowledge based on some measure are identified.

Knowledge presentation: In this step visualization and knowledge representation techniques are used to present the mined knowledge to the user.

1.3 Data Mining Algorithms & Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm (Adaptive Procedure), Nearest Neighbour method etc., are used for knowledge discovery from databases.

1.4 Classification

Classification is the frequently (most commonly) applied data mining mechanism, which explains a set of pre- classified

examples to develop a (procedure) model that can (identifies or categories) classify the population (Dataset) of records at large. During Fraud detection in network and Transactional (credit) risk applications are particularly well suited to this type of analysis. This way frequently applies by decision tree or neural network- based classification algorithms. The data (categorization) classification process involves learning (Trained previously) and then classification will apply on given data. Learning process can be done by analyzing data using classification Mechanism. In classification test data are used to find out (estimate) the accuracy of the classification rules which we applied. If the accuracy is acceptable the rules can be applied to the new data tuples.

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

1.5 Support Vector Machine

It is a supervised machine learning algorithm which can be utilizes for both classification and regression (Linearly) challenges. However, it is mostly used in classification problems. Here, we are explaining points (plot) each and every point (data item) as a point in n-dimensional space (where n is number of features you have) with the value of each and every points (plot) feature being the value of a particular coordinate system in given system.

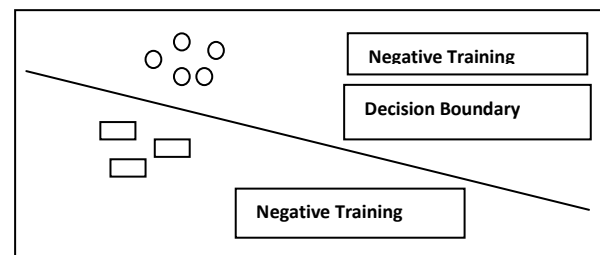


Figure 2: A linear Support Vector Machine [8]

1.6 Clustering

Clustering can be said as to find out of similar classes of objects. By using clustering mechanism we can further find out dense and sparse (n Dimensional Space) regions in object space and can discover overall distribution trends (pattern) and relation among many coordinate points (correlations) among data attributes. Classification by decision tree induction

- Partitioning Method
- Hierarchical Agglomerative methods
- Density based methods
- Grid-based methods
- Model-based methods

1.7 Prediction

Clustering Regression technique can be fitting for predication. Regression analysis can be used to create (procedure) model the relationship between one or more independent variables and dependent variables. In data mining independent (self-sufficient) variables are attributes already known and response (return) variables are what we

want to determine. But when we see in real- world problems are not simply (judge) prediction. For instance, data of sales, data of stock market prices, and many product failure rates in any industries or domain are very difficult to judge (predict) because they may depend on complex (or many variables) interactions of multiple predictor variables. Therefore, more composite ways (e.g., logistic regression, decision trees, or neural nets) may be required to forecast

Future values. The same model types can usually be used for both regression and classification mechanism. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees. By using Neural networks we can explain both classification and regression models. Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Simple Linear Regression: Simple linear regression is stats that allow us to encapsulate and read the relationships between two (steady) continuous (quantitative) variables:

- 1) First variable, denoted by x, is regarded as the predictor (judging elements), explanatory, or independent variable.
- 2) The other one, determine as y, is regarded as the outcome variable.

Multivariate Linear Regression: Here, we will judge (predict) the value of one or more outcomes from a set of attributes or properties (Predictors). It can also be used to find out the linear association between the predictors and outcomes. Predictors can be continuous or categorical or a combination of both.

2. LITERATURE SURVEY

2.1 Forecasting is Analyzing and Predicting the Future Behavior

In this paper Author concludes that forecasting (Future Analysis) is figure out and finding the future behavior of the selected data set. We will used past data to predict future data. It help in many ways in various domains such as controlling load balancing in network management, to explore future marketing campaigns in any industries like share market & FMCG, allocating or de-allocating resources and caching in networks and operating system , perfecting web pages for improve performance. There are limited number of researches have been done on Web site related forecasting.

3. PROBLEM IDENTIFICATION

A lot of research work has been done in the field. Authors have learned several things from this study (work). First, we will introduce some terms that are related with our topic such as, some brief description of Big Data, Data Warehouses, Data Mining and their classification and finally also will do the particular analysis.

4. BLOCK DIAGRAM & METHODOLOGY

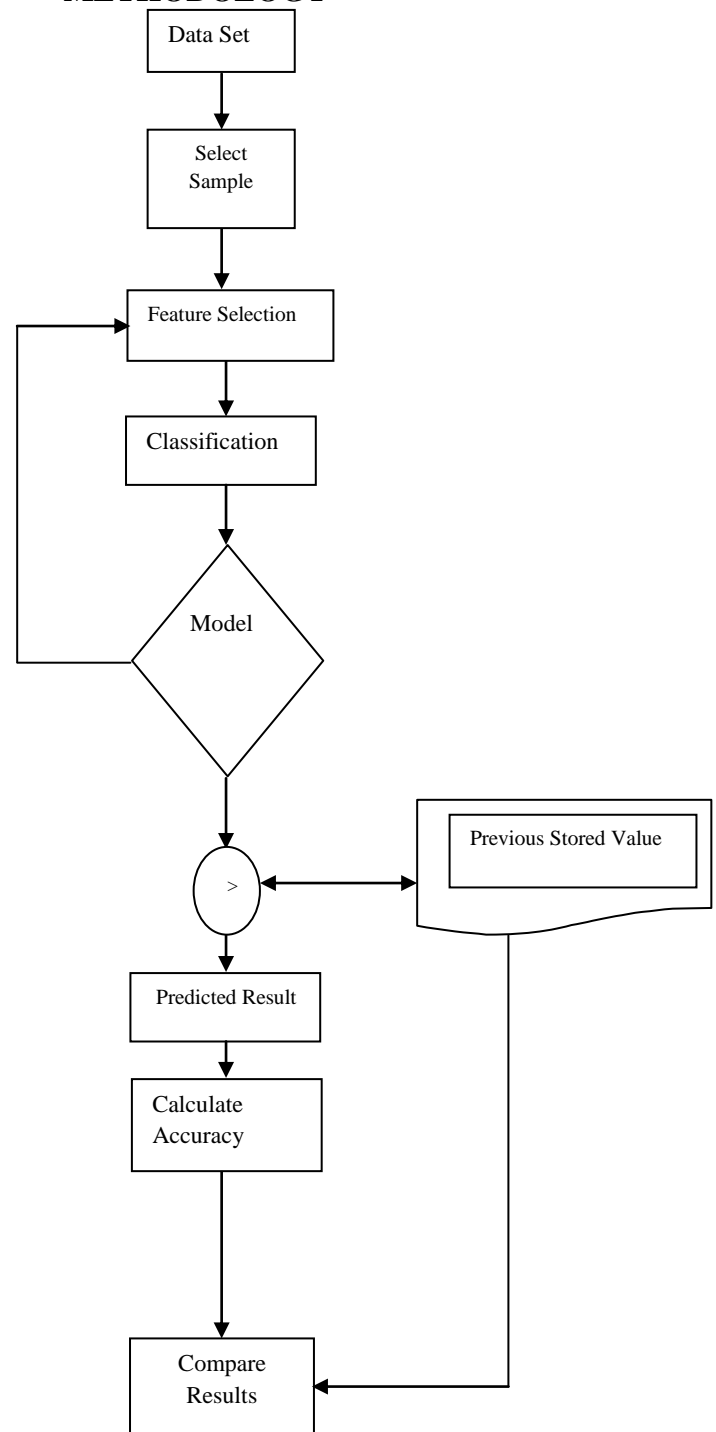


Figure 3: Block Diagram of Flow of Operation

4.1 Algorithm

1. Input / Load data set
2. Apply feature extraction
3. Received Extracted data as output
4. Generate Training and Testing data set (By applying techniques)

5. Apply any machine learning algorithm to training dataset
6. Build the Reduction Explanatory Predictor
7. Building Model using any machine learning model
8. Perform / Obtain validity check
9. Utilize the “test” set predictions to calculate all the performance metrics (Measure Accuracy and other parameters)

5. CONCLUSION

We studied many machine learning algorithm and we found that in day to day life or in internet era classification algorithm play vital roles to define and describe many things.

6. REFERENCES

- [1] Manisha rathi Regression modeling technique on data mining for prediction of CRM CCIS 101, pp.195-200,2010Springer-Verlag Heidelberg 2010.
- [2] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
- [3] Giudici Paolo, “Applied Data Mining-Statistical methods for business and industry” wiley, (2003) [5] Dash, M., and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis. 1:3 (1997) pp. 131-156. [6] Rencher C. Alvin, “Methods of Multivariate Analysis” 2nd Edition, Wiley Interscience, (2002).
- [4] Burges, C. (1998). A tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):955–974.
- [5] CiteSeer (2002). CiteSeer Scientific Digital Library. <http://www.citeseer.com>.
- [6] Duda, R. O. and Hart, P. E. (1973). Pattern Classification and Scene Analysis. John Wiley & Sons.
- [7] Greenbaum, A. (1997). Iterative Methods for Solving Linear Systems, volume 17 of Frontiers in Applied Mathematics. SIAM.
- [8] GLIM (2004). Generalised Linear Interactive Modelingpackage.<http://www.nag.co.uk/stats/GDGEsoft.asp>, <http://lib.stat.cmu.edu/glim/>.
- [9] Kubica, J., Goldenberg, A., Komarek, P., Moore, A., and Schneider, J. (2003). A comparison of statistical and machine learning algorithms on the task of link completion. In KDD Workshop on Link Analysis for Detecting Complex Behavior, page 8.
- [10] Lay, D. C. (1994). Linear Algebra and Its Applications. Addison-Wesley.