International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 3, May-Jun-2019, ISSN (Online): 2395-566X

Scaled Data Mining Algorithms for Frequent Item sets Mining over Large Scale Data

M. Sathish Kumar Assistant Prof. R. Saravanan
Dept. of Computer Science & Applications
SCSVMV University
Kanchipuram, Tamil Nadu, India

Abstract- **Information Analytics assumes a significant job in the basic leadership process. Bits of knowledge from such example examination offer immense advantages, including expanded income, cost cutting, and improved upper hand. Be that as it may, the shrouded examples of the incessant item sets become additional tedious to be mined when the measure of information increments over the time. Additionally, noteworthy memory utilization is required in mining the shrouded examples of the incessant item sets because of a substantial calculation by the calculation. Hence, an effective calculation is required to mine the shrouded examples of the incessant item sets inside a shorter run time and with less memory utilization while the volume of data increases over the time period. This paper reviews and presents a comparison of different algorithms for Frequent Pattern Mining (FPM) so that a more efficient FPM algorithm can be developed.**

Keywords- **Scaled Data Mining, Frequent Item sets, Large Scale Data, Data Mining, FPM Algorithm, Hidden Pattern**

I. INTRODUCTION

what's more, arranging huge informational index to distinguish concealed examples and set up connections to take care of the issues through information examination. It is additionally breaking down shrouded examples of information for classification into valuable data which is gathered and collected in like manner zones, Such as information distribution centers, for proficient investigation, information mining calculation, specialization business basic leadership and other data to at last cut expense and increment income. The significant advances associated with an information mining procedure are:

- Concentrate, change and burden information distribution center.
- Store and oversee information in multidimensional databases.
- Give information access to business examiner utilizing application programming.
- Present dissected information in simple reasonable structures, for example, chart.

II. DATA MINING PARAMETERS

In information mining affiliation principles are made by examining information for incessant examples, at that point utilizing the help and certainty criteria to find the most significant association with in the information. Other information mining parameter incorporates arrangement or way investigation, grouping, bunching and guaging. Succession or way examination parameters search for examples where one occasions prompts another later occasion a grouping parameter is an arranged rundown of sets of things and it is a common type of data structure found in large databases. A classification parameter looks for new pattern and might result in a change in the way the data is organized. Clustering parameter finds and visually document groups of fax that were previously unknown data items. The Knowledge Discovery in Databases (KDD) process having different stages they are.

- Selection
- Pre-processing
- Transformation
- Data mining
- Interpretation / Evaluation

Data mining involves six common classes task

- Anomaly deduction-identification of unusual data records
- Association rule learning- search relationship between variables
- Clustering – Discovering groups and structure in the data
- Classification – Generalizing known structure to apply new data
- Regression – Estimating the relationship among data or data sets
- Summarization – providing a compact representation of the data set including visualization and report generation

Data analysis play a major role in the decision making process, search pattern analysis benefits that is cost cutting, improved competitive and increased revenue. In the current framework, the concealed example of the regular itemsets become additional tedious mining process when the measure of information increments. Because of overwhelming calculation by the mining calculations required noteworthy memory utilization for mining the shrouded example of the incessant item sets. In my proposed framework a productive calculation is required to run the mining procedure in shorter term and less memory utilization when the volume of information increments.

Due to heavy computation by the mining algorithms needed significant memory consumption for mining the hidden pattern of the frequent item sets. In my proposed system an efficient algorithm is required to run the mining process in shorter duration and less memory consumption when the volume of data increases over the short time duration. The objective of my paper is to compare the significant algorithm, so that a more efficient algorithm can be developed.

III. EXISTING SYSTEM

In the current framework, the concealed example of the successive item sets become additional tedious mining process when the measure of information increments. Because of substantial calculation by the mining calculations required critical memory utilization for mining the concealed example of the continuous item sets.

IV.DRAWBACKS

The presentation of Frequent Pattern Mining (FPM) calculations from the parts of execution run time and memory utilization in mining the successive itemsets from an informational collection. The execution runs time of certain calculations for flat format information.

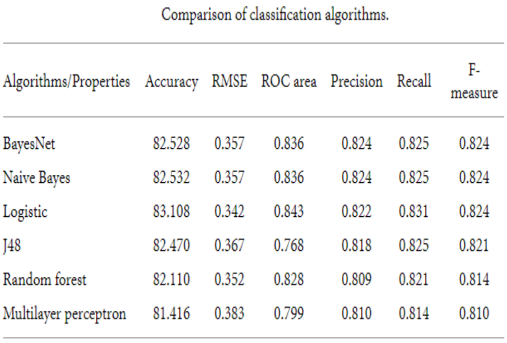
V. PROPOSED SYSTEM

In my proposed system an efficient algorithm is required to run the mining process in shorter duration and less memory consumption when the volume of data increases over the short time duration. Along these lines, a proficient calculation is required to mine the shrouded examples of the continuous item sets inside a shorter run time and with less memory utilization while the volume of information increments over the time span. This paper audits and shows an examination of various calculations for Frequent Pattern Mining (FPM) with the goal that a progressively productive FPM calculation can be created.

VI.RESEARCH METHODOLOGY DATA SOURCE

1. Comparison of Classification Algorithm.

Table 1 Comparison of Classification Algorithm

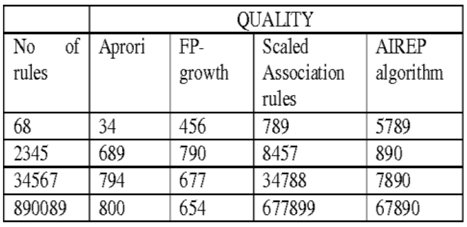


Comparison of classification algorithms.

Algorithms/Properties	Accuracy	RMSE	ROC area	Precision	Recall	F-measure
BayesNet	82.528	0.357	0.836	0.824	0.825	0.824
Naive Bayes	82.532	0.357	0.836	0.824	0.825	0.824
Logistic	83.108	0.342	0.843	0.822	0.831	0.824
J48	82.470	0.367	0.768	0.818	0.825	0.821
Random forest	82.110	0.352	0.828	0.809	0.821	0.814
Multilayer perceptron	81.416	0.383	0.799	0.810	0.814	0.810

2. Quality Comparison of Algorithms

Table 2 Quality of Comparison Algorithm



	QUALITY			
No of rules	Aprori	FP-growth	Scaled Association rules	AIREP algorithm
68	34	456	789	5789
2345	689	790	8457	890
34567	794	677	34788	7890
890089	800	654	677899	67890

3. Association Mining searches for frequent items in the data-set. In frequent mining usually the interesting associations and correlations between item sets in transactional and relational databases are found. In short, Frequent Mining shows which items appear together in a transaction or relation. Frequent mining is generation of association rules from a Transactional Dataset.

3.1 Support - It is one of the measure of interestingness. This tells about usefulness and certainty of rules. **5% Support** means total 5% of transactions in database follow the rule.

Support(A -> B) = Support_count(A ∪ B)

2.Confidence- A confidence of 60% means that 60% of the customers who purchased a milk and bread also bought butter.

Confidence(A -> B) = Support_count(A ∪ B) / Support_count(A)

If a rule satisfies both minimum support and minimum confidence, it is a strong rule.

3.Support_count(X) - Number of transactions in which X appears. If X is A union B then it is the number of transactions in which A and B both are present.

3.1Maximal Itemset- An itemset is maximal frequent if none of its supersets are frequent.

3.2 Closed Itemset- An itemset is closed if none of its immediate supersets have same support count same as Itemset.

3.3K- Itemset - Itemset which contains K items is a K-itemset. So it can be said that an itemset is frequent if the corresponding support count is greater than minimum support count.

- Lets say minimum support count is 3

- Relation hold is maximal frequent => closed => frequent

VII. METHODOLOGY

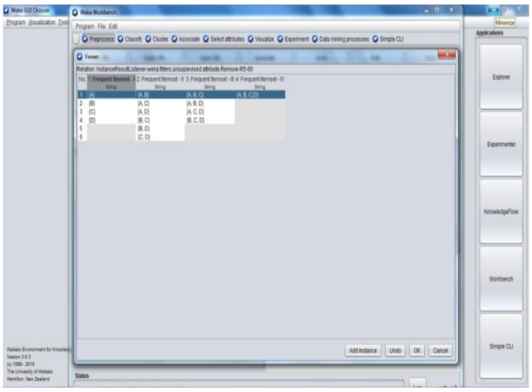
1. Data Sets

1. Frequent:
   {A} = 3; // not closed due to {A, C} and not maximal
   {B} = 4; // not closed due to {B, D} and no maximal
   {C} = 4; // not closed due to {C, D} not maximal
   {D} = 5; // closed item-set since not immediate super-set has same count. Not maximal
2. Frequent:
   {A, B} = 2 // not frequent because support count < minimum support count so ignore
   {A, C} = 3 // not closed due to {A, C, D}
   {A, D} = 3 // not closed due to {A, C, D}
   {B, C} = 3 // not closed due to {B, C, D}
   {B, D} = 4 // closed but not maximal due to {B, C, D}
   {C, D} = 4 // closed but not maximal due to {B, C, D}
3. Frequent:
   {A, B, C} = 2 // ignore not frequent because support count < minimum support count
   {A, B, D} = 2 // ignore not frequent because support count < minimum support count
   {A, C, D} = 3 // maximal frequent

{B, C, D} = 3 // maximal frequent

4.Frequent:

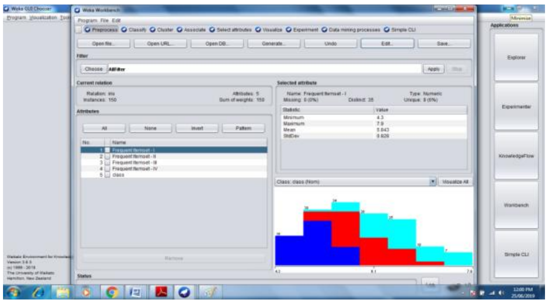
{A, B, C, D} = 2 //ignore not frequent



Relation: InstanceResultListener-weka.filters.unsupervised.attribute.Remove-R5-66

No.	1: Frequent Itemset - I String	2: Frequent Itemset - II String	3: Frequent Itemset - III String	4: Frequent Itemset - IV String
1	{A}	{A, B}	{A, B, C}	{A, B, C, D}
2	{B}	{A, C}	{A, B, D}	
3	{C}	{A, D}	{A, C, D}	
4	{D}	{B, C}	{B, C, D}	
5		{B, D}		
6		{C, D}		

Fig1 Association Rule Frequent Itemset



Filter: AllFilter

Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150

Selected attribute
Name: Frequent Itemset - I
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Attributes

No.	Name
1	Frequent Itemset - I
2	Frequent Itemset - II
3	Frequent Itemset - III
4	Frequent Itemset - IV
5	class

Class: class (Nom)

Fig. 2 Association Rule Frequent Itemset - Dataset

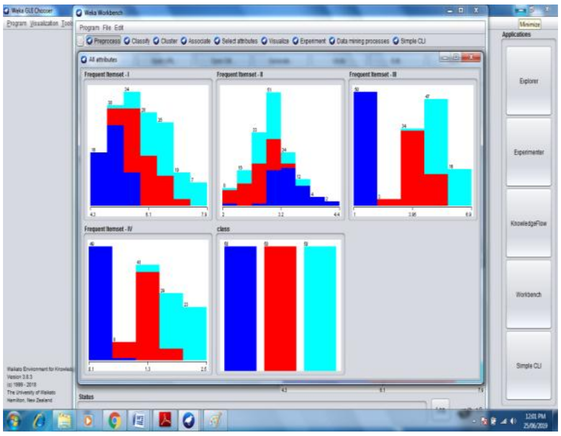


Fig.3 Association Rule Frequent Itemset Graphical Representation

1. Weka Tool- The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning toolkit introduced by Waikato University, New Zealand. At the time of the project's inception in 1992. WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. It can be run on Windows, Linux and Mac. It consists of collection of machine learning algorithms for implementing data mining tasks.

Data can be loaded from various sources, including files, URLs and databases. Supported file formats include WEKA‟s own ARFF format, CSV, Lib SVM‟s format, and C4.5‟s format. The second panel in the Explorer gives access to WEKA‟s classification and regression algorithms [12]. The corresponding panel is called "Classify" because regression techniques are viewed as predictors of "continuous classes". By default, the panel runs a cross validation for a selected learning algorithm on the dataset that has been prepared in the Pre-process panel to estimate predictive performance.

VIII. LITERATURE SURVEY

Agarwal RC, Aggarwal CC, Prasad VVV : In this paper we propose algorithms for generation of frequent itemsets by successive construction of the nodes of a lexicographic tree of itemsets. We discuss di erent strategies in generation and traversal of the lexicographic tree such as breadth 1rst search, depth 1rst search or a combination of the two. These techniques provide di erent trade-o s in terms of the I/O, memory and computational time requirements. We use the hierarchical structure of the lexicographic tree to successively project transactions at each node of the lexicographic tree, and use matrix counting on this reduced set of transactions for frequent itemsets.

We tested our algorithm on both real and synthetic data. We provide an implementation of the tree projection method which is up to one order of magnitude faster than other recent techniques in the literature. The algorithm has a well structured data access pattern which provides data locality and reuse of data for multiple levels of the cache. We also discuss methods for parallelization of the Tree Projection algorithm.

Aggarwal CC, Han J (eds): Therefore, an efficient algorithm is required to mine the hidden patterns of the frequent itemsets within a shorter run time and with less memory consumption while the volume of data increases over the time period. This paper reviews and presents a comparison of different algorithms for Frequent Pattern Mining (FPM) so that a more efficient FPM algorithm can be developed.

Bhuiyan MA, Hasan MA: This chapter will provide a detailed survey of frequent pattern mining algorithms. A wide variety of algorithms will be covered starting from Apriori. Many algorithms such as Eclat, Tree Projection, and FP-growth will be discussed. In addition a discussion of several maximal and closed frequent pattern mining algorithms will be provided. Thus, this chapter will provide one of most detailed surveys of frequent pattern mining algorithms available in the literature.

Agrawal R, Srikant R: This is a very long and complicated paper about taking a set of transactions (what the paper calls basket data) and finding association rules in them. For example, a marketing firm might want to ask “What percentage of people who bought X also bought Y?” Another question might be “What two items are most popular among people between ages 18 and 25.”The naive solution would be to do an exhaustive search across all possible subsets of items and count how many satisfy the predicate conditions we are looking for. This approach, although it would be efficient space-wise (only store the combinations we need) would waste a lot of time (creating all possible combinations). This paper presents a few algorithms that start with a seed itemset (one that already satisfies the Boolean predicates we wish to evaluate) and grow them into itemsets of maximal size.

Baralis E, Cerquitelli T, Chiusano S, Grand A: This paper proposes a parallel disk-based approach to efficiently supporting frequent itemset mining on a multi-core processor. Our parallel strategy is presented in the context of the VLDB-Mine persistent data structure. Different techniques have been proposed to optimize both data- and compute-intensive aspects of the mining algorithm. Preliminary experiments, performed on both real and synthetic datasets, show promising results in improving the efficiency and scalability of the mining activity on large datasets.

Chang V: In this paper, we propose an alternative service which uses the elastic capacities of Cloud Computing to escape the limitations of the desktop and produce accurate results more rapidly. The Business Intelligence as a Service (BIaaS) in the Cloud has a dual-service approach to compute risk and pricing for financial analysis. The first type of BIaaS service uses three APIs to simulate the Heston Model to compute the risks and asset prices, and computes the volatility (unsystematic risks) and the implied volatility (systematic risks) which can be tracked down at any time. The second type of BIaaS service uses two APIs to provide business analytics for stock market analysis, and compute results in the visualized format, so that stake holders without prior knowledge can understand.

Chee C-H, Yeoh W, Tan H-K, Ee M-S- This paper presents an integrated framework that comprises an automatic weighting method for assessing data quality (DQ) of the framework so as to better support the business intelligence (BI) usage. Specifically, we utilize business process modeling (BPM) notation and information product map and frame them into a hierarchical mapping structure. Furthermore, we develop and demonstrate an automatic weight-assignment method for evaluating critical dimensions (i.e., completeness and accuracy) of DQ of the integrated framework.

Through a plan science worldview, the adequacy of the structure and the related DQ weighting technique has been thoroughly approved by workforce the executives clients of a college. The system together with the DQ weighting strategy constructs client certainty by upgrading the detestability of a BI item. The programmed DQ weight task additionally gives better time proficiency in light of the fact that the heaviness of every datum trait is resolved consequently dependent on its utilization on the BI dashboard.

El-Hajj M, Zaiane OR: In order to allow a fair comparison of these algorithms, we performed an extensive set of experiments on several real life data sets, and a few synthetic ones. Among these are three new data sets,i.e. a super market basket dataset donated by Tom Brijs[9], a dataset containing click-stream data of a Hungarian on-line news portal donated by Ferenc Bodon [8], and a dataset containing Belgian traffic accident descriptions donated by Karoline Geurts[13].

Feddaoui I, Felhi F, Akaichi J: In this context, an algorithms number a frequent item sets and association rules extraction were presented. Uncommon element of these calculations is to age an enormous number of standards, making their abuse a troublesome undertaking. In this paper we will present another calculation for affiliation rules extraction. Proposed arrangement depends on two, in particular: visit item set extraction, and from these, it extricates affiliation rules.

Gullo F: Over the most recent couple of decades, information mining has been broadly perceived as a ground-breaking yet adaptable information investigation apparatus in an assortment of fields: data innovation in primis, yet in addition clinical prescription, human science, material science. In this specialized note we give an abnormal state outline of the most conspicuous assignments and techniques that structure the premise of information mining. The note additionally centers around probably the latest yet encouraging interdisciplinary parts of information mining.

IX.CONCLUSION

Our proposed technique keeps away from three of the fundamental disadvantages displayed by the standard mining calculations: creation of a high number of principles, disclosure of uninteresting examples and low execution. The outcomes demonstrate that the affiliation decide calculations that we assessed perform distinctively on our genuine world datasets than they do on the counterfeit dataset. used for Association Rule Algorithm along with FP Growth Algorithm research frequent itemset datasets.

BIBLIOGRAPHY

1. Agarwal RC, Aggarwal CC, Prasad VVV (2001) A tree projection algorithm for generation of frequent item sets. J Parallel Distrib Comput 61(3):350–371CrossRefzbMATHGoogle Scholar
2. Aggarwal CC (2014) An introduction to Frequent Pattern Mining. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 1–14Google Scholar
3. Aggarwal CC, Bhuiyan MA, Hasan MA (2014) Frequent Pattern Mining algorithms: a survey. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 19–64Google Scholar
4. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Paper presented at the proceedings of the 20th international conference on very large data bases, SantiagoGoogle Scholar
5. Baralis E, Cerquitelli T, Chiusano S, Grand A (2013) P-Mine: parallel itemset mining on large datasets. In: Paper presented at the 2013 IEEE 29th international conference on data engineering workshops (ICDEW), Brisbane Google Scholar
6. Chang V (2014) The business intelligence as a service in the cloud. Future Gener Comput Syst 37:512–534CrossRefGoogle Scholar
7. Chee C-H, Yeoh W, Tan H-K, Ee M-S (2016) Supporting business intelligence usage: an integrated framework with automatic weighting. J Comput Inf Syst 56(4):301–312Google Scholar
8. El-Hajj M, Zaiane OR (2003) COFI-Tree mining—a new approach to pattern growth with reduced candidacy generation. In: Paper presented at the workshop on frequent itemset mining implementations (FIMI’03) in conjunction with IEEE-ICDM, Melbourne Google Scholar
9. Feddaoui I, Felhi F, Akaichi J (2016) EXTRACT: new extraction algorithm of association rules from frequent itemsets. In: Paper presented at the 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), San Francisco Google Scholar.
10. Gullo F (2015) From patterns in data to knowledge discovery: what data mining can do. Phys Proc 62:18–22CrossRef Google Scholar.