The Nucleus ***57****, No. 4 (2020) 135-140*

https://doi.org/10.71330/thenucleus.2020.1084

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)
ISSN 2306-6539 (Online)

A Discrete Formulation of Unified Data Mining Model

D.M. Khan[1], A.U. Rehman[2], M.U. Rehman[3*], F. Shahzad[1] and N. Saher[1]

[1]*Department of Computer Science and IT, The Islamia University of Bahawalpur, PAKISTAN*

[2]*Department of Information Technology, KFUEIT, Rahim Yar Khan, PAKISTAN*

[3]*Department of Computer Science, KFUEIT, Rahim Yar Khan, PAKISTAN*

ABSTRACT

A race is going on to process the complex and huge amount of data. To achieve this, data analytics are proposing different models and methods. Parallel to this, rich research work has been done to simplify different mathematical models for the validation and for the acceptance level of calculated knowledge. In this paper, we propose a discrete formulation of a unified data mining model. It envisages that knowledge extraction is a multi-step process where different data mining processes such as clustering, classification and visualization are unified in a cascade way; that is, an output of a process is the input to another process which helps to achieve scalability and flexibility on a larger scale. Simultaneously, to prove whether our proposed model is valid or invalid, it is evaluated by discrete structure. For this, different mathematical formulations are formed to support the cause and then these mathematical formulations are evaluated to achieve the required target. Each mathematical formulation is examined in detail by using a simple technique called Truth Table and its Truth Values. Truth Table shows that evaluated mathematical formulations are valid and correct.

Keywords**:** *Unified Theory (UT), Unified Data Mining Theory (UDMT), Unified Data Mining Framework (UDMF), Unified Data Mining (UDM), Unified Mining Theory (UMT)*

1. Introduction

Artificial Intelligence (AI) and Data Mining are used to train machines in such a unique way that it transforms data into knowledge and it is also worthwhile to solve a lot of issues associated from pattern recognition to knowledge extraction. Since data mining is not limited to a particular field; hence, it covers a massive amount of multi-disciplinary fields such as Robotics, Frauds Detection, Medical/Disease Detection, Text Mining, Web Mining, Self Driving Cars, National Security and many more [1-3]. However, there are numerous challenges in the field of data mining research itself that should be addressed. These challenges are Mining Unbalanced, Complex and Multi-agent Data, Data mining in Distributed and Network Settings, Issues of Security and Privacy of Data Integrity in data mining [4, 5]. In this paper, the emphasis is landed on the unified data mining model using discrete formulation.

In a broader aspect, data mining falls under the field of AI. Data mining could be categorized into four different ways; i.e, supervised, semi-supervised, unsupervised and reinforcement learning. Supervised data mining utilizes the target field of the dataset, whereas an unsupervised approach discovers relationships among data without using any labeled class [6, 7]. Semi-supervised is a collaborative venture of clustering and classification; whereas, reinforcement learning is truly a hit and trial learning method. It becomes very easy to extract knowledge when supervised and unsupervised techniques are integrated [8, 9] but the other two differ regarding process and nature. Data mining experts, researchers and scientists are trying to develop a unified data mining theory that may answer very fundamental questions related to the exploration of useful information.

The reason behind formulating a unified data mining model is that the available data mining algorithms perform a single-step process such as clustering, classification, visualization, regression and association rule learning; therefore, these are unable to produce enough appreciable knowledge. Now it is evident that the knowledge discovery is a multi-step process; therefore, our proposed model emphasizes that data mining tasks such as clustering, classification and visualization should be unified to quest the issues related to the discovery of knowledge. A discrete mathematical approach is applied to validate the proposed model. Followings are the steps of the proposed model:

i. Calculate patterns of interest from pre-cleansed data.
ii. Define rules set for data; which are designed according to Data Owner (DO) requirements and implemented on selected patterns.
iii. Subsequent data is represented in the form of 2D/3D graphs and the graphical representation sets the foundation for the extraction of knowledge [10-12].

Table 1 demonstrates that different mechanisms are applied to achieve different criteria and specified algorithms are chosen according to DO requirements. These models lack in different manners and also there exists some kind of ambiguities as well; such as, different processes like clustering, classification and visualization are not part of the model and some processes are even repeated [13-16]. These mechanisms are basic and limited in terms of scope and some modification in any process, at any level, leads to a massive amount of mathematical or statistical complexities. Further, level of risk also exists for DO. After analyzing different mechanisms and criteria, the following limitations are figured out.

*Corresponding author: mujeeb.rehman@kfueit.edu.pk

i. Though, association of data clustering and data classification helps for decision-making rules set, yet unification of model and feature change is not taken into consideration.

ii. Unification of clustering and visualization provides 2D or 3D graphs which facilitates to afford an additional chance to check the results. None of the analyzed models try to implement this rule of thought at any level.

Table 1: A Summary of Data Mining Techniques.

Criteria	Mechanism Applied	Theory	Remarks
Coverage of Data Mining Tasks	Data partitioning, aggregation and transformation.	Pattern set, iterative and interactive data mining.	Following techniques are used with repeated procedures: Clustering and visualization. Clustering and classification. A theoretical model of inductive Database. Data mining framework is defined by using set theory.
Implementing Data Mining Algorithm	SQL (Structured Query Language) implementation with data mining algorithms.	Different mining algorithms are used.	Different data mining algorithms implementation is provided with SQL, no mathematical proof is provided, the consequence is not clear which may provide invalid results.
Easy Understanding of Model from User Perspective	SQL queries are used to process data with different data mining algorithms.	Probability is used to draw clusters.	Mining tasks are not well elaborated, multiple procedures are used for data users. Statistical/ mathematical calculation makes it even more complex for data users and data owners.
Overview of Unified Modelling Framework	Knowledge extraction is connected with mining algorithms. Authors failed to implement theories although they achieved most of the goals.	Developing clusters with different characteristics and using them to extract knowledge.	Object mining, model-based clustering, item-set mining, nonlinear dimensionality, and multi-resolution indexing are used as separate procedures. Error in any of these procedures will strike the overall system or may lead to non-relevant decisions.
Conclusion	Simple and basic.	There is severe complexity for measuring calculations, data users do not recommend these.	Simple and basic in terms of mathematical issues, ambiguity exists in terms of operation and reliability.

2. Proposed Model: Discrete Structure of a Unified Data Mining Model (DSUDM)

Here Data mining is provided as a knowledge extraction process that works as an intermediary layer between a dataset and application and transforms data into useable knowledge. It is generally believed that theories based on Unified Theory for Formulation (UTF) were replaced with a modified version of Unified Data Theory (UDT). Some new features are added and proposed in UDM to maximize the scope of UDT. It is further refined to newer versions of UDMT and UMDM, which enhance the flexibility and scalability of the proposed methodology [11]. The proposed model is illustrated in Fig. 1.

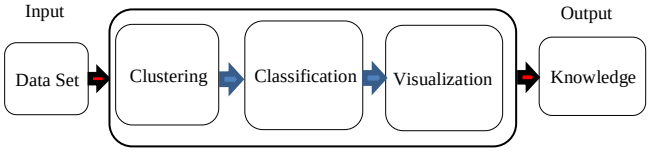

Fig. 1: A Unified Data Mining Model.

The proposed model is based on the unification of three data mining processes, i.e., clustering, classification and visualization as shown in Fig. 1. The assumption starts with the preparation of a dataset to the extortion of knowledge, which are given below:

Assumption 1: Formulate a universal dataset ‘D’, which is a basic input.

Assumption 2: From ‘D’, the clustering process is performed

Assumption 3: Formulate rules set for selected clusters for the same set of properties [18].

Assumption 4: To represent the relationship between unified data (after applying rule-set) and raw data (before applying rule-set), we plot data of each classifier from the rule-set on a 2D graph. This term is called data visualization [18].

Assumption 5: Sumrise knowledge from formulated visualized data.

3. Illustration of Proposed Model

In this section, the mathematical formulation of the proposed model is discussed in detail. The very basic concept behind this formulation is that different data formats such as text, number, tag and many others are expressed with the same notations [19]. A dataset called ‘D’ is used for knowledge extraction processes. This dataset contains raw data collected by using a different mechanism.

$$D = \{dataset\ of\ '\,n\,'\ attributes\}$$

$$D = \{d_1, d_2, d_3, \ldots d_n\} \quad (1)$$

Dataset ‘D’ is taken as input to UDM which includes 3 discrete processes such as clustering, classification and visualization. Depending upon organization’s requirements and nature of data [20], a number of different algorithms are selected. From dataset ‘D’, clusters of interest are nominated after applying the K-means data mining algorithm. The resulting set of clusters is expressed in a set called ‘C’.

$$C = \{set\ of\ clusters\ having\ '\,n\,'\ attributes\}$$

$$C = \{c_1 , c_2, c_3 \dots c_n\} \quad (2)$$

To classify these clusters, discrete data mining algorithms are applied and likewise different rules should be defined. These rules are further implemented that augment a new set 'R' into the system.

$$R = \{set\ of\ rules\ or\ classification\ \ of\ \ 'n'\ attributes\}$$

$$R = \{r_1, r_{2,}, r_3 \dots r_n\} \quad (3)$$

Classifying clusters with the help of rule-set leads to measurement of numeric values. These numeric values are plotted over 2D or 3D graphs; which help to express data in visual form and hence leads to extract knowledge. This evolution of UDM formulates another set 'V'.

$$V = \{set\ of\ visual\ data\ in\ form\ of\ 2d - 3d\ of\ \ 'n'\ attributes\}$$

$$V = \{v_1, v_{2,}, v_3 \dots v_n\} \quad (4)$$

The sets of equations shown above reveal that set 'V' is a subset of set 'R'; set 'R' is the subset of set 'C' and subsequently set 'C' is a subset of set 'D'. The process of evaluation with respect to user/organization and selected algorithms is explained in Fig. 2, where the data input is fed serially to cluster selection, data classification, data visualization, and finally the knowledge is extracted.

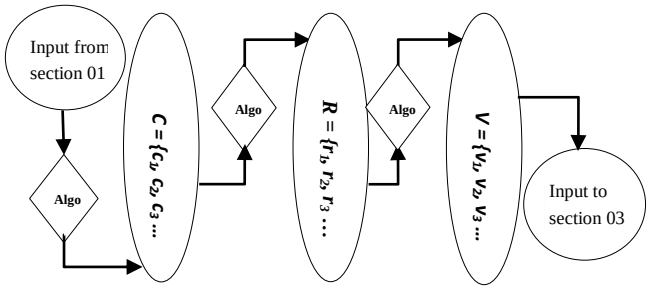

Fig. 2: Workflow of Unified Data Mining Model Processes.

Fig. 2 illustrates the proposed unified model of data mining, the abstraction of which is already depicted in Fig. 1. The mechanism elaborated in Fig. 2 is a multi-agent system [20], which is further divided into three associated processes. Each process is supported by an algorithm. In Fig. 2, algorithms are represented by a diamond shape and discrete processes such as clustering, classification and visualization of data are represented using oval shapes. To enhance proposed system support and also to provide a novelty in the proposed system, we emphasize an independent selection of algorithms. This opens a window for data owners, data users, or even for organizations, that they may select among the different ready-made algorithms or develop algorithms of their own. This also allows them to customize different pre-defined open-source algorithms [21]. Fig. 2 shows the data flow and data-set creation process in detail.

1. Basic data-input is provided to the selected algorithm which processes the data-set and hence results in a data-set of clusters. These clusters are either similar or dissimilar.
2. Data-set cluster is again processed by the selected algorithm which produces data-set of rules for the selected clusters for the same set of properties.
3. To visualize data in a more effective format, a data plotter of each classifier is demonstrated on the graph. In this regard, a relation is established among raw data and processed data. Since both algorithms and produced data-sets are concerned with the originality of data and extraction of knowledge through the proposed model which is demonstrated in set 'K'.

$$K = \{set\ of\ knowledge\ of\ \ 'n'\ attributes\}$$

$$K = \{k_1, k_{2,}, k_3 \dots k_n\} \quad (5)$$

We also formulate the evaluation criteria for the proposed model that helps to select 'knowledge' from its domain as shown in eq. (5). The Evaluation criteria are:

i. Compute the population of each cluster.
ii. Calculate the percentage of each parameter in a cluster.
iii. Determine the Minimal Description Length (MDL) value of each cluster.

3.1 Discrete formulation of equations 1, 2, 3, 4 and 5

The main advantages of the proposed model are scalability and flexibility which allow us to test different datasets by using different algorithms and also it supports the rule of refutation [5]. The proposed model supports all attributes and properties that are valid for itself. The datasets comprise of multiple objects and each object possesses its own properties and features [22]. Tarski's approach is used to prove the rule of thought, as mentioned elswhere [6, 8]. The set theory allows us to perform a different operation over sets and these operations are equal to logic operations of discrete mathematics [7]. By applying the above terminologies on datasets, we formulate a universal set from 'D', 'C', 'R', 'V' and 'K'. The universal set contains all of the elements of sets and is presented above as dataset members which are illustrated in eq. (6).

$$Dataset = \{D, \{C|a\mathcal{E}C\}, \{R|a\mathcal{E}R\}, \{V|a\mathcal{E}V\}, \{K|a\mathcal{E}K\}\} \quad (6)$$

While considering the above atomic formulas, discrete mathematics allows us to articulate new formulas as well. By applying logic connectives, satisfying any formula can define several other formulas to test whether a given argument is valid or not. The proved steps are:

i. Identify the premises and conclusion.
ii. Draw the truth table.
iii. Search for critical rows to prove whether given logic is valid or not [23].

These steps are illustrated in eq. (7).

$$p_1 \wedge\ p_2 \wedge\ p_3 \wedge\ p_4 \wedge p_5 \wedge\ p_6\ \rightarrow c \quad (7)$$

4. Results and Discussion

We use Boolean value True/False while proving the proposed model with the help of discrete mathematics. In order to evaluate our proposed theory in terms of its validity, we utilize 5 different sets 'D', 'C', 'R', 'V' and 'K' that have been discussed earlier; and it is illustrated in eq. (8) [23]. because we intend to prove a logical argument as shown in eq. (8). In table 2, the logical operations between 'D', 'C', 'R', 'V' and 'K' are performed

and its explanation is provided from eq. (1) to eq. (5). Furthermore, the performed operations deliver basics for premises and conclusion propositions that are the basic formulation of argument and hence are the main mechanism for proving any logic.

Table 2. Truth Table of Equation (8).

					Conclusion	Premises						
Sr.	D	C	R	V	K	D→C	C→R	R→V	V→K	(D→C) v (C→R)	(D→C) v (C→R) v (R→V)	(D→C) v (C→R) v (R→V)v(V→K)
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	0	1	1	1	0	1	1	1
3	1	1	1	0	1	1	1	0	1	1	1	1
4	1	1	1	0	0	1	1	0	1	1	1	1
5	1	1	0	1	1	1	0	1	1	1	1	1
6	1	1	0	1	0	1	0	1	1	1	1	1
7	1	1	0	0	1	1	0	1	1	1	1	1
8	1	1	0	0	0	1	0	1	1	1	1	1
9	1	0	1	1	1	0	1	1	1	1	1	1
10	1	0	1	1	0	0	1	1	0	1	1	1
11	1	0	1	0	1	0	1	0	1	1	1	1
12	1	0	1	0	0	0	1	0	1	1	1	1
13	1	0	0	1	1	0	1	1	1	1	1	1
14	1	0	0	1	0	0	1	1	0	1	1	1
15	1	0	0	0	1	0	1	1	1	1	1	1
16	1	0	0	0	0	0	1	1	1	1	1	1
17	0	1	1	1	1	1	1	1	1	1	1	1
18	0	1	1	1	0	1	1	1	0	1	1	1
19	0	1	1	0	1	1	1	0	1	1	1	1
20	0	1	1	0	0	1	1	0	1	1	1	1
21	0	1	0	1	1	1	0	1	1	1	1	1
22	0	1	0	1	0	1	0	1	0	1	1	1
23	0	1	0	0	1	1	0	1	1	1	1	1
24	0	1	0	0	0	1	0	1	1	1	1	1
25	0	0	1	1	1	1	1	1	1	1	1	1
26	0	0	1	1	0	1	1	1	0	1	1	1
27	0	0	1	0	1	1	1	0	1	1	1	1
28	0	0	1	0	0	1	1	0	1	1	1	1
29	0	0	0	1	1	1	1	1	1	1	1	1
30	0	0	0	1	0	1	1	1	0	1	1	1
31	0	0	0	0	1	1	1	1	1	1	1	1
32	0	0	0	0	0	1	1	1	1	1	1	1

After formulating different formulas, we use the first formula and its truth table which is depicted in Table 2; where truth values of "1/0" are used instead of "T/F"

$$(D \to C), (C \to R), (R \to V), (V \to K),$$
$$\{(D \to C) \text{v} (C \to R)\}, \{(D \to C) \text{v} (C \to R) \text{v} (R \to V)\},$$
$$\{(D \to C) \text{v} (C \to R) \text{v} (R \to V) \text{v} (V \to K)\} \quad \therefore K \qquad (8)$$

After applying different algorithms on the dataset, we determine whether the final result obtained, i.e., 'K' is true or not? Using argument principle in discrete mathematics, we determine and quantify critical rows in Table 2, i.e., only those rows where Boolean values of all premises are true. Here, we can see that for each critical row, the Boolean value of conclusion is true. An argument principle defined in discrete mathematics reveals that our results are valid for all

suppositions. Eq. (8) is a mathematical formulation, by which we plan to prove the proposed model. There are several approaches to prove any mathematical statement but we are using arguments because the given scenario is the same as that of Rules of Inference which is related to compound propositions or reasoning. Testimonies in the

science of mathematics are valid arguments that establish the validity of any mathematical equation. By an argument, we mean an order of mathematical statements that end with a statement usually known as a conclusion [26]. Eq.(8) evidently illustrates the seven premises and a conclusion. A valid argument means that the premises must follow the conclusion that is the actual formulation provided in eq. (8) and it is further elaborated in Truth-Table, i.e., Table 2. Now the resulting Table 2 is generated with a simple rule that has a total of 32 possible combinations using "2^n" which are the same as for premises. After analyzing Table 2 and searching the tautology, four different critical rows are found. Critical rows are highlighted in color. Now we must check each critical row with the conclusion (premise implies conclusion by definition) [24, 25], and hence proposed theory and its discrete mathematical formulation of unified data mining model is correct and valid.

Subsequently, all remaining formulations are evaluated and the resulting critical rows show tautology as represented in Table 3, which also proves our discrete formulation.

Table 3. Discrete Formulation for Proposed Model.

Sr.#	Discrete Formulation	Critical row values
1	(D → C) ^ (C → R) ^ (R → V) ^ (V → K) → K	4
2	(D → C) ^ (C → R) ^ (R → V) ^ (V → K) ^ {(D → C) v (C → R)} ^ {(D → C) v (C → R) v (R → V)} ^ {(D → C) v (C → R) v (R → V) v (V → K)} → K	4
3	(D → C) ^ (C → R) ^ (R → V) ^ (V → K) ^ {(D → C) → (C → R)} ^ {(D → C) → (C → R) → (R → V)} ^ {(D → C) → (C → R) → (R → V) → (V → K)} → K	4
4	(D → C) ^ (C → R) ^ (R → V) ^ (V → K) ^ {(D → C) → (C → R)} ^ {(D → C) → (R → V)} ^ {(D → C) → (V → K)} ^ {(C → R) → (R → V)} ^ {(C → R) → (V → K)} ^ {(R → V) → (V → K)} → K	4

In discrete mathematics, a number of evaluation criteria has been proposed like contradiction, mathematical induction, rules of inference, etc., we have chosen 'Rules of Inference' as evaluation criterion because the proposed model is pictorial likewise that of argument. Existing systems allow a single algorithm for mining, e.g., K-mean, Naïve Bayes, or Random Forest, etc. Users analyze the results of each deployed algorithm to get good results but the proposed model is unique in the sense that we are using different algorithms on different steps that allow achieving a higher accuracy rate.

4. Conclusions

In this research, we have proposed a discrete model which is verified and evaluated by a technique called Truth Table by utilizing the concept of argument. Most of the arguments prove the proposed theory as valid/correct which means that the proposed model can be used to extract knowledge by using different data mining processes and techniques. On the other hand, few other proportional logics were formed but their results do not get evaluated as valid arguments. This does not mean that the proposed model is invalid; rather it shows that particular arguments may not always be meaningful or in other words, different knowledge mining processes may calculate knowledge that is not required or either meaningless. It shows that different data mining techniques may provide different accuracy regarding the same data sets. So, the main impact of the proposed model provides a basis for many different fields of statistics for collecting, interpreting and presenting knowledge. The proposed model is useful for medical image processing and mining, data security, computer science, network security, computer vision, data science, etc., and all these applications are opened for implementation.

5. References

[1] D.M. Khan and N. Mohamudally, "An integration of K-means and decision tree (ID3) towards a more efficient data mining algorithm", *J. Comp.*, vol. 3, no. 12, pp. 76-82, 2011.

[2] D.M. Khan, N. Mohamudally and D. Babajee, "Investigating the Statistical Linear Relation between the Model Selection Criterion and the Complexities of Data Mining Algorithms", *J. Comp.*, vol. 4, no. 8, pp. 14-28, 2012.

[3] D.M. Khan and N. Mohamudally, "Model Selection Criterions as Data Mining Algorithms Selector The Selection of Data Mining Algorithms through Model Selection Criterions", *J. Comp.*, vol. 4, no. 3, pp. 102-114, 2012.

[4] D.M. Khan, N. Mohamudally and D. Babajee, "A unified theoretical framework for data mining", *Proc. Comp. Science*, vol. 17, pp. 104-113, 2013.

[5] Q. Yang and X. Wu, "10 challenging problems in data mining research", *Intl. J. Info. Tech & Decis. Mak.*, vol. 5, no. 04, pp. 597-604, 2006.

[6] D. Skillicorn, *Understanding complex datasets: data mining with matrix decompositions*: CRC press, 2007.

[7] R. Chattamvelli, "*Data mining algorithms*" Alpha science international, 2011.

[8] G. Li and H. Sheng, "Extracting features from gene ontology for the identification of protein subcellular location by semantic similarity measurement", In *Pacific-Asia Conf. Kn. Dis. and Data Min.*, pp. 112-118. Springer, Berlin, Heidelberg, 2007.

[9] Y.Y. Yao, "On modeling data mining with granular computing", *25th Ann. Intl. Comput. Soft. App. Conf. COMPSAC 2001*, pp. 638-643. IEEE, 2001.

[10] D.M. Khan and M. Nawaz, "A comparative study of single-step and multi-step data mining tools", *J. Comp.*, vol. 4, no. 10, pp. 26-41, 2012.

[11] D.M. Khan and N. Mohamudally, "The adaptability of conventional data mining algorithms through intelligent mobile agents in modern distributed systems", *Intl. J. Comp. Sci. Iss.*, vol. 9, no. 1, p. 38, 2012.

[12] D.M. Khan, N. Mohamudally and D. Babajee, "The Formulation of a Data Mining Theory for the Knowledge Extraction by means of a Multiagent System", *J. Comp.*, vol. 4, no. 8, pp. 29-38, 2012.

[13] S. Demri and E. Orlowska, "Logical analysis of indiscernibility", *Incomplete information: Rough set analysis*, pp. 347-380, Physica Heidelberg, 1998.

[14] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*: Springer Science & Business Media, 2012.

[15] R.R. Stoll, *Linear algebra and matrix theory*: A Book by Courier Corporation, 2013.

[16] M.T. Kane, "An argument-based approach to validity", *Psychol. bullet.*, vol. 112, no. 3, p. 527, 1992.

[17] D.M. Khan, F. Shahzad, N. Saher and N. Mohamudally, "Optimization and Analysis of Clusters", Sci. Int. Lahore, pp. 1959-1971, 2014

[18] D.M. Khan and N. Mohamudally, "The Relevancy of a Unified Data Mining Theory For the Big Data", *J. Appl. Environ. Biol. Sci.*, vol. 4, no. 8S, pp. 160-165, 2014.

[19] N. Mohamudally and D.M. Khan, "Application of a unified medical data miner for prediction, classification, interpretation and visualization on medical datasets: The diabetes dataset case.", In *Indus. Conf. on Data Mini.*, pp. 78-95. Springer, Berlin, Heidelberg, 2011.

[20] D.M. Khan, N. Mohamudally and D. Babajee, "Towards the Formulation of a Unified Data Mining Theory, Implemented by Means of Multiagent Systems (MASs)", *Adv. D. Min. Kn. Dis. App.*, pp. 03-42, 2012.

[21] B. Mahesh, "Machine Learning Algorithms-A Review", *Int. J. Sci. Res.*, vol. 9, pp. 381-386, 2020.

[22] R.A. Nisbet, "How to choose a data mining suite", *Data Mining Direct Special Report*, 2004.

[23] I. Lakatos, *Proofs and refutations: The logic of mathematical discovery*: Cambridge University Press, 2015.

[24] A. Broadie, "The practical syllogism", *Analysis*, vol. 29, no. 1, pp. 26-28, 1968.

[25] A. Popescu, "On Agency And Joint Action", *Studia Universitatis Babes-Bolyai-Philosophia*, vol. 65, no. 2, pp. 67-84, 2020.

[26] K.H. Rosen, "Discrete Mathematics and Its Applications", McGraw-Hill Education, 2012.