Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -1, April 2021, pp. 1-3 (3 pages)

A Survey of Evolutionary Algorithms for Data Mining

KAILASH PATIDAR, DR. DHANRAJ VERMA

Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore (M.P.)
Corresponding Author Email: kailashpatidar123 @gmail.com

Abstract— ***Evolutionary Algorithms are stochastic searching algorithms inspired by the process of neo-darwinian evolution algorithms. The inspiration for put on Evolutionary Algorithms to data mining is that these are a robust, adaptive search method that achieves a comprehensive search in the result space. Some of the popular evolutionary algorithms(EAs) is Ant Colony Optimization(ACO), Particle Swarm Optimization(PSO), Genetic Algorithms, Teaching Learning Based Optimization(TLBO), Cuckoo Search, etc. The key concept and principles used by the evolutionary algorithms(EAs) developed for resolving various data mining tasks, that is discovery of the classification rules, clustering, attributes selection and attribute creation.***

Index Terms— ***Data Mining, Evolutionary Algorithm, classification, clustering.***

I. INTRODUCTION

In the present scenario in every field there is huge amount of data gathered, processed and extracted for different purposes [1]. To acquire the data in the meaningful way is very important [2]. The main areas are health care, education, business enterprises and so on. If we think of data arrangement and data management in the way that meaningful extraction is possible then the data mining (DM) and machine learning algorithms may find to be useful for different purposes [3].

DM algorithm provides the way to extract the meaningful insights from the huge dataset. There are several DM algorithms which are useful in the direction of patter extractions for example association rule mining, clustering, classification, etc. [4]. The primarily operation needed in the arrangement of the data is data clustering. In which data can be arranged in similar groups of the same properties. K-means and fuzzy c-means algorithms are widely used clustering algorithms [5]. In the current scenario there is different complex solution where there is the need of proper thresholding and refinement [6]. In this case clustering alone may fail and evolutionary algorithms may be helpful with soft computing techniques. Some of the popular evolutionary algorithms(EAs) is Ant Colony Optimization(ACO), Particle Swarm Optimization(PSO), Genetic Algorithms, Teaching Learning Based Optimization(TLBO), Cuckoo Search, etc. [7]. Different machine learning algorithms are found to be useful in the extraction and data categorization process [8]. Some of the machine learning algorithms are support vector machine (SVM), logistic regression (LR), naïve Bayes (NB), K-Nearest Neighbor (KNN), etc.

Useful knowledge extraction is the special property of any information mining, which alludes to separating or "mining" learning from a lot of information or databases [8]. The procedure of discovering valuable examples or importance in crude information has been called knowledge discovery in databases (KDD) [19]. KDD gives a cleaning to the conflicting information. Information Mining additionally gives design, arrangement, perception and guideline partition. For comprehension the utility of information mining can classify as follows [9]:

- Regression is a calculation procedure that is regularly utilized for numeric expectation.
- Association returns affinities of an arrangement of records.
- Sequential example capacity looks for successive subsequences in an arrangement dataset, where a grouping record a recasting of occasions.
- Summarization is to make a reduced depiction for a subset of information.
- Classification maps an information thing into one of the predefined classes.
- Clustering distinguishes a limited arrangement of classifications to depict the information.
- Dependency displaying depicts critical conditions between variables.
- Change and deviation discovery is to find the greatest changes in the information by utilizing beforehand measured qualities.

Restorative finding is exceptionally subjective due to the clinical exploration and individual view of the specialists which may influence the determination. Various medical study have demonstrated that the result of the one patient can vary if the patient is investigated by distinct specialists doctor or even by the same specialists at the different periods of time in case of clustering. The supposed of restorative data mining remain to the focus shrouded leaning in the restorative field using data mining frameworks systems. It is possible to perceive models paying little heed to the way that we don't have totally understood the nice parts behind those models. In reality, even the models which are unnecessary can be found. Clinical stores containing a great deal of natural, clinical, and administrative data are logically having the chance to be open as friendly protection structures facilitate patient information for investigation and use objectives. Data mining frameworks associated with these information bases discover associations and models which are valuable in considering the experiences and information course of action and data arrangement.

II. PROCESS OF DATA MINING

The process of data mining is successive which require many steps to be followed which are as shown below in the form of a drawing (3).

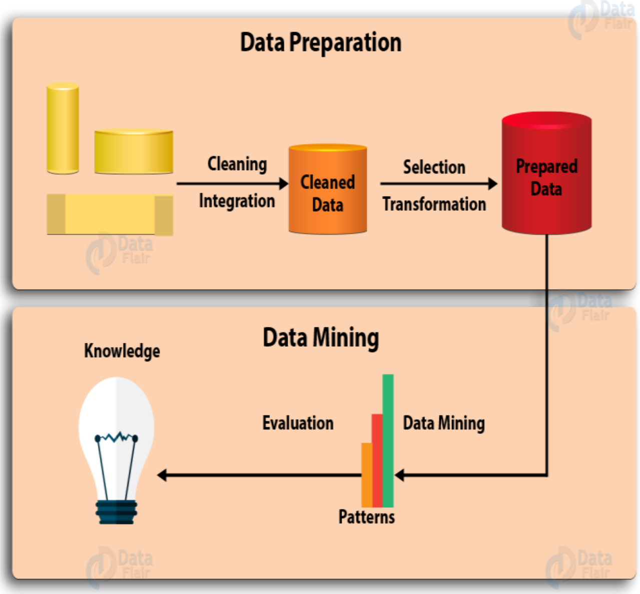

Figure 1: Data Mining Process Architecture

1. Extraction, transformation, and load operation data in the data warehouse system.
2. Store and achieve the information in a multidimensional data set framework system.
3. Give data access to business indicators and data innovation subject matter experts.
4. Evaluate the data by application programming software.
5. Present the data in the important format, like as a diagram, table or charts

III. DATA MINING APPLICATIONS

Data mining is a data analysis method that has been quickly improved and used in a large number of domains that were already using measurements. Examples of particular applications areas are:

- Data mining is an interdisciplinary field with wide and various applications
- There exist nontrivial gaps between data mining ethics and domain-explicit applications
- Retail business
- Commercial data analysis
- Biological data analysis
- Telecommunication industry

IV. EVOLUTIONARY ALGORITHM

Evolutionary algorithms are a experiential-based approach to solving problems that cannot be easily solved in polynomial time, such as classically NP-Hard problems, and anything else that would take far too long to comprehensively process. When used on their own, they are classically applied to combinatorial difficulties; however, genetic algorithms are repeatedly used in cycle with other methods, acting as a quick way to invention a somewhat optimal early place for another algorithm to work off of.

The evidence of an evolutionary algorithm is quite simple given that you are conversant with the procedure of natural selection. An EA contains four general steps: initialization, selection, genetic operators, and termination**.** These steps each parallel, roughly, to a particular surface of natural selection, and provide easy ways to modularize operations of this algorithm category. Simply put, in an EA, fitter members will survive and multiply, while unfit members will die off and not subsidize to the gene pool of further generations, much like in normal selection.

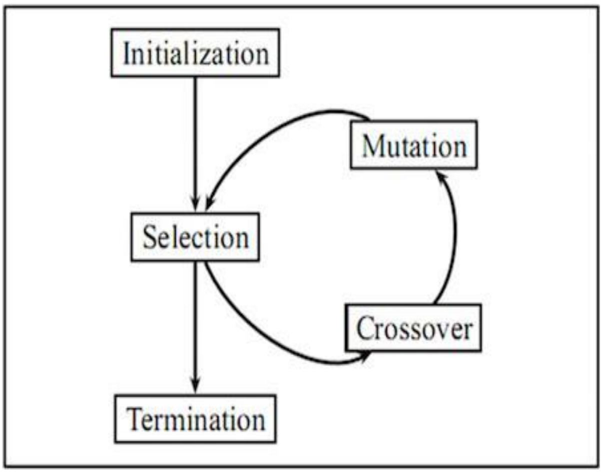

Figure 2: Process of natural selection

V. RELATED WORK

In 2020, Chug and Baweja [10] discussed the approach of how to retain the customers. They took the data of 150 customers and made the three clusters. They found the age an important parameter for the clustering. They discussed two techniques of clustering. In the first phase, they applied the k-means algorithm and for the second approach, they applied the agglomerative clustering. Their approach is helpful in segmentation for the existing customers, which may be helpful in grouping of psychographic data.

In 2020, Brown and Shi [11] discussed the different steps used to implement the clustering algorithm i.e., Fast Density-Grid. They used the Apache Spark with this so that they can parallelize it. They performed it in three stages like Grid Space Density, Determination by Densest Neighbor and Generation of cluster. There results may be helpful in parallel implementation and efficient even if the data exceeds some limit.

In 2020, Chebanenko et al. [12] proposed the Fuzzy system, which can be used to estimate the patient’s compliance for primary as well as further consultancy system. They performed the cluster analysis by using Fuzzy c-average algorithm on the patients’ data. According to this rate, they formed the three groups. There proposed method may be helpful for clinicians to discover cardiac patients in advance. The efficiency of cardiologists may be improved.

In 2020, Gong et al. [13] discussed about the visualization by which the validity of data clustering of electricity may improve. Their approach is found to be impactful in the improvement of the clustering of electricity data by their proposed a visual cluster analysis framework. They have suggested different characteristics of electricity as the future recommendation.

In 2020, Kang et al. [14] discussed the importance of cluster analysis in identifying the faults of automobile. They proposed a fault analysis model. It considered the various factors like deterioration failure, environmental factor and various human factor in their model. They applied a clustering algorithm and form a pedigree map. After pre-processing the data, they convert that into a form of ratio matrix. Further, they applied the k-value clustering. Their results showed that cluster analysis provides and analysis of driving habits. They showed that it may be good for auxiliary role and can be used for fault detection in later cases.

In 2020, Kesheng et al. [15] analyzed the behavior of 3,245 students at B grade universities. They used the network data from campus usage. They used the data mining for cleaning the data. They pre-process the data by using various methods like Kernel method, linear interpolation and spline method. They used the R language with the use of R-studio software. Their result grouped them into four groups and found around 350 students have the more internet usage. This data is helpful for student affair management, which provides support to counselor to improve the professional level.

VI. CONCLUSION

In this survey paper, we can find out that evolutionary algorithms (EAs) can able various troublesome difficult problems with minimum expense and high accuracy for multi-objective tasks. The paper additionally discusses how the evolutionary algorithms (EAs) produced results in the data mining and knowledge discovery field. The principle expect to utilize evolutionary algorithms (EAs) in data mining is as of their topographies which overcome the burden in moderate data mining methods and produces a best solutions. Evolutionary Algorithms (EAs) are utilized to decrease the trouble in discovering extraordinary and animating knowledge. As the development in the utilization of data mining algorithm in genuine space applications, a large portion of the algorithm cascades in the classification of multi-objective algorithm. Thusly the utilization of multi-objective Evolutionary Algorithms to work on the improve of data removal methods is a confident research part.

REFERENCES

1. Dubey A. K., Kumar A., Agrawal R., “An efficient ACO-PSO-based framework for data classification and preprocessing in big data”, Evolutionary Intelligence. 9, 1-4, 2020.
2. Dubey A. K., Gupta U., Jain S., “Computational Measure of Cancer Using Data Mining and Optimization”, International Conference on Sustainable Communication Networks and Application, Springer, Cham, pp. 626-632, 30 Jul 2019.
3. Agarwal R., Srikant R., “Fast algorithms for mining association rules”, In Proc. of the 20th VLDB Conference, pp. 487-499, 12 Sep 1994.
4. Han J., Pei J., Yin Y., “Mining frequent patterns without candidate generation”, ACM sigmod record. 29, 2, 1-2, 2000.
5. Jamil A., Salam A., Amin F., “Performance evaluation of top-k sequential mining methods on synthetic and real datasets”, International Journal of Advanced Computer Research, 7, 32,176-182, 2017.
6. Kumari I., Sharma V., “A review for the efficient clustering based on distance and the calculation of centroid”, International Journal of Advanced Technology and Engineering Exploration, 7, 63, 48-52, 2020.
7. Khandelwal A., Jain Y. K.., “Computational analysis of clustering techniques for the efficient cluster head selection”, International Journal of Advanced Technology and Engineering Exploration., 6, 60, 248-259, 2019.
8. Maghraby E. E., Gody A. M., Farouk M. H., “Noise robust speech recognition system using multimodal audio-visual approach using different deep learning classification techniques”, International Journal of Advanced Computer Research., 10, 47, 51-71, 2020.
9. Dubey A. K., Dubey A. K. , Agarwal V., Khandagre Y., “Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support”, CSI Sixth International Conference on Software Engineering pp. 1-6, 5 Sep 2012.
10. Chugh S., Baweja V. R., “Data Mining Application in Segmenting Customers with Clustering”, International Conference on Emerging Trends in Information Technology and Engineering, pp. 1-4, 24 Feb 2020.
11. Brown D., Shi Y., “A Distributed Density-Grid Clustering Algorithm for Multi-Dimensional Data”, 10th Annual Computing and Communication Workshop and Conference, pp. 0001-0008, 6 Jan 2020.
12. Chebanenko E., Denisova L., Serobabov A., “Intelligent Processing of Medical Information for Application in the Expert system”, Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, pp. 0085-0088, 14 May 2020.
13. Gong F., Bu .F, Zhang Y., Yan Y., Hu R. and Dong M. “Visual Clustering Analysis of Electricity Data Based on t-SNE”, IEEE 5th International Conference on Cloud Computing and Big Data Analytics, pp. 234-240, 10 Apr 2020.
14. Kang H., Zhao H. and Ai T., “Description of Aggregate System Clustering Algorithm and Its Application in the Analysis of Automobile Fault Law”, IEEE 5th Information Technology and Mechatronics Engineering Conference, pp. 1048-1051, 12 Jun 2020.
15. Kesheng L., Yikun N., Zihan L., Bin D., “Data Mining and Feature Analysis of College Students Campus Network Behavior”, 5th IEEE International Conference on Big Data Analytics, pp. 231-237, 8 May 2020.