J. Glob. Innov. Agric. Soc. Sci., 2015
ISSN (Online): 2311-3839; ISSN (Print): 2312-5225
DOI: xxxxxx
http://www.jgiass.com

A Review Paper on Various Data Mining Techniques

Hafiz Muaaz Hamid[1], Hafiz Muneeb Ahmad[2], Zahid Javed[3] and Tariq Shehzad[4]

Department of Computer Science, University of Agriculture, Faisalabad.

***Corresponding author's email:** muaaz_hamid2000@yahoo.com

Presently, a awfully great deal of knowledge keep in databases is increasing at an incredible speed. this needs a desire for brand new techniques and tools to help humans inmechanically and showing intelligence analyzing giant information sets to accumulate helpfulinfo. This growing want provides a read for a brand new analysis field known as knowledgeDiscovery in Databases (KDD) or data processing, which are a magnet for a attention from researchers in many alternative fields as well as info style, statistics, pattern recognition, machine learning, and information mental image. methoding} is that the process of discovering perceptive,fascinating, and novel patterns, yet as descriptive, apprehensible and prognostic models from large-scale information. during this paper we tend to overviewed totally different tasks includes in data processing. data processing involves the tasks like anomaly detection, classification, regression, association rule learning, account and cluster.

Keywords: data processing, classification, clustering, association rules

INTRODUCTION

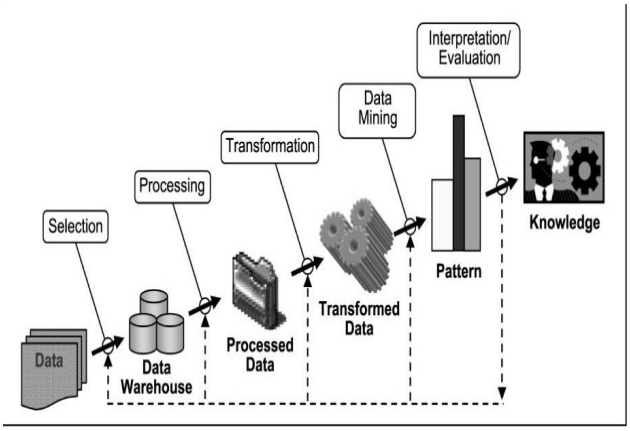
The last decade has practised a revolution in info accessibility and exchange of it through web. within the same strength additional business in addition as organizations began to gather knowledge associated with their own operations, whereas the info applied scientist are seeking economical mean of storing, retrieving and manipulating knowledge, the machine learning community centered on techniques that used for developing, learning and getting information from the info. data {processing} is that the process of analysing knowledge from completely different views and summarizing it into helpful info.

Data mining consists of extract, transform, and cargo dealing knowledge onto the info warehouse system, store and manage the info during a third-dimensional info system, by victimization application computer code analyse the info, give knowledge access to business analysts and data technology professionals, gift the info during a helpful format, sort of a graph or table. data processing involves the anomaly detection, association, classification, regression, rule learning, account and clump.

II. DATA MINING

Data mining is that the exploration and analysis of huge knowledge sets, so as to get important pattern and rules. The key plan is to search out effective thanks to mix the computer's power to method the information with the human eye's ability to notice patterns. The target {of knowledge|of knowledge|of information} mining is to style and work expeditiously with massive data sets. {data mining| data methoding} is that the part of wider process referred to as data discovery from info. [4]. {data mining|data methoding} is that the process of analysing knowledge from totally different views and summarizing the results as helpful info. It's been outlined as "the nontrivial method of characteristic valid, novel, doubtless helpful, and ultimately comprehensible patterns in data"

The definition {of data|of data|of information} mining is closely associated with another normally used term knowledge discovery [2]. data processing is AN knowledge base, integrated info, AI, machine learning, statistics, etc. several aras of theory and technology in current era are databases, AI, data processing and statistics may be a study of 3 robust massive technology pillars. {data mining|datamethoding} may be a multi-step process, needs accessing and making ready knowledge for a mining the information, data processing rule, analyzing results and taking applicable action. The data, that is accessed are often hold on in one or a lot of operational databases. In {data mining| data methoding} the information are often mined by passing numerous process.



Steps in Data Mining process

In data processing the information is deep-mined mistreatment 2 learning approaches i.e. supervised learning or unsupervised learning [5].

A. supervised Learning

In supervised learning (often additionally referred to as directed information mining) the variable sunder neath investigation will be split into 2 groups: informative variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the variable and informative variables the because it is completed in multivariate analysis. To proceed with directed data processing techniques the values of the variable should be notable for a sufficiently giant a part of the information set.

B. unsupervised Learning:

In unsupervised learning, all the variables area unit treated in same method, there's no distinction between dependent and informative variables. However, in distinction to the name rudderless data processing, still there's some target to realize. This target may be as information reduction as general or a lot of specific like bunch. The contrast between unsupervised learning and supervised learning is that the same that distinguishes discriminate analysis from cluster analysis. Supervised learning needs, target variable ought to be outlined which a adequate variety of its values area unitgiven unsupervised learning generally either the target variable has solely been recorded for too tinyvariety of cases or the target variable is unknown.

III. ISSUES IN DATA MINING

Data mining has evolved into a vital and active space of analysis attributable to the theoretical challenges and sensible applications related to the matter of discovering fascinating and antecedently unknown information from real-world databases. The most challenges to the infomining and therefore the corresponding concerns in coming up with the algorithms area unit as follows:

1. huge datasets and high spatiality.
2. Over fitting and assessing the applied math significance.
3. comprehensibility of patterns.
4. Non-standard incomplete knowledge and knowledge integration.
5. Mixed ever-changing and redundant knowledge.

IV. TASKS OF DATA MINING

Data mining as a term used for the precise categories of six activities or tasks as follows:

1. Classification
2. Estimation
3. Prediction
4. Affinity grouping or association rules
5. Clustering
6. Description and visualisation

The first 3 tasks - classification, estimation and prediction rules are samples of directed data processing or supervised learning. In directed data processing, the goal is to use the out there knowledge to create a model that describes one or a lot of specific attribute(s) of interest (target attributes or category attributes) in terms of the remainder of the out there attributes. successive 3 tasks – association rules, agglomeration and outline are sample of purposeless data processing i.e. no attribute is singled out because the target, the most goal is to determine some relationship among all attributes [6].

A. Classification

Classification consists of examining the options of a recently conferred object and distribution there to a predefined category. The classification task is characterized by the well-defined categories, and a coaching set consisting of reclassified examples. The task is to create a model which will be applied to unclassified knowledge so as to classify it. samples of classification tasks include: • Classification of credit candidates as low, medium or high risk • Classification of mushrooms as edible or toxic • Determination of that home phonephone lines are used for net access

B. Estimation

Estimation deals with ceaselessly valued outcomes. Given some computer file, we have a tendency to use estimation to return up with a price for a few unknown continuous variables like financial gain, height or master card balance. Some samples of estimation tasks include: • Estimating the amount of youngsters in a very family from the computer file of mothers' education • Estimating totalmenage financial gain of a family from the information of vehicles within the family • Estimating the worth of a bit of a true estate from the information on proximity of that land from a significantbusiness centre of town.

C. Prediction

Any prediction will be thought of as classification or estimation. The distinction is one in all stress. once data processing is employed to classify a subscriber line as primarily used for web access or a master card dealings as deceitful, we do not expect to be able to return later to envision if the classification was correct. Our classification could also be correct or incorrect, however the uncertainty is attributable to incomplete data only: enter the important world, the relevant actions have already taken place. The phone is or isn't used primarily to dial the native ISP. The mastercard dealings is or isn't dishonest . With enough efforts, it's attainable to visualize. prognosticative tasks feel totally different as a result of the records square measure classified in step with some expected future behaviour or calculable future worth. With prediction, the sole thanks to check the accuracy of the classification is to attend and see. samples of prediction tasks include: • Predicting the scale of the balance can|which can|that may} be transferred if a mastercard prospect accepts a balance transfer provide • Predicting that customers will leave at intervals next six months • Predicting that phonephone subscribers will order a value–added service like conference call or voice mail. Any of the techniques used for classification and estimation will be adopted to be used in prediction by mistreatment coaching examples whereverthe worth of the variable to be expected is already well-known, at the side of historical knowledgefor those examples. The historical knowledge is employed to make a model that explains the present discovered behaviour. once this model is applied to current inputs, the result's a prediction of future behaviour [7]

D. Association Rules

An association rule could be a rule which means certain association relationships among a collection of objects (such as "occur together" or "one implies the other") in an exceedingly info. Given a collection of transactions, wherever every dealing could be a set of literals (called items), associate degree association rule is associate degree expression of the shape X Y , wherever X and Y area unit sets of things. The intuitive that means of such a rule is that transactions of the info that contain X tend to contain Y. associate degree example of associate degree association rule is: "30% of farmers that grow wheat additionally grow pulses; a pair of all farmers grow each of those items". Here half-hour is termed the arrogance of the rule, and a couple of the support of the rule. The matter is to search out all association rules that satisfy user-specified minimum support and minimum confidence constraints.

E. Clustering

Cluster analysis are often used as a standalone data processing tool to achieve insight into the information distribution, or as a pre-processing step for alternative data processing algorithms in operation on the detected clusters. Several clump algorithms are developed and area unit classifiedfrom many aspects like partitioning strategi es, ranked strategies, density-based strategies, and grid-based strategies .Further knowledge set are often numeric or categorical. clump is that the task of segmenting a various cluster into variety of comparable subgroups or clusters. What distinguishes clump from classification is that clump doesn't believe predefined categories. In clump, there aren't any predefined categories. The records area unit sorted along on the idea of self similarity. clump is usually done as a prelude to another sort of data processing or modelling. as an example, clumpcould be the primary step in an exceedingly market segmentation effort, rather than making an attempt to return up with a one-size-fits-all rule for crucial what quite promotion works best for everycluster[6].

1) General Types Of Clusters:

- Well-separated clusters: A cluster could be a set of purposes in order that any purpose in a verycluster is nearest (or a lot of similar) to each alternative purpose within the cluster as compared to the other point that's not within the cluster.
- Center-based clusters A cluster could be a set of objects such Associate in Nursing object in a very cluster is nearest (more similar) to the "center" of a cluster, than to the middle of the othercluster. the middle of a cluster is commonly a center of mass.
- Contiguous clusters A cluster could be a set of purposes in order that a degree in a very cluster is nearest (or a lot of similar) to 1 or a lot of alternative points within the cluster as compared to any point that's not within the cluster.
- Density-based clusters A cluster could be a dense region of points, that is separated by in keeping with the low-density regions, from alternative regions that's of high density.
- Shared Property or abstract Clusters Finds clusters that share some common property or representa selected construct.

F. Description and mental image

Data mental image could be a powerful variety of descriptive data processing. it's notcontinuously straightforward to come back up with purposeful visualizations, however the correct image extremely are often value m association rules since the people in general area unitextraordinarily practiced at extracting that means from visual scenes. Information discovery goalsarea unit outlined by the supposed use of the system. There area unit 2 varieties of goals: (1) verification and (2) discovery. With verification, the system is proscribed to corroborative the user's hypothesis. With discovery, the system autonomously finds new patterns. the invention goal is anydivided into prediction, wherever the system finds patterns for predicting the long run behaviour of some entities and outline, wherever the system finds patterns for presentation to a user in humancomprehendible type.

V. CONCLUSIONS

Data mining involves extracting helpful rules or fascinating patterns from large historical information. Several data processing tasks are out there and every of them additional has several techniques. data processing is Associate in Nursing knowledge base, AI integrated info, machine learning, statistics, etc. data processing could be a sizable amount of incomplete, noisy, fuzzy, random application of the information found in hidden, regularity that don't seem to be known by individuals beforehand, however is doubtless helpful and ultimately intelligible data and information of non-trivial method during this paper we have a tendency to discusses some problems in data processing and activities used for data processing task.

REFERENCES

[1] Amandeep Kaur Mann, Navneet Kaur, Survey Paper on Clustering Techniques, IJSETR,2278 – 779.

[2] Pavel Berkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002.

[3] Oded Maimon, Lior Rokach, Data Mining AND Knowlwdge Discovery Handbook, Springer Science + Business Media.Inc, pp.321-352, 2005.

[4] Han, J., Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001

[5] K.Kameshwaran, K.Malarvizhi, Survey on Clustering Techniques in Data Mining, IJCSIT, Vol. 5, 2014, 2272-2276 [6] Aastha Joshi, Rajneet Kaur, A Review: Comparative Study of Various Clustering Techniques in Data Mining, IJARCSSE, Vol. 3, 2013, 2277 128X

[7] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, A Comparative Study of Various Clustering Algorithms in Data Mining,, International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012.

[8] Pradeep Rai, Shubha Singh, A Survey of Clustering Techniques, International Journal of Computer Applications, 2010.