Comparative study of Prediction Algorithm

Shraddha Painjane
Department of Computer Engineering
Shree L.R.Tiwari college of Engineering
Mira road, Mumbai, India
shraddhapainjane@gmail.com

Prof. Sachin Bojewar
Department of Computer Engineering
Vidyalankar Institute of Technology
Mumbai, India

Abstract— Finding information hidden in data is as theoretically difficult as it is practically important. Companies have been collecting data for decades, building massive data warehouses in which to store it. Even though this data is available, very few companies have been able to realize the actual value stored in it. Support vector machines (SVM) were originally designed for binary classification. ANN, another prediction approach is also the one used to classify large dataset. In this paper we compare the two algorithms used in prediction and do comparative analysis of it.

Keywords*—Machine learning,Supervised learning,prediction algorithm;*

*****

I. Introduction

Data Mining is defined as extracting information from large sets of data by several ways. In other words, it can also be said that data mining is the procedure of mining knowledge from large dataset.

Data mining is used in various fields of market which are listed below:

- Customer Profiling – Determining what kind of people will buy what kind of products can be done using data mining.
- Identifying Customer Requirements –Identifying the best products for different kinds of customer. Factors are considered for prediction to attract new customers.
- Cross Market Analysis - Association/correlations between product sales done using data mining.
- Target Marketing - Data mining helps to find groups of qualified customers who share the same characteristics such as income, spending habits, interests, etc.
- Determining Customer purchasing pattern - Data mining helps in determining customer purchasing pattern such as bread-milk purchase.

Data mining refers to mining or extracting knowledge from large sets of data by using several ways. Data mining can also be referred as knowledge mining but since this term seems to be too long, it is known as data mining itself. Many people accept the term knowledge discovery in database for data mining. It is also known as KDD.

Knowledge discovery comprises of an iterative sequence of the following steps:

1) Data Cleaning: removing irrelevant and noisy data.
2) Data Integration: Combination of multiple data sources.
3) Data Selection: Retrieval of data from database that is relevant to analysis task.
4) Data Transformation: Transformation or consolidation of data into forms that are appropriate for data mining by performing summary operations.
5) Data mining: Process of extracting data patterns using intelligent methods.
6) Pattern evaluation: Identifying some interesting patterns based on some measures representing knowledge.
7) Knowledge presentation: Mined knowledge is presented to the user using visualization and knowledge representation.

The data mining directly interacts with the user or a knowledge base. The data is studied and interesting patterns are extracted and presented to the user and/or can be stored in the knowledge base or database as knowledge. Since data mining uncovers the hidden patterns for evaluation, it is considered as an essential step in the entire process.

II. Prediction techniques

The process of creating, testing and validating a model that best predict the probability of an outcome is known as Predictive modelling. Predictive analytics software solutions contains a number of modelling methods such as artificial intelligence, statistics and machine learning.

Predictive analytics analyses current and historical facts to make predictions about the other unknown events or about the future. It comprises a variety of statistical techniques from data mining, machine learning and predictive modelling.

In business and market perspective, predictive models extracts patterns from historical and transactional dataset to identify risks and opportunities. Models unhide the relationships among many factors that allows assessment of risk or potential associated with particular sets of conditions and also executing candidate transactions by guiding decision making.The technical approach says that predictive analytics provides a predictive score for each individual i.e customer, healthcare patient, product SKU, employee, vehicle etc. It helps to determine, inform and influence organizational processes that sustains across huge number of individuals such as credit risk assessment, fraud detection, marketing, healthcare, manufacturing and government bodies including law enforcement.

Predictive analytics is a part of data mining that deals with extraction of information from data and using this data for prediction of trends and behavioral patterns. It is mostly seen that unknown events of interest are of the future, but predictive analytics can be applied to any type of unknown events related to present, past or future. For example, Identification of suspect after the crime is committed or credit card fraud as it occurs. Capturing relationships between predicted variables and explanatory variables from past occurrences and exploiting them to predict unknown outcomes relies on predictive analytics. However, the accuracy and usability of the results will greatly depend on the quality of assumptions and the level of dataset.

Predictive analytics can also be said as predicting at a more detailed level of granularity i.e generating probabilities of predictive scores for each individual organizational element. This separates it from forecasting. For example, Technologies that learns from experience or the dataset produced to predict the future behavior of individual or system in order to develop better decisions. In future industrial or marketing system, the value of predictive analytics will be to predict and prevent potential issues to achieve break down and be further integrated into analysis for further decision optimization.

III. Neural network

A neural network is a distributed, parallel information processing structure that consist of processing elements that can possess a local memory and can also carry out information processing operations. These processing elements are interconnected via unidirectional signal channels known as connections. Each processing element has a single output connection that fans out into many parallel connections as desired. Each connection carries the same signal i.e the processing element output signal. This processing element output signal can be of any derived mathematical type.

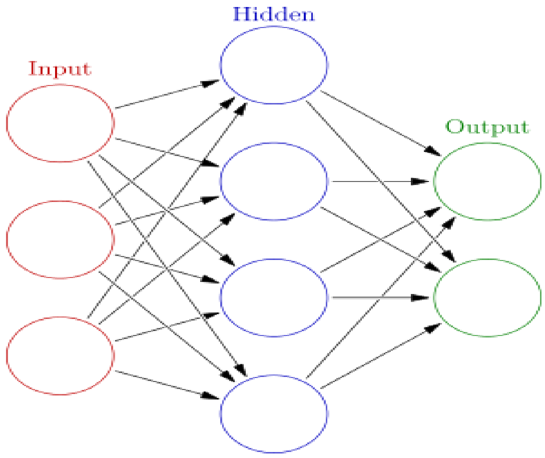
The processing of information that goes on within every processing element can be defined
Immediately and it also has a restriction that it must be completely local. It means that it must depend only on the current values of the input signals that arrives at the processing element via impinging connections and also on the values that are stored in the processing element's local memory. Neural network models have a similar description and explanation
Of the biological neural network of the human brain. Because of the size and the efficiency of the biological neural network, an artificial computer-based Neural network was introduced which would reflect only in a small fraction of the complexity and efficiency of human neural network.

The goal of the neural network is to solve problem abstractly in the same way the human brain would do. Modern neural network projects works with some few thousands to few millions neural units and millions of neural connections, which is lesser complex than the human brain in the factor of magnitude and also closer to computing power of the worm.

New patterns are stimulated in neural networks by new brain research. Another new approach is to use connections that would span much further and would link processing layers rather than being localized to neighbor neurons. Neural networks are based on real numbers, with the value of the core and of the axon typically being a representation between 0.0 and 1.

There is an interesting fact of the system that they are unpredictable in their success with self learning. Because of self learning, there is training data and after training some become great problem solvers but some doesn't perform well. In order to train these system well , several thousand cycles of interactions typically occurs.

Neural network works like a machine learning method that is it is a system that learns from data. This is used to solve a wide variety of task like speech recognition and computer vision which are complex when solved using ordinary rule based programming.



Back propagation, also known as propagation of error, is a common method of teaching artificial neural networks how a given task is to be performed. The layered feed forward Artificial neural networks use back propagation. This means that the artificial neurons are organized in layers which then sends their signals in forward direction and the errors are propagated backwards. The back propagation algorithm uses supervised learning in which we provide the inputs and outputs we want to the algorithm to compute and then the error is calculated. In this, error is nothing but the difference between actual and expected results. The only idea behind the back propagation algorithm is to reduce the error until the artificial neural network learns the training data.

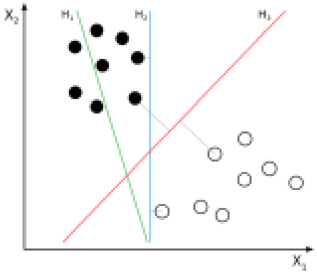
Summary of the technique:

1) A training sample is presented to the neural network.
2) The desired output is compared to the network's output of a particular sample. Error is calculated in each output neuron.
3) For each neuron, expected output is calculated from that sample and also scaling factor that determines how much higher or lower the output must be adjusted to match the desired output.

IV. SUPPORT VECTOR MACHINE (SVM)

In machine learning, support vector machines which are also known as support vector networks are supervised learning methods or models with associated learning algorithms that analyzes data used for regression analysis and classification. If given a set of training samples in which each marked as belongings to one or other group or categories, an SVM training algorithm establishes a model in which new examples are assigned to one or the other category, hence making it as a non-probabilistic binary linear classifier. An SVM model represents the examples as points in space which are mapped so that the examples of separate categories are divided by a clear gap that is as wide as possible. After that the new examples are then mapped into that same space and prediction is made to decide in which category do they belong to fall in that side of gap.

Supervised learning is possible only when the data is labelled. Hence when the data is not labelled, unsupervised learning approach is used which helps in to find natural clustering of the data to groups and then map new data to groups already formed. Support vector clustering is the clustering algorithm that provides an improvement to the support vector machines. This clustering algorithm is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.



A. *Limitations of SVM*

• The best choice of kernel for a any given problem is still a research problem.

• Speed and size (For large training dataset, it selects a small number of support vectors, which in turn minimizes the computational requirements during testing).

• The optimal design for multiclass SVM classifiers is still a research area.

V. SVM VS ANN

ANN (Artificial Neural Networks) and SVM (Support Vector Machines) are two well known strategies for supervised machine learning and classification.

(1) Rather than converging on global minima, Artificial neural network converges on local minima.
(2) If training goes on too long, Artificial neural network may often overfit considering the noise as part of the pattern.
(3) SVM's don't face any of these two problems.
(4) ANN are parametric models and SVM are non-parametric
(5) ANN size is fixed whereas SVM size varies.
(6) Advantage of artificial neural network over support vector machines are that ANN may have n number of outputs while SVM has only one.
(7) the neural network will make more sense because it is one whole, whereas the support vector machines are isolated systems.

VI. CONCLUSION

The expert studies have shown that SVM algorithms are computationally harder than neural network algorithms but results produced are better from SVM algorithms.

both ANN and SVM models had the ability to predict CEC within acceptable limits. ANN and SVM methods perform poorly in extrapolating maximum and minimum values of CEC data. ANN model provided better estimation in the testing period in comparison with SVM for CEC prediction. Before using both ANN and SVM modeling approaches for CEC prediction, it is suggested that these techniques may be used with the datasets from different regions as all machine learning approaches are data-dependent in Neural network perform poorly as the number of parameters are high and it take much time to process the information.

Neural network performs maximum experiments on the data nad hence gives results but support vector machine performs minimal experiments and give result in short time. Hence SVM are far better than neural networks.

Results in SVM are reproducible. That is data can be produced again and again

REFERENCES

[1] Schölkopf, Bernhard, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[2] I Kantardzic, Mehmed. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[3] Yao, Xin. "Evolving artificial neural networks." *Proceedings of the IEEE* 87.9 (1999): 1423-1447.

[4] Mao, Jianchang, and Anil K. Jain. "Artificial neural networks for feature extraction and multivariate data projection." *IEEE transactions on neural networks* 6.2 (1995): 296-317.

[5] Weber, Ben G., and Michael Mateas. "A data mining approach to strategy prediction." *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*. IEEE, 2009.