Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130612

Multi-Objective Sub-Linear Frequent Mining-Based Information Prediction in Biomedical Datasets using Big Data Analytics

G. Elangovan[1,*] *and G. Kavya*[2]

[1] Department of Computer Science & Engineering, Velammal Institute of Technology, Chennai, India.
[2] Department of Electronics & Communication Engineering, S.A. Engineering College, Chennai, India.

Received: 16 Feb. 2019, Revised: 19 Mar. 2019, Accepted: 11 May 2019
Published online: 1 Nov. 2019

Abstract: Recently, big data applications have been rapidly expanded into various industries. Healthcare is one of the industries that are seek to use big data platforms and mining. As a result, some large data analytics tools have been adopted in this field. Medical imaging, which is a pillar in diagnostic healthcare, involves a high volume of data collection and processing. The most challenging issue is common in sub-graph mining process to reduce the dimensionality of medical data set is minimized. In this paper, we propose a Multi-Objective Sub-Linear Frequent Mining (MOSLFM) to estimate the real values of outline processing in biomedical data, which is useful for computational complexity. This is repeated to find the minimum representation of the most frequent supplemental edges to be compatible with the sub-border margin. Sub-linear and sub-graph often use the mining process candidate generation model to find the biometric data set used to reduce the process. Projecting a high efficient progressing cluster partitioning method is used to determine the identified terms frequency in the biomedical dataset, so the process is simplified using lower complexity.

Keywords: Big Data, Data Mining, Map-Reduce, Frequent Pattern, Subgraph Mining

1 Introduction

In the big data of biomedical dataset, frequent sub-graph mining is essential to reduce the task of finding minimal dataset values, because the range of medical devices is vast and includes most healthcare document to specify the category of classes defined with supportive frequent data mining. Applying data mining systems to biomedical datasets helps to identify the patient information, which helps predict the time of process to conduct treatment. In high dimensional data process, the bio-medical industry has produced a lot of complex data about the patients, healing facilities assets, illnesses, analysis techniques, the patient's electronic records, etc. The map-reduce techniques can be used where the dimension and volume of data is higher. It has been used to identify the exact sub-group of data points which reduce the dimensionality, as well. The repeated information is classified as frequent information or outliers. Mining biomedical data can adopt frequency analysis. Various methodologies are devoted to frequent analysis and clustering the datasets. Every calculation has its advantages and disadvantages. Similarly the dataset on which we apply these systems may comprise missing qualities because all essential qualities for each record related to construction may not be predicted. Subsequently, the unavailable feature results in time complexity.

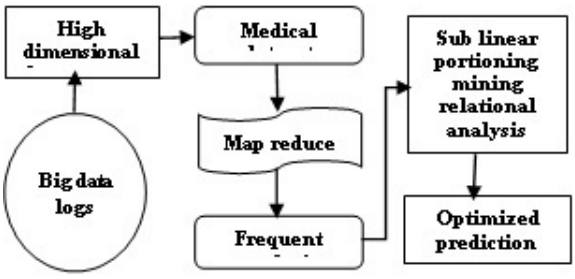


Fig. 1: The process of map- reduce in the biomedical dataset

* Corresponding author e-mail: ela@asia.com

The initial phases of addressing the high-dimensional dataset are obtained from biomedical data storage. Progression of Map- reduce for frequent analysis requires considering linear analysis qualities and partition process. Fig. 1 shows the process of basic map- reduce framework in biomedical dataset. This repeated scheme should then be standardized to make the restorative dataset practical for performing frequent subset classification using the current calculations of support and confidence value from the attributes. The cluster recognizes the repeated minimal supportive characteristics or highlights and reduces this specific class to the group later. This procedure is called dimensionality reduction or diminishment that require the current record into one of the class marks accessible by clusters. It is necessary to compare the candidates of different classes in a iterative manner to generate candidate set for any query. This procedure turns out to be substantially more intricate to probability. All the current frequent calculations dealing with missing qualities require the class names should have known partition in the linear probability they utilize and with which the missing attributes will be filled. In this case, prediction is the process of mining or generating a knowledge from the available samples towards any entity. If the frequents are missing, the first attempt is to distinguish the closest sub-sequence and fill the probability attributes in the medicinal records afterwards.

Medicinal data encounter a few challenges compared with regular datasets. The initial step is the instance of data gathering which necessitates characterizing the properties of attributes that identify the objective. This period of test data accumulation encounter some challenges when attempting to find the attributes. The likelihood of the nearness of missing qualities for a few attributes properties itself starts the primary test before adopting preprocessing. Addressing the frequent attributes is based on the management without the medicinal record that comprises of missing frequent qualities. The disappeared quality of value is to be filed with near case mean value. This is applied to the value of missing attribute where the restorative record may contain the estimations of the conditions which fundamentally serve in predicting the illnesses in a few circumstances. If we don't dispose the records comprising confidence states, the following system has to consider those records that need to be applied. In this case, the candidate selection should be performed in a iterative manner. Candidate generation never finds the relative sequence of similarity disclosure for unrelated value in attributes record. The most widely recognized way is utilizing the partitioned measure to perform attribution of these frequently-repeated values and to supplant these absent qualities using the reflective qualities. This requires an appropriate estimation because any probability of occurrence gives reliable outcomes with least time complexity.

This large work has a straightforward interface that enhances real itemset mining, planned co-candidate generation and massive sample calculations in circulation. Thus, this interface, as a large cluster of countries can implement the highest level of frequency analysis.

Deliberation of each transaction from the candidate enables us to express the straightforward calculations of repeated occurrences. The majority of frequent support values and confidence value calculations are included by applying a map-reducing task to each sequence value. Our contribution is to register a set of middle key terms of transaction, and has a decrease activity to each of the qualities that has a similar common key, the specific end goal, to produce higher efficiency in data proceeding with reduced complexity level. Utilizing a practical model is defined by candidate generation model using frequent analysis. The decrease of process activities enables us to repeat the recommendation adopting the circumstances of pointing the relational terms which is a frequent mining dataset with time relevance.

2 Related Work

The emphasis is given to find the repeating sub-graphs from graph databases using map- reduce, which is an exceptionally dynamic subject in essential data and flow data mining research [1]. Graphs provide a general method to show an assortment of relations among data and thereby finding frequent sub-graphs that have numerous applications in interdisciplinary research, such as substance informatics and bioinformatics. The flat and vertical disclosure models, which locate the associated sub-graphs that have an adequate number of edge-disjoint installing are in a single expansive undirected named little graph [2, 3] with dimensionality problem. The definition and calculations created for the graph exchange setting cannot be utilized to unravel the single-graph background.

As a response to this multifaceted dimensional quality, another deliberation enables us to explain the necessary calculations to reduce dimensionality problem [4]. They identify the relation between the users queries to generate recommendation to the users. An exponential number of conceivable sub-graphs creates the issue of frequent sub-graph mining. [5]. Maximal frequent mining has activated much enthusiasm because the measure of the maximal frequent sub-graphs set is substantially little compared to that of the set of frequent sub-graphs. Utilizing Hadoop is an open source execution of map- reduce to cause a progression of investigations on extensive scale with natural communities including a few circulations [6], clustering coefficient and breadth.

Frequent Itemset Mining (FIM) is a suitable apparatus for frequent co-concurrent things. Since its initiation, various noteworthy FIM calculations have been created to accelerate mining execution [7]. When the dataset measure is enormous, both the memory utilization and computational cost become restrictively high-priced. The dialect is intended for usability with no express parallelism while being amiable to effective parallel

execution on substantial clusters. Considering the big data, current database frameworks, such as social databases [8, 9] and protest databases that bomb as frequent sub-graph mining, are computational. However, conventional itemset mining procedures cannot deal with the extensive dataset [10]. Map- reduce system gives parallel calculation to deal with putting away and examining sizeable dimensional dataset on item groups in distributed computing [11]. Organized data and semi-organized data are compatable with graph portrayals.

Hybrid ontological and learning medical system, which supports the central leadership about patient treatment based on frequent analysis. This framework is identified with adaption to the instance of missing data [12].

The frequency of graphs shows its adequacy by working on a broad set of accurate patient well-being data from the Candidate Detection Model (CDM), connected to the first leadership situation of overseeing dozing pill. Natural and medicinal data are expanding at rapid rates, which out-pace Moore's law [13]. This is the subsequent effect of ongoing mechanical advancement and the exploratory demeanor of individual. That prompts researchers to answer more inquiries by conducting more examinations.

Compared with sub-graphs, designs offer and intense form of coordination that catches transitive associations between graph hubs (like the companion of a companion) which are incredibly regular in the current applications [14]. The information gathered from graph analysis are grouped by category relevance. The high dimensional dataset contains the raw information which does not produce rational frequent terms [15]. Hence, it is necessary to perform the dimensionality reduction while handling large scale datasets. The frequent terms increase the large scale intention of accessing the data. The reducing factor does not corresponds with the original extracted contents. The map-reduce concentrates on the repeated terms to reduce the data. The decisions cannot do the relational activity capabilities for frequent terms.

The major trouble in adapting to large data does not only lie in its large Volume because we may mitigate this issue by expanding our registering frameworks. Different graph-based approaches can be used to handle the case of dimensionality reduction. The web graphs have been used to identify the best web pages.

3 Proposed System

Identifying the frequent sub-graph mining using graph partition is the programming for task-reducing model and it is related to addressing high dimensionality problem to stimulate large application dataset processing. Frequent sub-graph mining is to join the edges to create a course of transitional values action in time series datasets. This work aims to reduce the frequent graph mining work that

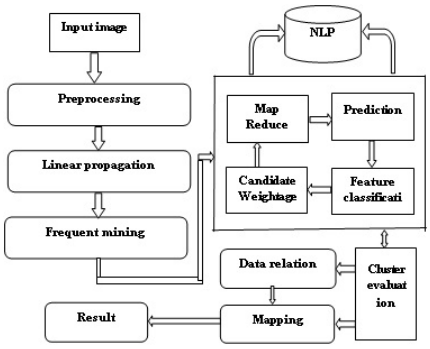


Fig. 2: Architecture diagram for MOSLFM

designs rules for each process consideration related to marginal edges with comparative values using dimensionality reduction in map- reduce. The frequent mining techniques can be used to identify the features of biomedical data towards any decision making process. Large graphs carry frequent extraction through dividing the sub-graph style. They are customarily parallelized and executed on a gigantic cluster group of resultant produces. The proposed intent analyses the Multi-Objective Sub-Linear Frequent Mining (MOSLFM) based on linear partitioned method with test case dataset are carried to predict the result from the biomedical sequence. The proposed method produces higher frequent mining accuracy result compared to other methods maintaining time preference.

Fig. 2 shows the proposed MOSLFM for frequent subgraphs implementation architecture which permits dimensional issues with no involvement with parallel and disseminated frameworks to make an effort. The proposed method utilizes data relation analysis with the candidate model to classify feature. Preprocessing is performed to remove the noisy and irrlevant data points. Our execution of map- reduces keeps running on a transaction to generate candidate model with specific values. This is very adaptable: a standard map- reduce calculation forms numerous sequence of attribute information with redundant state. The framework simply utilizes the open relational data using map- reduce programs which have been actualized. The upwards of subsequent frequent mining are executed on many mining-related clusters in each iteration to form clusters according to the relational measures.

The main objectives of the present paper are defined as follows:

(1) Map-reduce using multi-objective frequent sub-graph mining to extend regular candidate transaction algorithm: An instance for scenario between iterative of subsequence dataset is formed with reducing the importance of large graph generation process and the requirement of identifying regular frequent terms that are useful in the repeated prevention to diagnose biomedical DNA sequence dataset.

(2) Candidate selection using the maximal selection to extract the pattern that are iterated to evaluate the sub-graphs and the taxonomy is a framed probability of sequence. Finally, finds singularized maximal patterns forming iterative patterns.
(3) The relational cluster uses the spatio-temporal approach to reduce the sub-graph probability of iteration: by confidence value. Recognition in cycle samples is based on the information obtained by biological medicine. In the medical field, the average intermediate value fraction is used to get a variety of statistics about the patients. It is a series of tasks that often minimize and identify the causes of large datasets from large datasets.

In this proposed work, the data collected from the hospital information system, patient name, age, disease, location and county date are located in the database expressed with sufficient details for a patient including palliated laboratories of time series data sequences. The research analytical technologies often find the association according to frequency using association rules because the hospital has data collected from the information system.

3.1 Preprocessing

Preprocessing is performed to remove the redundant maps and item sets from the data set given.

3.2 Multi-objective Sub-linear Mining

Multi-objective sub-linear mining algorithm concentrates on mining the frequent attributes of the time series data. It also it plans to discover novel and valuable information with linear progression. This includes models that are more complicated than frequent candidate and successive instances. A multi objective selects the iteration on linear scanning calculation that goes for dissecting data from a database with frequent data. Results are typically different occurrences by petitioning the graph to the relevant nodes that require a total occurrence as shown in Table 1. Accordingly, the sub sequence frequent is a relation form that examines connections between transaction terms for the repeated count terms. We attempt to sort out this data in candidate generation and utilize the proposed sub-linear calculation to remove significant substructures from this candidate generation and thereby diminish the assignment of the process.

3.3 Linear Subgraph Frequent Sequence Pattern Mining

Frequent example mining is a fundamental advance within the time-spent relationship to predict sequences through splitting the sub-graphs based on iterative comparison. It has been selected according to the disclosure support values. Frequent subgraph patterns are formed by sets, subsequences, or substructures that appear in a dataset involving much re-occurrence compatible with sufficient reduced data. For instance, a subsequence, for selecting items, at that point, advances repeated terms followed by a selected repeated terms on the support chance that it frequently happens in a medical data history database. It is deferred to referred using the example below. The transactional term T is the term at each transaction of candidate model that detects the frequent analysis $T_s = \{H,B,Y,D,E\}$ at each transaction from obtained biomedical sequence. Table 2 provides the frequent candidate analysis

To generate a minimum support confidence using

$$M_s = \sum_{i}^{n} \frac{\text{average number of dataset}/2}{\text{total number of data}} \times \frac{\text{total number}}{\text{of transactions}} \tag{1}$$

the term transactional point of Term $T = \{T_1, T_2, T_3, T_4\}$ represented as subsequent term minimum (min) value 50% term with four transactions in average case is 2 supportive with maximal value 3 in between 2 as a minimum support.

The redundant term of candidate transaction is shown in Table 4 (Redundant confidence level).

To generate the total confidence between min-max term supportive combination of subsequence term

Table 1: Term transaction

Sequence term	Support Confidence
H	3
B	2
Y	2
D	1
E	1
V	1

Table 2: Sequence term support

Transaction T	Subsequent terms		
T_1	H	B	Y
T_2	H	Y	
T_3	H	D	
T_4	B	E	V

Table 3: Redundant Confident Value

Sequence term	Redundant confidence
$\{H,B\}$	1
$\{B,Y\}$	1
$\{H,Y\}$	2

Table 4: Redundant Confident Value

Sequence term	Confidence value
H	3
B	2
Y	2

Table 5: Confidence state of final reduction

Transaction	Support value	Confidence state
$\{H,Y\}$	2	$\{H,Y\},\{Y,H\}$

candidate as C_s.

$$C_s = \sum_{i}^{n} \frac{\text{a count of frequnet term}/2}{\text{candidate transaction count}} \times \text{total number of transactions} \tag{2}$$

by specific candidate reduction the dimensionality is reduced through analyzing the subsequence term from medical dataset organized by evaluating time complexity approach.

A substructure indicates various auxiliary structures, such as sub-graphs, sub-tree or sub lattices, which might be identified with substructures of data possibility using substructures of nodes graph information; it is known as an (often) early example. Frequent cases are found in the mining of contacts, as a basic part of the links and data found in several intricate lines.

3.4 Frequent Partition Candidate Generation

The proposed partition splits the sub sequences using the redundant-based strategy to remove the frequent data. It forms the uncovering and randomized support as well as confidence state. Two other segments are utilized to approve the run obtained from the first candidate values. Thus, this calculation makes the new set of the cluster mine tenets of frequent terms. It is rehashed till defining the client circumstances for cross approval of the affiliation rules.

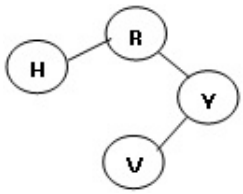
If candidate transaction $T = \{T_1, T_2\}$ be frequent data as $T_1 = \{H,B,Y,D,E,V\}$, $T_2 = \{H,B,Y,V\}$, the frequent set is considered as $\{E,F,G\}$ in both candidate pruning steps. The biomedical cancer sequence relations are identified by transaction resultant present in supportive measure of frequent data in each transaction. TF illustrated in the frequent dataset mining from each transaction illustrated in Fig. 3.

This calculation is also called as frequent partition candidate Association Rule Mining. This carries the minimum support of 2 confidence Levels in Table 4 The confidence state of final reduction is shown in Table 5.

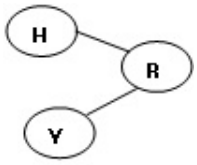
Confidences of the total support rule are $H \to Y$ and $Y \to H$. The partition mines the frequent values utilizing the linear-based subsequence probability. The proposed calculation is actualized with the engineered datasets. It delivered excellent execution and contrasted the most fascinating guideline mining calculation as well as non-repetitive calculation. Mapping transaction Map (key, value = candidate sequence set in transaction T_i): Let us consider frequency is F_i, distributed data is D_i, candidate generation is Cg_i, transaction is $E(i)$;

Input: Dataset Ds_i and F_r, $k-1 (k \geq 2)$.

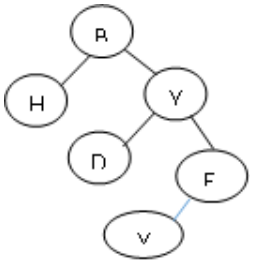
Step 1: Fetch F_r $k-1$ from D_i.



(a) Transaction T_1



(b) Transaction T_1



(c) Frequent from $T_1, T_2 \to TF$

Fig. 3: Frequent data set mining transaction

Construct to Build a hash tree for
Cg_i = generate linear task $(F_i - k - 1)$.
For each transaction $tr_i \in D_i$
Do carry the generation

Step 2: Candidate generation transaction set
$Cg_i(t)$ = subset $(Cg_i - k, T_i)$
For each transaction $i \in Cg_i(t)$ do
$E(i)\langle i, 1\rangle$
End
End

Step 3: Reduce the term for pattern sequence set
Reduce (key = item, value = count c):
For each key k do transaction Term T
For each value Val in k do
K-count $\pm$val
End

Step 4: if k val count $\geq$ min support count
$E_{\min}(i)\langle k, k - \text{count}\rangle$
End
End

Step 5: return F_s as frequent mining set

In this step, the vertical database is partitioned into equal-sized sub-sequences called remains support and distributed among several mappers. The subsequent evaluation is carried out by state of T at the time of transactions present in the position with the support confidence value.

Table 6: Subsequence frequent mining counts of confidence value

Subset frequent evaluation	Support measure	In state frequency value	State of confidence
$H \rightarrow Y$	2	3	0.66
$Y \rightarrow H$	2	2	0.99

Confidence state is measured by

$$K\text{-count} = \sum_{i=0}^{n} \frac{\text{support measure}}{\text{occurance} \in \text{transaction count}} \times 100 \quad \text{specified by each term} \vee \text{sequence representation} \tag{3}$$

Table 6 shows the sequence weight carried by medical dataset representation confidence values of frequent mining value separated by time series whose evaluation is based on the diagnosis level of the patient sequence readings.

Mappers extract the frequent singletons from their share with relative means average of cluster group by class and send them to the reducer weightage so that the class-by-reference value will be ordered. All frequent items are gathered in the reduce phase.

3.5 Linear Progress On The Frequent Dataset Cluster

In this stage, final projection of linear sub sequence values is frequently subject to average cluster value. They do not converge in clusters with minimum support and confidence values. Their frequent results are separated by the mean values of precision rate frequency of collecting information about the frequency in each transaction that has become unnecessary. Due to the data collection connection rules, diseases can be made in relation to these data types for frequent data sharing. Most values can be governed by further processing instruction using the details regarding the occurrence of these diseases at a particular time.

Algorithm C_k**:**

Sequence of frequent medical dataset of size k L_k:

Step 1: compute the size k frequent subsequence

Step 2: $L_1 = \{\text{identify frequent termsZ}\}$;

for $(k = 1;\ L_k! = n;\ k++)$

Compute to initialize for C_k

$C_k + 1 =$ frequent extracted from L_k organized;

Step 3: T contains number of transactions

Do for

Step 4: Represent the $C_k + 1$ that are contained in t as count

$L_k + 1 =$ frequent as $C_k + 1$ with C-minimum value min-support

End for

End for

Table 7: Comparison of precision rate

Methods/ dataset used	Impact of precision in%			
	Iso-Fsg	FFSM	IncGM + FFSM	MOSLFM
Pharmacy dataset	76.3	87.3	89.1	92.3
Product dataset	74.8	84.6	85.4	93.6
Hrt-ds dataset	73.2	85.5	86.8	94.2

Step 5: get Union k L_k;

The proposed mining method is often used to identify the result from subsequences to a large medical dataset. This research helps to implement large data-making decisions in clinical frequent diseases.

4 Result and Discussion

The proposed algorithm is often implemented to check MOSLFM performance rating measurements. This method of mining activity has been evaluated without assessing the effect of work on the structure of a knowledge-effecting database and minimizing the map change. The procedure is implemented by the framework to minimize the map. These techniques are used to evaluate the standard metrics. The implementation produces the scalability of map-reducing accuracy using the analysis of precious and recall rate.

The proposed algorithm produces higher efficiency and performance compared to the other dissimilar methods. The following Eqs. (4) and (5) show the precision and recall for finding frequent analysis of map-reducing accuracy calculated with true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values. The proposed system is compared with Isometric Fast sub-graph Mining (Iso-Fsg), Fast Frequent Sub-graph Mining (FFSM), Incremental Frequent Sub-graph Mining, and fast incremental approach (IncGM + FFSM)

$$\text{Precision values calculated by} = \frac{TP}{(TP+FP)} \tag{4}$$

$$\text{Recall value calculated by} = \frac{TP}{(TP+FN)} \tag{5}$$

The map-reducing accuracy is evaluated by analyzing the total number of records in

$$F\text{-measure (False reduction)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision + Recall})} \tag{6}$$

Fig. 4 shows the frequent sub-graph mining applied in various stages compared with the proposed system. The comparisons prove that the proposed system has higher-precision true values.

Table 7 proves the proposed implementation performance by producing true positive observation with

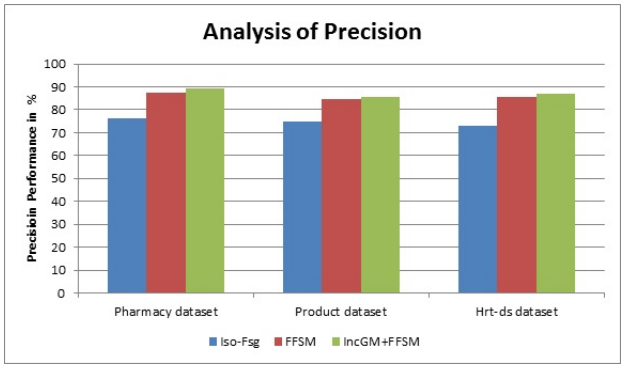


Fig. 4: Comparison of a precision rate

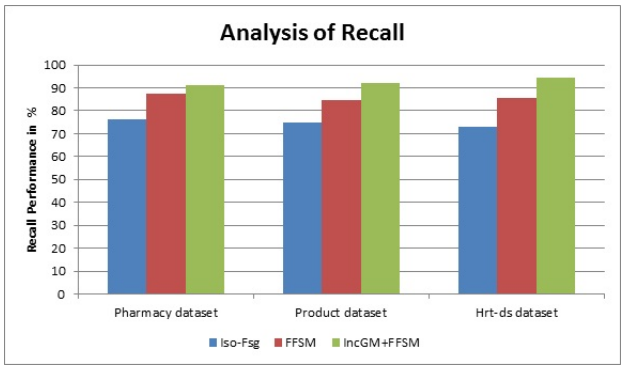


Fig. 5: Comparison of recall

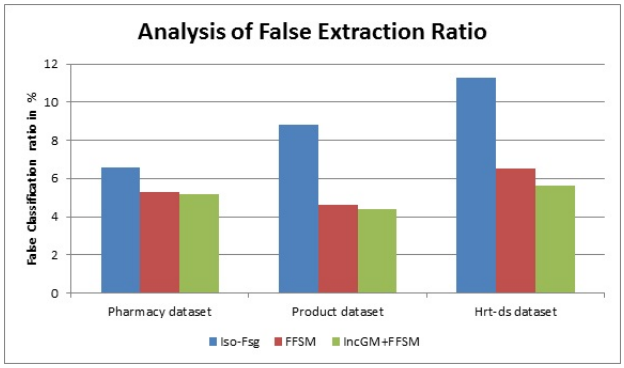


Fig. 6: Comparison of false extraction

frequent values in different dataset with dissimilar methods. The implementation of MOSLFM produces 92.3% higher precision performance than other methods.

Fig. 5 compares recall efficiency intent with other dissimilar methods. The frequent terms are iterated for total number of records in the dataset. The proposed system produces better true positive extraction rate than other methods.

Table 8 indicates the recall rate by producing various sub-graph frequent mining algorithms, the FFSM produces

Table 8: Comparison of recall

Methods/ dataset used	Impact of recall in%			
	Iso-Fsg	FFSM	IncGM + FFSM	MOSLFM
Pharmacy dataset	76.3	87.3	91.3	92.6
Product dataset	74.8	84.6	92.2	93.4
Hrt-ds dataset	73.2	85.5	94.6	95.2

Table 9: Comparison of false extraction

Methods/ dataset used	Comparison of false extraction in%			
	Iso-Fsg	FFSM	IncGM + FFSM	MOSLFM
Pharmacy dataset	6.6	5.3	5.2	4.4
Product dataset	8.8	4.6	4.4	4.3
Hrt-ds dataset	11.3	6.5	5.6	4.5

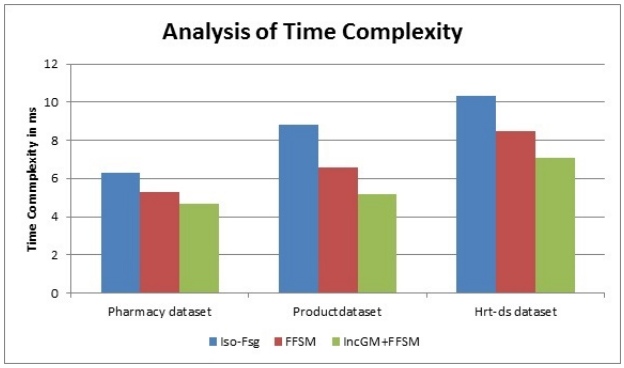


Fig. 7: Contrast of time complexity

87.3%, and IncGM + FFSM produce 91.3%. Compared with other methods, the MOSFLM has a great impact of 92.6% accuracy.

The Fig. 6 represents the ratio of false retrieval produced by different algorithms. The values obtained are compared and the proposed algorithm has achieved less false retrieval compare to other methods. The proposed system produces higher efficiency than other methods.

Table 9 shows the proposed system's false extraction to reduce the dimensionality up to 4.4% failed state compared to other methods.

Fig. 7 illustrates the execution of time complexity to analyze the frequent sub-graph with lower complexity. The contrast of time in each method varies the differential time to find the frequent mining process.

Table 10 shows the time process of evaluation to execute frequent mining process of evaluation. The differential dataset produces varied time to process the data. The proposed MOSFLM produces 4.2 ms lower time complexity than other methods.

The data is processed to time complexity (T_c). The proposed MOSFLM produces 4.2 ms lower time complexity than other methods.

Table 10: Contrast of time complexity

Methods/ dataset used	Impact of time complexity in ms			
	Iso-Fsg	FFSM	IncGM + FFSM	MOSLFM
Pharmacy dataset	6.3	5.3	4.7	4.2
Productdataset	8.8	6.6	5.2	4.7
Hrt-ds dataset	10.3	8.5	7.1	6.7

5 Conclusion

The frequent sub-graph mining from map-reducing frame work has become a crucial task for redundant mining process. The proposed multi objective sub linear frequent mining (MOSLFM) algorithm for the problem of dimensionality reduction on high dimensional data set has been evaluated under different parameters. This is conducted by finding the sub-graph frequent similarity values between data with candidate generation by iteration. It produces higher efficiency of frequent analysis precision value which is (92.3)% and the recall value is 92.6%; it is a better performance. Also, lower false rate with 4.4% with lower time complexity as 4.2 ms is compared to previous methods. The case study shows procedure of fixing missing values.

References

[1] J. Huan, W. Wang, and J. Prins. Spin: Mining maximal frequent sub-graphs from graph databases, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and data mining pp. 581–586 (2004).

[2] M. Kuramochi and G. Karypis, Finding frequent patterns in a large sparse graph, Data Mining Knowl. Discov, Vol. 11, No. 3, pp. 243–271 (2005).

[3] J. Wang, W. Hsu, M. L. Lee, and C. Sheng, A partition-based approach to graph mining, in Proc. 22nd Int. Conf. Data Eng., pp. 74–80 (2006).

[4] J. Dean and S. Ghemawat, Map reduce: simplified data processing on large clusters. Commun. ACM, Vol. 51, No. 1, pp. 107–113 (2008).

[5] L.T. Thomas, S.R. Valluri, and K. Karlapalem, Margin: Maximal frequent subgraph mining, ACM Trans. Knowl. Discov. Data, Vol. 4, No. 3 (2010).

[6] G.-P. Chen, Y.-B. Yang, and Y. Zhang, Map reduce-based balanced mining for closed frequent itemset, Proc. IEEE 19th Int. Conf. Web Serv., pp. 652–653. (2012).

[7] Atif Khan, John A. Doucette, and Robin Cohen, Insights into the value of machine learning. Validation of an ontological for patient treatment using a repository of patient data using a medical decision support system, ACM Trans. Intell. Syst. Technol, Vol. 4, No. 3, pp. (2013).

[8] George Tzanis.. Biological and Medical Big Data Mining. Int. J. Knowl. Disc. Bio info. Vol. 4, No. 01, pp. 42-56. (2014).

[9] S. Skiadopoulos, M. Elseidy, E. Abdelhamid, and P. Kalnis, Grami: Frequent subgraph and pattern mining in a single large graph, Proc. Very Large Database Endow., Vol. 7, pp. 517–528 (2014).

[10] X. Xiao, W. Lin, and G. Ghinita, Large-scale frequent subgraph mining in map- reduce, Proc of Int. Conf. Data Eng., pp. 844–855. (2014).

[11] X. Jin, B. W. Wah, X. Cheng, and Y. Wang. Significance and challenges of big data research. Big Data Research, Vol. 2, No. 2, pp. 59-64 (2015).

[12] O.Y. Al-Jarrah, P.D. Yoo, S. Muhaidat, G.K. Karagiannidis, and K. Taha, Efficient machine learning for big data: A review. Big Data Research, Vol. 2, No. 3, pp. 87–93 (2015).

[13] Hoang Nguyen, Wei Liu and Fang Chen, Discovering Congestion Propagation Patterns in Spatio-Temporal Traffic Data. IEEE Transactions on Big Data, Vol 3, No. 2, pp. 169-180 (2017).

[14] Dr. Neelesh Jain, A Review on Big Data Environment on Different Frameworks, Techniques, and Tools & quot, International Journal of Core Engineering Management, Vol. 3, No. 3, pp. 24–30 (2016).

[15] José María Luna, Francisco Padillo, Mykola Pechenizkiy and Sebastián Ventura, Apriori Versions Based on Map reduce for Mining Frequent Patterns on Big Data IEEE Transactions on Cybernetics, Vol. 48, No. 10, pp. 2851–2865 (2018).

G. Elangovan received B.E. Degree in Computer Science and Engineering from Anna University, Chennai. Then, he received M.E Degree in Computer Science and Engineering from Bannari Amman Institute of Technology, Anna University, Coimbatore. In addition, he is pursuing his Ph.D. in Information and Communication Engineering at Anna University, Chennai. He serves as an Assistant Professor at the Department of Computer Science and Engineering at Velammal Institute of Technology, Chennai. Furthermore, his research interest includes Big Data Analytics, Theoretical Computer Science and Cloud Computing.

G. Kavya received her B.E. Degree in Electronics and Communication Engineering from Government College of Engineering, Salem under Madras University. Then, she completed her M.E in Electronics Engineering from Madras Institute of Technology, Anna University, Chennai and her Ph.D. in Electronics Engineering from Sathyabama University, Chennai, India 2015. She is highly involved in teaching and research. She has been teaching Engineering students for 19 years. She serves as Professor of Electronics and Communication Engineering at S.A.Engineering College, Chennai. Her professional interests includes VLSI, Wireless Communication, Embedded Systems, Telemedicine and Data Analytics. She is a member in various professional societies, such as IEEE, IETE ISTE and IAENG. She has published nearly 22 papers in well-reputed journals, conference proceedings and magazines. She is the Principal Investigator of a research project funded by Indian Space Research Organization (ISRO), Government of India.