IJRSET 2015 SPL Volume 2, Issue 10 Pages: 23-26
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

Data Mining Techniques – A Survey

[1] N. Tamilselvi, [2] S. Saranya, [3] P. Usha,
[1, 2, 3] Assistant professors,
[1, 2, 3] Department of computer science,
[1, 2, 3] Dr. N. G. P. Arts and Science College.

Abstract:-

Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Data mining plays an important role in various fields of human life. Classification techniques are supervised learning techniques that classify data item into predefined class. It is one of the most useful techniques in data mining to build classification models from an input data set and these techniques commonly build models that are used to predict future data trends.

Keywords: - Classification, Mining Techniques, Algorithms.

1. INTRODUCTION

Data mining consists of several techniques that can be used to extract relevant information from data. Data mining has several tasks such as association rule mining, classification and prediction, and clustering. Classification is one of the most useful techniques in data mining to build classification models from an input data set.
Model construction: describing a set of predetermined classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
- The set of tuples are used for model construction: training set.
- The model is represented as classification rules, decision trees, or mathe matical formulae.
- Accuracy rate is the percentage of test set samples that are correctly classified by the model.
- Test set is independent of training set, otherwise over fitting will occur.

The goal of the predictive models is to construct a model by using the results of the known data and is to predict the results of unknown data sets by using the constructed model. The models are

- Decision Trees
- Artificial Neural Networks
- Support Vector Machine
- K-Nearest Neighbor
- Navie-Bayes

Before classification technique preprocessing techniques are applied for efficient results.

Decision Tree:

Decision tree learning goal is to create a model that predicts the value of a target variable based on several input variables. Decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal node denotes a test on an attribute,

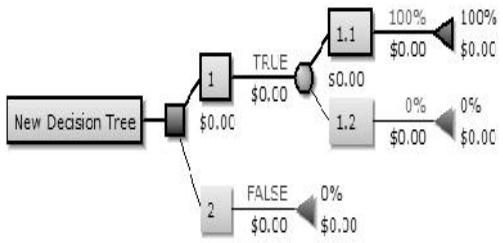
each branch represents the outcome of a test, and each leaf node holds a class label. The topmost node in a tree is the root node. Each interior node corresponds to one of the input variables and there are edges to children for each of the possible values of that input variable Decision tree learning is one of the most successful techniques for supervised classification learning. In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

(x,Y)=(x1, x2, x3…….xk, Y)

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **x** is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.A decision tree consists of 3 types of nodes:

1. Decision nodes - commonly represented by squares
2. Chance nodes - represented by circles
3. End nodes - represented by triangles



2. TYPES

Decision trees are classified into two categories they are

- **Classification tree.**
- **Regression tree**

The specific decision-tree algorithms are:

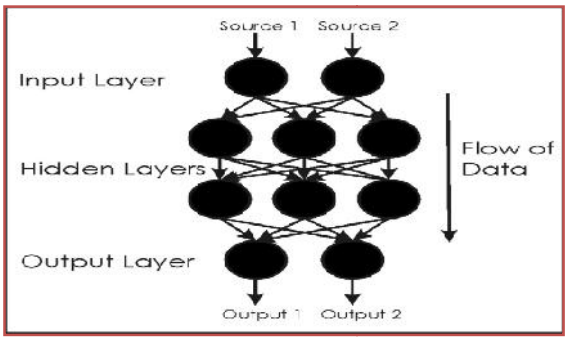
- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (Chi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
- MARS: extends decision trees to handle numerical data better.

Artificial Neural Networks (**ANNs**) are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.Neural networks are similar to biological neural networks in the performing of functions collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which individual units are assigned. The term "neural network" usually refers to models employed in statistics, cognitive psychology and artificial intelligence. Neural network models which command the central nervous system and the rest of the brain are part of theoretical neuroscience and computational neuroscience.

3. NETWORK FUNCTION

An ANN is typically defined by three types of parameters:

- The interconnection pattern between the different layers of neurons
- The learning process for updating the weights of the interconnections
- The activation function that converts a neuron's weighted input to its output activation.



Mathematically, a neuron's network function f(x) is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables.

4. NEURAL NETWORK ARCHITECTURE

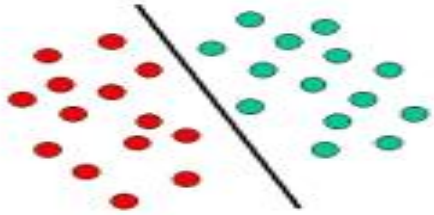
Layers: The basic architecture consists of three layers of neuron input, hidden, and output layers. In feed-forward networks, the signal flow is from input to output units, strictly in a feed-forward direction. The data processing can extend over multiple units, but no feedback connections are present. Recurrent networks contain feedback connections. Contrary to feed-forward networks, the dynamical properties of the network are important.

In some cases, the activation values of the units undergo a relaxation process such that the network will evolve to a stable state in which these activations do not change anymore.

In other applications, the changes of the activation values of the output neurons are significant, such that the dynamical behavior constitutes the output of the network. There are several other neural network architectures, depending on the properties and requirement of the application.

Support Vector Machine:

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED.



The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function, so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, the method is likely to give poor performances.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

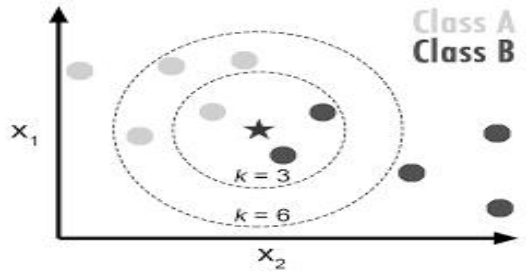
5. K-NEAREST NEIGHBORS

The **k-Nearest Neighbors algorithm** is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is

the average of the values of its k nearest neighbors.

- KNN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The KNN algorithm is among the simplest of all machine learning algorithms.
- Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.
- The neighbors are taken from a set of objects for which the class or the object property value is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.



The algorithm can be summarized as:

- A positive integer k is specified, along with a new sample.
- We select the k entries in our database which are closest to the new sample.

CONCLUSION

This paper gives a general introduction of data mining, the process of discovering interesting knowledge from large amounts of data stored in information repositories. It also discusses background on data mining and methods to integrate uncertainty in data mining such as K-means algorithm. It is also shown that data mining technology can be used in many areas in real life including biomedical and DNA data analysis, financial data analysis, the retail industry and also in the telecommunication industry. One of the biggest challenges for data mining technology is managing the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. Future research will involve the development of new techniques for incorporating uncertainty management in data mining.

REFRENCES

[1]. P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scientific, 1996.

[2]. P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp. 9-15, New York, Aug. 1998.

[3]. Christopher J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, v.2 n.2, p.121-167, June 1998.

[4]. Fayyad, Usama, Gregory Piatetsky Shapiro, Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, 1996.

[5]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, London: Academic Press, 5, 2001.

[6]. Kantardzic, Mehmed, Data Mining: Concepts, Models, Methods, and Algorithms, New York: John Wiley & Sons Inc publishes, 2003.

[7]. P. Domingos and G. Hulten. "Mining highspeed data Streams" In Knowledge Discovery and Data Mining, 2000, pp 7180.

[8]. M. B. Harries, C. Sammut, and K. Horn. "Extracting hidden context. Machine Learning", 32(2): 1998. pp 101–126

[9]. Micheline Kamber" Data Mining: Concepts and Techniques", Second Edition Jiawei Han University of Illinois at Urbana-Champaign ISBN 13: 978-1-55860-901-3