ISSN:0975-9646

Review Techniques of Knowledge mining

Reema Thareja,Charu Sharma,Ridhi Gupta,Prerna Goel,Kriti singhal

Department of Computer Science, Shyama Prasad Mukherjee College,University Of Delhi,
Punjabi Bagh, Newdelhi, India.

***Abstract*— "Knowledge mining is the extraction of useful, often previously unknown information from large databases or data sets." In this paper, we have described the need of knowledge mining and how knowledge mining technology can be effectively applied in extracting useful information from available huge amount of data. Knowledge mining is principally concerned with the quantitative synthesis and visualization of research results and findings.**

***Keywords*—** Knowledge mining, Framework, Linear Regression, Classification, Clustering, Decision trees, Entropy.

1.INTRODUCTION

Like a biological organism evolves, a non-cellular structure such as an organization does too. The difference between them is merely the term- organic. On the other hand, if an organization does not have an organic component it won't be existing. They both have to evolve and you might have heard the phrase- "survival of the fittest" given by Darwin. So the question is how a non-cellular, non-living organization evolves? How does it remain the "fittest"? An organization has to face fierce competition to maintain and upgrade its level in the market. How does it do that? The answer is simple but not easy- it has to innovate continuously. They have huge piles of data lined up in databases. It is their job to use this data to the organization's benefit. They need something more powerful than just sorting to make this data useful knowledge. So people need to create a platform in which newly covered ideas can be synthesized into knowledge for continuous innovation.

With an increase in the amount of data, there is a need to extract useful knowledge. Data in this huge amount cannot be managed by humans, so intelligent systems are required for maintaining & updating data and for extracting meaningful knowledge from it. Major research is being done to make some tools that can discover useful and large patterns from the database.

DATA —> PATTERNS[4]

The extracted pattern is not only strong but useful if it represents and understands user's goals [4].

For deriving new concepts, new patterns from the data and finding a more simpler solutions to the problem, we must have some previous knowledge with the data. Otherwise it will never find new paths.

DATA + PREVIOUS KNOWLEDGE + GOAL
—> NEW KNOWLEDGE[4]

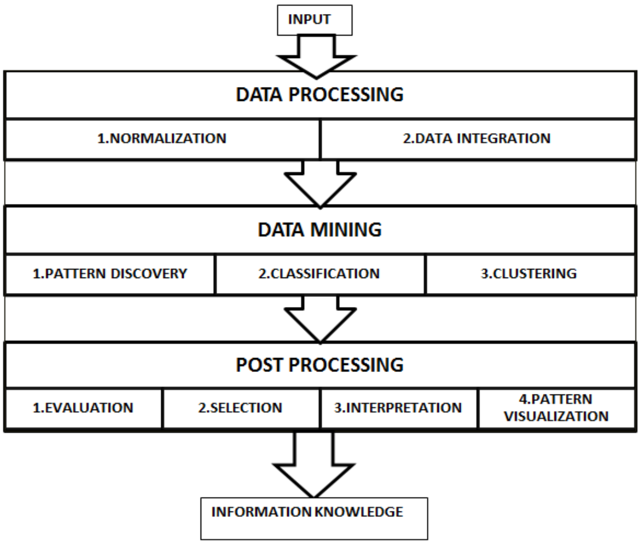
There are various regions where this knowledge can be used Therefore data mining system must be capable of generating different kinds of knowledge from a given data source. These programs are arranged in toolboxes which mostly need manual start up. It becomes very time consuming and laborious. To overcome this, many data mining tools are being clubbed into one but they need to invoked by high level knowledge generation language. Knowledge visualization is also very important since now a day people are trying to understand data mining results.

To address above mention tasks, knowledge mining is used. So knowledge mining is *deriving new knowledge from the database using prior knowledge by developing wide range of data analysis methods*. The algorithms should not only be effective but easy to understand and implement.

Knowledge discovery in databases(KDD) is accomplished in several steps. The pattern extraction phase of KDD is known as data mining and it can take on several types**.**

2.FRAME WORK OF KDD

The various steps of KDD are:



Data Selection

A proper human interaction is required in the processing of KDD. The way in which the subset and the data set are to be selected requires knowledge of the domain from which the data is to be taken. The motive is to remove the unwanted information from our dataset which helps in reducing the search space during data mining process.

Pre Processing

There are cases when the databases suffer from loss of data or incorrect data. During pre-processing outliers are removed if they are not required and techniques are chosen for handling of missing data and the normalization of data is done.

For example, following formula can be used for normalizing.

$$\{Zi=xi-\min(x)/\max(x)-\min(x)\}.$$

Where x = (x1,...,xn) and zi is the ith normalized data point.

Transformation

Transformation helps in reducing various varieties of data elements and at the same time preserving the quality of the data. At this step our data becomes organized and new attributes are defined.

Data Mining

Different data-mining techniques or models can be used depending on the expected outcome for example the decision tree model, which we will cover shortly.

Evaluation

In the last step documentation and interpretation of the outcomes is done. The knowledge retrieved is transformed in a more user friendly format.

TECHNIQUES USED IN KNOWLEDGE MINING

Linear regression

Regression is a function used to predict a number like age, height, weight. You can predict for example: children's weight given his/her height, age etc.

To start the regression task, you should have a data set and the target values should be known. If you want to predict children heights, then you may want to have observed data for children over a period of time.

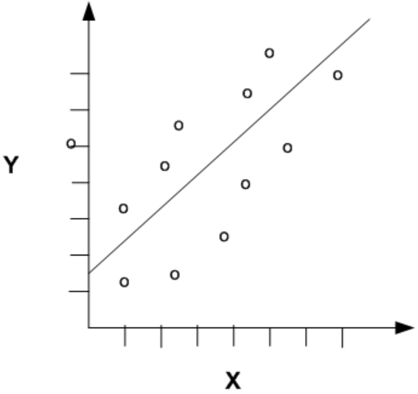
As it a function, there must be some parameters. So in order to predict your "number", you should determine the value of the parameters that causes the function to best fit for the data set. It shows that regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x_1 , x_2 , ..., x_n), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (e).[1]

$$\mathbf{y = F(x,\theta) \; + e} \; .$$

You should be able to find the best parameter values so as to minimize the error.

Now, ***linear regression*** is the simplest form of regression with just one predictor (x).

This graph shows the linear relationship between x and y.



[1]

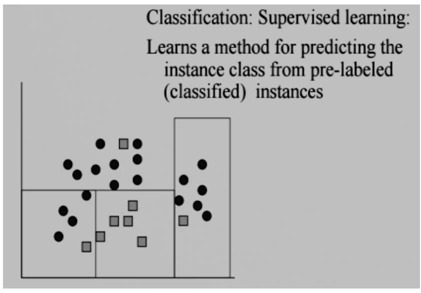
In this case, our equation changes to $y=\Theta_2x+\Theta_1$ and our parameters (coefficients) are:

Slope(Θ_2): angle between straight line and data set.

Y intercept(Θ_1): point where x=0.

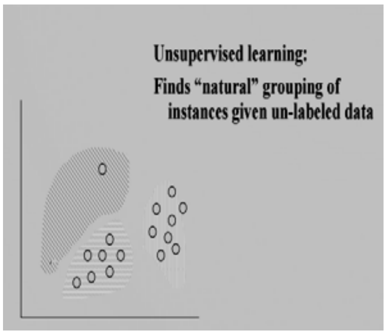
Clustering

The process of grouping physical or abstract objects into classes of similar objects is called clustering. Classes should be made from the data. This is different from classification in a way that in classification, classes are already defined.


Classification: Supervised learning:
Learns a method for predicting the instance class from pre-labeled (classified) instances

[2]

Fig 1: Classification

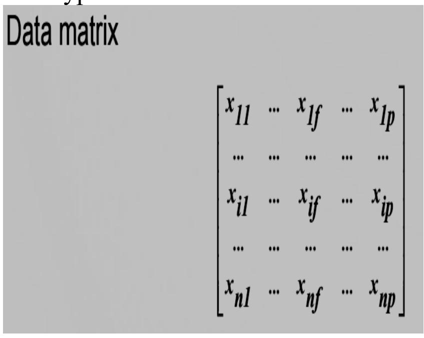

Unsupervised learning:
Finds "natural" grouping of instances given un-labeled data

[2]

Fig 2: Clustering

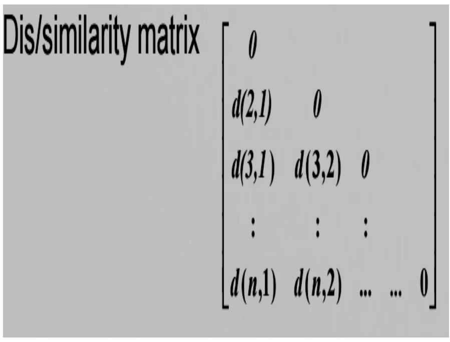
Clustering helps to understand the natural grouping in data. Clustering patients according to the symptoms they show can be a valid example. Identifying groups of houses according to their house type, value, and geographical location[2] is also an example of clustering [3].

There are two types of data structures used in clustering:


Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

[2]


Dis/similarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

[2]

Where d(i,j) is the distance function.

Attributes like Interim scaled factors, Twofold factors, Factors of blended sorts are utilized as a part of grouping. Grouping techniques can be further isolated into-

- *'Partitioning algorithms*: Construct various partitions and then evaluate them by some criterion.

- *Hierarchy algorithms*: Create a hierarchical decomposition of the set of data (or objects) using some criterion.
- *Density-based*: based on connectivity and density functions.
- *Grid-based*: based on a multiple-level granularity structure
- *Model-based*: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.'[2]

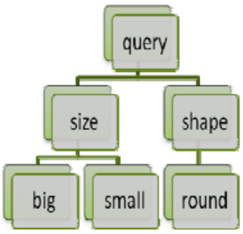
It consists of assigning items in a collection to target categories or classes.it is the most widely used KDD approach[3]. There are many classification techniques. Decision tree technique is explained here thoroughly.

Classification

It consists of assigning items in a collection to target categories or classes.it is the most widely used KDD approach [3]. There are many classification techniques. Decision tree technique is explained here thoroughly.

Decision tree builds classification or regression models in the form of a tree structure.

Example:

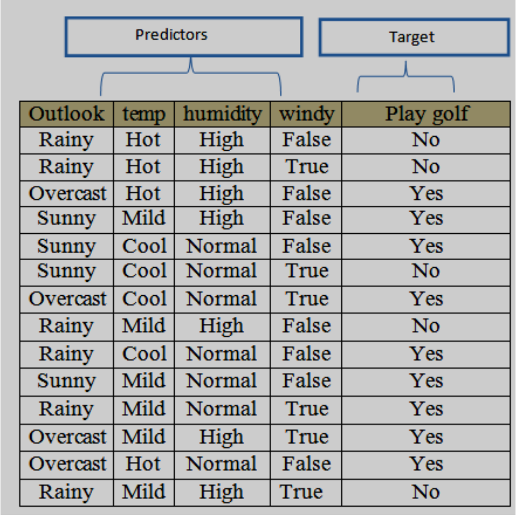


It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy).

Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called **root node**.

Decision trees can handle both categorical and numerical data.


Predictors
Target

Outlook	temp	humidity	windy	Play golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

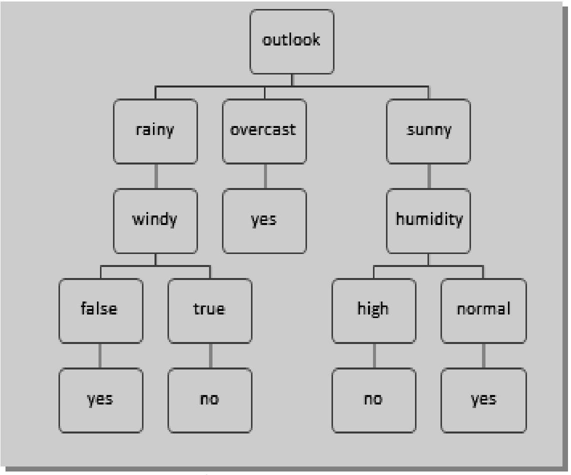


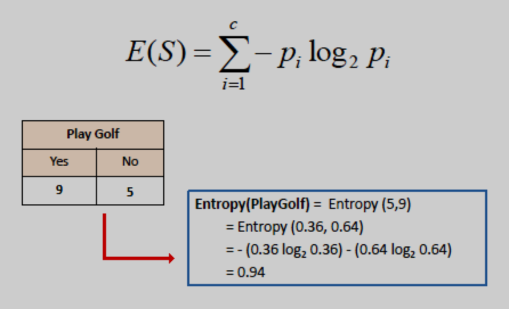
Fig:Decision tree

Entropy

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with homogenous values. ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attribute:



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

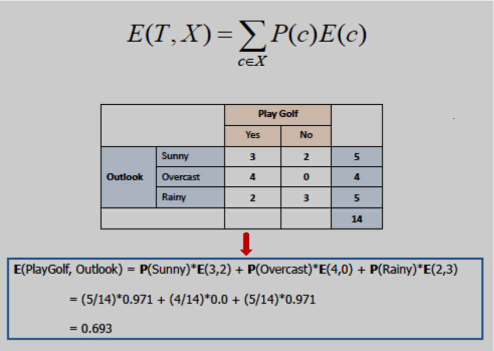
Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

b) Entropy using the frequency table of two attributes:



$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain(i.e., the most homogeneous branches)

3. Future Scope And Conclusion

Data mining that is the process of Knowledge Discovery becomes one of the best tools on advanced statistics, machine learning, Artificial Intelligence, Pattern Matching and Computation Capabilities in The Business Field. In the future, it is very likely that Knowledge Mining becomes predictive analysis (agosta, 2004) Data Mining's applications that will enrich human life in various fields such as Business, Education, Medical field, Scientific field , Politics, include

-Detecting Eco-System Disturbances.

-Distributed Data Mining. Distributed algorithm is developed for association analysis such as parallel decision tree construction.

Knowledge Mining is an interdisciplinary research area.

Thus, in the future, KM development may need integration with different technologies and demand more methodologies to solve KM problems.[6]

Management of knowledge resources has become a strong demand for development. Discovering the useful knowledge has also significant approach for management and decision making. The techniques/methods of knowledge mining which are discussed in this paper can be used for the same.

References

[1]. Oracle® Data Mining Concepts, 11g Release 1 (11.1), Part Number B28129-02

[2]. http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

[3]. https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm

[4]. Knowledge mining: A proposed new direction by Ryszard S. Michalski.

[5]. http://www.saedsayad.com/decision_tree.html

[6]. http://www.ijser.org/paper/Data-Mining-and-Its-Applications-A-Review.html