A NOVEL APPROACH FOR MINING FREQUENT ITEMSETS IN MEDICAL IMAGE DATABASES

G.Parthasarathy¹, D.C.Tomar²

¹Research Scholar, Dept. of CSE, Sathyabama University, Chennai, India
²Professor, Dept. of IT, Jerusalem College of Engineering, Chennai, India.

Received on 09-08-2016
Accepted on 05-09-2016

Abstract

Frequent Itemset Mining is evaluating the rules and relationships within the data items are optimizing it, in the large medical image databases (for e.g. Images, Docs, AVI files etc). It is one of the major problems in DM (Data mining) domain. Finding frequent item set in the large set is one of the computational complexities in mining. To improve the efficiency and performance of the mining frequent item set algorithm, the key term is to apply pruning techniques which reduces the search space and its complexity of the algorithm. Here we proposed a robust technique of pruning called pruning for uncertain data’s. Here our methodology is used to mine the data sources of uncertain data model. We have analyzed and implemented all well known algorithmic models for mining frequent item sets for both binaries and uncertain data’s. Our experimental results show that FPgrowth performance is high for binary data sets where our method performs at high rate of accuracy for uncertain data sets.

Keywords- Data mining, MPpruning, FPgrowth, Pruning, Classification.

I. Introduction

In recent years, data mining has reached in all the domain for storing, processing, retrieving data streams and mining frequent data sets is said to be learning. At an instance the real world problem is defined explicitly by arrival time and time stamp for explicitly. The algorithm for mining data stream should course in single pass for the characters of data streams. According the survey FPMAX* is fastest one for all data sets. This algorithm process with two processes, initialization with scanning datasets for constructing FP-Tree.

By radiologist analysis of medical image is time consuming. So there is an need to analyze large set of medical image with more accuracy and reliability there is a need for latest technology software tool which can reduce the process of the radiologist. The proposed machine learning algorithm used to train an small set of medical images to classify a
When in a store where buyers (customers) comes and buys some items, how likely she would buy the specific items or how various buyers buy the items together?. Recently the researchers started to analysis the rules for data where the existence for the particularity is uncertain. For example consider the medical database MP with various data sets such as multimedia files, docs etc. with a black box algorithm BA. BA takes the input .avi files, .png, and .jpeg files are deliverables be the appropriate files with the suitable formats (say images or video files) along with their corresponding likelihood values. The output of BA may depend on various key parameters such as frame rate, fps (frames per second), size, format etc. The result may be tabulated and is consider as one of the transaction. Let \( I(k) = \{i_1, i_2, i_3, ..., i_k\} \) be the set of the item sets called items. A subset \( X \) by which \( X \subseteq I \) is called as item set. Item set with \( K \) item is called \( K \)-item set.

Analysis of these types of data sets may be used to classify the multimedia files to a specific field. For example, a file with high likelihood values for high frame rate, size belongs to the video files and a file with low likelihood values for size, frame rate belongs to image files.

**Table 1: Example of multimedia files database.**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Video</th>
<th>Audio</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>.avi-30%</td>
<td>.3-40%</td>
<td>.jpeg-50%</td>
</tr>
<tr>
<td>R2</td>
<td>.eg-40%</td>
<td>.wmv-45%</td>
<td>.png-45%</td>
</tr>
<tr>
<td>R3</td>
<td>.mkv-58%</td>
<td>-</td>
<td>.tiff-55%</td>
</tr>
<tr>
<td>R4</td>
<td>.3gp-25%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R5</td>
<td>.divx-80%</td>
<td>.codec-45%</td>
<td>-</td>
</tr>
<tr>
<td>R6</td>
<td>.4-56%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R7</td>
<td>.3gpp-10%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**II. Related Work**

This paper has been developed for digging out and review huge knowledge about the data provided by Deep web. The deep web submits to data sources with backend databases that are only accessible through the query forms and Sentiment analysis; where online reviews and blogs can be dig for recognize writers’ views and feelings on topics of current interest. The prototypes in differences between the values for the same entity are grouped by inheriting differential rules. In this solution, a statistical hypothesis test had been used to identify important variation in values.
of the final quantitative attributes between two data sources. The pruning method is used to identify differential rules and pruned if their behaviours are predicted by their complementary ancestor rules. In this algorithm, a hash table is used to store the identified differential rules according to their profile representations and applied four travel related deep web data sources [1].

Since youngxin proposed the very optimal mining algorithm using depth-first search method to acquire entire probabilistic frequent closed itemsets. To shrink the search space and ignore unnecessary computation, probabilistic pruning and bounding methods [4]. The problem of pruning technique using voronoi diagrams to reduce the number of distance calculation is proposed in this paper and improved the effectiveness [5].

All the previous studies on purning technique analysis a data model to reduce the redundant of data. To improve the efficiency of the U-Apriori algorithm; in this paper propose a data trimming technique to ignore irrelevant candidate support increments performed in the Subset-Function.

In which, the input and output cost has been reduced [6]. In this paper, we studied the discovery of frequent patterns and association rules from probabilistic data under the Possible World Semantics and which discovers frequent patterns in bottom-up and top-down approaches [7].

### III. Medical Image Pruning Architecture

In this part we plan a methodology to extract frequent model of medical objects. The framework can be described as a series of methods. It will dig out the medical objects and its frequency of availability from the large Database and builds a model medical l Objects Datasets. The next method is to first we characterize the frequent order catalog in form of numerical illustration where each object in the transaction is represented then the next step tree using the numerical illustration of each transaction dataset. The final step finds frequent medical model with their respective support count by intersecting its numerical ordered list.

![Fig.1. Medical Prune Methodology.](image-url)
IV. Techniques and Algorithm

Theorems:

**Theorem 1: Subset Frequent**

**Definition:** A subset which is of restricted belongs to any frequent itemset is a maximal frequent item set.

**Theorem 2: Superset Infrequent**

**Definition:** A subset of any infrequent item set is not a frequent item set.

**Theorem 3: Superset Frequent**

Definition: A superset of any frequent item set is a frequent item set.

**Theorem 4: Subset Equivalence**

Definition: Let Z & X be the itemsets associated with node n and x. If x is left to the node of n, then Z ⊆ X

Then Z ⊆ X is defined as Sub (Z) = Sub(X), prune all the children of n.

**Theorem 5: Subset Infrequent**

**Definition:** A subset which is of restricted belongs to any frequent item set is a not a maximal frequent item set.

Pruning Techniques:

Here we use three existing pruning techniques used by FPMax*[2] namely

- Subset infrequency pruning [1,2].
- Superset frequency pruning [1,2].
- Subset equivalence pruning [1, 2].

A. **Subset infrequency pruning:**

Consider the node n is in search space tree, then each item x in conn_tail(n) become the member of free_tail(n), count the support of the item sets head(n) U {x}.

```
Begin
    If (Count (head (n) ||x) & Sup (head (n) ||x))
        Add=free_tail (n);
        head (n)=1
        x=Val (add);
        Assign=head (n)*x;
        Apply theorem2
End
```

Note: **Theorem2:** A subset of any infrequent item set is not a frequent item set.
B. Superset frequency pruning:
Look ahead Pruning is also called as superset pruning, considering node $x$, if item sets $\text{head}(x) \cup \text{tail}(x)$ is frequent. Then all the children nodes of $x$ should be pruned according to the theorem 1. There are two methods in existing methodology one is to count the support (breadth first search) and another method is to check for any supersets (depth first search)[2].

C. Subset equivalence pruning (SEP):
SEP is said to be true when it satisfies and obeys the theorem 4.

Note: Let $Z \& X$ be the item sets associated with node $n$ and $x$. If $x$ is left to the node of $n$, then $Z \subseteq X$. Then $Z \equiv X$ is defined as $\text{Sub}(Z) = \text{Sub}(X)$, prune all the children of $n$.

Proof:
Let $X$ be any item set associated with the node $n$, node $x$ be left to the node $n$, there must be the item $i \in Z \setminus x$ and $i \in X$.

Because $\text{sub}(Z) \equiv \text{sub}(X)$, then any transaction $\text{Trans}$ which contains $X$ must contain $i[2]$.

MP pruning:
MPP is said to be true when it satisfies and obeys the all the theorem 1-5.

Proof:
Let $V$ be any item set associated with the node $n$, node $x$ be left to the node $n$, there must be the item $i \in Z \setminus x$ and $i \in X$.

Because $\text{sub}(Z) \equiv \text{sub}(X)$, then any transaction $\text{Trans}$ which contains $V$ must contain $i[2]$.

For example
Fig 4 shows the MP prune algorithm, whereas figure shows the SEP algorithm. Consider the below stated example, set of elements be $a, v, c, d$ for which

$\phi \equiv \forall(\text{theorems})$

$v \sqsubseteq (a,c), c \sqsubseteq (a,d), d \sqsubseteq (a,v)$ and

$\text{sub}(v) = \text{sub}(a) \cup \text{sub}(c), \text{sub}(c) = \text{sub}(a) \cup \text{sub}(d), \text{sub}(d) = \text{sub}(a) \cup \text{sub}(v)$

![Fig 2: SEP method.](image-url)
Mining Frequent Item set:

To support the count of the items, we need to access the records in the database. The database can be stored in the memory on the following ways:

Table 2: Database Representation.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Representation</th>
<th>Database</th>
<th>Identifier</th>
<th>Scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Horizontal</td>
<td>Transaction Db</td>
<td>Row_id</td>
<td>Single pass</td>
</tr>
<tr>
<td>2</td>
<td>Vertical</td>
<td>Transaction DB</td>
<td>Trans_id</td>
<td>Single pass</td>
</tr>
<tr>
<td>3</td>
<td>Bit-vector</td>
<td>2D DB’s</td>
<td>2d_bitvector</td>
<td>Double pass</td>
</tr>
</tbody>
</table>

FP max mining algorithm is composed of two main steps namely FP Tree construction and mining frequent item sets based on FP Tree. FP Tree is constructed of mainly two parts: namely 1) Header table & 2) Prefix tree [2].

Header Table:

Header table contains all the items in the DB [2]. All link of the header table is set to zero (0). Whenever an item is added to the tree, corresponding entry is updated in the tree.

Prefix Tree: Prefix tree consists of the entire list of items in Header Table and their support in the prefix tree [2].

Table 3: Transaction items after pre-processing.

<table>
<thead>
<tr>
<th>SID</th>
<th>ITEMS</th>
<th>FREQUENT ITEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A,B,C,D</td>
<td>A,C,B,D</td>
</tr>
<tr>
<td>S2</td>
<td>C,D,E</td>
<td>C,D</td>
</tr>
<tr>
<td>S3</td>
<td>E,F,A,B</td>
<td>E,F,A,B</td>
</tr>
<tr>
<td>S4</td>
<td>C,D,E</td>
<td>C,D</td>
</tr>
</tbody>
</table>
**D. Medical image classification process**

1. Medical image acquisition

2. Medical image enhancement (ie) preprocessing

3. Feature Extraction

4. Classification

The acquisition of medical images by creation of prototype model of a database for having the medical images. (ie)

The medical images directly get from the medical systems. To improve the quality of medical image we need to do preprocessing. From the cleaned image the essential feature will digging out the classification algorithm used. To classify the medical image association rule mining is used. The Association rule mining scope is to introduce associations between items in a transactional medical databases.

**Algorithm:**

Function \( O = MPprune(T, M, C, Min\_sup) \)

\( T = \text{input\_parameter}, M: \text{FPTree}, C: \text{MFI \ for \ } T, Min\_sup: \text{CFI \ for \ } min\_sup(\text{MPprune}); \)

Output: \( O; \)

\( \text{If } T \rightarrow \rightarrow \text{path}(p) \text{ then} \)

\( \text{Insert}(p) \rightarrow M; \)

\( \text{Else} \)

\( \text{For } i: \text{T.header} \)
If(MPprune \equiv \forall(\text{Theorem}))

Continue;

\textbf{If} \text{Sub}(i) < \text{min}_\text{sup};

\text{Set } Y = \text{T.base} \parallel i;

\text{If}(\text{sub}(i) \equiv Y)

\text{Add } Y \text{ to } C;

\text{If}[T[i] \neq \text{NULL}) \text{ then}

\text{Free}_\text{Tail} = \{\text{Frequent Item set for } T[i]\}

\text{Else}

\text{Sort_desc(Free}_\text{Tail});

\text{End if}

\text{End if}

\text{End if}

\text{Else if} (MPprune \equiv \forall(\text{Theorem})) \text{ then}

\text{Construct FPTree}

\text{Goto Insert(FPTree)→M;}

\text{Merge FPTree with } M

\text{End if}

\text{End}

\text{End if}

\text{End}

V. Results & Performance measure of the algorithm

We analyze the performance and running time of the algorithm on binary and uncertain data sets. The common dataset taken is connecting datasets and it was downloaded from [3]. The experimental results shows that our proposed methodology can efficient reduce the search space. Here the cost is very low when compared to subset equivalence pruning technique because in our methodology we are focusing our novelty over the frequently used item sets rather focusing on CFI(closed frequent Item sets) and MFI (Maximal frequent Item sets). Here our proposed methodology is to check whether the item sets has superset and their support is equal or not. Here we can store the
MP item sets which are lower than the threshold values. Thus the search space of our proposed methodology is at little smaller when compared to other of CFI & MFI. Hence the whole efficiency of the proposed work is at improved rate, especially for uncertain data sets [1, 2, 5, 6, and 9].

**Table 4: Example of Item sets sampling-MPprune.**

<table>
<thead>
<tr>
<th>F. ID</th>
<th>ITEMS</th>
<th>FREQUENT ITEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>A,B,C</td>
<td>A,C,B</td>
</tr>
<tr>
<td>F2</td>
<td>C,D,E,F,G</td>
<td>C,D</td>
</tr>
<tr>
<td>F3</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>F4</td>
<td>A,B,G</td>
<td>A,B,G</td>
</tr>
<tr>
<td>F5</td>
<td>B,G</td>
<td>B,G</td>
</tr>
<tr>
<td>F6</td>
<td>A,B,C</td>
<td>A,B,C</td>
</tr>
<tr>
<td>F7</td>
<td>A,G,B</td>
<td>A,G</td>
</tr>
<tr>
<td>F8</td>
<td>A,B,I</td>
<td>B,I</td>
</tr>
</tbody>
</table>

**Fig.5. Implementation results of SP prune.**

**Fig.6. Performance Vector of MPprune**
VI. Conclusion

In this paper we extend MP prune for uncertain data. We proved that the proposed method has better accuracy when compared to other methods. Finally we analyzed the performance of frequently used data over binary and uncertain data sets. In binary data sets FPMax* plays better performance than other algorithms while in uncertain data sets MP prune algorithm outperforms than existing other algorithms in most of the cases that we examined. When the support threshold is of 6% - 26% we noticed that our algorithm is similar as same as FPMax*. The above method increase the accuracy and optimize decision time allowing a detailed analysis of huge number of medical image in the short time.

References


Corresponding Author:
G.Parthasarathy*,
Email: amburgps@gmail.com