Volume 6, Issue 7, July 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study of Data Mining Algorithms for Efficiency and Effectiveness

Pranjali P. Ghode
Department of Computer Science and Engineering,
Sipna College of Engineering and Technology, Amravati, Maharashtra, India

Abstract— Data mining is the process of extracting or mining knowledge from large amount of data. There are many databases and warehouses available all around the world. The major task of data mining is to utilize the information or knowledge from the database. Data mining has a large family of different algorithms and the scope of research is rapidly increasing to improve accuracy. In this paper, we deal with algorithmic aspects of data mining. So, we study the mining algorithms and analyze these algorithms to evaluate their performance.

Keywords— Data mining, mining algorithms, databases.

I. INTRODUCTION

Data mining refers to extracting interesting information or patterns from large information repositories such as: relational database, data warehouses, XML repository. It is a process of inferring knowledge from such huge data. The most important usage: Customer Segmentation in Marketing, Shopping Cart analyses, Management of Customer Relationships, Web usage Mining, Text Mining.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detailed transactional data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

II. RELATED WORK

In the data mining algorithms, there are various studies and proposals presented in literature till date. The study and review on some researches related to the pattern mining algorithms are presented below.

K-Means [1] was proposed by MacQueen in 1967. K-means is one of the most famous data mining algorithm. Also, EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin. Next, CHAID is a fundamental decision tree learning algorithm. It was developed by Gordon V Kass in 1980. Next, GSP[2] is one of the first algorithm for discovering sequential patterns proposed by Srikant et al. (1992). In 1993, the Apriori Algorithm was proposed by Agrawal & Srikant [3]. Next, AprioriTID was proposed in the same article. Also, they proposed FPGrowth_association_rules algorithm [4]. In 1994, Ng & Han introduced the algorithm CLARANS. In 1996, Cheeseman & Stutz proposed AUTOCLASS. In 1998, MNR algorithm was proposed by Kryszkiewicz. In 1998, Sander et al. proposed the algorithm GDBSCAN which uses the same two parameters as algorithm DBSCAN. Next, AprioriClose algorithm was proposed by Pasquier et al.(1999). Next, Gibson proposed STIRR algorithm and Ganti proposed CACTUS algorithm which looks for hyper-rectangular clusters (called interval regions). Next, Karypis proposed CHAMELEON hierarchical agglomerative algorithm which utilizes dynamic modeling in cluster aggregation.

In 2000, Zaki [5] proposed RuleGen algorithm for discovering sequential and Han [6] proposed FPGrowth algorithm for discovering frequent itemsets. Next, (Tan et al., KDD 2000; Tan, Steinbach & Kumar, 2006, p.469) proposed Indirect association rules algorithm. In 2001, Pei et al. proposed H-Mine algorithm and Zaki proposed Eclat for discovering frequent itemsets. Also, Zaki proposed SPADE algorithm which discovers all frequent sequential patterns. Next, Pei proposed PrefixSpan [7] algorithm for discovering sequential patterns. Next, the SSVM was proposed by Lee, Y.J. and O.L. Mangasarian (2001) [8]. Lee and Mangasarian (2001) proposed Reduced Support Vector Machines (RSVM) [9]. In 2002, Ayres proposed SPAM algorithm and Zaki proposed Charm algorithm. Also, the Ertoz proposed algorithm SNN which blends a density based approach. In 2003, Yan proposed CloSpan pattern-growth algorithm [10] and Tzvetkov proposed TSP [11] algorithm for discovering sequential patterns. The DCI_Closed algorithm [12] is one of

the fastest algorithms. It was proposed by Lucchese et al. (2004). In 2005, Borgelt proposed Relim algorithm. Y. Liu, W-K. Liao[13] proposed IGB algorithm. Next, Fournier-Viger[14] proposed Sporadic association_rules algorithm for mining perfectly sporadic association rules. Next, L. Szathmary[15] proposed Apriori Inverse algorithm. Next, Liu proposed Two-Phase algorithm and Bastide, Y., Taouil[16] proposed LAPIN algorithm which mines sequential patterns.

In 2006, Zaki, M.J.[17] proposed the Hirate-Yamana algorithm for discovering frequent sequential patterns and L. Szathmary has proposed Charm-MFI algorithm for discovering frequent maximal itemsets in a transaction database. Next, Yun Sing Koh[18] proposed Closed_association_rules algorithm for mining "closed association rules", which are a concise subset of all association rules. Next, UApriori is an algorithm which was proposed by Chui et al. (2007). Also, Wang proposed BIDE [19] algorithm for discovering sequential patterns. Next, Zart is an algorithm for discovering frequent closed itemsets and their corresponding generators in a transaction database. Next, Szathmary proposed AprioriRare algorithm [20]. In 2008, Gao proposed FEAT algorithm and Laszlo Szathmary[21] proposed FHSAR algorithm. In 2009, Yen proposed CloStream algorithm and Chang proposed EstDec algorithm. In 2010, Deng & Xu proposed VME algorithm and Fournier-Viger proposed CMRules algorithm.

III. PROPOSED METHOD

A. Proposed System

We propose a framework which will guide users in the selection of suitable data mining algorithms for their data mining work. This framework will classify the data mining algorithms and will show their characteristics.

B. Proposed System Design

In proposed system, the problem regarding selection strategy of data mining algorithms and the complexity of input itemset is creating problem for the developer and researchers. There is requirement of the framework which is able to provide the information regarding complexity but in data mining algorithms, complexity will not indicate the efficiency and particular algorithm. So, we need to include principles and techniques of every algorithm so that selection strategy is become simple. So, we have proposed following framework for big analysis of present data mining algorithms and all the algorithms are categorised into

1) Frequent item set mining
2) Association rule mining
3) Sequential pattern mining
4) Classification
5) Clustering algorithm

All above five categories are analysed on the basis of the principles used by algorithms, the techniques adopted by algorithms, and the complexities of the algorithms.

So sample framework is shown below.

Table I: Sample framework

Sr. No.	Algorithms	Principle	Technique	Complexity
A. Frequent Itemset Mining				
B. Association Rule Mining				
C. Sequential Pattern Mining				
D. Classification				
E. Clustering Algorithms				

A. Filling the data in the proposed framework

There are 168 algorithms that we have found from different research papers and we have categorised these algorithms on the basis of various parameter like principles, techniques & time complexity. We have collecting information from all sources like different books, journals.etc. With the help of above information, we have formed a tentative framework which looks as-

Table II: framework filled with different data mining algorithm

Sr.No.	Algorithms	Principle	Technique	Complexity
A. FREQUENT ITEMSET MINING				
1	**Apriori**	Apriori Principle(Candidate Generate& Test)	Join &Prune Method	O(MN+ (1-R^M)/(1-R))
2	**APRIORITID**	Apriori Principle(Candidate Generate& Test)	Performing Intersection Method(The Intersection of Two Sets of Transaction Ids)	O(MN+ (1-R^M)/(1-R))
3	**Apriori Hybrid**	Apriori Principle(Candidate Generate& Test)	Join with Prune method and Transaction Id	O(n+1)
4	**SAM**	Split And Merge Scheme	Perfect Extension Pruning	O(nlogn)
5	**Improved Apriori Based On Transaction Compression**	Apriori Principle(Candidate Generate& Test) with Expand	Preprocessing is Done on the Database to Remove Redundancy	O(n+1)
6	**OOO Algorithm**	Apriori Principle(Candidate Generate& Test)	New Strategy for Accessing Database Items	O(n+1)
7	**Improved Algorithm for Weighted Apriori**	Apriori Principle(Candidate Generate& Test)	Large Amount of Items Divided Categorized into Grps	O(n/2)
8	**ImprovedApriori Based on Matrix**	Apriori Principle(Candidate Generate& Test)	Binary Matrix Reduce the Database Scans	O(n^2)
9	**Fast Update 2 (Fup2)**	Apriori-Based Algorithm	Candidate Generate & Test	O(n/2)
10	**Sliding Window Filtering (SWF)**	Apriori-Based Algorithm.	Partition the Database	O(n+1)
11	**Fpgrowth**	Fp Growth Principle	Projected Database Based Data Mining Techniques	O(n^2)
12	**Direct Hashing And Pruning (DHP)**	Apriori-Based Algorithm.	Using Hashing Technique	O(n^2)
13	**Sampling Algorithm**	Sampling Scheme	Pick any Random Sample for Checking Frequency of whole Database at Lower Threshold Support	O(n^2)
14	**Eclat**	Apriori Principle(Candidate Generate& Test)	Depth-First Search &Use Intersection of Transaction Ids List	O(n^2)
15	**Hmine**	Pattern-Growth Approach	Uses the Hyperlink Pointers & Divide and Conquer Method	O(MN+ (1-R^M)/(1-R))
16	**DCI_CLOSED**	Frequent Closed Itemsets	Closure Climbing	O(nlogn)
17	**Charm_Bitset**	Bottom -Updepth-First Browsing of a Prefix Tree	Candidate Generate & Test	O(n+1)
18	**CHARM_MFI**	Bottom -Updepth-First Browsing of a Prefix Tree	Performing Post-Processing after Discovering Frequent Closed Itemsets	O(n^2)

19	**DSM-FI**	Single Pass Algorithm	SFI-Forest Method	O(n^2)
20	**DIFFSET**	Vertical Data Representation	Depth-First Search Algorithm with Comparing Neighbour Element	O(n+1)
21	**DEFME**	Depth-First Minimal Pattern Mining	Fast Minimality Checking	O(n+1)
22	**Pascal**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(n+1)
23	**Zart**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(n^2)
24	**APRIORIRARE**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(n^2)
25	**APRIORIINVERSE**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(n+1)
26	**ESTDEC**	Approximate Algorithm.	Depth-First Search Algorithm	O(n+1)
27	**Uapriori**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(n+1)
28	**Two-Phase**	Apriori-Based Algorithm	Merge Technique	O(n+1)
29	**Hui-Miner**	High Utility Itemset Mining	High Utility Itemsets Without Candidate Generationusing Pruning Strategy	O(n+1)
30	**VME**	Apriori-Based.	To Compute the Gain of an Itemset via Union Operations on Product Id_Nums	O(n+1)
31	**Msapriori**	Apriori-Based Algorithm.	Depth-First Search Algorithm	O(MN+ (1-R^M)/(1-R))
32	**Fpmax**	Fpgrowth	Depth-First Search Algorithm with Comparing Neighbour Element	O(n^2)
33	**Pincersearch**	Bottom -Up Search	Two-Way-Search Based on MFS	O(n^2)
34	**MMFI**	Ldquonbnrdquo (Node By Node) Method	Two-Way-Search Based on MFS	O(n^2)
35	**Reverse Apriori**	Reverse Approach	Two-Way-Search Based On MFS	O(n^2)
36	**DTFIM**	Distributed Trie-Based Frequent Itemset Mining	Two-Way-Search Based On MFS	O(n^2)
37	**Genmax**	Mining Maximal Itemsets	Two-Way-Search Based On MFSconsidering Big Number	O(N^2)
38	**Fp-Close**	Fp-Growth	Depth First Search & Closure Climbing	O(N^2)
39	**Closet**	Fp-Growth	Closure Climbing	O(n^2)
40	**Closet+**	Fp-Growth	Upward Checking	O(n^2)

**N - No. of transactions, M – Threshold no. of transactions,R - No. of Unique Items*
n= Number of iteration,n=No. of Items

B. ASSOCIATION RULE MINING

1	GENETIC ALGORITHM	Win's Principal Of Survival Of The Fittest In Natural Genetics	GA Process	O(n^2)
2	SETM	Like AISAlgo.	Sql Join Operation	O(n^2)

3	AIS	Apriori-Based Algorithm.	Candidate Generation	O(n^2)
4	MCISI	Mining Imperfectly Sporadic Rules	Novel Itemset-Tidset Search Space	O(n^2)
5	IGB	Informative And Generic Basis of Association Rules	Novel Generic Base Approach	O(n^2)
6	Sporadic_Association_RulesApriori-Inverse	Downward-Closure Principle	Discovering Sporadic Rules by ignoring all Candidate Itemsets above a Maxi-Mum Support Threshold.	O(nlogn)
7	MMR	Minimum Condition Maximum consequence	Using Representative AssociationRules (RR)	O(nlogn)
8	Indirect_Association_Rules	Indirect Associations	Using HI-Mine	O(nlogn)
9	FHARM	Based on Completely Hiding Sensitive Association Rule	Performed Heuristic Approachesfor Avoiding Hidden Failures	O(nlogn)
10	TOPKRULES	Rule Based	Performed Rule Expansions & Several Optimization	O(nlogn)
11	TNR	Rule Based and Tree Respresentation	Use Search Procedure to Find Top-K Non-Redundant Rules	O(nlogn)
12	CD(Count Distribution	Data Parallelism Algorithms	Candidate Pruning Techniques	O(n+M)
13	PDM (Parallel Data Mining)	Parallel Data Mining	Direct Hashing Technique	O(n+M)
14	Dma (Distributed Mining Algorithm)	Data Parallelism Paradigm	Candidate Pruning Techniques and Communication Message Reduction Techniques	O(n+M)
15	Ccpd (Common Candidate Partitioned Database)	Common Candidate Partitioned Database	Short-Circuited Subset Checking Method	O(n+M)
16	IDD (Intelligent Data Distribution)	Task Parallelism	Bin-Packing Technique	O(n+M)
17	DD (Data Distribution)	Task Parallelism	Candidate Pruning Techniques	O(n+M)

**N - No. of transactions, M – Threshold no. of transactions, R - No. of Unique Items*
n= Number of iteration, n=No. of Items

C. SEQUENTIAL PATTERN MINING

1	Aprioriall	Apriori Based	Candidate Generation & Pruning Technique	O(n)
2	Apriorisome	Apriori Based	Candidate Generation	O(n)
3	Dynamicsome	Apriori Based	Candidate Generation	O(n+1)
4	Prefixspan (Prefix-Projected Sequentialpattern Mining)	Prefix MontonProperty&Pattern Growth Principle	Pseudo Projection Technique& Divide Conquer Strategy	O(nlogn)
5	I-Prefixspan Algorithm	Projection Based &Pattern Growth Principle	Spanning Algorithm	O(n+1)

6	P-Prefixspan Algorithm	Projection Based& Pattern Growth Principle	Spanning Algorithm	O(n+1)
7	C-fm-Prefixspan Algorithm	Pattern Growth Principle	Spanning Algorithm	O(n+1)
8	DRL-Prefixspan Algorithm	Projection Based	Spanning Algorithm	O(n+1)
9	C-Prefixspan Algorithm	Projection Based	Spanning Algorithm	O(n+1)
10	Sequential Pattern Mining with Length-Decreasing Support (Slpminer)	Projection Based	Support Vector Base Technique	O(n^2)
11	SPIRIT (Sequential Pattern Mining withregular Expression Constraints)	Apriori-Like Approach	Using Constraint-Based Pruning Followed by Support-Based Pruning	O(n+M)
12	GSP(Generalized Sequential Pattern)	Apriori-Like Approach	Candidate Generate& Test & Divide and Search Space Technique.	O(MN+ (1-R^M)/(1-R))
13	SPAM (Sequential Pattern Mining)	Apriori-Like Approach	Depth First Strategy &Use Bitwise Operation	O(n+M)
14	CM-SPAM	Apriori-Like Approach	Co-Occurrence Pruning	O(n^2)
15	Lapin (Last Position Induction)	Apriori-Like Approach	Candidate Sequence Pruning &Database Partitioning	O(n+1)
16	Last Position Induction Sequentialpattern Mining (Lapin-Spam)	Apriori-Like Approach	Same Principles As Spam With the Exception of The Methods for Candidate Verification And Counting	O(n+1)
17	CLASP	Mining Frequent Closed Sequential Patterns	Using Method Dfs-Pruning and Checkavoidable	O(n+1)
18	CM-CLASP	Mining Frequent Closed Sequential Patterns	Pruning Technique	O(n+1)
19	Clospan (Closed Sequential Pattern)	Apriori-Like Approach	Candidate Generate &Test Using, Backward Subpattern and Backward Superpattern Pruning to Prune Redundant Search Space	O(n+1)
20	CMDS (Closed Multidimensional Pattern Mining)	Based on Closed Multidimensional Pattern Mining	CMDS Technique	O(n+1)
21	PAR-CSP (Parallel Closed Sequential Pattern Mining) Algorithm	Apriori-Like Approach	Load-Balancing Scheme	O(n+1)
22	Bide+ (Bi-Directional Extension)	Apriori-Like Approach	Back Scan Pruning & Scan-Skip Optimization	O(n+1)

23	MAXSP	Pattern Growth Principle	Use Pseudo Projection and The Process of Searching for the MaxmalBackword Extension of a Prefix	O(n^2)
24	VMSP (Vertical Mining Of Maximal Sequential Patterns)	Based on Vertical Mining of Maximal Sequential Patterns	Used By Efficient Filtering of Non-Maximal Patterns (EFN),Forward-Maximal Extension Checking (FME) &Candidate Pruning With Co-Occurrence Map (CPC) Method	O(n^2)
25	FEAT	Frequent Sequence Generators Mining	Forward Prune & Backward Prune	O(n^2)
26	FSGP	Frequent Sequence Generators Mining	Pruning Technique	O(n^2+1)
27	GOKRIMP	Minimum Description Length Principle	Dependency Test which only Chooses Related Events for Extending a Given Pattern	O(n+1)
28	TKS	Apriori-Like Approach	CandidatePruning withPrecedenceMap	O(n+1)
29	TSP_NONCLOSED	Based On The PrefixspanAlgorithm.	Multi-PassSearch Space TraversalStrategy	O(n^2)
30	CMRULES	Apriori-Like Approach	It First FindsAssociations Rules Between Items To Prune The Search Space To Items That Occur Jointly In Many Sequences.	O(n^2)
32	RULEGROWTH	Mining Sequential Rules Common To Several Sequences	Pattern-Growth Approach	O(n+1)
33	RULEGEN	Rule Based	Rule Generation Tech.	O(n+1)
34	TRULE GROWTH	Based on The Rulegrowth Algorithm	Mining SequentialRules Commonto Several Sequences with a Sliding-Window	O(n+1)
35	TOPSEQRULES	Based on The RulegrowthAlgorithm	The Save Procedure	O(n+1)
36	SeqDim	Apriori-Like Approach	BUC Method	O(n+1)
37	Freespan (frequent Pattern-ProjectedsequentialPattern Mining)	Pattern Growth Principle	Divide And Search Space Technique	O(n+1)
38	SBLOCK	Mining Frequent ClosedSequences with the Help of Stack as Memory	Worked on Concept of Using Separate Memory Blocks	O(n^2)

		Block		
39	SLPMINER	Pattern Growth Principle	Projection-Based approach UsesLength-Decreasing Support	O(n+1)
40	WAP-MINE	Pattern Growth Principle	Tree-Structure Mining Technique	O(n+1)
41	MFS (Maximal Frequent Sequences)	Apriori-Like Approach	SuccessiveRefinement Approach	O(n+1)
42	MFS+	Incremental -Based Algorithms	Successive Refinement Approach & Compute The Updated Set of Frequent Sequences Given the Set of Frequent Sequences Obtained from Mining the Old Databse	O(n+1)
43	INCSPAN	Incremental -Based Algorithms	Worked on Concept of Using Separate Memory Blocks	O(nlogn)
44	MILE	Incremental -Based Algorithms	It Recursively Utilizes The Knowledge of Existing Patterns to Avoid Redundant Data Scanning	O(nlogn)
45	INDEXED BIT MAP (IBM)	Apriori-Like Approach	Consists of Two Phases; First, Data Encoding andcompression,Second ,Frequentsequence Generation	O(n+1)
46	PSP	Apriori-Like Approach	Perform Retrieval Optimizations	O(n+1)
47	CCSM(Cache-Based Constrained SequenceMiner)	Apriori-Like Approach	Similar to Gsp& Using K-Way Intersections of Id-Liststo Compute The Support Of Candidates	O(Mn+ (1-R^M)/(1-R))
48	Msps(Maximal Sequential Patterns Usingsampling)	Apriori-Like Approach	Pruning & Signature Technique	O(Mn+ (1-R^M)/(1-R))
49	RE-HACKLE(Regular Expression-Highly Adaptiveconstrained Local Extractor)	Apriori-Like Approach	Re-Hackle Approach	O(n+1)
50	WINEPI	Apriori-Like Approach	Uses WINEPI Approach IEA Sliding Window	O(n+1)
51	ISM	Incremental And Interactive Sequence Mining.	Spade Approach	O(n)
52	ISE	Incremental -Based Algorithms	Candidate Sequence Pruning &Bottom-Up Search	O(n)
53	FASTUP	Incremental -Based Algorithms	Uses the Generating -Pruning Method	O(n)
54	SUFFIXTREE	Incremental	Suffixtree Technique	O(n)

		-Based Algorithms		
55	INCSP	Incremental -Based Algorithms	Candidate Sequence Pruning	O(n)
**N - No. of transactions, M – Threshold no. of transactions, R - No. of Unique Items n= Number of iteration, n=No. of Items*				
D. CLASSIFICATION				
1	CHID (Chi-Squared Automatic Interaction Detector)	Adjusted Significance Testing Principle	Decision Tree Learning Algo.	O(nlogn)
2	Quest(Quick Unbiased, Efficient, Statistical Tree)	Univariate And Linear Combination Splits	Ten-Fold Cross-Validation	O(n)
3	Naive Bayes	Probabilistic Classifierbased On Applying Bayes' Theorem	Instance-Based Learning	O(nlogn)
4	CART (Classification And Regression Tree)	Use Gini Diversity Index	Post Pruning	O(nlogn)
5	ID3(Iterative Dichotomiser 3)	Used Gini Index	If Then Rules	O(n)
6	C4.5	Greedy Algorithm	Divide & Conquer Manner &Uses A Single-Pass Algorithm Derived From Binomial Confidence Limits	O(n+1)
7	C5.0/SEC 5	Uses Information-Based Criteria	Depth First Construction,Pre-Pruning	O(n)
8	HUNT'S	Greedy Algorithm	Divides And Conquers Approach	O(n+1)
9	K-NN	Instance-Based Learning	Use Efficient Indexing Techniques	O(Dn)
10	LMNN	Machine Learning Algorithm	Optimizes The Matrix {M} With The Help Of Semidefinite Programming	O(n+1)
11	Neighbourhood Components Analysis	Maximizes A Stochastic Variant Of The Leave-One-Out Knn Score On The Training Set.	Non-Parametric Learning Method	O(n+1)
12	Backpropagation Algorithm	Based On Neural Networks	Feed-Forward Computation	O(n+1)
13	SVM (Support Vector Machine)	Binary Classifier	Pair Wise Classification Used	O(L(M^2))
14	SSVM (Smooth Support Vector Machine)	Smooth Support Vector Machine	Uses A Smooth Unconstrained Optimization	O(L(M^2))
15	RSVM (Reduced Support Vector Machine)	RSVM (Reduced Support Vector Machine)	Uses A Smooth Unconstrained Optimization	O(L(M^2))
16	GSVM (Granular Support Vector Machine)	Based On Statistical Learning & Granular Computing	GSVM Modeling Approach	O(L(M^2))

17	LSVM (Lagrangian Support Vector Machine)	Based On An Implicit Lagrangian Of The Dual Of A Simple Reformulation Of TheStandard Quadratic Program Of A Linear Support Vector Machine	Simple Iterative Approach	O(Lm)+O(L(M^2))
D-Dimensional, N=Data Points M=Number of Attributes, L= Data, O(Lm)-Matrix Vector multiplication				
E. CLUSTERING ALGORITHMS				
1	SLINK	Single Link Clustoring	Linkage Metric	O(n^2)
2	CLINK	Complete-Linkage Clustering	Hierarchical Cluster Analysis Based Upon A Distance Matrix	O(n^2)
3	COBWEB	Hierarchical Clustering	Model-Based Learning	O(n+1)
4	CURE (Clustering Using Representatives)	Hierarchical With Distance-Based Clustering	Sampling	O(n+1)
5	CHAMELEON	Hierarchicalclusters Of Arbitraryshapes	Linkage Metric	O(n^2)
6	EM(Expectation–Maximization)	Based On Approximating Maximum Likelihood	Iterative Optimization	O(n^2)
7	SNOB	Mixture Model In Conjunction With The Mml Principle	Probabilistic Approach	O(n^2)
8	AUTOCLASS	Probabilistic Model-Based Clustering	Used Expectation–Maximization	O(n+1)
9	MCLUST	Mixture Model Clustering	Probabilistic Approach	O(n+1)
10	PAM (Partition Around Medoids)	Partition Around Medoids	Centroid Based Technique (Partition Method)	$O(k(n-k)^2)$
11	CLARA(Clustering Large Applications)	Clustering Large Applications	Representative Object Based Technique (Partition Method)	O(K(40+K)+K(N-K))+(Time)
12	CLARANS (Clustering Large Applications Based Upon Randomized Search)	K-Medoids Methods, Spatial Database Clustering Algorithms.	Based OnRandomized Search + Focusingtechniques With Spatial Access Methods	O(n^2)
13	DBSCAN (Density Based Spatial Clustering Of Applications With Noise)	Density-Based Clustering	Density-Based Connectivity	O(nlogn)
14	OPTICS (Ordering Points To Identify The Clustering Structure)	Density-Based Clustering	Density-Based Connectivity	O(nlogn)
15	DBCLASD (Distribution-Based Clustering Of Large Spatial Databases)	Density-Based Connectivity	Density-Based Connectivity	O(3(n^2))

16	DENCLUE (Density-Based Clustering)	Density-Based Clustering	Density Functions	O(Log D)
17	GDBSCAN	Density-Based Connectivity	Density Based Clustering In Spatial Databases	O(nlogn)
18	DESCRY	Density-Based Clustering With Large Data Set	Agglomerative Method Used In Thepreclustering Step & Similarity Metrics Of Interest.	O(Nmd) Time,High -Dimensional Data Sets, O(N Logm)Low Imensional Data Sets
19	BANG	Grid-Clustering Algorithm	The Patterns Are Grouped Into Blocks And Clustered With Respect To The Blocks By A Topological Neighbor Search Algorithm.	O(n)
20	ROCK (Robust Clustering Algorithm For Categorical Data)	Hierarchical Clustering	Sampling	O(n+1)
21	SNN (Shared Nearest Neighbors)	Graph-Based Algorithm	Shared Nearest Neighbor (Snn) Clustering Approach	O(n+1)
22	CACTUS (Clustering Categorical Data Using Summaries)	Co-Occurrence For Attribute-Value Pairs	Summarization, Clustering and Validation	O(Cn)
23	STIRR (Sieving Through Iterated Reinfircement)	Based On Non-Linear Dynamic System	Co-Occurrence Phenomenon	O(n+1)
24	AMOEBA	Hierarchical Clustering Based On Spatial Proximity Using Delaunaty Diagram	Amoeba Strategy	O(nlogn)
25	DIGNET	Incremental Unsupervised Learning.	K-Means Cluster Representation Without Iterative Optimization.	O(n+1)
26	BIRCH (Balanced Iterative Reducing And Clustering Using Hierarchies)	Hierarchical With Distance-Based Clustering	Multiphase Clustering Tech.	O(n)
27	CLASSIT	Hierarchical Clustering	Model Based Clustoring Method	O(n+1)
28	K-MEANS	Vector Quantization	Partition Algo.	O(Nkt)
29	K-MEDOIDS	Based On K-Means Algorithm And The Medoidshift Algorithm	Centroid Basedtechnique (Partition Method)	O(n+1)
30	K-MODELS	Unsupervised Learn ing Algorithms	Representative Object Based Technique (Partition Method	O(Pk)
31	WAVECLUSTER	Signal Processing	Grid-Base Methods	O(n)

32	PDD(Principal Direction Divisive Partitioning)	Singular Value Decomposition	Binary Divisive Partitioning	O(n+1)
33	CLIQUE (Clustering In Quest)	Density Based And Grid-Based	Partitions	O(n)
34	STING(A Statistical Information Grid Approach)	Grid-Based Algorithm	Top-Down Approach	O(n)
35	PIC(Power Iteration Clustering)	Graph-Based Algorithm	Iterative Matrix –Vector Multiplication.	O(n+1)
36	SM(Shi And Malik)	Spectral Clustering	Class Of Techniques Which Relies On The Eigen Structure Of A Similarity Matrix	O(n^3)
37	KVV(Kannan,Vempalaandvetta)	Spectral Clustering	A Heuristic Method	O(n^3)
38	JNW(Jordan–Ng–Weiss)	Spectral Clustering	Partitioning Algorithm	O(n^3)
39	FUZZY C MEANS	Soft Clustering	Fcm Method	O(n^3)
K=No. of Clustors, n=Data Points, n=Number of Populated Grid Cell, N = Objects, P = The Number of Points, T = Number of Iterations, C=Centre				

IV. CONCLUSION

In this project, we have proposed framework for assessing data mining techniques. This framework will help users to make better decision for selecting suitable data mining technique. Use of this framework will help the user to analyse the performance of data mining techniques based on execution time. This framework will give an insight to the data miners. After using this framework to assess various categories of data mining algorithms, we have found out following points -

1. Within the category of Frequent itemset mining.—
- SaM is the fastest algorithm.
- Apriori, AprioriTID, MSApriori, HMine are the slowest algorithms.

2. Within the category of Association rule mining.—
- Sporadic association rules, Apriori Inverse, MMR, Indirect association rules, TopKRules, TNR, FHARM are the fastest algorithms.
- Genetic Algorithm, SETM, AIS , MCISI, IGB are the slowest algorithms.

3. Within the category of Sequential pattern mining.—
- IncSpan, MILE and PrefixSpan are the fastest algorithms.
- GSP(Generalized Sequential Pattern) is the slowest algorithm.

4. Within the category of classification.—
- CHID (CHI squared Automatic Interaction Detector), Naive Bayes and CART (Classification and regression tree) are the fastest algorithms.
- QUEST (Quick Unbiased, Efficient, Statistical Tree), ID3(Iterative Dichotomiser 3), C5.0/Sec 5 are the slowest algorithms.

5. Within the category of clustering –
- DENCLUE is the fastest algorithm.
- fuzzy C Means, SM , KVV & JNW are the slowest algorithms.
  DENCLUE algorithm is the fastest due to the use of the principle of density based clustering. Also, fuzzy C Means, SM & KVV are the slowest due to the use of the principle of Spectral based clustering.

6. After examining all these techniques individually, it has been seen that fp growth based algorithms are faster than Apriori based algorithms.

7. Overall, DENCLUE, MILE, SaM are the fastest algorithms. Apriori, GSP (Generalized Sequential Pattern) are the slowest algorithms.

V. FUTURE SCOPE

Our proposed framework involves various criteria like principles, techniques, time complexity of algorithms. This framework is very much useful to the users. Whereas in future, we can add more criteria to this framework. These criteria can be memory space, effectiveness etc.

More algorithms can be added to this framework which will make it richer.

REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

[2] Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min. Knowl. Discov. 8(1): 53-87 (2004)

[3] Finding Frequent Item Sets by Recursive Elimination Christian Borgelt. Workshop Open Source Data Mining Software (OSDM'05, Chicago, IL), 66-70. ACM Press, New York, NY, USA 2005

[4] Mohammed Javeed Zaki: Scalable Algorithms for Association Mining. IEEE Trans. Knowl. Data Eng. 12(3): 372-390 (2000) eclat

[5] Zaki, M.J., Gouda, K.: Fast vertical mining using diffsets. Technical Report 01-1, Computer Science Dept., Rensselaer Polytechnic Institute (March 2001) 10 declat

[6] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang. "H-Mine: Fast and space-preserving frequent pattern mining in large databases". IIE Transactions, Volume 39, Issue 6, pages 593-605, June 2007, Taylor & Francis.

[7] Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal: Discovering Frequent Closed Itemsets for Association Rules. ICDT 1999: 398-416 apriori close

[8] Claudio Lucchese, Salvatore Orlando, Raffaele Perego: DCI Closed: A Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets. FIMI 2004

[9] Mohammed Javeed Zaki, Ching-Jiu Hsiao: CHARM: An Efficient Algorithm for Closed Itemset Mining. SDM 2002.

[10] Zaki, M.J., Gouda, K.: Fast vertical mining using diffsets. Technical Report 01-1, Computer Science Dept., Rensselaer Polytechnic Institute (March 2001) 10 DECHARM

[11] L. Szathmary (2006). Symbolic Data Mining Methods with the Coron Platform. CHARM MFI

[12] Arnaud Soulet, Françoi s Rioult (2014). Efficiently Depth-First Minimal Pattern Mining. PAKDD (1) 2014: 28-39 DefMe

[13] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. ACM SIGKDD Explorations Newsletter, 2(2), 66-75. PASCAL

[14] L. Szathmary, A. Napoli, S. O. Kuznetsov. ZART: A Multifunctional Itemset Mining Algorithm. Laszlo Szathmary, Amedeo Napoli, Sergei O. Kuznetsov In: CLA, 2007. ZART

[15] Laszlo Szathmary, Amedeo Napoli, Petko Valtchev: Towards Rare Itemset Mining. ICTAI (1) 2007: 305-312 APRI RARE

[16] Yun Sing Koh, Nathan Rountree: Finding Sporadic Rules Using Apriori-Inverse. PAKDD 2005: 97-106 APR INVERSE

[17] Yun Sing Koh, Nathan Rountree: Finding Sporadic Rules Using Apriori-Inverse. PAKDD 2005: 97-106 CloStream

[18] Joong Hyuk Chang, Won Suk Lee: Finding recent frequent itemsets adaptively over online data streams. KDD 2003: 487-492 estDec

[19] C. Kit Chui, B. Kao, E. Hung: Mining Frequent Itemsets from Uncertain Data. PAKDD 2007: 47-58 UApriori

[20] Y. Liu, W.-K. Liao, A. N. Choudhary: A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets. PAKDD 2005: 689-695

[21] M. Liu, J.-F. Qu: Mining high utility itemsets without candidate generation. CIKM 2012, 55-64 HUI MINER

