A Comparative study of Techniques in Data Mining

Manika Verma¹, Dr. Devarshi Mehta²

¹Asst. Professor, Department of Computer Science, Kadi Sarva Vishwavidyalaya, Gandhinagar, India
²Associate Professor, GLS Institute of Computer Technology, Ahmedabad, India

Abstract- Data Mining encompasses tools and technique for the “Extraction or Mining of knowledge from huge repository of data”. In Data Mining various techniques that used are Association Rule Mining, Sequential Pattern Mining, Clustering, and Classification. Varieties of algorithms are developed for each of these techniques. This paper present comparison of algorithms developed for different techniques. It also includes comparison of various algorithms available for sequential pattern mining like GSP, FreeSpan, PrefixSpan and tools available for implementation of Data Mining Algorithm.

Keywords-- Data Mining, Sequential pattern mining.

I. INTRODUCTION

Valuable information is hidden inside the repository of data. Since, the speed at which data is generated is much faster than it can be processed and made sense, this information often remains buried and untouched. It is impossible for individuals to find valuable information hidden behind data without technological resources.

Data Mining encompasses tools and technique for the “extraction or mining of knowledge from large amounts of data” (Han and Kamber, 2001). It is the process of discovering knowledge by extracting previously unknown valid actionable information or hidden patterns from large data base.

The importance of data mining has been established for business applications, criminal investigations, biomedicine, and counter-terrorism. Most retailers, for example, employ data mining practices to uncover customer buying patterns – Amazon.com uses purchase history to make product recommendations to shoppers. Data mining can be applied wherever there is an abundance of data available for and in need of analysis.

Data Mining is used for extracting various interesting patterns. Example for few of the patterns are i) Along with item A, majority of time item B is also sold (Association mining) ii) After item A is sold, majority of time item B is also sold.(Sequence Pattern mining) iii) Those customer having age between 25-30, and having salary between 20,000 to 40,000 tends to buy Mobile phones ranging from 30,000 to 40,000 (Classification) iv) Students having same characteristics, like those scoring below average in Mathematics (Clustering). This paper shows the comparison between various techniques and algorithm developed for these techniques. [1][28]

II. TECHNIQUES IN DATA MINING

Techniques that are used in data mining describe the type of mining and data recovery operations.

Following Tree demonstrate few key techniques and algorithm for each key technique that are used for discovering interesting patterns. [2]
CLASSIFICATION OF DATA MINING ALGORITHM [6][26][2][27]

Figure 1
A Map Of Data Mining Algorithms Offered In The Spmf (Open-Source Data Mining Software) [38]
### III. TECHNIQUES, ALGORITHMS AND LIMITATIONS

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Algorithm Name</th>
<th>Introduction</th>
<th>Founded By and Founded in Year</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSOCIATION RULE MINING</td>
<td>Apriori Algorithm</td>
<td>Apriori Algorithm find frequent itemsets from a transaction dataset and derive association rules.[29]</td>
<td>Developed by Agrawal and Srikant (1994) [29]</td>
<td>Candidate generation generates large numbers of subsets. Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset ( S ) only after all ( 2^{</td>
</tr>
<tr>
<td>REGRESSION</td>
<td>Support Vector Machines</td>
<td>In machine learning, support vector are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data.[10]</td>
<td>Cortes and Vapnik, (1995) [30]</td>
<td>One of the initial drawbacks of SVM is its computational inefficiency SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes). [11]</td>
</tr>
<tr>
<td>CLUSTERING</td>
<td>k-means Algorithm</td>
<td>The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, ( k ). [31]</td>
<td>Lloyd (1957, 1982), Forgy (1965), Friedman and Rubin (1967), and McQueen (1967)[25]</td>
<td>Sensitive to initialization. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. K-means has problems when clusters are of Different Sizes, Densities and Non-globular shapes. Problems with Outliers Empty Clusters Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always.[4]</td>
</tr>
<tr>
<td>CLASSIFICATION</td>
<td>C4.5</td>
<td>C4.5 is an algorithm used to generate a decision tree. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.[32]</td>
<td>Ross Quinlan (1997)[3]</td>
<td>Does not work well with small training data set Small variation in data can lead to different decision trees (especially when the variables are close to each other in value)[3]</td>
</tr>
</tbody>
</table>
IV. SEQUENTIAL PATTERN MINING ALGORITHM [12]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Approach</th>
<th>Founded By and Founded in Year</th>
<th>Advantages</th>
<th>Disadvantage</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSP</td>
<td>Apriori-based</td>
<td>Proposed by Agrawal and Srikan (1994) [29]</td>
<td>Reduces search space[33]</td>
<td>Scans the database multiple times, Generate a huge set of CandidateSequences[33]</td>
<td>Efficiency is less[33]</td>
</tr>
<tr>
<td>FreeSpan</td>
<td>Pattern-Growth</td>
<td>Proposed by Jiawei Han, Jian Pei, BehzadMortazavi-Asl, Qiming Chen, UmeshwarDayal, Mei-Chun Hsu(in 2000)[34]</td>
<td>FreeSpan projects a large sequencedatabase recursively into a set of small projected sequence databases based on the currently mined frequent sets, the subsequent mining is confined to each projected database relevant to a smaller set of candidates.[34]</td>
<td>If a pattern appears in each sequence of a database, its projected database does not shrink every time as in Prefix Span.[35]</td>
<td>Efficiency is better than GSP but worse than PrefixSpan.[34]</td>
</tr>
<tr>
<td>PrefixSpan</td>
<td>Pattern-Growth</td>
<td>Proposed by Jian Pei, Jiawei Han, BehzadMortazazi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, UmeshwarDayal, Memberand Mei-Chun Hsu (in 2001)[24]</td>
<td>No candidate sequence needs to be generated, Projected databases keep shrinking. [33]</td>
<td>Major cost of PrefixSpan: constructing projected databases. [33]</td>
<td>Efficiency is much better as compared to GSP and SPADE. [33]</td>
</tr>
<tr>
<td>CloSpan</td>
<td>Pattern-Growth</td>
<td>Proposed by Xifeng Yan, Jiawei Han, RaminAfshar ( in 2003)[23]</td>
<td>A closed sequential pattern s: there exists no superpattern s’ such that s’⊂ s, and s’ and s have the same support [22]</td>
<td>Detect Closed Subsequences only. [21]</td>
<td>Reduces the number of (redundant) patterns but attains the same expressive power Using Backward Subpattern and Backward Superpattern pruning to prune redundant search space CloSpan will outperform SPADE and SPAM when the patterns to be mined are long and database is large [22]</td>
</tr>
</tbody>
</table>

V. TOOLS AVAILABLE FOR IMPLEMENTATION OF DATA MINING ALGORITHM

1. A Sequential Pattern Mining Framework:
   - SPMF is an open-source data mining library written in Java.
     a. It is distributed under the GPL v3 license.
   - It offers implementations of 52 data mining algorithms for:
     i. sequential pattern mining,
     ii. association rule mining,
     iii. frequent itemset mining,
     iv. sequential rule mining,
     v. clustering
The source code of each algorithm can be integrated in other Java software. The current version is v0.94 and was released the 12th August 2013. [8]

2. **Weka 3: Data Mining Software in Java:**
   a. Weka is a collection of machine learning algorithm for data mining tasks written in Java.
   b. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.
   c. It is also well-suited for developing new machine learning schemes. [7]

3. **Orange:**
   a. Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting.
   b. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques.
   c. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform Qt framework.
   d. Orange is distributed free under the GPL. [9]

4. **RapidMiner:**
   a. Rapid Miner is an open-source system for data mining that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.
   b. RapidMiner is written in the Java programming language.
   c. RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.
   d. RapidMiner functionality can be extended with additional plugins. The Rapid Miner Extensions marketplace provides a platform for developers to create data analysis algorithms and publish them to a broader community.
   e. RapidMiner is distributed under the AGPL open source license and has been hosted by SourceForge where it is rated the #1 business analytics software. [17][14]

5. **KNIME:**
   a. KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform.
   b. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept.
   c. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.
   d. KNIME is written in Java and based on Eclipse and makes use of its extension mechanism to add plugins providing additional functionality.
   e. KNIME's core-architecture allows processing of large data volumes that are only limited by the available hard disk space (most other open source data analysis tools are working in main memory and are therefore limited to the available RAM). E.g. KNIME allows analysis of 300 million customer addresses, 20 million cell images and 10 million molecular structures.
   f. Additional plugins allows the integration of methods for Text mining, Image mining, as well as time series analysis.
   g. KNIME integrates various other Open-Source-projects, e.g. machine learning algorithms from Weka, the statistics package R, as well as LIBSVM, JFreeChart, ImageJ, and the Chemistry Development Kit.
h. KNIME is implemented in Java but also allows for wrappers calling other code in addition to providing nodes that allow to run Java, Python, Perl and other code fragments.[16]

6. Apache Mahout
   a. Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification.
   b. Many of the implementations use the Apache Hadoop platform.
   c. Mahout also provides Java libraries for common math operations (focused on linear algebra and statistics) and primitive Java collections.
   d. Mahout is implemented in Java
   e. Mahout is a work in progress. [19]

7. jHepWork
   a. jHepWork is an interactive framework for scientific computation, data analysis and data visualization designed for scientists, engineers and students.
   b. Since 2013, it was renamed to SCaViS project.
   c. The program is designed for interactive scientific plots in 2D and 3D and contains numerical scientific libraries implemented in Java for mathematical functions, random numbers, statistical analysis, curve fitting and other data mining algorithms.
   d. jHepWork is an attempt to create a data-analysis environment using open-source packages with a coherent user interface and tools competitive to commercial programs.
   e. The idea behind the project is to incorporate open-source mathematical and numerical software packages with GUI-type of user interfaces into a coherent program in which the main user interface is based on short-named Java/Python classes.

   f. jHepWork is multiplatform since it is written in Java, thus it runs on any operating system where the Java virtual machine can be installed.
   g. jHep Work is written in Java, Jython. jHepWork uses high-level programming languages, such as Jython (Python implemented in Java), JRuby (Ruby implemented in Java), but Java coding can also be used to call jHepWork numerical and graphical libraries.
   h. While the program falls into the category of open source software, it is not completely free for commercial usage. [18]

8. Rattle
   a. Rattle GUI is a free and open source software (GNU GPL v2) package providing a graphical user interface (GUI) for Data Mining using the R statistical programming language.
   b. Rattle is written in R.
   c. Rattle provides considerable data mining functionality by exposing the power of the R Statistical Software through a graphical user interface.
   d. Rattle can be used for statistical analysis, or model generation.[20]

VI. CONCLUSION

The paper basically focuses on algorithms in data mining, their purposes and limitations. For identifying sequences various sequential pattern mining algorithms have been developed. The paper also focuses on purpose of various sequential patterns mining algorithm, their advantages and disadvantages. Various tools are available for data mining. The programming languages in which the tools are developed and the purposes for which tools are developed have been discussed in paper.

REFERENCES


[22] “Sequential Pattern Mining Tutorial”
[23] XifengYan Jiawei Han , RaminAfshar, CloSpan: Mining Closed Sequential Patterns in Large Datasets, CiteSeer,2003
[24] Jian Pei , Jiawei Han , BehzadMortazavi-asl , Helen Pinto , Qiming Chen , UmeshwarDayal , Mei-chun Hsu. CloSpan: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. CiteSeer,2001
[26] “Cluster Analysis Tuotrial”
[31] “JarostawStepaniuk”, “Rough Granular Computing in knowledge discovery and data mining”
[33] “Sequential Pattern Mining Tutorial” www.is.informatik.uni-duisburg.de
[34] JiaweiHan,JianPei,BehzadMortazavi-asl,FreeSpan:Frequent Pattern-Projected Sequential Pattern Mining
[35] Manish Gupta,Jiawei Han, Approaches for pattern Discovery using Sequential Data Mining
[37] Mohammed J.Zaki, “SPADE:An Efficient Algorithm for Mining Frequent Sequences”