iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43149

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Association Rule Mining using FP-Growth and An Innovative Artificial Neural Network Techniques

Pankaj Kumar Gond[1], Aditya Shukla[2], Satish Sahani[3], Neha Gond[4], Dr. Harvendra Kumar[5]
[1]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*
[2]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*
[3]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*
[4]*B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India*
[5]*Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India*

Abstract: ***With the prodigious proliferation of ginormous-scale data depots, a need for incorporating the empirical techniques of Data Mining (DM) with the effectiveness of commensurate systems to intuitively manage cosmic volumes of data has now risen. To quell these obstructions of managing data efficiently, in this document we present a new algorithm based on ANN for DM activities, which overcomes the problems in the current available algorithms of mining apropos of their execution time and interestingness and to prove its efficiency the new algorithm will be compared to a popular mining algorithm: FP Growth.***
The following sections of the paper are sorted as – section I exhibits the introduction, section II exhibits the previously done related works, section III is exemplifies of the methodologies, section IV exhibits the experimental setup of the analogous work, section V inculcates the experimental result and finally section VI depicts the conclusion of the respective work.
Keyword: Data Mining, Association Rule Mining, FP Growth algorithm, Artificial Neural Networks.

I. INTRODUCTION

DM is a component of a wider process called KDD or knowledge discovery from databases. It involves researchers from a large assortment of approaches, including *c*omputer scientists and *s*tatisticians, as well as those working in fields such as *m*achine learning (ML), *a*rtificial intelligence (AI), *i*nformation retrieval and *p*attern recognition (PR). In this process, before a data set can be mined it first has to be cleaned. This method supplants fallacies, ensures resoluteness and takes obscured values into account. It may use convert, edit and evaluate simple statistical techniques or it may use highly sophisticated data analysis. What is new for data miners is the employment of these techniques on vast quantities of data. In general, it uses quite simple statistical techniques, though it may use highly sophisticated data analysis. This paper, focus on an *a*ssociation rule mining (ARM) algorithms named FP-growth, which is being juxtaposed with an optimised algorithm which is based on the ANN concepts. The comparisons are carried out in between the different algorithms on the basis of their execution time. Even, interestingness is obtained from the comparison of FP Growth algorithm, and Optimized algorithm. Thus to test impactfulness of the ANN approach on the DM algorithm on its execution time and the interestingness is observed and juxtaposed with FP growth algorithm.
ANN is a loosely modelled system based on the human *b*rain. The field goes by many names, such as connectionism, parallel *d*istributed processing (PDP), neuro-*c*omputing, natural intelligent systems, ML algorithms, and ANN. It has the flair to account for any functional dependency. Optimised algorithm basically follows an UL algorithm in which an ANN is volitional in a way to perpetrate task of data encoding along with data *d*ecoding to reconstitute input.

II. RELATED WORK

By concentrating at the foibles of the stereotyped "support-confidence" ARM framework and the worriment of mining negative ARs, the concept of *i*nterestingness measure is introduced. Junrui Yang and Lin Xu [1], analysed the commonly used *i*nterestingness measures, and merged the *i*nterestingness measure based on the discrepancy idea and the cosine measure to obtain a new interestingness measure model. By using this model in mining ARM based rules, the positive and negative ARs can be effectively mined and segregated. Moreover the algorithm is further verified by using a certain grocery store transaction dataset in comparison with the existing positive and negative ARM algorithm and the mining results using the stereotyped "support-*c*onfidence" framework. The results verifies the potencies of the improved interest model and the algorithm using the model, and contribute to the overall improvement of the mining accuracy of the algorithm.

©IJRASET: All Rights are Reserved

DM is the process of extracting hidden and useful patterns and information from data. It is a technology that helps businesses to prognosticate future trends and deportments, allowing them to make proactive and knowledge driven decisions. In this series, Anshu in [2] shows the procedure of DM and how it can help decision makers to make better decisions. It also presents a detailed description of DM techniques and their applications in various fields. Some of the major techniques such as *c*lassification, *c*lustering etc., helps in finding the *p*atterns to decide upon the *f*uture trends in businesses to grow. Different techniques can be used for different purposes as per their pros and cons. DM has been widely used in the business field, and ML can perform data *a*nalysis and pattern *d*iscovery, thus playing a key role in DM application. Teng Xiuyi and Gong Yuxia in [3] expounds the definition, model, *d*evelopment stage, types and commercial application of ML, and emphasizes the role of machine learning in DM. As per them, the ML technology in DM can be applied in many industries, including financial industry, retail industry, insurance industry, telecommunication industry and so on. For example, in the financial industry, the financial analysts use DM to build prediction models to intuit the patterns that have caused market fluidity in history, thereby improving the ability of predicting market fluidity. Similarly, in the retail industry, sales people can build predictive models through DM to understand who is most likely to respond to commonality, thereby increasing sales. When enterprises apply DM technologies, it is believed that enterprise fully cognize the advantages and disadvantages of various technologies and methods, and select appropriate technologies for specific environments and tasks. Frequent pattern mining is an important topic that is needed for the *I*nternet of *T*hings (IoT) applications frequently. Many IoT applications have been refined in which continuous streaming data is used. In continuation with this, Halil İbrahim DEDE et al., in [4] performed a comparison study based on pursuance evaluation of FP-stream algorithm and WMFP-SW algorithms, which are developed for frequent pattern analysis in IoT, using a novel real world dataset. The dataset was created by collecting 6 different sensor values over a 3 month period in an established testbed. As per the earlier studies, both the algorithms are efficient enough in extracting FPs but the experiments done shows that WMFP-SW algorithm extracts more FPs then FP-stream algorithm when minimum support is increasing. Regarding the results of execution time, the FP stream is always faster than the WMFP-SW. Moreover, the number of common FPs and the total number of FPs per algorithm decrease as the minimum support value increases.
Generally, the power systems are disturbed by the presence of electromagnetic interference and crosstalk between the transmission link layers in the transmission and dissipating process. Though it is easy to produce transmission distribution faults, in sincere graft to embellish the productiveness of fault diagnosis, a method of fault diagnosis for power systems based on ANN algorithms was proposed by Kai Xu in [5]. In this, the multi sensor quantization fusion method is used to carry out the electricity. The transmission dissipating signal in the power transmission link layer is hauled out from the power system, and the transmission dissipating signal is decomposed and the ARs are excavated. Furthermore, the spectral analysis model is used to citrate the spectral characteristics of the transmission information of the power system, and the fault diagnosis and fault type identification are carried out according to the spectrum difference. The power system fault features are then further tiered and identified by an ANN algorithm to realize the optimal diagnosis of power system fault. The simulation results show that the method was more accurate and more streamlined in the fault diagnosis of the power system.

III.METHODOLOGIES

ARM is an event for determining organizations, patterns, and relationships between sets of objects on a dataset. The law of association is of the form M ⟶N [support, confidence]. The support and confidence are the primary measures that are used for assurance of the rule. AR is said to be strong if it satisfies both the minsupp (minimal support) and the minconf (minimal confidence) that is defined by the user. These ARs are easy to establish due to the small database but become more complex as the database transforms. Some general values and concepts are required in order to better understand DM in large data sets. A set contains n items and is called an n-itemset. So set {A, B} is a set of 2- itemset. Based on the frequency of itemsets, the number of active functions is calculated. So here, a well-known mining algorithms i.e., FP growth is taken along with an Optimised algorithms using ANN are considered for the experiments of analysing the performance of mining ARs.

A. Artificial Neural Networks

ANNs are biologically inspired that is, they are made of neurons that perform in the manner that it is analogous to the most elementary functions of the biological neurons. These elements are then organized in a way that may be related to the anatomy of the brain. ANN were introduced as early as 1960. After a lapse of nearly two decades, interests in ANN has grown rapidly over the past few years. Professionals from diverse fields like engineering, physiology, physics and others are intrigued by the potential offered by this technology. Researchers are moving ahead on the theoretical as well as the applications front. Suddenly, this technology has raised the possibility of realizing several attributes of human learning and intelligence.

The models suggested can now capture some features of the brain such as high fault tolerance, dexterity to recognize objects under varied positions, high degree of robustness and ability to learn.

The fields which are revived in the light of ANN, include:

1) Cognitive sciences, Vision.
2) Pattern recognition.
3) Learning paradigms.
4) Artificial intelligence.

Intensive research is going on in all the fields mentioned above and many models of ANN have been proposed. Though many of the algorithms presented do not present rigorous proofs of convergence, their utility have been demonstrated by the applications. Since the models have internal distributed representation of knowledge in terms of connection strength it becomes difficult to decipher the knowledge in conventional norms. Models are either of self-organizing type or learn under supervision.

Generally a neuron can realize basic logic function and hence we can think it as switching device. Though the parallel architecture of ANN has an appeal for parallel processing, the hardware development in this field is very limited.

B. FP Growth Algorithm

FP-growth algorithm is an association analysis algorithm proposed by Han Jiawei and others in 2000. The algorithm uses divide and conquer strategy: compress the database providing frequent itemsets into a FP-Tree, and keep the association information between itemsets. When compared with one of the most common mining algorithms Apriori, its advantages are as follows: first, it does not produce candidate itemsets; and second, it does not need to traverse data sets frequently. FP-growth algorithm generates FP-Tree data structure to find frequent itemsets, reduce the number of scanning transaction sets, and can efficiently mine data.

The workflow of FP-growth algorithm is as follows: First, it scans the transaction set for the first time, and list the number of times each item appears in the transaction set. Then according to the experience, the minimum support is determined. Then the process starts as, items that do not meet the minimum support are deleted from each transaction of the transaction set, and the items meeting the minimum support are arranged in reverse order according to the number of times, so as to obtain a frequent item set. Secondly, the root node of FP-Trees are set to null. Then the modified transaction set is scanned again, and a path is established from the root node to the leaf nodes as per the order of frequent itemsets. If the items in the transactions are already in the FP-Tree, then the corresponding number of items is uplift by 1, and different items are added after it, that is, the same items are shared and the coefficient is uplift by 1. Finally, after this the step 2 is repeated again to get the FP-Tree, header table and node linked list. According to the item header table, we can find the conditional pattern base of FP-Tree, and then we can find the conditional FP-Tree. Finally, the above process is repeated to find the frequent itemsets of the corresponding sub conditional pattern base, as shown in Fig 1.

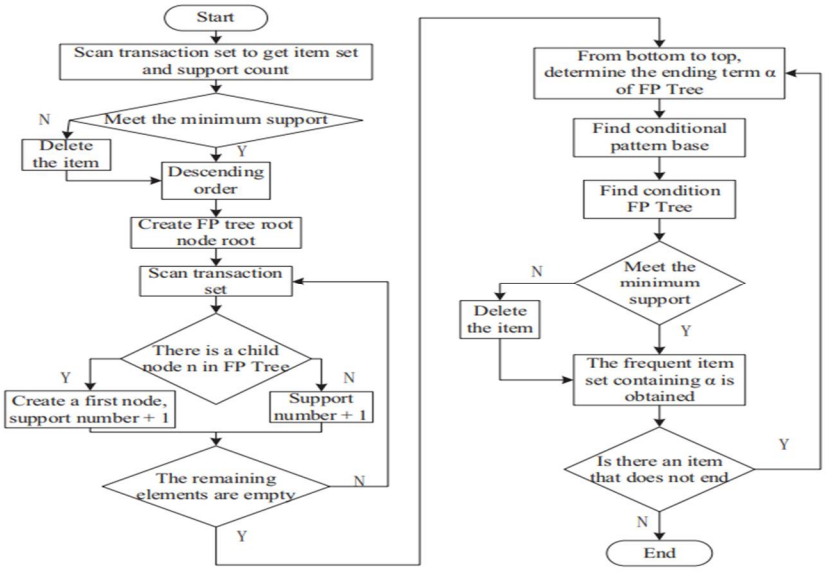


Fig 1: Establishment of FP-Tree

IV. EXPERIMENTAL SETUP

For this experiment, an Intel(R) Core(TM) i3-8130U CPU @2.20GHz with 8.00 GB RAM or above has been used. The operating system was Microsoft Windows 10 Enterprise. While comparing the performance of the two algorithms FPGrowth, and Optimised algorithm using ANN, no other application has being used implementation for feasible analysis of the implementation time of algorithm. The datasets used for the experiment are the shop transactions dataset, and the parameters of the selected datasets is having 120 number of unique dataset in total 520 number of transactions.

V. EXPERIMENTAL RESULTS

For analysing the performance of different algorithms this section is classified into two parts as implementation-time analysis.

A. Data Visualization

Shop transactional dataset:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
frankfurter	rolls/buns	soda	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
chicken	tropical fruit	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
butter	sugar	fruit/vegetable juice	newspapers	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fruit/vegetable juice	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
packaged fruit/vegetables	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

B. Rules Generation

	Unnamed: 0	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	140	frozenset({'citrus fruit', 'tropical fruit', '...	frozenset({'other vegetables'})	0.005694	0.193493	0.004474	0.785714	4.060694	0.003372	3.763701
1	1675	frozenset({'curd', 'domestic eggs'})	frozenset({'whole milk'})	0.006507	0.255516	0.004779	0.734375	2.874086	0.003116	2.802763
2	1615	frozenset({'butter', 'curd'})	frozenset({'whole milk'})	0.006812	0.255516	0.004881	0.716418	2.803808	0.003140	2.625286
3	3510	frozenset({'yogurt', 'tropical fruit', 'whippe...	frozenset({'whole milk'})	0.006202	0.255516	0.004372	0.704918	2.758802	0.002787	2.522974
4	578	frozenset({'yogurt', 'tropical fruit', 'root v...	frozenset({'whole milk'})	0.008134	0.255516	0.005694	0.700000	2.739554	0.003616	2.481613
5	1071	frozenset({'yogurt', 'butter', 'other vegetabl...	frozenset({'whole milk'})	0.006406	0.255516	0.004372	0.682540	2.671221	0.002735	2.345125

C. Implementation time Analysis

In this section, different Implementation Time is observed for different support values of two algorithms and to observe their performance. The lower the Implementation time, means the higher the performance. Implementation time analyses of the datasets are shown in the following tables.

Table II: Implementation time (second) analysis of the algorithms for different Support values for both datasets.

Dataset	Min Support Value	FP Growth Algorithm	Optimized Algorithm
Shop Transactions Dataset	0.002	48.21	43.09
	0.004	46.63	40.2
	0.006	41.8	37.20
	0.008	35.73	31.13
	0.010	32.31	26.35

©IJRASET: All Rights are Reserved

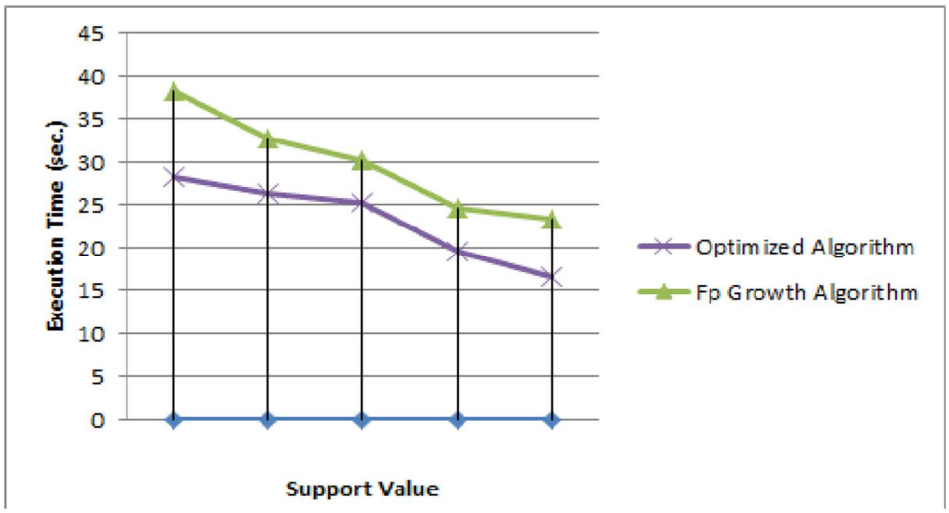


Fig. 2: Implementation time analysis for different Support values of Shop transations dataset.

The ovservation from fig 2, shows that the implementation time of the FP-Growth algorithm takes more time compare to optimised algorithm. Here algorithms using the ANNs concepts perform better according to implementation time. In the next part, we are analysing the interestingness of rules generated by FP Growth and optimised algorithms.

D. Interestingness Analysis

Interestingness defines the acceptability of a rule.

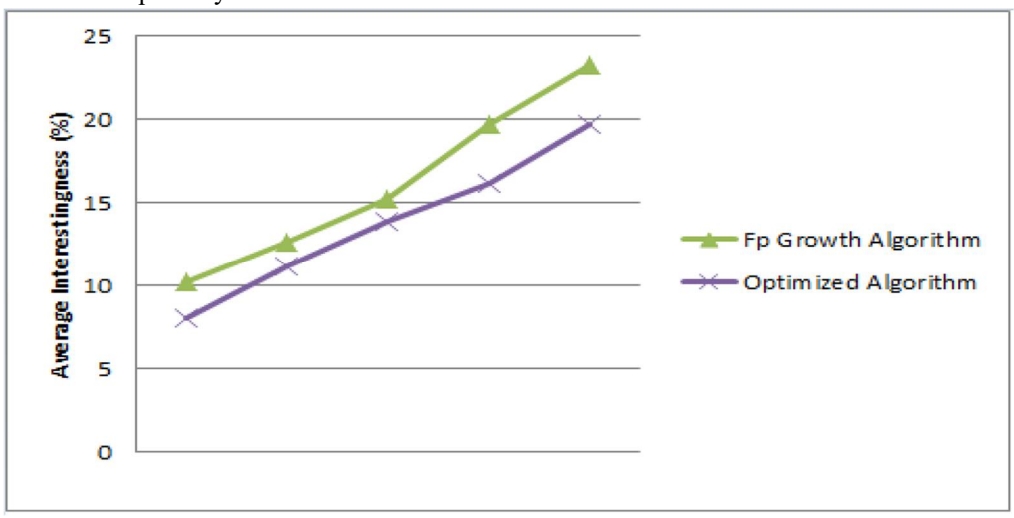


Fig. 3: Average Interestingness analysis of rules for different Support values for shop transaction dataset

Observation form Fig 3, categories the average interestingness of shop transaction dataset using FP growth and optimised algorithm. The average interestingness is higher for the FP Growth compare to Optimised algorithm.

VI. CONCLUSION

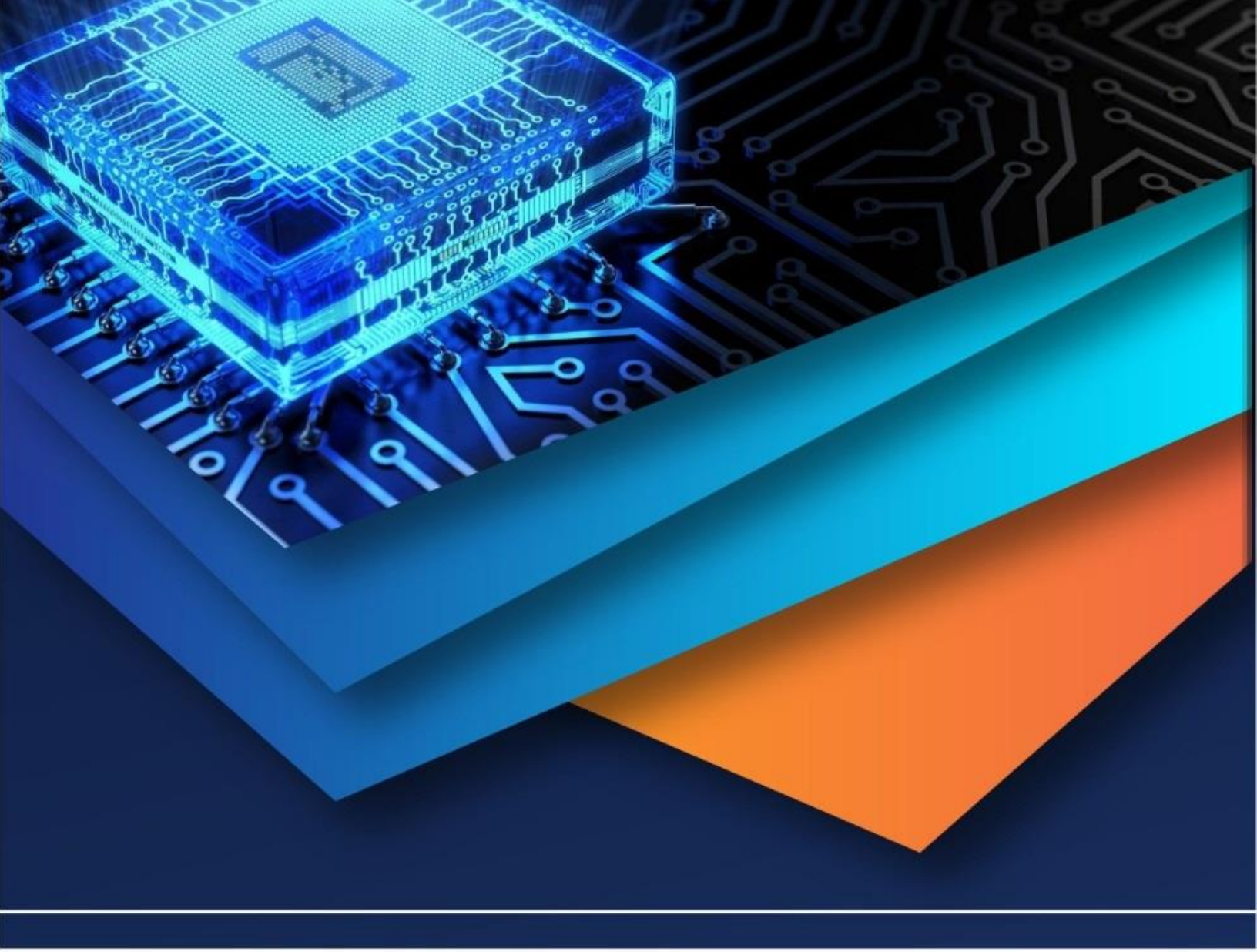
The observation in this paper focusses on the implementation time of FP-Growth algorithm, and the optimised algorithm which is made to explore application of ANNs in DM. The observation shows that the optimised algorithm takes lesser implementation-time as compare to FP Growth. The average Interestingness value obtained by FP growth is more than the optimised algorithm, this results shows that the optimised algorithm generates lesser number of association rules, means optimised algorithm generates only the potentially strong association rules. The results shows that the optimised algorithm has better in terms of implementation-time, efficiency and interestingness then FP-Growth algorithms.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Junrui Yang and Lin Xu, "A novel interestingness measure based on a fusion model for association rules mining", MATEC Web of Conferences, pp. 1-6, 2021.
[2] Anshu, "Review Paper on Data Mining Techniques and Applications", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 7, issue 2, pp. 22-26, March 2019.
[3] Teng Xiuyi and Gong Yuxia, "Research on Application of Machine Learning in Data Mining", IOP Conference Series: Materials Science and Engineering, pp. 1-5, 2018.
[4] Halil İbrahim Dede et al., "Comparison of Frequent Pattern Mining Algorithms in Internet of Things", 28th Signal Processing and Communications Applications Conference (SIU), pp.1057-1060, 2021.
[5] Kai Xu, "Fault Diagnosis Method of Power System Based on Neural Network", International Conference on Virtual Reality and Intelligent Systems, pp. 172-175, 2018.
[6] Mohammadreza Ramzanpour and Simone A. Ludwig, "Association Rule Mining Based Algorithm for Recovery of Silent Data Corruption in Convolutional Neural Network Data Storage", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 3057-3064, 2020.
[7] Sempe Leholo et al., "Solar Energy Potential Forecasting and Optimization using Artificial Neural Network: South Africa Case Study", Institute of Electrical and Electronics Engineers, pp. 533-536, 2019.
[8] Manomita Chakraborty et al., "Data Mining Using Neural Networks in the form of Classification Rules: A Review", 4th International Conference on Computational Intelligence and Networks (CINE), pp. 332-337, 2020.
[9] Biswajit Panja, et al., "Crime Analysis Mapping, Intrusion Detection - using Data Mining", IEEE Technology and Engineering Management Conference (TEMSCON), pp. 255-259, 2020.
[10] Tianyu Xia et al., "Data Association Rules Mining Method Based on RBF Neural Network Optimization Algorithm", International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 433-436, 2020.
[11] X. Y. Kong et al., “Fault diagnosis method of a distribution network based on cloud-edge computing architecture and wavelet neural network”, Distribution & Utilization, vol.37, pp. 17-23, 2020.
[12] Y. H. Huang et al., "Research on weak point analysis of distribution networks based on FP-Growth algorithm", Electrical Measurement & Instrumentation, pp. 1-7, 2020.
[13] Hossein Riahi-Madvar et al., "Derivation of optimized equations for estimation of dispersion coefficient in natural streams using hybridized ANN with PSO and CSO algorithms", IEEE Access, vol.8, pp.156582-156599, 2020.
[14] Jin Bo Chen et al., "A Log Analysis Technology Based on FP-growth Improved Algorithm", International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), pp.1096-1101, 2021.
[15] Wisnu Arya Dipa and Wikan Danar Sunindyo, "Software Defect Prediction Using SMOTE and Artificial Neural Network", International Conference on Data and Software Engineering (ICoDSE), pp. 1049-1054, 2021.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)