International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-4, Issue-3, March 2018 Pages 28-30

A Survey of Data Mining Tasks

Sanjana K S

***Abstract*- Data Mining is the process of discovering patterns in large data sets and establish relationships to solve problems through data analysis. The main goal of data mining is to identify patterns and to transform them into a more understandable structure for further analysis. Data mining process pares the overall task of finding patterns from data into a set of well defined subtasks. Data mining uses sophisticated algorithms to find patterns and evaluate the possibility of a future event. There are fundamentally different types of tasks these algorithms address. In this paper we make an effort to briefly explain these funadamental tasks .**

***Index Terms*— Classification, clustering, Data Minig , KDD , Regression,**

I. INTRODUCTION

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data [1]. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discovering new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge [1]. These data mining algorithms fundamentally address different data mining tasks.

II. DATA MINING TASKS

A. CLASSIFICATION AND CLASS PROBABILITY ESTIMATION

attempts to predict, for each individual in a population, which of a(small) set of classes this individual belongs to. Usually the classes are mutually exclusive. For a classification task, a data mining procedure produces a model that, given a new individual, determines which class that individual belongs to.[2] Classification can be performed both on structured and unstructured data. A scoring model applied to an individual produces, instead of a class prediction, a score representing the probability that the individual belongs to class.[2]

Classification Algorithms:

1) Logistic Regression: In this algorithm, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. This is more useful for understanding the influence of several independent variables on a single outcome variable.[3]
2) Naïve Bayes: Naïve Bayes algorithm, is based on Bayes' theorem with strong independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.[4]
3) Stochastic Gradient Descent: Also known as incremental gradient descent ,Stochastic Gradient Descent is a stochastic approximation of the gradient descent optimization and iterative method for minimizing an objective function that is written as a sum of differentiable functions.[5]
4) K-Nearest Neighbours does not attempt to construct a general internal model, but simply stores instances of the training data. This algorithm is simple to implement ,robust to noisy training data, and effective if training data is large.[3]
5) Decision tree : Decision tree repetitively divides the working area into sub part by identifying lines. [6]
6) Random Forest: This algorithm operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees . [7]
7) Support vector machines : They are linear discriminants. The data points are seperated by the widest bar and the center line through the bar is the linear descriminant.

B. REGRESSION

attempts to estimate or predict, for each individual, the numerical value of some variable for that individual. A regression procedure produces model that, given an individual, estimates the value of the particular variable specific to that individual[2]

Regression algorithms:

1) Simple linear regression model: It is a statistical method that enables users to summarize and study relationships between two continuous variables.[8]
2) Lasso regression: LASSO stands for Least Absolute Selection Shrinkage Operator wherein shrinkage is defined as a constraint on parameters. The goal of lasso regression is to obtain the subset of predictors that minimize prediction error for a quantitative response

Sanjana K S, Department of computer science and engineering, Siddaganga Institute of Technology, Tumkur, India,

variable. It is basically a shrinkage and variable selection method.[8]

3) Logistic regression: One of the major upsides is this popular algorithm is that it can include more than one dependent variable which can be continuous or dichotomous . This algorithm provides a quantified value to measure the strength of association according to the rest of variables.[8]
4) Support Vector Machines: SVM is another most powerful algorithm with strong theoretical foundations based on Vapnik-Chervonekis theory. This algorithm can be leveraged both for classification or regression challenges. SVM algorithms uses epsilon –insensitivity loss function to solve regression problems.[8]

C. SIMILARITY MATCHING

tries to recognize similar individuals based on the information known about them. If two entities are similar in a way, they share other characteristics as well.[9] Similarity is the underlying principle for making product recommendations. Similarity measures underlie certain solutions to other data mining task, such as classification, regression, and clustering.[2]

D. CLUSTERING

attempts to group individuals in a population together by their similarity, but not driven by any specific purpose. Clustering is useful in preliminary domain exploration to see which natural groups exist because these groups in turn may suggest other data mining tasks or approaches[2].
Clustering does not predict an outcome or target variable but can be used to improve predictive model.

Clustering algorithms:

1) K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.[10]
2) Fuzzy c-means allows one piece of data to belong to two or more clusters. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point.[11]
3) Hierarchical clustering seeks to build a hierarchy of clusters. Two basic strategies for hierarchical clustering clustering are:
    a) Agglomerative : A “bottom up” approach; each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
    b) Divisive: A top-down approach; all observation start in one cluster ,and splits are performed recursively as one moves down the hierarchy.

E. CO-OCCURRENCE GROUPING

attempts to find associations between entities based on transactions involving them. [2] When two items co-occur , there is an association between the two entities as indicated by the grouping agent. The agent can be a consumer purchasing items in a drug store [12] When more than one agent also associates those two items, that is another instance which strengthens that association. The collection of all the co-occurring elements form a framework from which to mine associations, be them in the form of clusters, association rules, or transitive associations and can be found within systems which highlight for the analyst those unknown facts.

F. PROFILING

also known as behavior description attempts to characterize the typical behavior of an individual , group, or population. Profiling is often used to establish behavioral norms for anamoly detection applications such as fraud detection and monitoring for intrusions to computer systems. [2] For example, if we know what kind of purchase a person typically makes on a credit card, we can determine whether a new charge on the card fits that profile or not.

G. LINK PREDICTION

attempts to predict connections between data items, usually by suggesting that a link should exist and possibly also estimating the strength of the link. Link prediction is common in social networking systems. For example, for recommending movies to customers one can think of a graph between customers and the movies they’ve watched or rated. Within the graph, we search for links that do not exist between customers and movies, but that we predict should exist and should be strong. These links form the basis for recommendations. [2]

H. DATA REDUCTION

attempts to take a large set of data and replace it with a smaller set of data that contains much of the important information in the larger set. The smaller dataset may be easier to deal with or to process. Moreover the smaller dataset may better reveal the information. For example, a massive dataset on consumer movie-viewing preferences may be reduced to a much smaller dataset revealing the consumer taste preferences that are latent in the viewing data. [2]

REFERENCES:

[1]Oracle –Data mining concepts

[2] Data Science for Business- Foster provost and Tom Fawcett.

[3]https://analyticsindiamag.com/7-types-classification-algorithms/

[4]https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[5]https://en.wikipedia.org/wiki/Stochastic_gradient_descent

[6]https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

[7]https://en.wikipedia.org/wiki/Random_forest

[8]https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

[9]https://www.kdnuggets.com/2015/01/fundamental-methods-data-science-classification-regression-similarity-matching.html

[10]https://en.wikipedia.org/wiki/K-means_clustering.

[11]https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm

[12]Journal of Technology Research -Co-occurrence analysis as a framework for data mining by Jan W. Buzydlowski Holy Family University, volume 6 ,January 2015