International Journal of Electrical Engineering and Technology (IJEET)
Volume 12, Issue 7, July 2021, pp. 16-22, Article ID: IJEET_12_07_003
Available online at https://iaeme.com/Home/issue/IJEET?Volume=12&Issue=7
ISSN Print: 0976-6545 and ISSN Online: 0976-6553
https://doi.org/10.34218/IJEET.12.7.2021.003

PROPOSED HIGHLY ROBUST AND EFFICIENT DATA-MINING MODEL FOR KNOWLEDGE DATA DISCOVERY PROCESS

Abhishek Gupta [1], Cypto [2], Dr. Arun Sahayadhas [3], Dr. P. Karthikeyan [4]

[1] Department of Computer Science and Engineering,
Vels Institute of Science, Technology and Advanced Studies (VISTAS), Tamilnadu, India

[2] Department of Computer Science and Engineering,
Madras Institute of Technology (MIT), Tamilnadu, India

[3] Department of Computer Science and Engineering,
Vels Institute of Science, Technology and Advanced Studies (VISTAS), Tamilnadu, India

[4] Department of Production Technology,
Madras Institute of Technology (MIT), Tamilnadu, India

ABSTRACT

Today we are living in the age of data-science where we have to handle plenty amount of heterogeneous data-sets. Sometimes it can be as simple as comma separated value (CSV) files where as it can be complex like images also. Apart from that even if it is too simple like csv it's very tedious task to grab the statistics of the same that can be utilized for the research purposes. So, we have devised a statistical model to grab better statistics on the same. This model is used to capture statistics based on csv, excels, images etc. using classification and/ or clustering algorithm using a control statement/ keyword. Furthermore, the same can be used for the purpose of devising the association rule-mining using Apriori algorithm.

Key words: Data-Mining, Classification, Clustering, Association Rule Mining, Apriori Algorithm.

Cite this Article: Abhishek Gupta, Cypto, Arun Sahayadhas, P. Karthikeyan, Proposed Exploratory Data Analysis Model to Statistically Analyze the Data-Set for Research Purposes, *International Journal of Electrical Engineering and Technology (IJEET)*, 12(7), 2021, pp. 16-22.
https://iaeme.com/Home/issue/IJEET?Volume=12&Issue=7

1. INTRODUCTION

For the purpose of knowledge discovery data-mining is used, which performs mining on the available data-sets, that have been collect either primary or secondary data collection methods.

So, this whole process from data gathering or say data-collection till knowledge discovery is known as "data-mining SDLC (Software Development Life Cycle)".

Diagram, to show data processing steps in data-mining SDLC: -

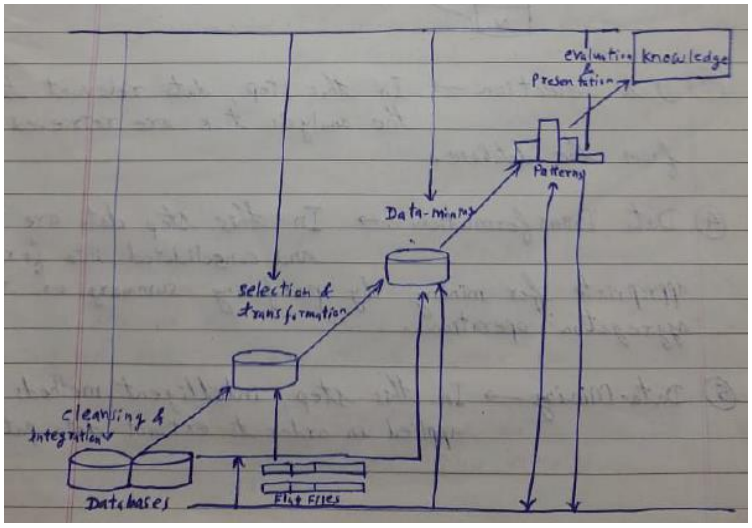

Figure 1 Data-Mining SDLC

Explanation of each step followed in the diagram: -

1. Data cleansing: - This step, removes noise and in-consistent data from data-set being examined.
2. Data integration: - In this step multi-source's data or in other words data coming from various sources are combined.
3. Data selection: - In this step, data relevant to the analysis task are retrieved from the database.
4. Data transformation: - In this step data are transformed and consolidated into forms appropriate for mining by performing summary or the aggregation operation.
5. Data mining: - In this step, intelligent methods are applied in order to extract data patterns.
6. Some examples of intelligent methods are: -
   a) Association: - It refers to uncovering relationship among data
   b) Classification: Decision tree – It maps data into predefined or say predetermined groups or classes. It is represented by multiple ways: -
      i) Decision tree
      ii) Classification (IF-Else) rule etc.
   c) Clustering: - It is very similar to classification except that groups are not predefined in this we put data having similar characteristics to one group, and dissimilar data to another group.
   d) Regression: - It is used to map a data-item to a real valued prediction variable.
   e) Pattern evaluation: - In this step, data patterns are evaluated. To identify the truly interesting patterns, represent knowledge based on interesting measures.
   f) Knowledge presentation: - In this step, knowledge is represented by the help of various visualization and knowledge representation techniques are used to present mined knowledge to users.

Classification

- "Designing and training the model structure" to perform classification and prediction data-mining tasks.
- Given a collection of records (training set)

Each record contains a set of “attributes”, one of the attributes is “class”.

- Find (“learn”) a model for the class attribute a function of the values of the other attributes.
- Goal: - “Previous unseen records should be assigned a class as accurately as possible”.
- Learning: - We can think of at least 3 different problems being involved in learning

1. Memory
2. Averaging
3. Generalization

Example: - Imagine that i am trying to predict whether my neighbour is going to drive into work, so i can ask for a lift.

So, we have observed, whether she drives into work seems to depend on the: - Temperature, expected precipitation, day of week, what she is wearing.

We observe our neighbour for 3 day and drive data-set: -

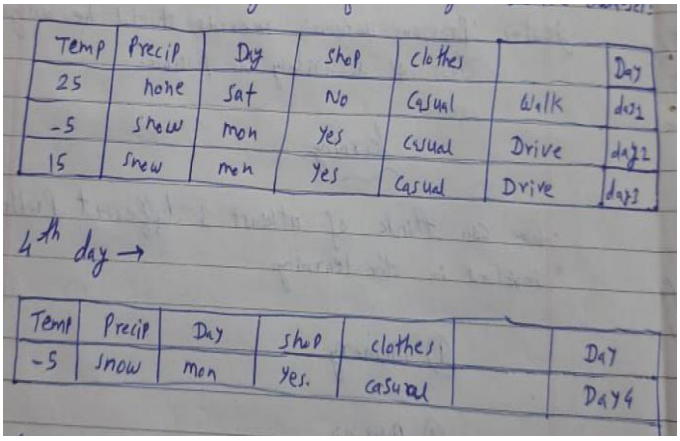

Figure 2 Training and test data-set

Now we have to predict, via above classification => its “drive”, (memory concept). So, it can be done easily

So, this can be done through the below one, which is self-explanatory in itself: -

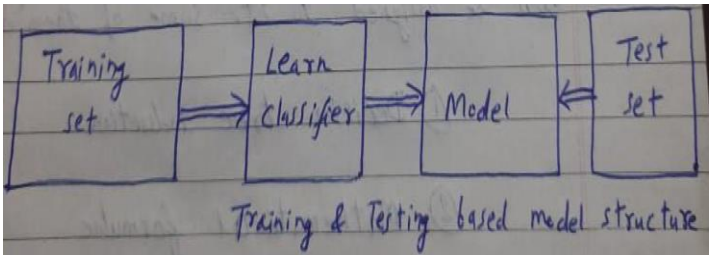

Figure 3 Training and testing based model structure

Clustering

Requirements of Clustering

1. High-dimensionality: - It should not only be able to handle the low dimensional data but also high dimensional data.
2. Ability to deal with noise data.
3. Interpretability and comprehensiveness.
4. Scalability.

5. Ability to deal with different kind of attributes.
6. Discovery of cluster with attribute shape.

Cluster-Analysis Techniques/ Methods in Data-Mining

1. Partitioning method: It is self-explanatory in itself: -

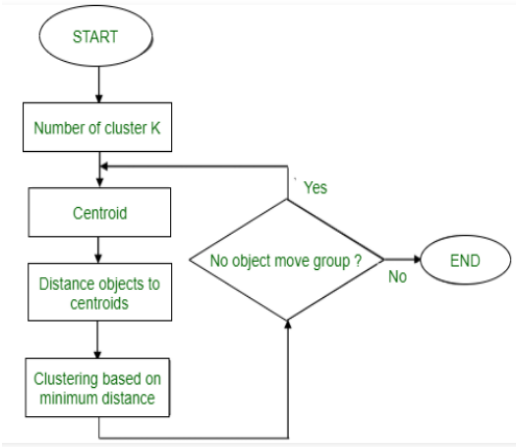

Figure 4 Partitioning Method

2. Hierarchical method: It is self-explanatory in itself: -

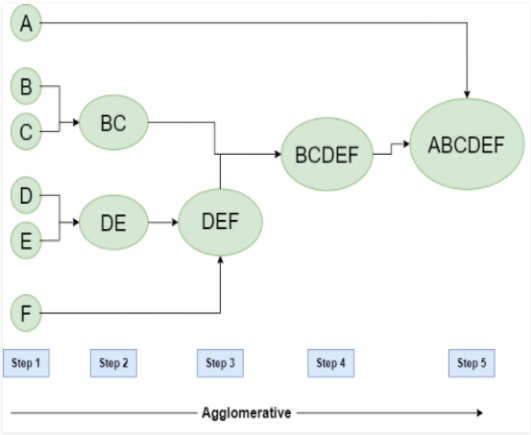

Figure 5 Hierarchical method

3. Density-based method: It is self-explanatory in itself: -

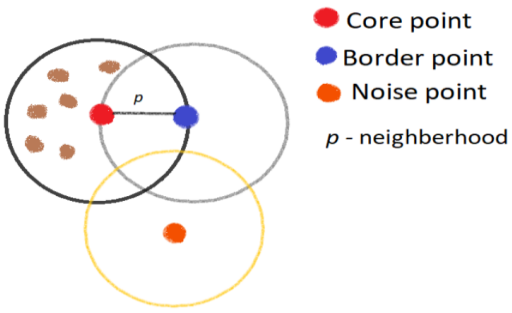

Figure 6 Density-based method

4. Grid-based method: It is self-explanatory in itself: -

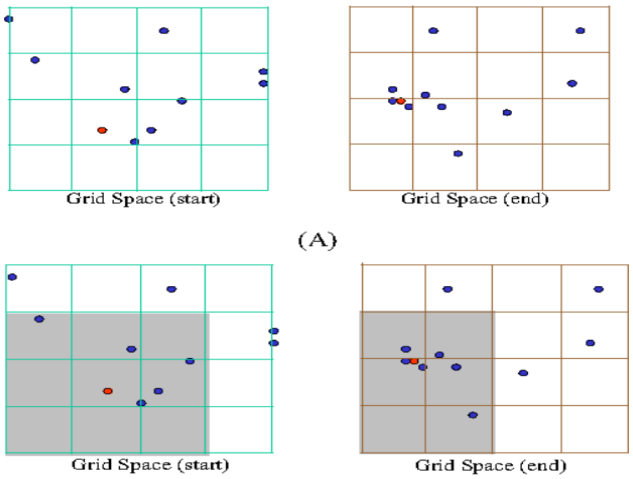

Figure 7 Grid-based method

5. Model-based method: It is self-explanatory in itself: -

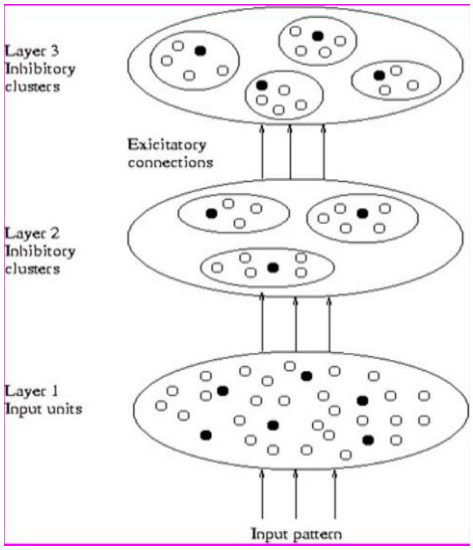

Figure 8 Model-based method

6. Constraint-based method: It is self-explanatory in itself: -

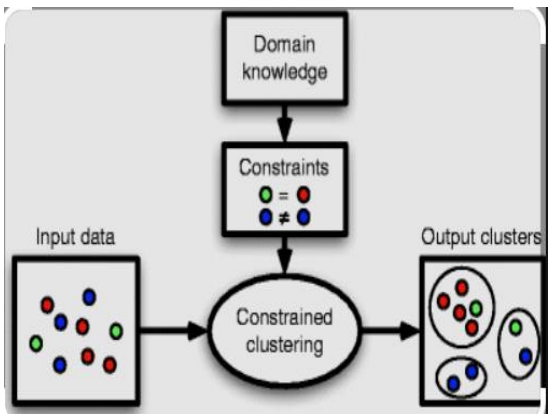

Figure 9 Constraint based method

Association Rule Mining using Apriori Algorithm using Example

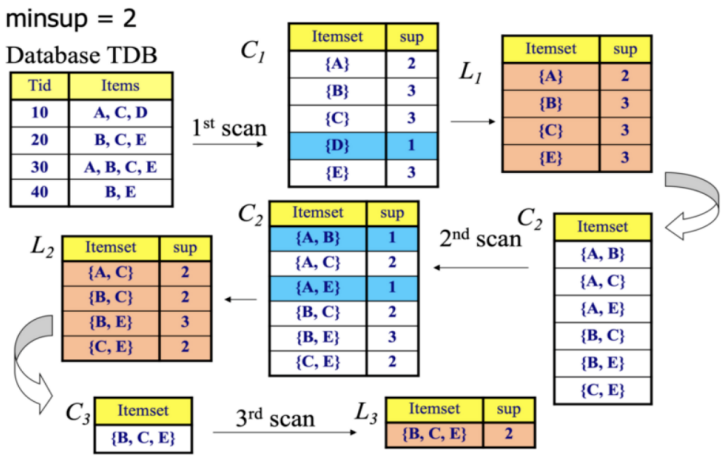

Figure 10 Apriori algorithm

2. PROPOSED TECHNIQUE

Below we have shown the diagram of the proposed model: -

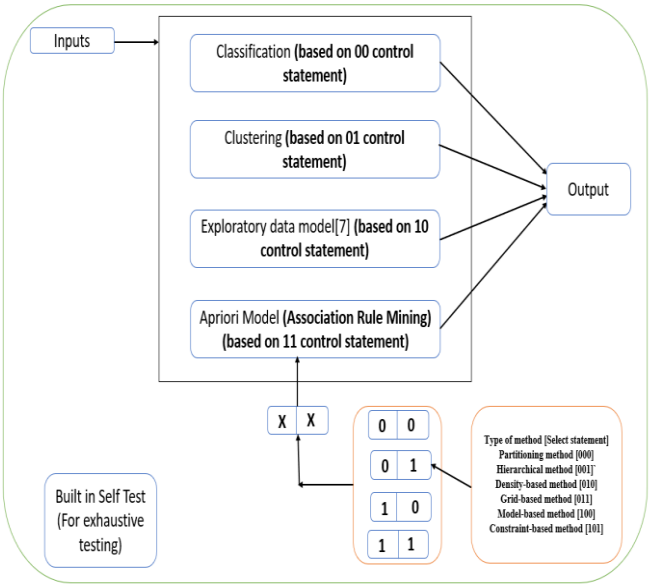

Figure 11 Proposed model

In this model we have used the 2-bit control statement to precisely select the model to work upon. Furthermore, we have used the 3-bit control statement for selection of the type of method out of six clustering methods.

3. RESULTS AND DISCUSSIONS

In this research paper we have proposed a model to perform the classification and clustering on the incoming data-set based on the control unit to have better insights on the data-set so that the data can be mined for the better outcomes. Furthermore, for having more significant results we have proposed to use the EDA model described in [7], which can give precise results for the comma separated files. Along with that we have proposed to use the Apriori algorithm for the association rule mining with top support and confidence matrices.

4. CONCLUSION

In this research paper we have worked on data-mining/data-analysis for gathering the significant insights using Knowledge Data Discovery (KDD) process where classification and/ or clustering algorithms has been used for the mining of data and Apriori algorithm has been used for association rule mining.

CONTRIBUTORS

Abhishek Gupta: - He is permanent employee of Accenture Services Pvt. Ltd. as Application Development Senior Analyst. Apart from that he is also part-time Ph.D. Student of VELS University. He is the idea generator and paper writer for this paper.

Cypto: - He is working as an assistant professor in DMI College of engineering Chennai and he is also doing his research in Anna University in the area of Deep Learning under the guidance of Dr. P. Karthikeyan. Cypto and Abhishek has written the code for the same.

Dr. Arun Sahayadhas: - He is permanent employee of VELS University most precisely VISTAS and Guide/ Mentor of Abhishek Gupta.

Dr. P. Karthikeyan: - He is working as an Assistant Professor in Department of Production Technology, Madras Institute of Technology, Chennai and Guide/ Mentor of Cypto.

REFERENCES

[1] "Proposed Techniques to Design Speed Efficient Data Warehouse Architecture for Fastening Knowledge Discovery Process" in *IEEE International Conference held by "University of California, Irvine*", 19 February 2021.

[2] "Proposed Techniques to Optimize the DW and ETL Query for Enhancing data warehouse efficiency", *IEEE International Conference on Computing, Communication & Security (ICCCS - 2020),* IIT Patna INDIA, Oct-2020.

[3] "A Comprehensive Survey to Design Efficient Data Warehouse for Betterment of Decision Support Systems for Management and Business Corporates", *International Journal of Management (IJM),* IAEME Publication, Volume 11, Issue 7, July 2020, pp. 463-471.

[4] "Review on Data warehousing Tools, Techniques and Terminologies", *International Journal of Multidisciplinary Educational Research book,* Volume 8, Issue 11(3), November 2019.

[5] "A Complete Reference for Informatica Power Center ETL Tool", in *International Journal of Trend in Scientific Research and Development (ijtsrd),* ISSN: 2456-6470, Volume-3 | Issue-2, February 2019, pp.1063-1070.

[6] "Design of speed, energy and power efficient reversible logic based vedic ALU for digital processors", *2012 Nirma University International Conference on Engineering (NUiCONE),* 2012, pp. 1-6, doi: 10.1109/NUICONE.2012.6493259.

[7] "Proposed Exploratory Data Analysis Model to Statistically Analyze the Data-Set for Research Purposes", *International Journal of Electrical Engineering and Technology (IJEET),* 12(7), 2021, pp. 11.