A SURVEY - BASED ON METHODS AND TECHNIQUES OF DATA MINING

A.Vishnu Priya
M.E. Computer science and engineering
Tamilnadu college of engineering
Vishnupri36@gmail.com

Dr.K.Krishnaswari
Professor and Head of the department
Tamilnadu college of engineering
krishnaswar@gmail.com

Abstract—Data mining is a powerful and a new field having various techniques. It converts the raw data into useful information in various research fields. Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. While the last few years knowledge discovery tools have been used mainly in research environments, sophisticated software products are now rapidly emerging. Here in this paper, explained about the knowledge discovery database (KDD) and also survived the various methods of data mining and its techniques.

Index Terms—Data mining, information prediction, raw data, Knowledge discovery in databases, surveys.

I. INTRODUCTION

Data Mining or Knowledge Discovery is needed to make sense and use of data. Knowledge Discovery in Data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1]. Data mining consists of more than collection and managing data; it also includes analysis and prediction. People are often do mistakes while analyzing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. There are several applications for Machine Learning (ML), the most significant of which is data mining. Numerous ML applications involve tasks that can be set up as supervised. In the present paper, we have concentrated on the techniques necessary to do this. In particular, this work is concerned with classification problems in which the output of instances admits only discrete, unordered values.

Figure 1 Data Mining and the KDD Process(Source: Fayyad, et.al., 1996)

Based on figure 2, KDD process consists of iterative sequence methods as follows [7, 9]:
1. Selection: Selecting data relevant to the analysis task from the database
2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources
3. Transformation: Transforming data into appropriate forms to perform data mining
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; Extracting data patterns
5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing
redundant or irrelevant patterns; Translating the useful patterns into terms that human understandable

II. Related work:

The survey based on the applications of data mining to organizational knowledge management for effective capturing, storing and retrieving, and transferring knowledge. The review concentrate on categories into four main groups: (i) knowledge resource; (ii) knowledge types and/or knowledge datasets; (iii) data mining tasks; and (iv) data mining techniques and applications used in KM.

Table 1 shows the tabulation on data mining and its techniques

<table>
<thead>
<tr>
<th>Authors</th>
<th>Knowledge Resources</th>
<th>Knowledge Types</th>
<th>DM Tasks</th>
<th>DM techniques / Applications</th>
<th>and Transfer</th>
<th>Knowledge Measurement</th>
<th>Data Mining Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lavrac et al. (2007)</td>
<td>Healthcare Public Health Data • The healthcare providers database • The outpatient healthcare statistic database • The medical status Database</td>
<td>Classification and clustering</td>
<td>Clustering Methods: • Agglomerative Classification; • Principal Component Analysis; • The Kolmogorov-Smirnov Test; • The Quantile Range Test and Polar Ordination Classification on C4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hwang et al. (2008)</td>
<td>Healthcare (Clinical Diagnosis) • Knowledge Conversion</td>
<td>Dependence Modeling</td>
<td>Data Mining Tool IBM Intelligent Miner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen, Lu &amp; Sheu (2009)</td>
<td>Financial • Knowledge Sets – strings of data, models, parameters, and reports • Knowledge Sharing Processes to a Corporate Bond Classification</td>
<td>Classification Clustering</td>
<td>HybridSO FM/LVQ Classifier for Bond Rating, The Self-Organizing Feature Map (SOFM) The Learning Vector Quantization (LVQ) • Ontology of Knowledge Management and Knowledge Sharing • Financial Knowledge Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Title</td>
<td>Classification</td>
<td>Theory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Zhu &amp; Pan (2010)</td>
<td>Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering</td>
<td>Decision Tree Neural Network Early Warning and Proactive Control Systems (EW&amp;PC)</td>
<td>Knowledge Sharing Knowledge Graph Knowledge Flows (KFs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (2010)</td>
<td>Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering</td>
<td>Decision Tree Neural Network Early Warning and Proactive Control Systems (EW&amp;PC)</td>
<td>Knowledge Sharing Knowledge Graph Knowledge Flows (KFs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cantu &amp; Ceballos (2010)</td>
<td>Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering</td>
<td>Decision Tree Neural Network Early Warning and Proactive Control Systems (EW&amp;PC)</td>
<td>Knowledge Sharing Knowledge Graph Knowledge Flows (KFs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu et al. (2010)</td>
<td>Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering</td>
<td>Decision Tree Neural Network Early Warning and Proactive Control Systems (EW&amp;PC)</td>
<td>Knowledge Sharing Knowledge Graph Knowledge Flows (KFs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparative Study of Different Algorithms Implemented to Reduce Scalability Issue

### A) Clustering methods

<table>
<thead>
<tr>
<th>Title of the paper</th>
<th>Conclusion</th>
<th>Drawbacks</th>
<th>Performance</th>
<th>Future work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.A Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering [2011]</td>
<td>Contains advantages of both memory and model based CF Scalable Accurate</td>
<td>Sparsity problem</td>
<td>Mean Absolute Error</td>
<td>Ultra-cluster rates smoothing can be implemented in future</td>
</tr>
<tr>
<td>2.A Collaborative</td>
<td>Low MAE than</td>
<td>Mean Absolute</td>
<td>Cold start problem</td>
<td>Cold start problem</td>
</tr>
</tbody>
</table>

---

112
<table>
<thead>
<tr>
<th>Title of the paper</th>
<th>Conclusion</th>
<th>Drawbacks</th>
<th>Performance</th>
<th>Future work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hybrid Recommender System with Temporal Information [2012]</td>
<td>Scalable &amp; Sparsity and cold start problems are covered</td>
<td>Privacy issue is not handled</td>
<td>Mean Absolute Error and Recommendation time</td>
<td>Geographical information can be considered and Improving time stamping for better performance RS for N-screen service</td>
</tr>
<tr>
<td>2. An Improved Profile based CF Scheme with Privacy [2011]</td>
<td>Improved performance And accuracy More scalable</td>
<td>Sparsity problem Accurancy and quality are not so good compared to CF scheme</td>
<td>Mean Absolute Error</td>
<td>For more accurate results combining the proposed schemes, such as clustering or data reduction techniques combine for enhancement.</td>
</tr>
<tr>
<td>3. RF-REC: Fast and Accurate Computation of Recommendations based on Rating Frequencies [2010]</td>
<td>Highly scalable More accurate Easy to implement High predictability Predictions at constant time</td>
<td>Poor performance and Higher model building time</td>
<td>Root Mean Square Error</td>
<td>Improve quality of diversity and novelty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Filtering Recommendation Algorithm Based on User Clustering and Item Clustering [2010]</th>
<th>traditional CF Quality improvement More accurate and more scalable than traditional</th>
<th>Error</th>
<th>m</th>
<th>m is to be handled</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Collaborative Filtering Recommendation Algorithm Based on Cluster [2011]</td>
<td>Combination of CF and clustering to improve scalability and sparsity problem</td>
<td>Cold start</td>
<td>Only algorithm is proposed</td>
<td>For better performance, the cold start and privacy is to be taken into consideration</td>
</tr>
<tr>
<td>4. Scaling-up Item-based Collaborative Filtering Recommendation Algorithm based on Hadoop [2011]</td>
<td>Minimizing communication cost Scalable Efficient performance</td>
<td>Scalability to be handled more effectively</td>
<td>Speed up efficiency</td>
<td>On algorithm side, optimize allocation and execution Process to be included On application side, scalability is to be improved</td>
</tr>
</tbody>
</table>

B) Other methods
III CONCLUSION

Data mining is a “decision support” process in which we search for patterns of information in data. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc. The commercial, educational and scientific applications are increasingly dependent on these methodologies. Decision trees are a reliable and effective decision making technique which provide high classification accuracy with a simple representation of collected KDD. This paper surveys different data mining techniques that can be used to efficiently and accurately capture user behavior. The paper also presents guidelines that show which techniques may be used more efficiently according to the task implemented by the application.

References


[29] Siavash Ghodsi Moghaddam, Ali Selamat, “A Scalable Collaborative Recommender Algorithm Based on User Density-Based Clustering”, 3rd international conference on Data Mining and Intelligent Information Technology Applications (ICMiA), IEEE 2011, pp. 246-249


A. Vishnu Priya completed her B.E(computer science and engineering) in 2014 from K.S.R. College of Engineering under Anna University,Coimbatore. Currently she is pursuing M.E.( computer science and engineering) in Tamilnadu College of Engineering under Anna University,Chennai.

Mrs.Dr.K.Krishneswari completed her B.E (computer science and engineering) in 2000 from tamilnadu college of engineering under Bharathiyar University, Coimbatore.M.E.(software engineering) in 2005 from sri ramakrishna engineering college under Anna University, Chennai. Ph.D degree from Anna University, Coimbatore. She is working as an assistant professor and head of the department of computer science and engineering at tamilnadu college of engineering, Coimbatore. She is a member of CSI, ISTE. Her research areas include Image processing, Pattern recognition, Biometrics and having 13 years of teaching experience in engineering colleges and also have published 30 papers in journal.