Data Mining

A Research Oriented Study Report

By :-
Akash Saxena
121030931

Organization of the Presentation

- General Overview of Data Mining Concept
- A detailed study of Association Rule Mining Concepts and Algorithms

General Overview

Basic Data Mining Introduction

A View Into The World of Data Mining

- Introduction
- What is Data Mining?
- Why Data Mining?
- Various Techniques of Data Mining

INTRODUCTION

- Data Abundance
- Services for Customer
- Importance of Knowledge Discovery

What is Data Mining?

" It is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic. The pattern discovered must be meaningful, in that they lead to some advantage, usually economic advantage. The data is invariably present in substantial quantities."

In easier terms….

Data mining is about extracting useful knowledge from large databases.

Further simply, Data miner analyzes historical data, discovers some patterns in them and these help human analyst (semi-automatic) or automated decision making tool to predict an apt outcome for a future situation.

Why Data Mining?

- Automation has beaten Manual Human Analyst efforts
- Increasing Competition in the market
- Accurate, Fast, Unexpected predictions having enormous effect on economy.

Data Mining Application Areas

- Business Transactions
- E-commerce
- Scientific Study
- Health Care Study
- Web Study
- Crime Detection
- Loan Delinquency
- Banking

Data Mining Techniques

- Association Rule Mining
- Cluster Analysis
- Classification Rule Mining
- Frequent Episodes
- Deviation Detection/ Outlier Analysis
- Genetic Algorithms
- Rough Set Techniques
- Support Vector Machines

Association Rule Mining

Going Specific

What will I cover…

- Basic terminology in Association DM
- Market Basket Analysis
- Algorithms
  - Apriori Algorithm
  - Partition Algorithm
  - Pincer Search Algorithm
  - Dynamic Itemset Counting Algorithm
  - FP Tree growth Algorithm

Basic Association DM Terms

- Support

  It is the percentage of records containing an item combination compared to total number of records.

- Confidence

  It is the support of an item combination divided by support for a condition. We actually measure how confident can we be, given that a customer has purchased one product, that he will also purchase another product.

- Association Rule

  It is a rule of the form X=›Y showing an association between X and Y that if X occurs then Y will occur. It is accompanied by a confidence % in the rule.

Basic Association DM Terms

- Itemset

  It is a set of items in a transaction. K-itemset is a set of ‘k’ number of items.

- Frequent Itemset

  It is an itemset whose support in a transaction database is more than the minimum support specified.

- Maximal Frequent Itemset

  It is an itemset which is frequent and no superset of it is frequent.

- Border set

  It is an itemset if it is not frequent but all its proper subsets are frequent.

Basic Association DM Terms

- Downward Closure Property

  Any subset of a frequent itemset is frequent.

- Upward Closure Property

  Any superset of an infrequent itemset is infrequent.

Market Basket Analysis

It is an analysis conducted to determine which products customers purchase together. Knowing this pattern of purchasing traits of customer can be very useful to a retail store or company.

This can be very useful as once it is known that customers' who buy product A are likely to buy product B, then company can market both A and B together. Thus making purchase of one product target prospects of another.

Market Basket Analysis

- For Example consider the following database:

Transaction	*Products*
1	Burger, Coke, Juice
2	Juice, Potato Chips
3	Coke, Burger
4	Juice, Ground Nuts
5	Coke, Ground Nuts

Seeing this, there is no visible obvious rule or relationship between items in the buying patterns of the customers.

Move ahead to see mining of relationships…

Example explanation

	Burger	*Juice*	*Coke*	*P chips*	*G Nuts*
Burger	2	1	2	0	0
Juice	1	3	1	1	1
Coke	2	1	3	0	1
P Chips	0	1	0	1	0
G Nuts	0	1	1	0	2

Above table shows how many times was one item purchased with other item. Central diagonal shows how items were purchased with themselves so we ignore it.

Giving view of terms in this example

Association Rule: “If a customer purchases Coke, then he will probably purchase a burger” is an association rule associating Coke and Burger.

Support: This rule has a support of 40%.

records in which both burger and coke occur together = 2

total number of records = 5

support = 2/5 * 100 = 40%

Confidence: Above rule has a confidence of 66%

support for combination (Coke+Burger) is 40%

support for condition (Coke) is 60%

confidence = 40/60 * 100 = 66%

Giving view of terms in this example

Itemset: {Coke, Burger} is a 2-itemset containing 2 items. {Coke} is a 1-itemset.

Frequent Itemset: If Minimum support be 2 then {Coke}, {Juice}, {Burger}, {G Nuts}, {Coke, Burger} are frequent itemsets.

Maximal Frequent Itemset: {Coke, Burger} is a maximal frequent itemset. This is confirmed by the fact that both its subsets viz. {Coke}, {Burger} are frequent too.

Apriori Algorithm

This is a level wise algorithm developed by Dr R Aggarwal & Dr R Srikant.

A set of frequent 1-itemsets is found. Then it is used to generate frequent 2-itemsets and these 2-itemsets are used to generate 3-itemsets and so on.

It has two parts:

-Joint Step Candidate Generation Process

-Pruning Process

Apriori Algorithm continued…

Input: Database D of transactions, Minimum Support Threshold σ

Output: L, set of all Frequent itemsets in D.

Initialize: k=1, C_1=all 1-itemsets

Read Database D to count support of C_1 to determine L_1

L_1= {frequent 1-itemsets}

k=2   // k represents pass number

```
While ($L_{k-1} \neq \emptyset$) do
Begin
        $C_k=\emptyset$
        For all itemsets $l_1 \epsilon L_{k-1}$do
        For all itemsets $l_2 \epsilon L_{k-1}$do
        if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge$ ………….
        $\wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$
        then $c= l_1[1], l_1[2], l_1[3],\ldots\ldots, l_1[k-1], l_2[k-1]$
         $C_k= C_k \cup \{c\}$

        for all $c \epsilon C_k$
        for all (k-1)-subsets of d of c do
                if $d$ ἐ $L_{k-1}$
                then $C_k= C_k / \{c\}$
```

```
For all transactions t ϵ D do
Increment count of all candidates in $C_k$ that are contained in t.
$L_k$ = All candidates in $C_k$ with minimum support
K=k+1
End

Answer: $U_k$ $L_k$
```

Pincer Search Algorithm

This algorithm is one of the fastest Apriori based algorithm which implement horizontal mining. It was developed by Dao I Lin and Z M Kedem.

It uses the Apriori Method but makes it more efficient with the use of concept of Maximal Frequent Itemset thus combining both bottom-up (for frequent itemset generation) and top-down approach (for searching MFS).

```
$L_0 = \emptyset$, $k=1$, $C_1 = \{\{i\} \mid i \in I\}$; $S_0 = \emptyset$
MFCS = {all items}; MFS = $\emptyset$
Do until $C_k = \emptyset$ and $S_{k-1} = \emptyset$
    read the database and count support for $C_k$ &
    MFCS
MFS = MFS U {frequent itemsets in MFCS}
$S_k$ = {infrequent itemsets in $C_k$}
    if $S_k \neq \emptyset$
        for all itemsets s ϵ $S_k$
                for all itemsets m ϵ MFCS
                if s is a subset of m
                MFCS = MFCS \{m}
                for all items e ϵ s
                if m \ {e} is not a subset of any itemset
    in MFCS
                MFCS = MFCS U {m \ {e}}
```

```
For all itemsets c in $L_k$
If c is a subset of any itemset in current MFS
Delete c from $L_k$
```

```
Generate candidates from $C_{k+1}$ from $C_k$
If any frequent itemset in $C_k$ is removed from $L_k$
Then for all itemsets l ϵ $L_k$
    for all itemsets m ϵ MFS
        if first k-1 items in l are also in m
                for i from k+1 to |m|
        $C_{k+1} = C_{k+1}$ U {{l.item$_1$ ,…., l.item$_k$, m.item$_k$}

For all itemsets in $C_{k+1}$
If c I not a subset of any itemset in current MFCS
Delete c from $C_{k+1}$
K=k+1
```

Answer : MFS

Thank You