Sudan University of Science &Technology

College of Graduated Studies

Distributed Frequent Itemset Mining

تنقيب العناصر الأكثر تكراراً الموزعة

A dissertation Submitted in partial Fulfilment of the requirements for MSc degree in computer science

By

Huda Jamal Abdel-hammed Musa

Supervisor

Dr. Mohamed Elhafiz Mustafa Musa

May 2014

DEDICATION

To my family

To my teachers

To my husband

To my best friends

ACKNOLEDGMENT

First and foremost, I would like to express my sincere thanks to my thesis supervisor**Dr. Mohammed Elhafiz Mustafa**for providing me their precious advices and suggestions. This Thesis wouldn't have been a success for me without their comments and suggestions.

Next, I would like to express my family: my father **Jamal Abdelhammed Musa,** my mother **Elham Abdulla Abbas,**my brothers and my sisterswithout their support I would never had dreamt of pursuing higher studies.

AlsoI would like to expressmy husband **Mustafa Mohammed Ahmed** for their unconditional love and support in every part of my life.

Special respect and thanks to my teacher **Wafa Fisal** for support and help.

Also I would like to thank **Sudan University of Science and Technology**for providing me such a graceful opportunity to become a part of its family.

Lastly I would like to thanks all the persons those are related to the thesis directly or indirectly.

ABSTRACT

Association rule mining is an important technique to discover hidden relationships among items in the transaction. The problem is that association rules are generated by first mining of frequent itemsets in distributed datasets does not gain the best and most accuracy rules.The goal of the thesis is to experimentally finding the most frequent itemsets from distributed data sources which is first phase of association rules generation. Firstly, the global frequent itemsetare generated from global dataset. Secondly, the global datasetare divided into three sites, and then generating the local frequent itemsets from each site. A comprehensive search for the best way to combine the local itemset has been conducted. In this search we find that the union of smallest and biggest of itemsets intersected with the middle always gives result which is equivalent to global itemsets.

المستخلص

إستخلاص قواعد الربط (Association Rules)تعتبر أسلوباً مهماً لإكتشاف العلاقات الخفية بين العناصر. تكمن المشكلة في أن قواعد الربط المستخرجة من أول عملية تعدين Mining لمجموعة العناصر المتكررةFrequent itemset في البيانات الموزعة لا تعطي القواعد المراد تكوينها بصورة دقيقة.الهدف من هذه الدراسة هو إيجاد مجموعة العناصر الأكثر تكرارا من مصادر البيانات الموزعة والتي تعتبر المرحلة الأولى من مراحل إستخراج قواعد الربط. أولا يتم توليد مجموعة العناصر الأكثر تكراراً لمجموعة البيانات الشاملة Global dataset.ثانياً يتم تقسيم مجموعة البيانات الشاملة إلى ثلاثة مجموعات جزئية،ومن ثم يتم توليد مجموعة العناصر المكررة في كل مجموعة. فى هذه الدراسة تم إجراء بحث شامل عن أفضل طريقة لدمج مجموعات العناصر المتكررة من كل المجموعات، فوجدنا أن إتحاد أصغر مجموعة عناصر Smallest local frequent itemsets مع أكبر مجموعة عناصرBiggest local frequent itemsets تقاطعمجموعة العناصر الوسطى (Middle local frequent itemsets) دائماً تعطي عناصر متكررة مكافئة لمجموعة العناصر المتكررة في المجموعة الشاملة.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRDUCTON

Chapter1

1.1 Introduction

Association rule mining (ARM) is an active data mining research area. Most ARM algorithms focus on sequential or centralized environments where no external communication is required. Although nowadays there is huge data in distributed database and no standard approach to build efficient association rule mining in these data.

1.2 Problem

Modern organizations are geographically distributed. Typically, each site locally stores its ever-increasing amount of day-to-day data. Using centralized data mining to discover useful patterns in such organizations' data isn't always feasible because merging datasets from different sites into a centralized site incurs huge network communication costs. Data from these organizations are not only distributed over various locations but also vertically fragmented, making it difficult if not impossible to combine them in a central location. Most Distributed Association rule mining (DARM) algorithms don't have an efficient message optimization technique, so they exchange numerous messages during the mining process. Distributed data mining has thus emerged as an active sub-area of data mining research.

1.3 Objectives

Distributed ARM system aims to generate rules from different database spread over various geographical sites. Hence, they require external communications throughout the entire process. DARM algorithms must reduce communication costs so that generating global association rules costs less than combining the participating sites' datasets into a centralized site.

1.4Methodologies

We have two main steps:

1.4.1.1 Local Rules Generating

Each site generates the frequent itemsets. Then it will be used to generate association rules that satisfy minimum confidence.

1.4.2 Global Rules Refining

After generating the local frequent itemsets and the rules at each site, generates the globally frequent itemsets.

1.5 Scope

We can generate Association Rule from any datasets that distributed among various sites to discover the most frequent itemsets.

1.6 Thesis Structure

This thesis contains four chapters as follows:

Chapter 2 presents the background of the data mining. It covers in detail about the data mining, association rule mining and distributed association rule mining. In addition discuses some related works on distributed association rule mining. Chapter 3 contains proposed system, the experiments and the results. Chapter 4 discusses the conclusion and Future work.

CHAPTER 2

ASSOCIATION RULE MINING

Chapter 2
AssociationRuleMining

2.1 Introduction

In this chapter a background of data mining and association rule mining is discussed. This chapter also covers in detail about the distributed association rule mining algorithms. Also it provides of some related works.

2.2 Data Mining

2.2.1 Backgrounds

There are basically two most important reasons that data mining (DM) has attracted a great deal of attention in the recent years. First, our capability to collect and store the hugeamount of data is rapidly increasing day by day. The second reason is the need to turn such data into useful information and knowledge. The knowledge that is acquired through the help of data mining can be applied into various applications like business management, retail and market analysis, engineering design and scientific exploration.[1]

There are many definitions for data mining:

- Data mining (sometimes called data or Knowledge Discovery in Database KDD) is the process of analyzing data from different perspective and summarizing it into useful information. [3]
- Data mining or Knowledge Discovery in Database (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handling large amount of information. [5]

  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. Technically, data mining is the process of finding correlations or patterns among many fields in large databases. Data mining tools and techniques are used to generate information from the data that we have stored in our repositories over the years.

2.2.2 Data Mining Tasks

The process of mining is often controlled by the requirements of the users. The user may be a business analyst or may be a marketing manager. Different users have different need of information. Depending on the requirements we can use different data mining tasks.[2]

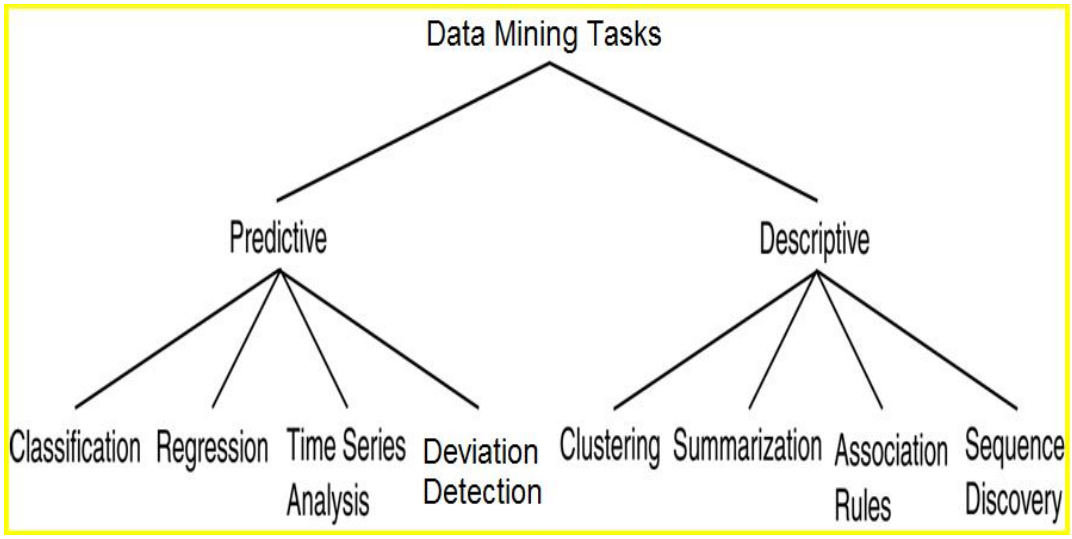


Figure 2.1: Data Mining Tasks

2.3 Association Rules Mining

Association rule mining is an interesting data mining technique. That is used to find out interesting patterns or associations among the data items stored in the database. Support and confidence are two measures of the interestingness for the mined patterns.

Databases or data warehouses may store a huge amount of data to be mined. Mining association rules in such databases may require substantial processing power. A possible solution to this problem can be a distributed system. Moreover, many large databases are distributed in nature which may make it more feasible to use distributed algorithms. Major cost of mining association rules is the computation of the set of large itemsets in the database. Distributed computing of large itemsets encounters some new problems. One may compute locally large itemsets easily, but a locally large itemsets may not be globally large. [2]

Many parallel or distributed ARM algorithms were designed for shared memory parallel environments. Based on the nature and implementation of each algorithm, we can divide the existing algorithms into two groups: parallel ARM and DARM.

2.3.1 Parallel ARM

We can categorize parallel ARM algorithms as data-parallelism or task-parallelism algorithms. In the former, the algorithms partition the datasets among different nodes; each site performs the task independently but must access the entire dataset. [5]

The main challenges associated with parallel data mining include minimizing I/O, minimizing synchronization and communication, effective load balancing, effective data layout, deciding on the best search procedure to use. The parallel algorithms are Count Distribution, Candidate Distribution and Hybrid Count and Candidate Distribution. [6]

2.3.2 Distributed ARM

DARM discovers rules from various geographically distributed datasets. However, the network connection between those datasets isn't as fast as in a parallel environment, sodistributed mining usually aims to minimize communication costs.

Distributed ARM algorithms involve distributed association rule learning, collective decision tree learning, distributed hierarchical clustering, other distributed clustering algorithms, collective Bayesian network learning, collective multi-variate regression. [7]

2.3.2.1 FDM (Fast Distributed Mining of association rules):

FDM mine rules from distributed datasets partitioned among different sites. In each site, FDM finds the local support counts and prunes all infrequent local support counts. After completing local pruning, each site broadcasts messages containing all the remaining candidate sets to all other sites to request their support counts. It then decides whether large itemsets are globally frequent or not. Then generates the candidate itemsets from those globally frequent itemsets. [5]

Generally FDM has the following distinct features:

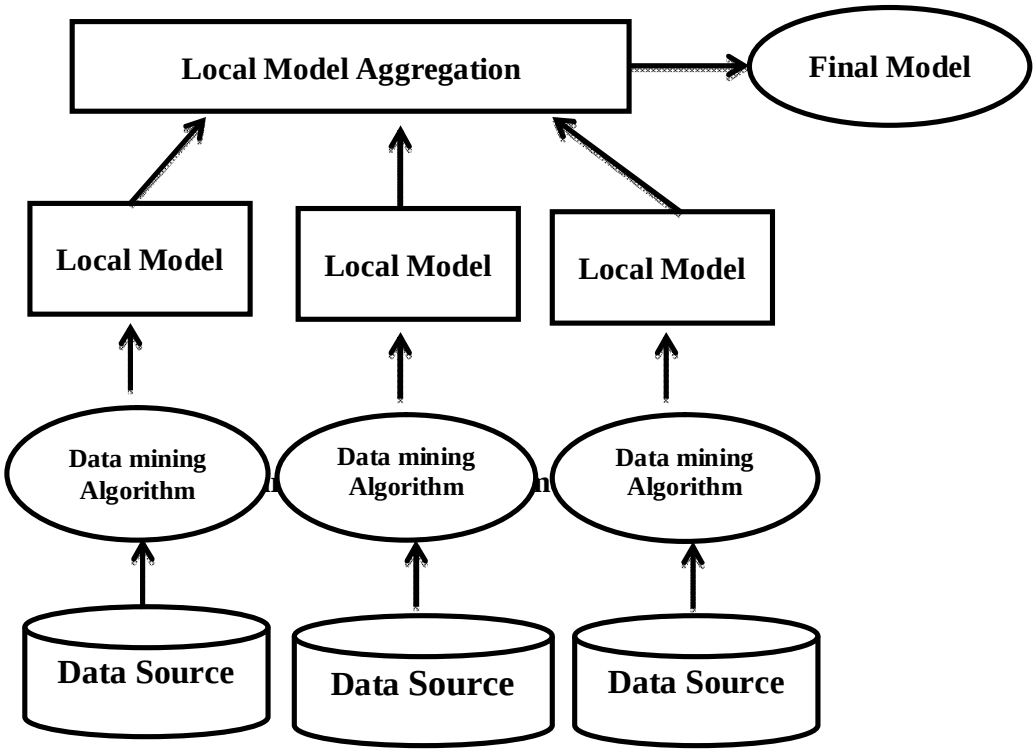
1. Some relationships between locally large sets and globally large ones are explored to generate a smaller set of candidate sets at each iteration and thus reduce the number of messages to be passed.

2. After the candidate sets have been generated, two pruning techniques, local pruning and global pruning, are developed to prune away some candidate sets at each individual site.

3. In order to determine whether a candidate set is large, this algorithm requires O (n) messages for support count exchange, where n is the number of sites in the network.

2.4 Distributed Data Mining (DDM)

When data mining is undertaken in an environment where users, data, hardware and the mining software are geographically dispersed, it is called distributed data mining. Thus distributed data mining refers to the mining of distributed datasets. The datasets are stored in local databases hosted by local computers which are connected through a computer network. Data mining takes place at a local level and at a global level where local data mining results are combined to gain global findings. Distributed data mining is often mentioned with parallel data mining in literature. While both attempt to improve the performance of traditional data mining systems they assume different system architectures and take different approaches. In distributed data mining computers are distributed and communicate through message passing. In parallel data mining a parallel computer is assumed with processors sharing memory and or disk. Computers in a distributed data mining system may be viewed as processors sharing nothing. This difference in architecture affected in algorithm design, cost model, and performance measure in distributed and parallel data mining. Typically, such environments are also characterized by heterogeneity of data and multiple users. DDM offers techniques to discover knowledge in distributed data. [3] A typical DDM framework is shown in figure 2.2.



2.5 Literature Review

In Market Basket Analysis If we think of the universe as the set of items available at the store, and then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associatedor purchased together. These patterns can be represented in the form of association rules.

For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule 2.1 below:

Computer =>antivirus software [support = 2%; confidence = 60%](2.1)

Rule **support** and **confidence** are two measures of rule interestingness. They respectively reflect the **usefulnes**s and **certainty** of discovered rules.

A support of 2% for Association Rule (2.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.

A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. [7]

2.5.1 Frequent Itemsetsand Association Rules

Let I= {*I1*, *I2*…*Im*} be a set of items. Let *D*, the task-relevant data, is a set of database transactions where each transaction *T* is a set of items such that *T is in* I. Each transaction is associated with an identifier, called TID.

An association rule is an implication of the form *A*=>*B*, where *A* is in I , *B* is in I , and *A* and *B* are disjoint. The rule *A*=>*B* holds in the transaction set *D* with support *s*, where *s* is the percentage of transactions in *D* that contain *A* union *B*.This is taken to be the probability, *P* (*A* union *B*).

The rule *A*=>*B* has **confidence***c* in the transaction set *D*, where *c* is the percentage of transactions in *D* containing *A* that also contain *B*.This is taken to be the conditional probability, *P* (*B*|*A*). That is,

Support(*A*=>*B*) = *P* (*A* union *B*) (2.2)

Confidence(*A*=>*B*) = *P* (*B*|*A*): (2.3)

Rules that satisfy both a minimum support threshold (*min sup*) and a minimum confidence threshold (*min conf*) are called strong.

From Equation (2.3) we have:

Confidence(*A*=>*B*) = *P*(*B*|*A*) = *support* (*A* union *B*) / *support* (*A*). (2.4)

2.5.2 Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Elk requires one full scan of the database.[1]

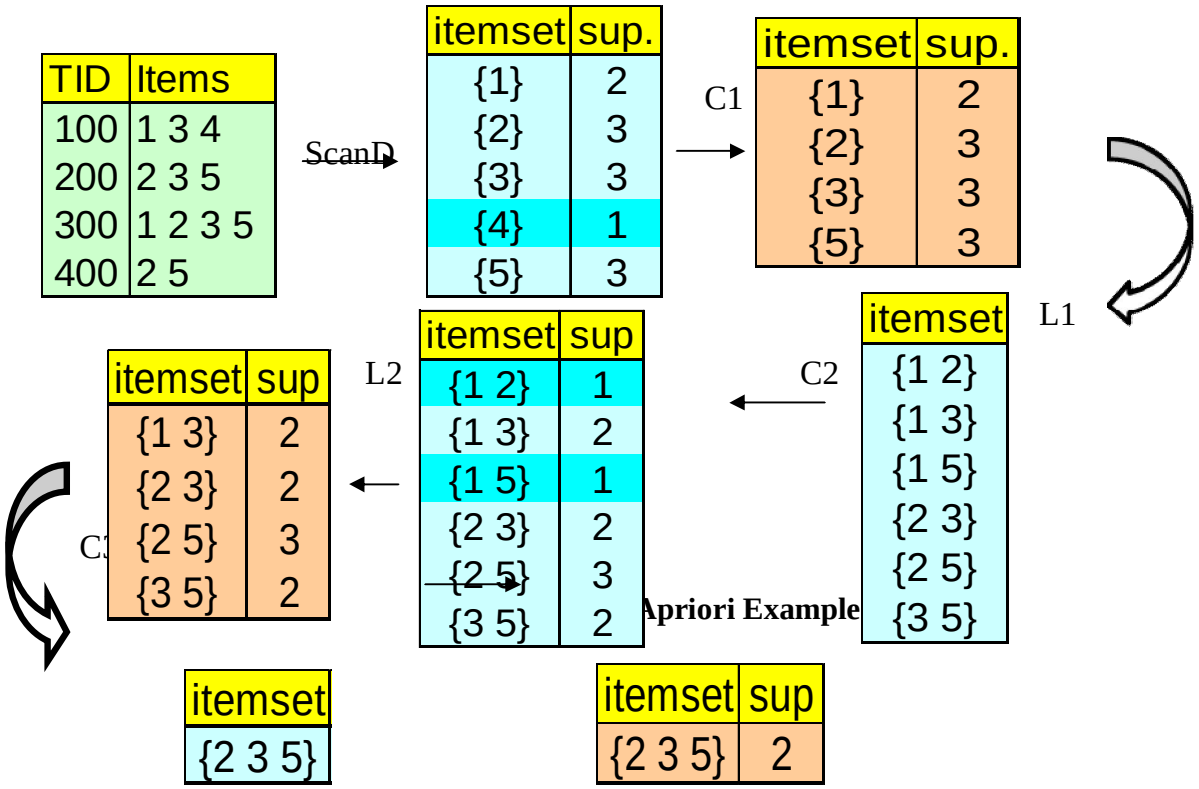
2.5.2.1 Example

Let Set of items: I= {1, 2, 3, 4, 5}.

Transactions: D = {t100, t200, t300, t400}.

Support of an itemset: Percentage of transactions which contain that itemset.

Large (Frequent) itemset: Itemset whose number of occurrences is above a threshold.



2.6 Related Works

Many algorithms have been proposed to find frequentitemsets from a very large datasets. The number of datasets scans required for the task has been reduced from a numberequal to the size of the largest itemsets in Apriori, totypically just a single scan in modern ARM algorithms such as Sampling. When data is saved in a distributed datasets, a distributed data mining algorithm is needed to mine association rules. It has been addressed by some researches and number of distributed algorithms has been proposed. [4]

The partition algorithm is based on apriori algorithm. It consists of two phases. Firstly partitions the data into a number of non-overlapping partitions. For each partition, all frequentitemsets are found. These are referred as local frequent itemsets.A local frequent itemset may or may not be frequent with respect to the entire dataset D. Any itemset that is potentially frequent with respect to D must occur as afrequent itemset in at least one of the partitions. Therefore all local frequent itemsetsare candidate itemsets with respect to D. The collection of frequent itemsets from

allpartitions forms the global candidate itemsets with respect to D.Finally the algorithmunion all the local frequent itemsets to generate global frequent itemsets. It reduces the number of complete database scans up to two and hence improves the performance of mining algorithm. [10].

Sampling algorithm (mining on a subset of a given data) is also based on apriori algorithm. The basic idea of the samplingapproach is to pick a random sample S of the given data D, and then search forfrequent itemsets in S instead of D. In this way, we trade off some degree of accuracyagainst efficiency. The sample size of S is such that the search for frequent itemsets inS can be done in main memory, and so only one scan of the transactions in S inrequired overall [10]. Sampling can reduce I/O costs by drastically shrinkingthe number of transaction to be considered. It can speed up the mining processby more than an order of magnitude.in another hand, because we are searching for frequent itemsets in S rather than in D,it is possible that some of the global frequent itemsets was missed. [15]

E. Ansari, G.H. Dastghaibifard, M. Keshtkaran, H.Kaabi presented a new distributed Trie-based algorithm (DTFIM) to find frequent itemsets. This algorithm isproposed for a multi-computer environment. They added an idea from FDM algorithm for candidategeneration step.The point of this algorithm is that every site keeps a copyof Trie locally, and they synchronize their data so that all localTrie copies are the same at the end of each stage. After localsupport is counted, all sites share their support counts and determine the global support counts, in order to removeinfrequent itemsets from their local Trie. These resultsshow Trie data structure can be used for distributedassociation rule mining not just for sequential algorithms. [12]

CHAPTER 3

PROPOSED SYSTEM

Chapter 3

Proposed System

3.1 Introduction

This chapter describes the proposed system, the dataset, the experiments and the results.Figure 3.1 is an overview of the proposed system for distributed association rules mining. The chapter also reports and discusses the experiments' results.

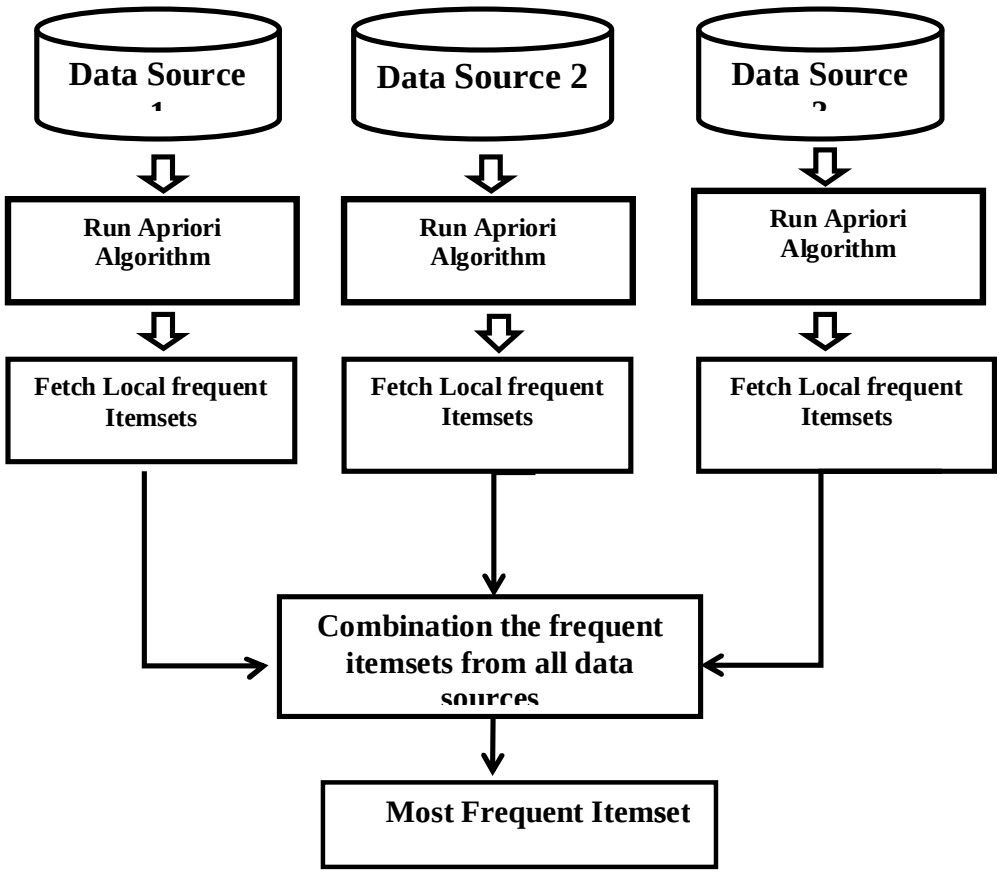


Figure 3.1: The Proposed System Structure

3.2Experiments Dataset

The dataset have been downloaded from the university of California site (UCI);this data was extracted from the census income database. It goal is to predict whether income exceeds 50.000$/year. Table 3.1 summarizes details of the dataset and Table 3.2 describes dataset attributes.

Table 3.1: Dataset Statistics

Dataset Characteristics:	Multivariate
Attribute Characteristics:	Categorical, Integer
Number of Instances:	48843
Number of Attributes:	13
Area:	Social
Date Donated	1996-05-01

Table 3.2: Dataset Attributes Description

Attribute	Description
Age	Continuous
Work Class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Martial-Status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Female, Male
Gain	Continuous
Loss	Continuous
Hours-per-week	Continuous
Country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
Salary	>50K, <=50K

3.3 Preprocessing Stage

We use WEKA tool to generate the association rules from sites.

The proposed system is divided into two phases. First generate local frequent itemsets for each site. Second Local frequent itemsets from each site are combined to generate global frequent itemsets.

3.3.1Generate local frequent itemsets

At each site we apply apriori algorithm. Figure 3.2 shows the datasets uploaded in WEKA and figure 3.3 shows the association rule that collected from the dataset.

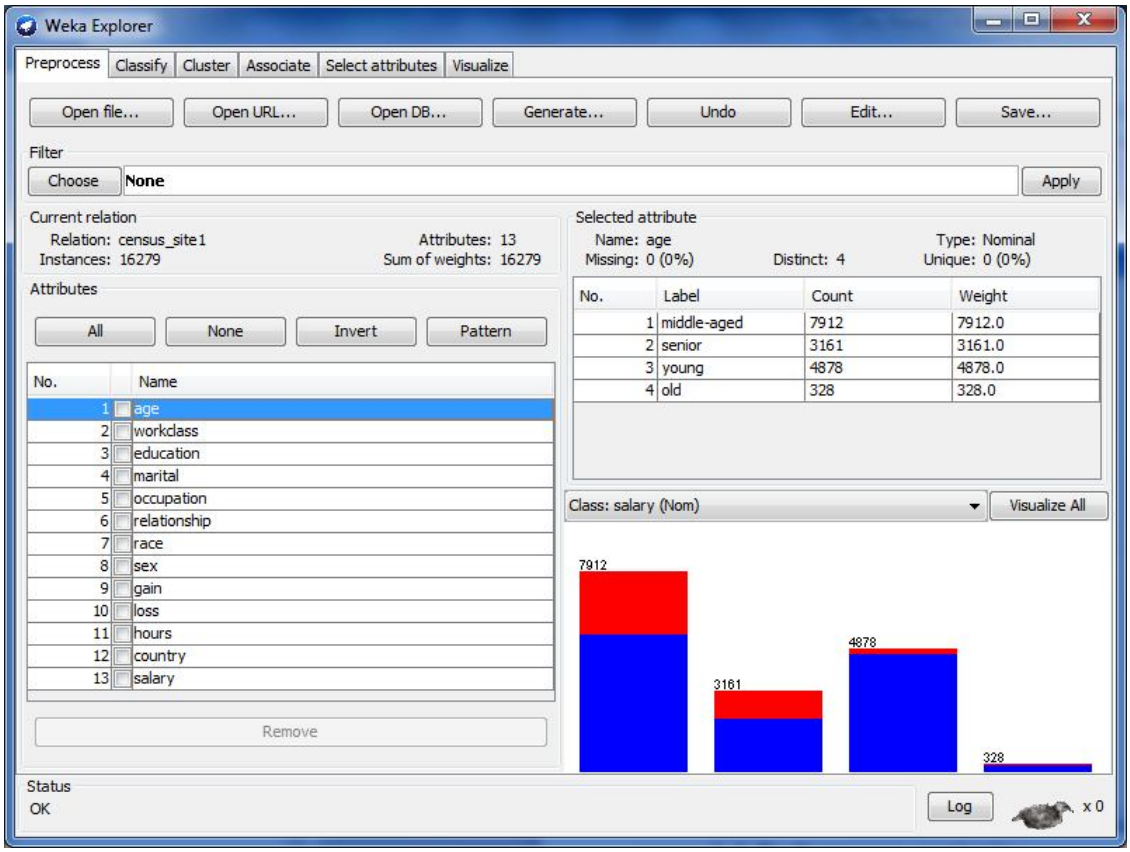


Figure 3.2: Dataset at Site1

Attributes at Site1are visualized in Figure 3.3.

Figure 3.3: Visualize Attributes at Site1

3.3.2 Generate Global frequent itemsets

After generates the local frequent itemsets from each site, we combined them to generates the most frequent itemsets. Table 3.3shows the total number of records, minimum support and frequent itemsets for global CENSUS dataset.

Table 3.3: Global CENSUS Dataset

	Total Rows	MinSup	Frequent Itemsets
CENSUS	48,843	0.2	3

Table 3.4 shows the results after divided the CENSUS dataset into 3 sites.

Table 3.4: CENSUS Dataset (3 Sites)

	Total Rows	MinSup	Frequent Itemsets
Site1	16,280	0.2	3
Site2	16,280	0.2	4
Site3	16,280	0.2	3

Figure 3.4 and Figure 3.5 shows the result in details.

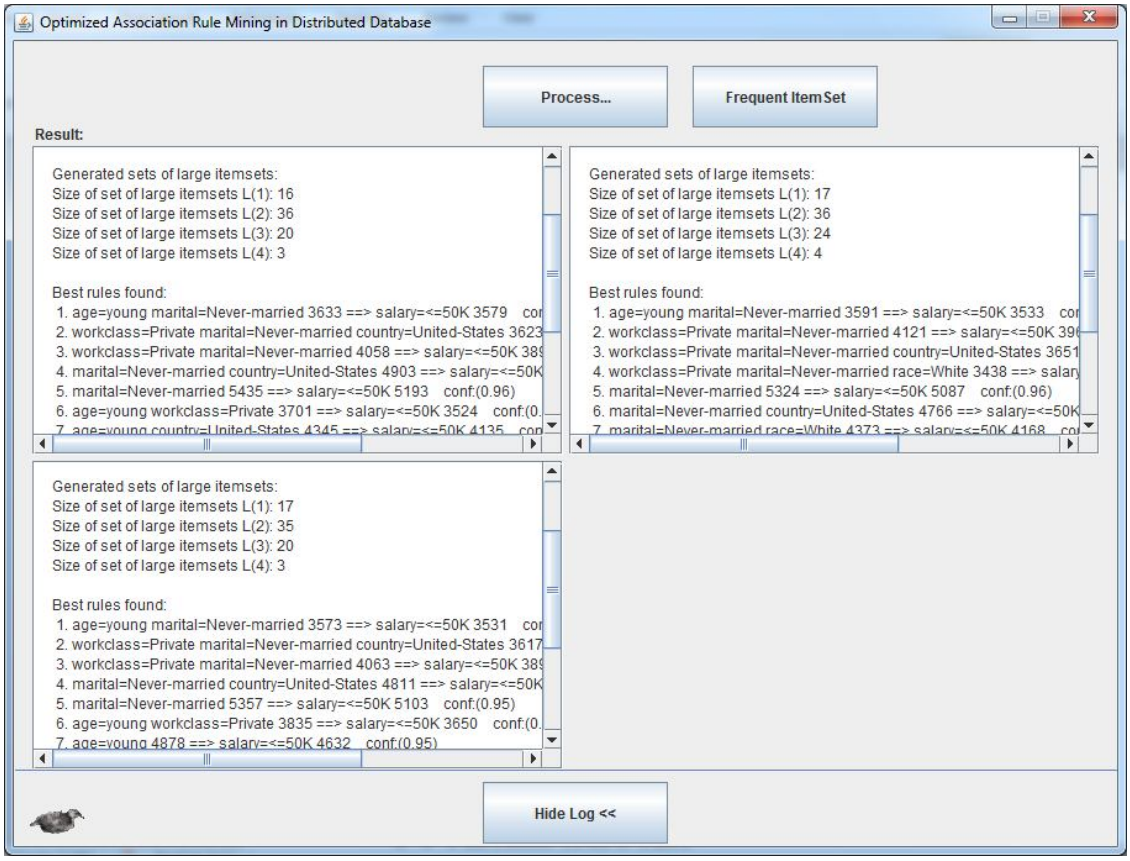


Figure 3.4: The Local Frequent Itemsets and Association Rules in Site1, Site2, Site3

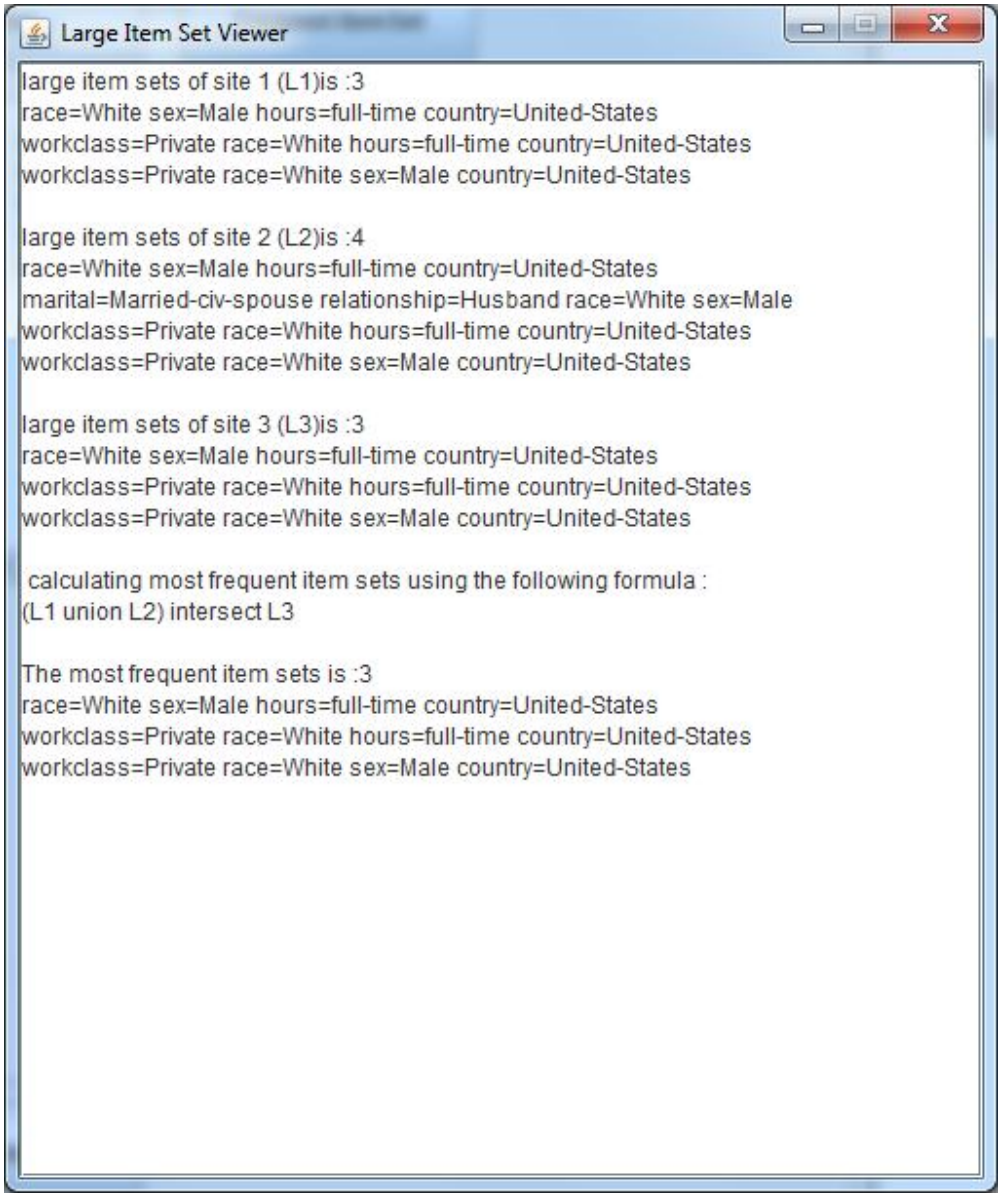


Figure 3.5: The Details of Frequent Itemsets

3.4 The Results

To generate the most frequent itemset firstly we divide the datasets into 3 sites S1, S2, S3. Then we generate the large itemsets from each site L1, L2, L3. Lastly we combine the large itemsets by using the proposed rule. Table 3.5 shows the total number of records and frequent itemsets at each site in CENSUS, CAR, NURSERY, SAMPLE_MODELING datasets.

Table 3.5: The Local Frequent Itemsets in Site1, Site2, Site3

Datasets	No of Records	Global Frequent Itemset	Local Frequent Itemsets in Site1	Local Frequent Itemsets in Site2	Local Frequent Itemsets in Site3
CENSUS	**48,000**	**3**	**3**	**4**	**3**
CAR	**1728**	**11**	**16**	**29**	**30**
NURSERY	**12960**	**16**	**12**	**8**	**13**
SAMPLE_MODELING	**75,000**	**5**	**5**	**1**	**5**

Union and intersection the local frequent itemsets was applied to generate the most frequent itemset, union gives frequent itemset greater than the actual frequent itemset. And intersection gives frequent itemset less than the actual frequent itemset. See table 3.6.

Table 3.6: Union all and Intersect all Local Frequent Itemsets

Datasets	Global frequent itemsets	Union all	Interest all
CENSUS	**3**	**4**	**3**
CAR	**11**	**64**	**0**
NURSERY	**16**	**20**	**11**
SAMPLE_MODELING	**5**	**5**	**1**

Because this problem we proposed a rule to generate frequent itemset that equal the actual frequent itemset in global dataset. See table 3.7.

Table 3.7: The Proposed Rule

No of Distributed Datasets	Large Itemsets	Mining Frequent Itemsets
3	L1,L2,L3	$(L1 \cup L3) \cap L2$

Such that L1 is the large itemset in site1, L2 is the large itemset in site2and L3 is the large itemset in site3.

We union the maximum number of large itemset (L3) and the minimum number of large itemset (L1), then intersect the result with the third large itemset (L2) to generates the most frequent itemset.

(MAX large itemset∪ MIN large itemset) ∩ Third large itemset.

Table 3.8 shows the results of frequent itemsets after apply the proposed rule.

Table 3.8: Rules applied over datasets

Dataset	Global frequent itemsets	Union all	Interest all	(L1∪l3) ∩ l2	(L1∩ l3) ∪ l2	(L1∩l2) ∪ l3	(L1 ∪l2) ∩ l3	(L2∪ l3) ∩ l1	(L2∩ l3) ∪ l1
CENSUS	3	4	3	3	3	4	3	3	3
CAR	11	64	0	11	16	30	16	0	27
NURSERY	16	20	11	16	29	18	11	11	18
SAMPLE	5	5	1	5	5	5	5	1	5

In alldatasets the proposed rule generate the truth frequent itemsets. But other rules generate number of frequent itemsets greater or less than the actual frequent itemsets. Table 3.9 shows the differencesof results.

Table 3.9: Differences of Frequent Itemsets, +I means combined local itemsets is greater than global itemsets by I and –I means less by I. while ∅ indicate same numbers

Dataset	Union all	Interest all	(L1∪l3) ∩ l2	(L1∩ l3) ∪ l2	(L1∩l2) ∪ l3	(L1 ∪l2) ∩ l3	(L2∪ l3) ∩ l1	(L2∩ l3) ∪ l1
CENSUS	+1	∅	∅	∅	+1	∅	∅	∅
CAR	+48	-16	∅	∅	+14	+5	-16	+11
NURSERY	+9	∅	∅	+17	+7	-5	∅	+7
SAMPLE	∅	-4	∅	∅	∅	∅	-4	∅

CHAPTER 4

CONCULOSION AND FUTURE WORKS

Chapter 4
Conclusion and Future Work

4.1 Conclusion

In this thesis we have discussed a new approach to obtained frequent itemsetsfrom distributed data sources. Firstly,we generate the global frequent itemset from global dataset. Secondly,we divide theglobal dataset into three sites,and then we generate the local frequent itemsets from each site. A comprehensive search for the best way to combine the local itemset has been conducted. In this search we find that the union of smallest and biggest of itemsets intersected with the middle always gives result which is equivalent to global itemsets. The experiment of this thesis has been conducted on four different datasets. These datasets have different sizes and attribute types.

4.2 Future Work

Some of the future work that could done to find more result on the topic of this thesis could be:

- Doing more experiments for more than 3 sites with different sizes.
- Generating a tool that allows users to obtained frequent itemsets from distributed datasets. And embedding this tool in one of the famous data mining software like weka.
- Suggesting a way for generate the global frequent itemsets from datasets that are not uniformly distributed.

References

[1] A.O Ogunde and A.S Sodiiya, “Improved cost models for agent-based association rule mining in distributed databases,” Computer Science Series, 9th Tome 1st Fasc, 2011.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Database," Conf. Very Large Databases (VLDB 94), Morgan Kaufmann, 1994,pp. 407-419.

[3] V. S. Rao and S. Vidyavathi, “Distributed Data Mining and Mining Multi-Agent Data,” in (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 04, 1237-1244, 2010.

[4] S. Paul, “An Optimized Distributed Association Rule Mining Algorithm in parallel and distributed data mining with XML data for improved response time,” in International Journal of Computer Science and Information Technology, April. 2010.

[5] G.K Gupta, “Introduction to Data Mining with Case Studies,” Prentice Hall, 2006.

[6] M. Z. Ashrafi, D. Taniar and K. Smith, “ODAM. An Optimized Distributed Association Rule Mining Algorithm,” in IEEE distributed systems online, march 2004.

[7] M. Chen, J. Han and P. Yu, “Data Mining: An Overview from Database Perspective,” in IEEE Transactions on Knowledge and Data Engineering, 1996.

[8] Frequent Itemsets mining datasets repository. Available at: http://fimi.cs.helsinki.fi/data/.[may 21, 2013].

[9] J. Hipp, U. Guntzer and G. Nakhaeizadeh, “Algorithms for Association Rule Mining,” A General Survey and Comparison, SIGKDD Explorations, 2000.

[10] Jiawei Han und Micheline Kamber, “Data Mining – Concepts and Techniques,” Chapter 5.2.

[11] Baptiste Jeudy, “Optimization of Association Rule Mining Queries,” in Intelligent Data Analysis, Volume 6, 2002.

[12] E. Ansari, G. H. Dastghaibifard, M. Keshtkaran, “DTFIM: Distributed Trie-based Frequent Itemset Mining,” 2003.

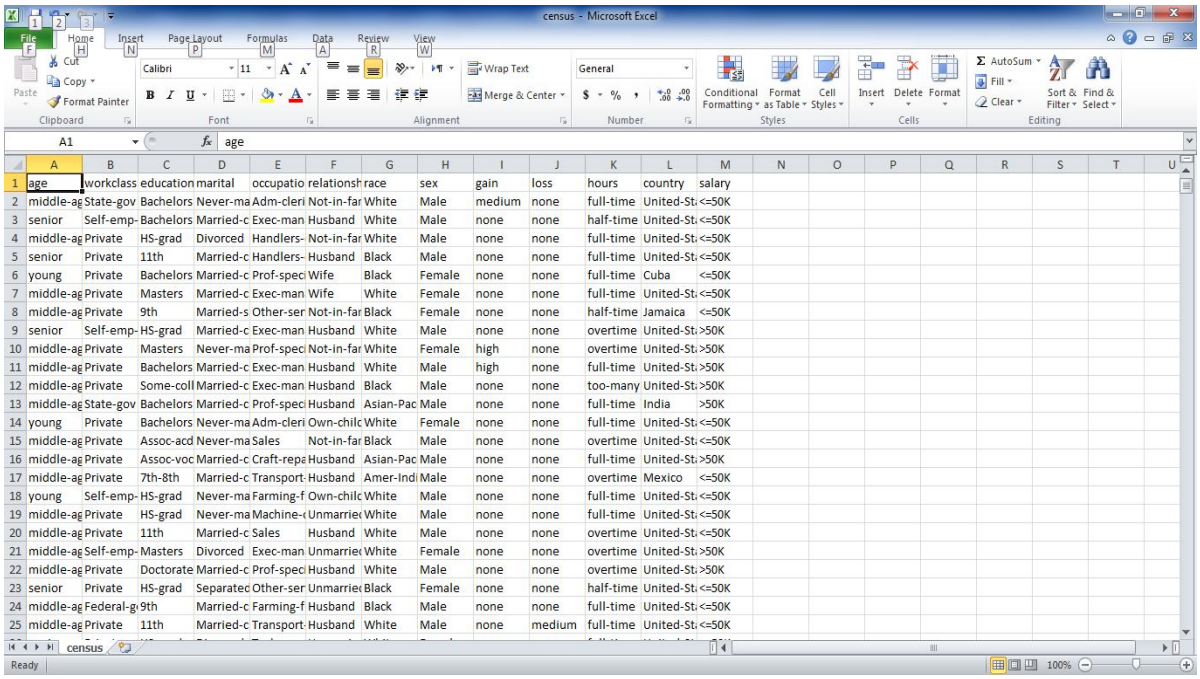
[13] Pankaj Kandpal, “Association Rule Mining in Partitioned Databases: Performance Evaluation and Analysis,” Indian Institute of Information technology, Allahabad, July 2007.

[14] Andrea Pietracaprina and Dario Zandolin, “Mining Frequent Itemsets using Patricia Tries,”University of Padova, 2001.

[15] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Wei Li, Mitsunori Ogihara, “Evaluation of Sampling for Data Mining of Association Rules,” NSF Research Initiation Award(CCR-9409120) and ARPA contract F19628-94-C-0057.

Appendix

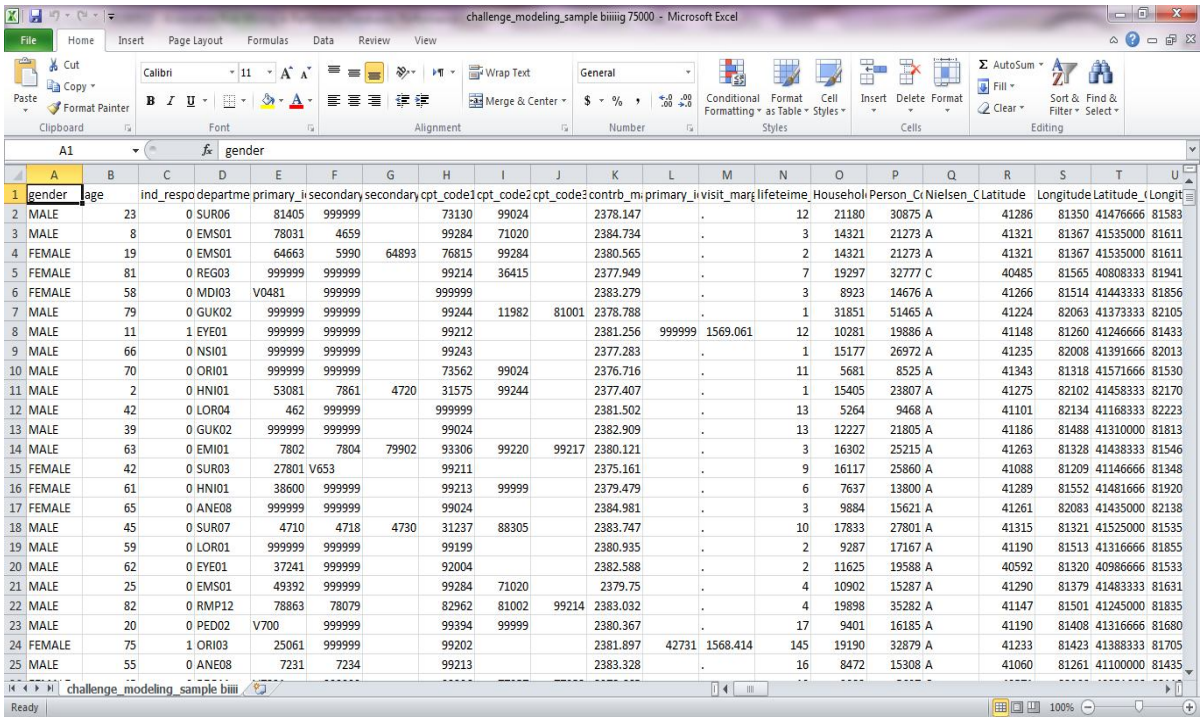
Appendix



census - Microsoft Excel

age	workclass	education	marital	occupatio	relationsh	race	sex	gain	loss	hours	country	salary
middle-ag	State-gov	Bachelors	Never-ma	Adm-cleri	Not-in-far	White	Male	medium	none	full-time	United-St	<=50K
senior	Self-emp-	Bachelors	Married-c	Exec-man	Husband	White	Male	none	none	half-time	United-St	<=50K
middle-ag	Private	HS-grad	Divorced	Handlers-	Not-in-far	White	Male	none	none	full-time	United-St	<=50K
senior	Private	11th	Married-c	Handlers-	Husband	Black	Male	none	none	full-time	United-St	<=50K
young	Private	Bachelors	Married-c	Prof-speci	Wife	Black	Female	none	none	full-time	Cuba	<=50K
middle-ag	Private	Masters	Married-c	Exec-man	Wife	White	Female	none	none	full-time	United-St	<=50K
middle-ag	Private	9th	Married-s	Other-ser	Not-in-far	Black	Female	none	none	half-time	Jamaica	<=50K
senior	Self-emp-	HS-grad	Married-c	Exec-man	Husband	White	Male	none	none	overtime	United-St	>50K
middle-ag	Private	Masters	Never-ma	Prof-speci	Not-in-far	White	Female	high	none	overtime	United-St	>50K
middle-ag	Private	Bachelors	Married-c	Exec-man	Husband	White	Male	high	none	full-time	United-St	>50K
middle-ag	Private	Some-coll	Married-c	Exec-man	Husband	Black	Male	none	none	too-many	United-St	>50K
middle-ag	State-gov	Bachelors	Married-c	Prof-speci	Husband	Asian-Pac	Male	none	none	full-time	India	>50K
young	Private	Bachelors	Never-ma	Adm-cleri	Own-child	White	Female	none	none	full-time	United-St	<=50K
middle-ag	Private	Assoc-acd	Never-ma	Sales	Not-in-far	Black	Male	none	none	overtime	United-St	<=50K
middle-ag	Private	Assoc-voc	Married-c	Craft-repa	Husband	Asian-Pac	Male	none	none	full-time	United-St	>50K
middle-ag	Private	7th-8th	Married-c	Transport	Husband	Amer-Indi	Male	none	none	overtime	Mexico	<=50K
young	Self-emp-	HS-grad	Never-ma	Farming-f	Own-child	White	Male	none	none	full-time	United-St	<=50K
middle-ag	Private	HS-grad	Never-ma	Machine-(	Unmarrie	White	Male	none	none	full-time	United-St	<=50K
middle-ag	Private	11th	Married-c	Sales	Husband	White	Male	none	none	overtime	United-St	<=50K
middle-ag	Self-emp-	Masters	Divorced	Exec-man	Unmarrie	White	Female	none	none	overtime	United-St	>50K
middle-ag	Private	Doctorate	Married-c	Prof-speci	Husband	White	Male	none	none	overtime	United-St	>50K
senior	Private	HS-grad	Separatec	Other-ser	Unmarrie	Black	Female	none	none	half-time	United-St	<=50K
middle-ag	Federal-g	9th	Married-c	Farming-f	Husband	Black	Male	none	none	full-time	United-St	<=50K
middle-ag	Private	11th	Married-c	Transport	Husband	White	Male	none	medium	full-time	United-St	<=50K

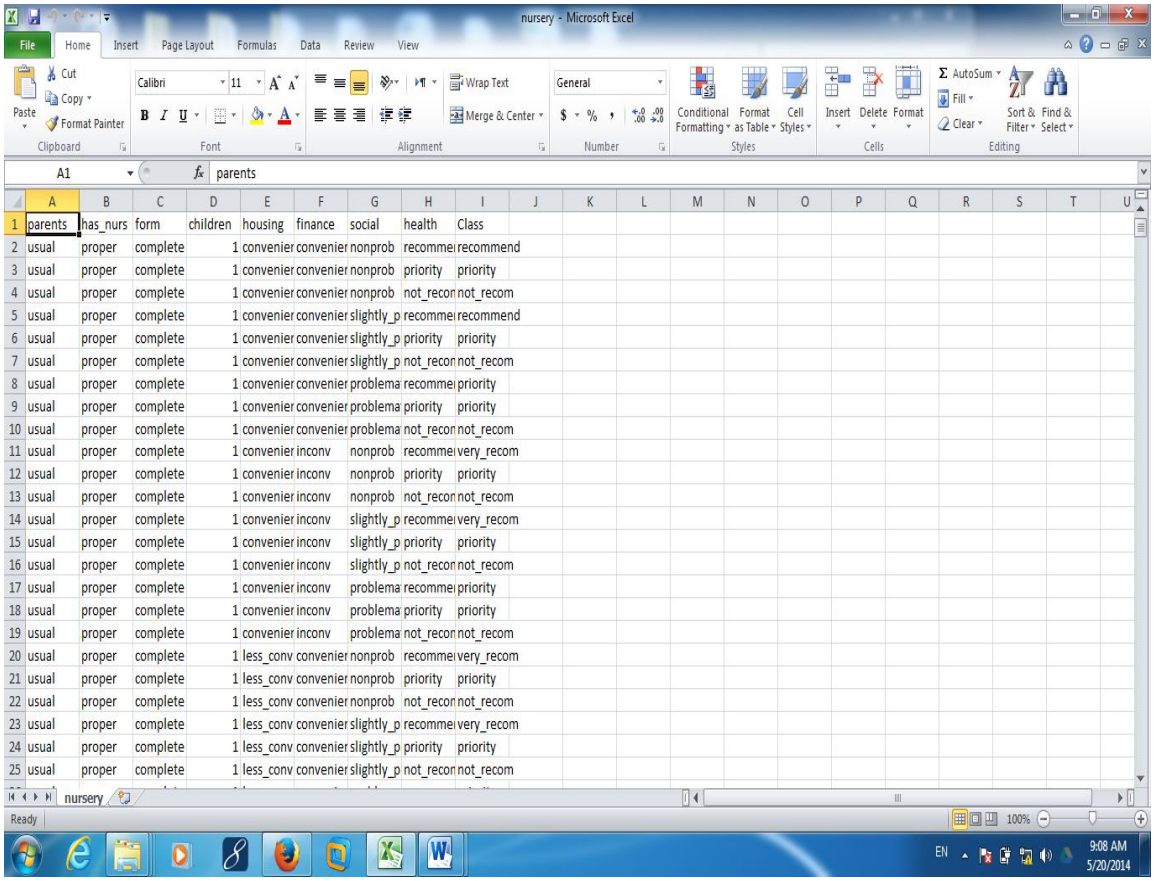
1. CENSUS Dataset



challenge_modeling_sample biiiiig 75000 - Microsoft Excel

gender	age	ind_respo	departme	primary_i	secondary	secondary	cpt_code1	cpt_code2	cpt_code3	contrb_m	primary_i	visit_mars	lifeteime_	Household	Person_C	Nielsen_C	Latitude	Longitude	Latitude_(	Longit
MALE	23	0	SUR06	81405	999999		73130	99024		2378.147		.	12	21180	30875	A	41286	81350	41476666	81583
MALE	8	0	EMS01	78031	4659		99284	71020		2384.734		.	3	14321	21273	A	41321	81367	41535000	81611
FEMALE	19	0	EMS01	64663	5990	64893	76815	99284		2380.565		.	2	14321	21273	A	41321	81367	41535000	81611
FEMALE	81	0	REG03	999999	999999		99214	36415		2377.949		.	7	19297	32777	C	40485	81565	40808333	81941
FEMALE	58	0	MDI03	V0481	999999		999999			2383.279		.	3	8923	14676	A	41266	81514	41443333	81856
MALE	79	0	GUK02	999999	999999		99244	11982	81001	2378.788		.	1	31851	51465	A	41224	82063	41373333	82105
MALE	11	1	EYE01	999999	999999		99212			2381.256	999999	1569.061	12	10281	19886	A	41148	81260	41246666	81433
MALE	66	0	NSI01	999999	999999		99243			2377.283		.	1	15177	26972	A	41235	82008	41391666	82013
MALE	70	0	ORI01	999999	999999		73562	99024		2376.716		.	11	5681	8525	A	41343	81318	41571666	81530
MALE	2	0	HNI01	53081	7861	4720	31575	99244		2377.407		.	1	15405	23807	A	41275	82102	41458333	82170
MALE	42	0	LOR04	462	999999		999999			2381.502		.	13	5264	9468	A	41101	82134	41168333	82223
MALE	39	0	GUK02	999999	999999		99024			2382.909		.	13	12227	21805	A	41186	81488	41310000	81813
MALE	63	0	EMI01	7802	7804	79902	93306	99220	99217	2380.121		.	3	16302	25215	A	41263	81328	41438333	81546
FEMALE	42	0	SUR03	27801	V653		99211			2375.161		.	9	16117	25860	A	41088	81209	41146666	81348
FEMALE	61	0	HNI01	38600	999999		99213	99999		2379.479		.	6	7637	13800	A	41289	81552	41481666	81920
FEMALE	65	0	ANE08	999999	999999		99024			2384.981		.	3	9884	15621	A	41261	82083	41435000	82138
MALE	45	0	SUR07	4710	4718	4730	31237	88305		2383.747		.	10	17833	27801	A	41315	81321	41525000	81535
MALE	59	0	LOR01	999999	999999		99199			2380.935		.	2	9287	17167	A	41190	81513	41316666	81855
MALE	62	0	EYE01	37241	999999		92004			2382.588		.	2	11625	19588	A	40592	81320	40986666	81533
MALE	25	0	EMS01	49392	999999		99284	71020		2379.75		.	4	10902	15287	A	41290	81379	41483333	81631
MALE	82	0	RMP12	78863	78079		82962	81002	99214	2383.032		.	4	19898	35282	A	41147	81501	41245000	81835
MALE	20	0	PED02	V700	999999		99394	99999		2380.367		.	17	9401	16185	A	41190	81408	41316666	81680
FEMALE	75	1	ORI03	25061	999999		99202			2381.897	42731	1568.414	145	19190	32879	A	41233	81423	41388333	81705
MALE	55	0	ANE08	7231	7234		99213			2383.328		.	16	8472	15308	A	41060	81261	41100000	81435

2. Sample_Modeling Dataset



	A	B	C	D	E	F	G	H	I
1	parents	has_nurs	form	children	housing	finance	social	health	Class
2	usual	proper	complete	1	convenier	convenier	nonprob	recommer	recommend
3	usual	proper	complete	1	convenier	convenier	nonprob	priority	priority
4	usual	proper	complete	1	convenier	convenier	nonprob	not_recon	not_recom
5	usual	proper	complete	1	convenier	convenier	slightly_p	recommer	recommend
6	usual	proper	complete	1	convenier	convenier	slightly_p	priority	priority
7	usual	proper	complete	1	convenier	convenier	slightly_p	not_recon	not_recom
8	usual	proper	complete	1	convenier	convenier	problema	recommer	priority
9	usual	proper	complete	1	convenier	convenier	problema	priority	priority
10	usual	proper	complete	1	convenier	convenier	problema	not_recon	not_recom
11	usual	proper	complete	1	convenier	inconv	nonprob	recommer	very_recom
12	usual	proper	complete	1	convenier	inconv	nonprob	priority	priority
13	usual	proper	complete	1	convenier	inconv	nonprob	not_recon	not_recom
14	usual	proper	complete	1	convenier	inconv	slightly_p	recommer	very_recom
15	usual	proper	complete	1	convenier	inconv	slightly_p	priority	priority
16	usual	proper	complete	1	convenier	inconv	slightly_p	not_recon	not_recom
17	usual	proper	complete	1	convenier	inconv	problema	recommer	priority
18	usual	proper	complete	1	convenier	inconv	problema	priority	priority
19	usual	proper	complete	1	convenier	inconv	problema	not_recon	not_recom
20	usual	proper	complete	1	less_conv	convenier	nonprob	recommer	very_recom
21	usual	proper	complete	1	less_conv	convenier	nonprob	priority	priority
22	usual	proper	complete	1	less_conv	convenier	nonprob	not_recon	not_recom
23	usual	proper	complete	1	less_conv	convenier	slightly_p	recommer	very_recom
24	usual	proper	complete	1	less_conv	convenier	slightly_p	priority	priority
25	usual	proper	complete	1	less_conv	convenier	slightly_p	not_recon	not_recom

3. Nursery Dataset