IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 4 Issue 5 May, 2016

An Overview of Data Mining Models (Descriptive and Predictive)

Kwame Boakye Agyapong, Dr. J.B Hayfron-Acquah, Dr. Michael Asante.

Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.

Email: opanin007@yahoo.com, jbha@yahoo.com,mickasst@yahoo.com

ABSTRACT

Data mining compromises talented ways to expose hidden designs within huge volumes of data. These hidden designs can possibly be used to forecast forthcoming performance. This paper discusses the various data mining models in order to gain a major understanding of the various data mining algorithms and the way these can be utilized in various business applications and the way these algorithms can be used in the descriptive and predictive data mining modeling. The descriptive data mining tasks characterize the general properties of the data present in the database, while in contrast predictive data mining technique perform inference from the current data for making prediction. This overview briefly introduces these two most important techniques that perform data mining task as Predictive and Descriptive. Between this predictive and descriptive they consist of their own method as Classification, clustering, summarization, association, etc., all are briefly introduced here. This overview presents most important predictive and descriptive data mining techniques by which most of the mining task are performed. This predictive and descriptive data mining task includes classification, clustering, regression, summarization, and some other techniques all of which are briefly describe in this overview which performs mining task correctively.

Keywords: Algorithms, Technique, Classification, Clustering, Data mining, Descriptive, Predictive, Summarization.

1. INTRODUCTION

1.1 Data Mining

There is a massive amount of data accessible in the Information Production. This data is of no importance until it is transformed into useful material.wikipedia.org(2007). It is necessary to analyze this massive amount of data and remove useful information from it. Removal of information is done with other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. We would be able to use this information in many applications such as Fraud Detection, Market Investigation, Production Control, Science Exploration, after the foregoing process are over.

Data mining satisfy its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. This goal of data mining can be satisfied by modeling it as either Predictive or Descriptive nature. The Predictive model works by making a prediction about values of data, which uses known results found from different datasets. The tasks include in the Predictive data mining model includes classification, prediction, regression and analysis of time series. The Descriptive models mostly identify patterns or relationships in datasets. It serves as an easy way to explore the properties of the data examined earlier and not to predict new properties.

In a classification technique, you typically have historical data called labeled examples and new examples. Each labeled example consists of multiple predictor attributes and one target attribute that is a class label. While unlabeled examples only consist of the predictor attributes. The goal of classification is to construct a model using the data from history and accurately predicts the new class of examples.

1.2 The Scope of Data Mining

Data mining originates its name from the resemblances between probing for treasured business information in a large database, for instance, discovering linked products in gigabytes of store scanner data, and pulling out a mountain for a vein of treasured ore. The two processes above will like to find out where exactly the treasured can be found. If the database given is satisfactory in size and quality, then the data mini9ng technology can produce new prospects by given these competences.

Behaviors and automated prediction of trends. Data mining mechanizes the procedure of discovering predictive information in large databases. Questions that conventionally requisite widespread hands-on investigation can now be answered straight from the data rapidly. Targeted marketing is a classic example of a predictive problem. Data mining customs data on past advertising mailings to categorize the goals probably to get the best out of return on investment in forthcoming mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated detection of formerly unknown patterns. Data mining tools swing through databases and categorize formerly hidden designs in one step. An instance of design detection is the investigation of retail sales data to categorize apparently dissimilar products that are often bought together. Other design detection problems include noticing deceitful credit card transactions and recognizing irregular data that could symbolize data entry inputting errors.

Data mining techniques can produce the benefits of mechanization on prevailing software and hardware platforms, and can be applied on new systems as prevailing platforms are elevated and new products established. Data mining tools can investigate large databases in few minutes if they are applied on high performance parallel processing systems. Processing faster means that users can mechanically trial with additional models to apprehend compound data. High swiftness makes it hands-on for users to investigate massive amounts of data. Databases that are bigger, always in turn produce better-quality forecast.

1.3 Data Mining Models

There are two main data mining models types. These are: Predictive and Descriptive. The descriptive model recognizes the designs or relationships in data and discovers the properties of the data studied. For instance, Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is like classification however the groups are not predefined, but then again are well-defined by the data alone. It is also referred to as unsubstantiated learning or subdivision. It is the wall off or splitting up of the data into collections or clusters. The clusters are well-defined by learning the performance of the data by the domain experts. The term splitting up is used in very precise framework; it is a process of separation of database into split grouping of related tuples.

Predictive analytics has been defined by Delen&Demirkan (2013) as to have data modeling as a prerequisite when making authoritative predictions about the future using business forecasting and simulation. These address the questions of "what will happen?" and "why will it happen?" A different study by Lechevalier, Narayanan, &Rachuri (2014), defines Predictive analytics as a tool that "uses statistical techniques, machine learning, and data mining to discover facts in order to make predictions about unknown future events," in investigating a domain-specific framework for Predictive analytics in manufacturing. The predictive model makes forecast about unidentified data values by using the identified values. For instance, Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are meant to forecast the forthcoming state of the data. Forecast is the process of investigating the existing and previous states of the attribute and forecast of its forthcoming state. Classification is a method of plotting the target data to the predefined clusters or classes. The regression includes the book learning of purpose that map data element to actual valued forecast variable. In the time series analysis, the value of an attribute look at it as differs over time. In time series analysis the distance measures are used to define the resemblance between different time series, the

structure of the line is studied to define its deeds and the past time series plot is used to forecast forthcoming values of the variable.

1.4 Descriptive Models

Summarization is the process of giving the recap information from the data. The association rule discovers the connection amongst the diverse traits. Association rule mining is a two-step process: Finding all frequent item sets and Generating strong association rules from the frequent item sets. According to Mortenson, Doherty, & Robinson (2014), descriptive analytics recaps and transforms data into expressive information for reporting and one-to-one care but also allows for thorough examination to answer questions such as "what has occurred?" and "what is presently bang up-to-date?" SAP, (2014) also described descriptive analytics as control panel applications that support development implementation in sales and procedures administration, allowing for real-time tracking.

Summarization can be observed as squeezing a given set of dealings into a smaller set of designs while recollecting the supreme likely information. Summarization is a common and authoritative though often time-consuming method to examining large datasets. For example, suppose one wants to examine census data in order to appreciate the association amongst level of education and salary in Ghana. A very dense summary of the census can be observed by plotting the average salary by education level. This summary will be adequate for some resolutions, but others may need more time to realize a healthier empathetic of the data. A simple example would be to embrace the standard deviation information sideways with the averages. In addition, it may be more see-through, for instance, to breakdown the average salaries by age group, or to eliminate distant salaries. In general, the summarization involves both identifying overall trends and important exceptions to them, it does not lead to forecast.

Sequence data is alternative tool used as Descriptive model. Consumer bargain hunting sequences, medical treatment data, and data associated to natural tragedies, science and engineering procedures data, stocks and markets data, telephone occupation designs, weblog click streams, package execution sequences, DNA arrangements and gene appearance and assemblies data are some instances of sequence data. The Sequence data normally links data but does not forecast in to the future though decision can be taken after it has been expressed. Let's consider K= {ki, kii, kiii … kn} be a set of items. An item-set M is a subset of items i.e. M ⊆ K. A sequence is an ordered list of item-sets (also called elements or events). Items inside a component are unordered and we would list them alphabetically. An item can occur at most once in a component of a sequence, but can occur several times in dissimilar components of a sequence. The number of instances of items in a sequence is called the length of the sequence. l-sequence is a sequence with length l. E.g., s={<a(bd)(bcde) >} is a sequence which consists of 5 distinct items and 4 elements. Length of the sequence is 8. Sequence database is also a group of sequences stored with their identifiers. We say that a sequence k is a subsequence of m, if k is an "estimate" of m, resulting by removing components and or items from m. E.g. {<a(c)(bd)f>} is a subsequence of k.

There can also be Multidimensional Sequential Pattern Mining as expressed by Pinto, Han, Pei, Wang, Chen, &Dayal, (2001). Let's consider pattern P1= {use a 100-hour free internet access bundle⇒donate to 20 hours'/month bundle⇒elevated to 50 hours per month bundle⇒elevated to unlimited bundle from an Internet Service Provider(ISP) like Vodafone in Ghana. This pattern may hold for all customers below age of 25 who are males. For other consumers, design P2= {use a 100-hour free internet access bundle⇒elevated to 50 hours per month bundle} may hold. Clearly, if successive design mining can be related with consumer group or other multi-dimensional information, it will be more operative since the confidential designs are often more valuable. Pinto et al. (2001) again propose a mixing of well-organized sequential design mining and multi-dimensional investigation procedures (Seq-Dim and Dim-Seq) and implanting multi-dimensional information into sequences and mine the whole set by means of a uniform sequential design mining technique (Uni-Seq). A multi-dimensional sequence database has the schema: RID; record identifier, Pi, Pii … Pm; attributes, and S is the sequence. The derived formula only

helps to link the items but does not help one to forecast into the future though decision can be taken out of that.

Cluster analysis is another type of Descriptive model which groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. If the similarity (or homogeneity) within a group, or the difference between groups, is great, the "better" or more distinct the clustering. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.

From the foregoing it can be established that the descriptive model recognizes the designs or relationships in data and discovers the properties of the data studied. It does not always forecast to the future as would be seen in the Predictive model.

1.5 Predictive Model

Time-series methods are also parts of Predictive analytics, making use of methods such as moving averages, exponential smoothing, autoregressive models, linear, non-linear and logistic regression Souza, 2014. "What is a time-series database?" A time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, utility studies, inventory studies, yield projections, workload projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and medical treatments. A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

1.6 Trend Analysis

A time series involving a variable Y, representing, say, the daily closing price of a share in a stock market, can be viewed as a function of time t, that is, Y = F(t). Such a function can be illustrated as a time-series graph, as shown in Figure 2, which describes a point moving with the passage of time.

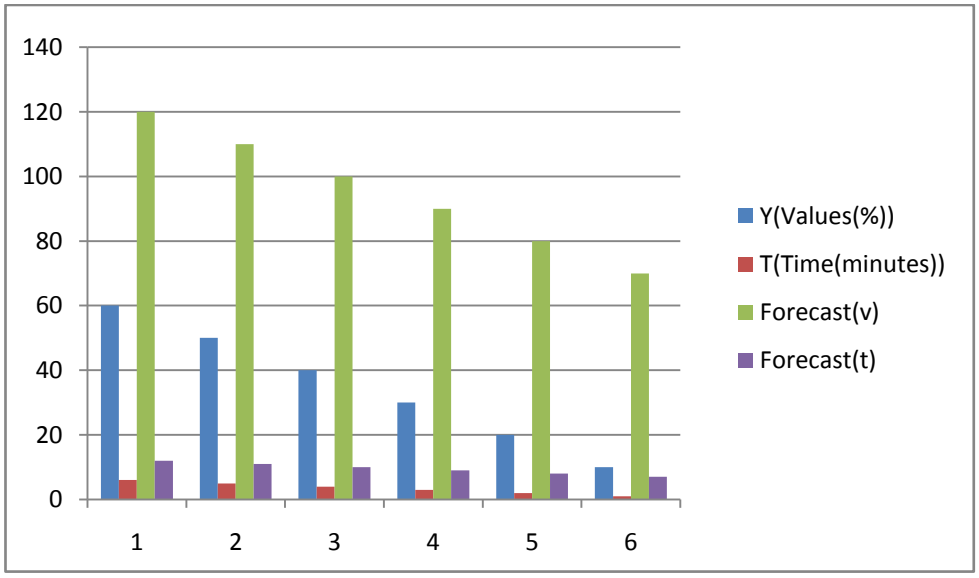

Fig.1. A point moving with the passage of time.

"How can we study time-series data?" The two main goals in time-series analysis are: (1) modeling time series (i.e., to gain insight into the mechanisms or underlying forces that generate the time series), and (2) forecasting time series (i.e., to predict the future values of the time-series variables). From figure 1, we can use the previous time and value to forecast the future values and time as well. With that the two main goals of time series are met.

Classification is one kind of predictive modeling. More precisely, classification is the method of passing on new objects to predefined groups: We should build model such as decision tree to forecast labels for future unlabeled records when a set of labeled records are given. Classification algorithms typically contain two phases:

Training Phase: In this phase, a model is built from the training instances.

Testing Phase: In this phase, the model is used to allocate a label to an unlabeled test instance. Some common application domains, in which the classification problem arises, are as follows:

- Medical Disease Diagnosis: Currently, the use of data mining techniques in medical technology has become greater than before increasing adhesive friction. The structures may be pull out from the medical records, and the class labels agreeing to whether or not a patient may give a lift to a disease in the future. In these cases, it is necessary to make disease forecasts with the use of such information.
- Customer Target Marketing: Since the classification problem relates eye variables to a objective class, this technique is enormously popular for the problem of customer target marketing. In such cases, eye variables relating the customer may be used to forecast their buying interests on the base of former training examples. The target variable may encrypt theCustomer's buying interest.
- Biological Data Analysis: Biological data is repeatedly symbolized as discrete orders; in which it is necessary to forecast the possessions of particular orders. Sometimes, the biological data is similarly conveyed in the form of networks. Therefore, classification procedures can be applied in a diversity way in this scenario.
- Social Network Analysis: Many forms of social network investigation, such as collective classification, associate labels with the primary nodes. These are then used in order to forecastthe labels of other nodes. Such submissions are very valuable for forecasting valuable properties of actors in a social network. The discussions of classification methods above also use the prevailing data to forecast into the future.
- Supervised Event Detection: In many sequential circumstances, class labels may be related to time stamps conforming to rare events. For instance, an intrusion activity may be symbolized as a class label. In such cases, time-series classification methods can be very valuable.
- Multimedia Data Analysis: It is repeatedly necessary to make classification of large volumes ofmultimedia data such as photos, videos, audio or other more compound multimedia data. Multimedia data investigation can repeatedly be thought-provoking, because of the difficulty of the primary feature space and the semantic break between the feature values and agreeing implications.
- Document Categorization and Filtering: Many applications, such as newswire services, necessitate the classification of huge facts of documents in real time. This application is indicated to as document categorization, and is an important area of research in its own right.

1.7 Regression Tree

Regression is another Predictive data-mining model also known as is a supervised learning technique. This method investigates of the reliance of some attribute values, which is at the mercy of the values of other attributes mostly existing in same item. The progress of a model can forecast these attribute values for new cases. The difference between regression and classification is that regression deals with numerical or continuous target attributes, whereas classification deals with discrete or categorical target attributes. In other words, if the target attribute comprises continuous (floating-point) values, a regression method is essential. The most common form of regression is linear regression, in which a line that best fits the data is calculated, that is, the line that minimizes the average distance of all the points

from the line. This line becomes a predictive model when the value of the dependent variable is not known; its value is predicted by the point on the line that corresponds to the values of the independent variables for that record. Let's consider an algorithm of the Regression tree

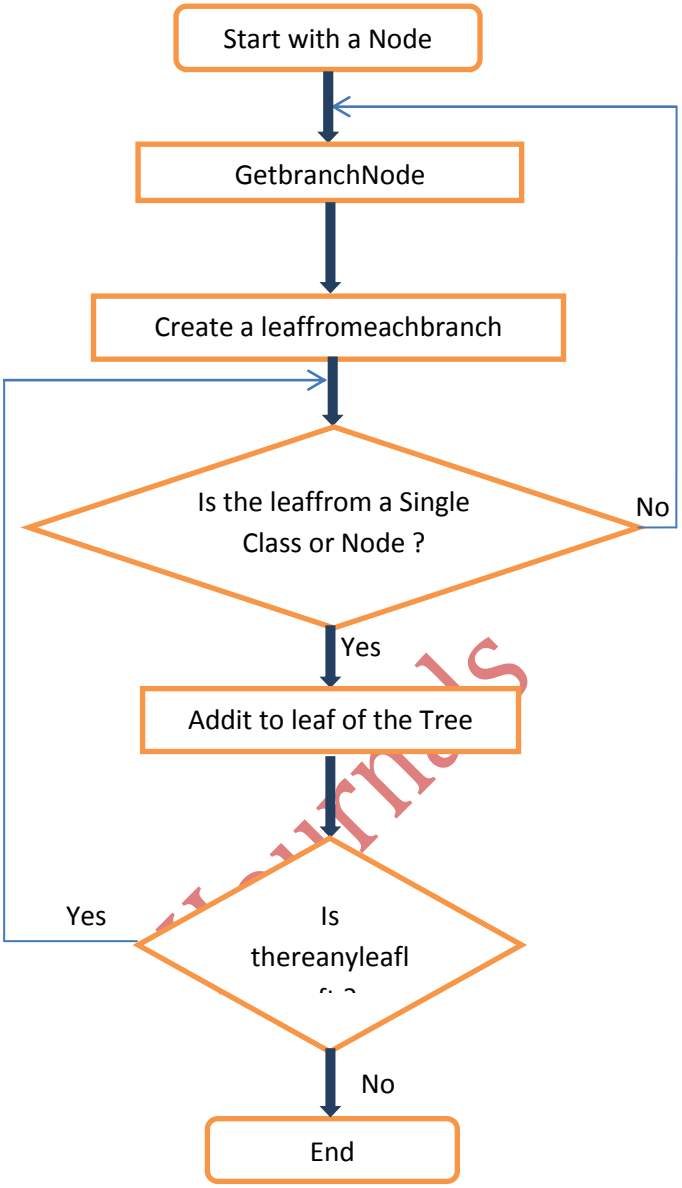

Fig.2. An Algorithm for Regression

1.7.1 *Explain the algorithm for Regression in figure 2*

1. Begin with data at root node.
2. Choose a trait and express a logical test on that trait. The Choice is based on "split-points", which must be assessed for all traits.
3. Generate a child node on each result of the assessment, and transfer subclass of instances filling that result to the equivalent child node.
4. Recursively measures each child node. The recursion ends for a child node if it is "pure" (i.e., all examples it contains are from a single class or "nearly pure" i.e., most examples it contains are from the same class.). The child node at which a recursion stops is called a leaf of the tree.

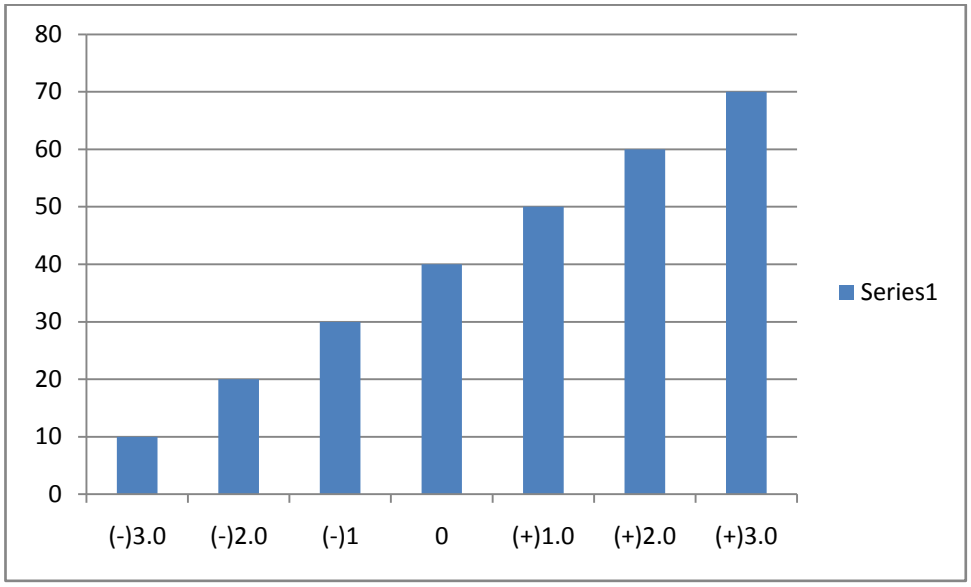

Fig. 3 illustrates as glucose increases, Diabetes also increases.

Figure 3 illustrates that as glucose increases, the probability of contracting Diabetes increases linearly on the scale. The figure 3 is implemented as regression tree and is displayed on the scale. The values in Y-axis depicts level of glucose whilst the X-axis, is for the rate at which Diabetes increases. Hence the regression tree can be used for forecasting.

2. CONCLUSION

From the analysis, the purpose of Predictive model is to determine the future outcome rather than current behaviour. Its output can be categorical or numeric value. For example, given a prediction model of credit card transactions, the likelihood that a specific transaction is fraudulent can be predicted. Another Predictive model discussed, Regression is a supervised learning technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the development of a model that can predict these attribute values for new cases. The second model of data mining, Descriptive is normally used to generate frequency, cross tabulation and correlation. Descriptive model can be defined to discover interesting regularities in the data, to uncover patterns

and find interesting subgroups in the bulk of data. Summarization as discussed maps data into subsets with associated simple descriptions. It is therefore advisable to use the Predictive model to the Descriptive model since the latter does not forecast into the future.

3. REFERENCES

[1]. Mortenson, M. J., Doherty, N. F., & Robinson, S. (2014). OperationalresearchfromTaylorism to terabytes:aresearch agenda for the analyticsage. European Journal of OperationalResearch, 583-595.

[2]. SAP. (2014, 05 31). SAP HANA Marketplace. Retrievedfrom SAP : http://marketplace.saphana.com SAP. (2014, 05 31). SAP HANA partner race`. Retrievedfrom SAP : http://global.sap.com/germany/campaigns/2012_inmemory/partner-race/race.epx S

[3]. Delen, D., &Demirkan, H. (2013). Data, information and analytics as services. Decision Support Systems, 359- 363.

[4]. 4.Lechevalier, D., Narayanan, A., &Rachuri, S. (2014). Towards a Domain-Specific Framework for PredictiveAnalytics in Manufacturing. 2014 IEEE International Conference on Big Data (pp. 987-995). Gaithersburg: National Institute of Standards and Technology.

[5]. Souza, G. C. (2014). 2014. Business Horizons, 595-605.

[6]. 6J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu.” Transactions on KnowledgeDiscovery in Data”, Volume 5,

Issue 3, pages 16:1-24, August 2011, ACM Press.

[7]. Data Mining. March 2007. <http://en.wikipedia.org/wiki/Datamining/>

[8]. Data Mining and Knowledge Discovery with Evolutionary Algorithms, A.A. Freitas, Springer-Verlag, 2002.

[9]. Dietterich TG (1997) Machine learning: Four current directions. AI Mag 18(4):97–136

[10]. Oliveira, S.R.M., Zaiane, O.R.: Toward standardization in privacy preserving data mining. In:ACM SIGKDD 3rd Workshop on Data Mining Standards, pp. 7–17 (2004)

