Data Mining from Machine Learning Perspective

Veni Gupta[1] Amita Kapoor[2]
[1,2]Shaheed Rajguru College of Applied Sciences for Women, University of Delhi

Abstract— Today, a huge amount of data is generated every hour, in website logs, in camera CCTVs footage, social networking sites and so on. The superfluity of data has made it important and difficult to discover useful knowledge from the data. Data scientists are using data mining to interpret this knowledge and apply it into some new useful and innovative technology with the help of machine learning. Both, data mining and machine learning are two interdisciplinary branches, which combine science as well as statistics. In today's world they both play an important role in making the modern Technology more useful and adaptive. This paper provides a brief review of the present day data mining and machine learning techniques, their uses and applications.

Keywords: Data Clustering, Machine Learning, Cellular Cloning Fraud and its Detection

I. INTRODUCTION

In past few years, a huge amount of data is being generated and collected. As the technology is advancing and the use of internet is increasing, a huge amount of the data is collected worldwide every day. The use of bar codes, the computerization of many transactions, the growing e-commerce has resulted in a very large amount of the data. Thus, there is a need of efficient technology to study this data and analyze it to discover the knowledge hidden in it. The new tools and techniques are required to intelligently transform this data into useful knowledge and apply that knowledge. Data mining is also known as "Knowledge Discovery in Databases" and has recently emerged as an efficient technique to extract out the knowledge and information from a wide range of databases (temporal, spatial, and relational). Data mining is the non-trivial extraction of implicit previously unknown and useful information [1]. Data mining can be viewed from different angles and perspectives like databases, machine learning, statistics, artificial intelligence etc. Research in data mining has many dimensions, for example the foundations of data mining, data mining and machine learning algorithms, mining temporal, spatial and multimedia data, data and knowledge representation from data mining, soft computing, pattern recognition, neural networks and many more. Data mining is required to handle different types of data, from relational databases to databases containing data of complex structure, like images and videos. Data mining can also be done on temporal, spatial and multimedia databases. Data mining algorithms should be efficient and scalable to discover proper knowledge from databases. The knowledge discovered from databases is abstracted at different levels depending on the data where data mining is applied. Data mining is one of the methods employed for Knowledge discovery in databases (KDD) i.e. the process of discovering useful knowledge from the data[2]. Figure 1 shows the flowchart of using data mining in KDD. Data mining techniques finds its applications in a wide range of applications:

1) Marketing and Financial Investment: Many marketing companies' uses data mining to observe the customers' needs and behavior to launch a new product or to invest in an existing product. Also used to stock indexes.
2) Fraud Detection: Today a lot of online transactions are taking place. Data mining is used to detect credit card fraud, mobile SIM cloning by detecting the unusual behavior from the usual communication patterns [3]. HNC Falcon and Nestor prism methods are used for the same [2].
3) Scientific Domain: Used in astronomical observations, genomic data and biological data [3]. Even in CERN experiment data mining techniques are being employed to identify anomalous atomic events.
4) Manufacturing: Used for the optimization of the process and for easy troubleshooting. Like, the CASSIOPEE trouble shooting system was used by three European airlines to diagnose the problems for Boeing 737[2].

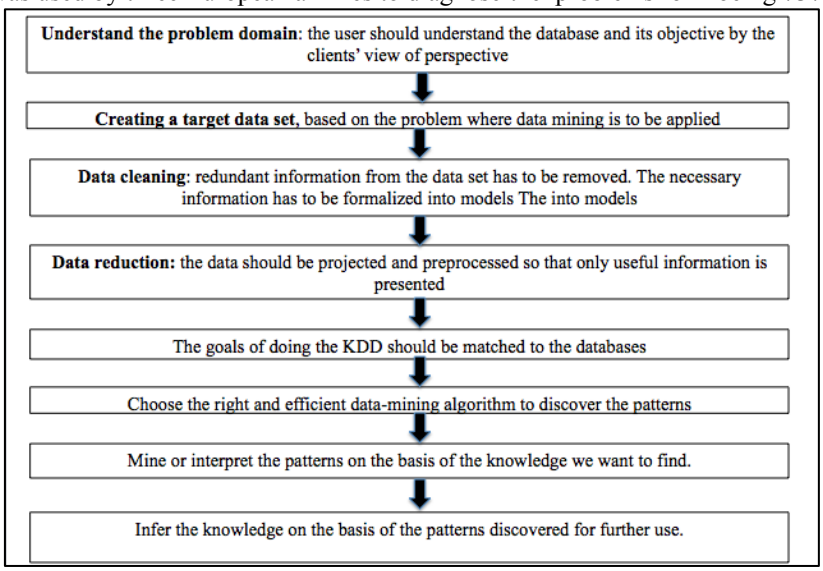

Fig. 1: the KDD Process

Below, we briefly describe most widely used data mining methods viz. classification, clustering, regression, summarization, dependency modeling.

A. Data Classification

Data classification is the process where the whole data is searched, studied and then according to different objects the data is mapped into different classes. In simple words, data classification is assigning an object to a certain class based on its similarity to previous examples of other objects. It can be done with reference to original data or based on a model of that data. A classification model is constructed by considering a database E also called as training set. The database E consists of many tuples, which in itself have many attributes, in very large databases each tuple is known by a class identity or label associated with it. The main objective of data classification is to make a classification model on the basis of the whole dataset. The data set is divided into different classes on the basis of similarities between them. These models are also known as classification rules on the basis of which further data can also be divided. Data classification usually includes decision tree models [1].Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Data classification can be used to classify transactions as genuine or fraudulent – e.g. credit card usage, insurance claims, cell phone calls etc., or to classify prospects as good or bad customers.

B. Regression

Regression is the process of constructing a function that maps a data item to a real valued prediction variable. Regression is used to fit an equation in the database. The simplest form of regression is the linear regression which uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given Value of x. advanced techniques, such as multiple regressions, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation. Regression technique can be used to predict whether a patient will survive on the basis of specific diagnosis test included, it can be used in predicting which good will be in more Demand on the basis of customers' interests.

C. Data Clustering

Data clustering is the process of classifying or grouping physical or abstract objects into classes of similar objects [1]. It is also known as unspecified or unsupervised classification. Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes.

The most common approach in data clustering is using the distance-based approaches. In distance based approaches all data points are given in advance and are scanned frequently, the class to which the data it belongs is not specified, the algorithm clusters them into different classes based on how one input is different from another (distant). It is basically known as conceptual clustering rather than physical clustering as the data is not clustered on the basis of the geometrical distances between them but on the logical class objects. Clustering is mainly based on the probability analysis as we are predicting things. For example probability-based trees are used to construct the clusters. These trees are not of balance height for the skewed data.

D. Dependency Modeling

Dependency modeling is basically concerned about finding a model, which describes functional dependencies between the variables [2].There are two levels of dependency models

- Structural Level: It tells the locally dependent variables.
- Quantitative Level: It specifies how much strong the dependency among the variables is.

Probabilistic dependency networks are used to describe dependency models. They are usually used in areas like development of probabilistic medical experts based on the databases, information retrieval

E. Summarization

Summarization in essence is generalization of data. It abstracts the data from a low level to high level. It gives us a compact description of the data. For example a huge amount of data collected on the population of different cities can be summarized on the basis of their average, variance, median and mode. More advanced methods can be derivation of summary rules, discovery of functional relationships between variables.

Summarization or data generalization can have different levels of abstraction and can be viewed from different angles.

F. Machine Learning

Machine learning (ML) is the study of how computer realizes the human learning mechanism [4]. In ML we employ the computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests [5]. Data mining and machine learning both together are used to analyze huge amount of data. In machine learning there are three categories of learning algorithms. First is supervised learning, also known as regression or classification, this model approximates the mapping between the input and output of the given data. In this the network is trained using training data sets. Training data sets include both the input and desired output results. These are usually fast and accurate algorithms. They are able to generalize the new data into its desired output sets. [6]. Second is unsupervised learning, here ML tries to find

hidden structures in unlabelled data. The desired output results are not specified in the training data sets. It can cluster the input data in classes only on the basis of their statistical properties. In this the learning process attempts to find the classes to which an input data belongs to [6]. Third is reinforcement learning, here a policy is searched for an agent to take actions that maximize the cumulated rewards in a given environment.[6]

There are a large number of machine learning techniques, below we briefly describe some of them.

1) Rule Induction

Rule induction (RI) in machine learning constructs a decision tree or decision rules from the target set with some classification [5]. RI is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data. RI is an emerging field in machine learning as it derives a set of rules from a given data. A decision tree is translated into a set of rules. The rules are normally stated in disjunctive normal form. The tree is applied to a set of data to evaluate its accuracy in classifying new examples. RI only creates mutually exclusive classes

2) Neural Networks

A Neural Network (NN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information [5]. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing working in unison to solve specific problems. Neural networks learn by experience. An NN neural is configured for a specific application, such as pattern recognition or data classification, through a learning process. They consist of input nodes, output nodes and n- number of hidden nodes. The input node receives signals and computes the output by transferring the signals through various nodes. An activation function is used for the output function. Neural networks works on classification. A good NN can classify or predict with a high degree of accuracy. Next we present some of the successful case studies where the techniques of data mining and ML have been used effectively and efficiently.

G. Method Of Fetal Assessment : University of Arkanas for Medical Sciences

A non-invasive system called the Squid Array for Reproductive Assessment (SARA) is developed that is used to gather the fetal heartbeat and to predict whether the baby is healthy or risk-prone [7]. SARA operates using magneto-encephalography (MEG) and super conducting quantum interference device (SQUID) technologies, which captures the heartbeats of the fetus being observed. Sara captures the heartbeats of the fetus in sine wave form, the knowledge of data mining and machine Learning is then applied to predict the health of the baby. As first we have the data in sine wave form, using data mining this data is constructed into understandable chart information of heartbeat. Data is standardized by subtracting the heartbeat per minute value from its mean, so that it does not give false results. After standardization of data, data is grouped into clusters and then decision tree is made. Each cluster contains heartbeat in a waveform. Dynamic time warp algorithm is used to compare the data. DTW compares the two waveforms on the basis of their distances. Distance of each curve against the representative heart waveform is captured. If the deviation is more than the standard curve baby is high risk prone otherwise he is normal.

II. CELLULAR CLONING FRAUD AND ITS DETECTION

Each cellphone transmits two unique identification numbers that are mobile identification number (MIN) and Electronic Serial Number (ESN). These numbers are transmitted without encryption, therefore they are easily hacked. A hacker then enters the others MIN and ESN number into his cell phone and whatever activity he does is charged through the others cell phone. So this was becoming a serious condition. Fawcett and Provost [8] used data mining and machine learning to prevent cellular cloning[8]. Their framework for fraud detection consists of the profilers that capture the typical behavior of an account when it is in use, and then identifies if the user shows any variation from this typical behavior. The profilers are connected with the detector, which learns how to detect the fraud on the basis of the output of the profilers.First stage of data mining includes studying the different calls and searching for the indicators of the fraud. RL program is used for the searching of rules with certainty factors above a user-defined threshold. The second stage includes construction of a detector that studies the output from the profilers. A third stage has a detector that compares the output of profilers for one stage with the other previous stages. A linear threshold unit (LTU) is selected after observing the data. It enables the judgment to be good and fast. In the last stage detector analyses the user's behavior based on many indicators and generates an alarm if it detects a fraud activity.

A. Future Scopes

The machine learning and data mining can be used in prediction of the weather conditions. Like, a disaster forecasting mechanism can be made, which can tell about hurricanes, earthquakes, and typhoons from the previous pattern collected and by analyzing the current weather conditions. Such an intelligent system can save millions of lives by predicting disasters before they strike. A general purpose robot or a computer can be made that learns from the experiences, such a robot can learn to clean, cook, work in farm or in industries.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, and Philip Yu, Data Mining: An Overview from a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, 8 (1996), 6: 866-883.

[2] Fayyad Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.

[3] Mitra, Sushmita, Sankar K. Pal, and PabitraMitra. "Data mining in soft computing framework: a survey." IEEE transactions on neural networks 13.1 (2002): 3-14.
[4] Yuntian, Wu. "Based on Machine Learning of Data Mining to Further Explore." Computer Science and Information Processing (CSIP), 2012 International Conference on. IEEE, 2012.
[5] Bose, Indranil, and Radha K. Mahapatra. "Business data mining—a machine learning perspective." Information & management 39.3 (2001): 211-225.
[6] Jin, Yaochu. "Pareto-based multi-objective machine learning." Hybrid Intelligent Systems, 2007. HIS 2007. 7th International Conference on. IEEE, 2007.
[7] Copeland, Wes, and Chia-Chu Chiang. "A method for fetal assessment using data mining and machine learning." Information Security and Intelligence Control (ISIC), 2012 International Conference on. IEEE, 2012.
[8] Fawcett, Tom, and Foster J. Provost. "Combining Data Mining and Machine Learning for Effective User Profiling." KDD. 1996.