DATA MINING: PROCESS, TOOLS, TECHNIQUES AND APPLICATIONS

G. ABIRAMI

Department of computer science
Mangayarkarasi College of arts and science for women

ABSTRACT:
This paper deals with detail study of Data Mining its techniques, tasks and related Tools. Data Mining refers to the mining or discovery of new information in terms of interesting patterns, the combination or rules from vast amount of data. It helps in classifying, segmenting data and in hypothesis formation. With such a vast amount of data, there is need for powerful technique for better interpretation of these data. Including commercial and open source, many program available to perform data mining. Data mining tools predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human.

1. INTRODUCTION
Data mining is an innovative system which requires some of exclusive abilities and understanding. Currently there may be no trendy framework wherein to perform information mining projects. This approach that the achievement or failure of facts mining task is fairly depending on the unique man or woman or team sporting it out and a success practice cannot necessarily be repeated throughout the organization. Data mining wishes a well known method so as to help translate enterprise problems into information mining obligations, advocate appropriate records changes and facts mining strategies, and provide method for comparing the effectiveness of the results and documenting the enjoy sample looking i.e. Information mining, locating presentation, finding interpretation and locating assessment. A typical KDD procedure is proven in Figure.1 [4].

Figure 1: A typical Knowledge Discovery process
2 DATA MINING PROCESS
The CRISP-DM (Cross Industry Standard Process for Data Mining) project addressed elements of those problems by way of defining a technique version which provides a framework for carrying out records mining tasks that is independent of each the enterprise zone and the generation used. The CRISP-DM manner version pursues to make massive data mining tasks, less costly, more reliable, extra repeatable, greater possible, and faster.

2.1 Data Mining Industry needs a Standard Process Model
The statistics mining industry is presently at the chasm (Moore, 1991) between early marketplace and predominant flow market (Agrawal, 1999). Its commercial success remains no longer guaranteed. If the early adopters fail with their data mining tasks, they may not blame their own incompetence there in the use of information mining well but assert that information mining does not work. In the market, there is nonetheless to a degree the expectancy that facts mining are a push-button era. However, this is not authentic, as maximum practitioners of information mining recognize. Data Mining is a complicated technique requiring various equipment and special human beings. A procedure version can assist to understand and control the interactions along this complex procedure.

2.2 The CRISP-DM Methodology
CRISP-DM builds on preceding tries to define knowledge discovery. This section offers a top level view of the CRISP-DM methodology. More detailed records may be determined in (CRISP, 1999)

![Figure 1: Four Level Breakdown of the CRISP-DM Methodology for Data Mining](image)

2.3 The Generic CRISP-DM Reference Model
The CRISP-DM reference model for statistics mining provides an overview of the existence cycle of a records mining task. It incorporates the stages of a challenge, their respective tasks, and their outputs. The life cycle of a records mining undertaking is damaged down in six levels that are proven in Figure 2. The collection of the phases isn't always strict. The arrows indicate simplest the maximum important and common dependencies among phases, however in a particular task, it depends at the outcome of each segment Which phase, or which unique challenge of a phase, must be completed subsequent.
2.3 CRISP-DM in Action
Response modeling is a method to enhance the performance and effectiveness of mailing movements in marketing. It lets in to increase the response price even as reducing the charges of a marketing campaign. By generalizing the understanding we have already got on our potentialities with the assist of facts mining strategies, we're capable of predict the likelihood of capability clients to reply to our mailings.

3. DATA MINING TOOLS
Here are parts of the table with the energetic equipment in [9] as License code: CO - industrial, OS - open source

<table>
<thead>
<tr>
<th>Tool</th>
<th>Company</th>
<th>License</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. Ants</td>
<td>H. Ants Analytics</td>
<td>CO</td>
<td>family of data mining tools with a focus on business applications</td>
</tr>
<tr>
<td>ADAPA</td>
<td>ADAPA</td>
<td>CO</td>
<td>Develops the ADAPA decision engine which is a transactional activity, unique.</td>
</tr>
<tr>
<td>Cohesion SPAD</td>
<td>Cohesion</td>
<td>CO</td>
<td>company provides data mining, decision support, and decision support for business.</td>
</tr>
<tr>
<td>23K - Data in Knowledge</td>
<td>23K</td>
<td>OS</td>
<td>additional tools for EA and training, tool IK for using under development, free academic version, see Akala9, all developments since 2004</td>
</tr>
<tr>
<td>DataApplied Data Applied</td>
<td>Data Applied</td>
<td>CO</td>
<td>web service for DataAnalysis, SaaS</td>
</tr>
<tr>
<td>DataDexterity</td>
<td>DataDexterity</td>
<td>CO</td>
<td>web tools for fuzzy matching, applications on CRM, spam analysis, fraud detection</td>
</tr>
<tr>
<td>GhostDive</td>
<td>TOS Podd</td>
<td>CO</td>
<td>multi model support</td>
</tr>
<tr>
<td>IBM/SPSS Modr</td>
<td>IBM</td>
<td>CO</td>
<td>business intelligence, now in cooperation with IBM, Predictive Analytics software (PASW), SPSS is an IBM company since 2009</td>
</tr>
<tr>
<td>Revolution Enterprise</td>
<td>Revolution</td>
<td>OS/CO</td>
<td>based on open source software R with many additional tools for big data (e.g. Hadoop) support and database coupling, some commercial parts also free for academic use</td>
</tr>
<tr>
<td>Tally Predictive Modeling State (TSS)</td>
<td>Tally</td>
<td>CO</td>
<td>includes famous separate tools CART, MARS, TreeNet, Random Forests</td>
</tr>
<tr>
<td>SAS Enterprise Miner</td>
<td>SAS Enterprise</td>
<td>CO</td>
<td>one of the world’s leading tools, enterprise support</td>
</tr>
</tbody>
</table>

4. DATA MINING TECHNIQUES
4.1 Classification

The potential to expect a pupil’s overall performance is very crucial in instructional environments. Students’ instructional performance is based totally upon various elements like non-public, social, mental and other environmental variables. A very promising tool to reap this objective is the use of Data Mining. Data mining strategies are used to perform on massive quantity of records to discover hidden styles and relationships helpful in selection making. Data Mining may be used in academic subject to beautify our know-how of learning method to consciousness on figuring out, extracting and evaluating variables associated with the gaining knowledge of process of students as defined by means of Alas el-Hales [2]. Mining in academic surroundings is known as Educational Data Mining.

4.2 Clustering

Clustering is figuring out comparable companies from unstructured statistics. Clustering is the mission of grouping a fixed of objects in a any such way that item in same group are more much like every other than to those in other agencies. Once the clusters are decided, the objects are labeled their corresponding clusters, and commonplace capabilities of the items in cluster are summarized to form a class description. For instance, a financial institution may also cluster its customer in to numerous companies primarily based on the similarities in their income, age, intercourse, house and so forth, and the command traits of the clients in a collection can be used to explain that group of clients.

4.3 Regression

In results- and size-oriented surroundings, the policymakers who oversee higher education, the dad and mom who pay for it, and the scholars who make university selections search for proof of institutional best to differentiate institutions and manual decision making. This evidence includes objective final results measures: Do college students examine new records or competencies? Do they graduate? Are they ultimately a hit in addition education or careers? But subjective measures also indicate institutional fine: Do college students have a wealthy and rewarding college revel in? Do they believe they have learned and grown? Are they satisfied?

4.3.1 Data

Data for this evaluation are drawn from a scholar opinion survey at a public studies university in spring 2000 (ACT, 2000). The survey became administered to students in a consultant sample of undergraduate lessons and either finished in magnificence or allotted in class for of entirety prior to the subsequent meeting. Specifically, fifty eight% of the respondents are woman in comparison to 50% of the pupil body, and seventy two% are higher elegance me in comparison to fifty eight% of the pupil body. Exploring the results of gender and university elegance on pleasure is beyond the scope of this evaluation; however neither emerges as a widespread predictor. The pupil frame as a whole is extra ethnically various than many public universities, but ethnicity additionally fails to become a large predictor.
4.3.2 Regression Analysis
Stepwise regression provides one method of figuring out specific experiences which are related to general pleasure. Approximately one hundred forty gadgets have been entered, quite fewer than the whole variety of objects on the survey because WHAT SATISFIES STUDENTS? 255 some similar objects had been blended and some had been removed due to the fact they have been Applicable to small subsets of survey respondents or had very huge amounts of missing records.

4.4 Outer
Understanding the biology of ailment-inflicting, or pathogenic, organisms will have a wonderful impact on enhancing the best of human lifestyles. With the development of excessive throughput genome sequencing tasks, biologists have amassed huge amounts of raw biological sequences that are publicly to be had. In order to gain a better know-how of the shape and function of such sequences, one crucial task going through the biology network is to correctly classify these sequences into corresponding purposeful households.

Figure 1. The Gram-negative bacteria cell.

For example, the presence of a string of amino acid residues in a protein that bureaucracy a structure called a Trans membrane a-helix is indicative of a protein resident on the inner membrane. Such a-helix structures have a completely feature collection and cutting-edge algorithms for their detection in
a given protein collection are very correct. However, the fundamental outer membrane proteins, feature of Gram poor micro organism, do no longer consist of Tran’s membrane a-helices, but rather form anti parallel b-strands that shape.

4.4.1 Support Vector Machine Based Classification
SVMs expect all statistics to be represented as vectors in a few feature areas. Given a classified set of training data from two instructions (the wonderful and the bad magnificence), SVMs discover a hyper plane that correctly separates the education records of the 2 exceptional lessons at the same time as maximizing the distance of both elegance from the hyper aircraft (maximizing the margin).

4.5 Sequential Patterns
The undertaking of sequential pattern mining is an information mining mission Specialized for analyzing sequential statistics, to discover sequential styles. More precisely, it includes coming across exciting subsequences in a fixed of sequences, wherein the interestingness of a subsequence can be measured in phrases of various criteria together with its occurrence frequency, length, and income.

4.6 Prediction
Today, within the United States, about one in eight ladies over their lifetime has a risk of developing breast cancer. An analysis of the maximum recent information has shown that the survival charge is 88% after five years of prognosis and 80% after 10 years of diagnosis. The discovery of the survival charge or survivability of a certain disease is possible through extracting the expertise from the statistics related to that disease. One of those records sources is SEER (Surveillance Epidemiology and End Results), which is a unique, dependable and vital resource for investigating the special factors of most cancers. The SEER database combines patient-level information on cancer web page, tumor pathology, level, and purpose of loss of life. Relationship of the association
4.7 Association Rules
Fault detection is a key problem in the development of superior spacecrafts. Although several detection strategies consisting of restriction-sensing, simulation and professional structures have been utilized for this cause, they’ve frequently omitted small anomalies within the house-keeping statistics (HK facts) and some of them have caused deadly damages to the overall missions as a result. One cause for the problem is that traditional fault detection methods commonly require a wonderful a priori know-how on the device behavior for each spacecraft. It ought to be observed here that, in many real past fault cases, next guide inspections of the saved statistics found out that a few signs had regarded long earlier than the deadly faults befell. Therefore, its miles quite herbal to think of making use of the data mining techniques to the saved HK information to collect robotically some beneficial information for anomaly detection.

5. DATA MINING APPLICATION
Various fields use statistics mining technology because of fast get admission to of facts and valuable facts from enormous amount of statistics. Data mining technologies had been implemented correctly in many areas like advertising and marketing, telecommunication, fraud detection, and finance, scientific and so forth. Some of the application is indexed under.

A. Financial Data Analysis
B. Retail Industry
C. Telecommunication Industry
D. Biological Data Analysis
E. Other Scientific Applications

6. CONCLUSIONS
In this paper, we've discussed element examine of information mining with diverse research like responsibilities, gear and techniques. The implementation of statistics mining strategies will allow users to retrieve meaningful information from actually included data. These techniques provide style
of packages for industries like retail, Telecommunication, Bio-clinical and many others. These tools are expecting destiny tendencies and behaviors, allowing enterprise to make an Proactive and gift know how in the shape that is without problems understood to human.

REFERENCE

Lecturer, Computer Technology, V.A.P.M.Almala, Latter, India1 Lecturer, Electronics and Telecommunication, V.A.P.M.Almala, Latter, India 2 Assistant Professor, Electronics and Telecommunication, Sandipani Technical Campus F.E., Latter, India3