Chapter 26
A Taxonomy of Data Mining Problems

Nayem Rahman
Portland State University, Portland, USA

ABSTRACT

Much of the research in data mining and knowledge discovery has focused on the development of efficient data mining algorithms. Researchers and practitioners have developed data mining techniques to solve diverse real-world data mining problems. But there is no single source that identifies which techniques solve what problems and how, the advantages and limitations, and real-life use-cases. Lately, identifying data mining techniques and corresponding problems that they solve has drawn significant attention. In this paper, the author describes the progress made in developing data mining techniques and then classify them in terms of data mining problems taxonomy to help assist practitioners in using appropriate data mining techniques that solve business problems. This will allow researchers to expand the body of knowledge in this discipline. This article proposes a data mining problems taxonomy based on data mining techniques being used. Prominent data mining problems include classification, optimization, prediction, partitioning, relationship, pattern matching, recommendation, ranking, sequential patterns and anomaly detection. The data mining techniques that are used to solve these data mining problems in general fall under top 10 data mining algorithms.

INTRODUCTION

In this early 21st century organizations capture all business activities in some computer storage systems. Data are also gathered or used from the systems owned by others. Over a period of time data growth has increased significantly in organizations. Emergence of the Internet, social networking tools (e.g., Twitter, Facebook and LinkedIn), and online shopping sites allows for capturing huge data volume related to business. In 2014 the US government had mandated release of huge volume of privacy-protected Healthcare and Medicare data which could be used by researchers, policy makers, business organizations and general public for analysis and decision making (US Govt. Health and Human Services Of-

fice, 2014). With the advent of commodity hardware (to process big data), computer processing power (thanks to Moore's Law), maturity of computer engineering, software engineering, network bandwidth and increasingly low cost of data storage companies are able to capture, process, transform and store large volume of data. Organizations find business value in the data and have come to rely on with their decision-making process.

Traditional Business Intelligence (BI) tools consists of reports, interactive query and Online Analytical Processing (OLAP) all of which can provide intelligence as to what happened in the past. In the past reporting was based on what happened. These days business would like to understand what is going to happen now and in the future using predictive analytics (for example). With the increase in fraudulent activities there is a desire to detect it immediately – using anomaly detection (credit card fraud). Here data mining techniques and algorithms come into picture and play a prominent role in providing solutions to complex business problems.

Data mining techniques are used to discover previously unknown and valuable interesting patterns and relations in large data sets (Phyu, 2009). Given that the amount of data has been growing in enterprise data warehouses, it is apparent that data-driven decision making will help organizations achieve competitive advantage. However, the challenge of using this data to achieve business success is dependent on efficient data mining methods (Wu et al., 2014) that help in extracting hidden knowledge and translating that into business values.

In this study, the author presents the data mining problems that are aligned with providing solution to business problems. These data mining problems include anomaly detection (Ogwueleka et al., 2011), prediction, classification, pattern recognition (Jain, 2010), sequence discovery, data visualization (Shaw et al., 2001) and recommendation system. The author also presents the data mining techniques used to solve data mining problems. These are Bayesian networks, neural networks, decision trees, association rules, clustering, support vector machines, logistic regression, and k-nearest neighbors.

Last two decades most research was conducted on the theoretical and computational process of data mining and knowledge discovery (Shaw et al., 2001). Now is the time to evaluate those data mining techniques and classify them from data mining problem taxonomy perspective. This will allow users to choose an optimal algorithm to solve a particular data mining problem. This paper discusses data mining problem taxonomy issues by presenting a complete taxonomy of data mining problems from the context of real world applications.

LITERATURE REVIEW

Most business organizations have huge volumes of data. Knowledge workers with no special tools cannot read and analyze the data. It requires processing of data using ETL tools, business intelligence (Schlesinger & Rahman, 2015; Rahman & Iverson, 2015) and data mining tools (MacLennan et al. 2009) in order to make informed business decisions. ETL tools, database SQL and reporting tools can do this job to some extent. These tools are not enough to analyze data and find interesting patterns. For instance, the data mining findings that those who buy diapers also happen to purchase beers is an unrelated and complex pattern to identity (Padmanabhan and Tuzhilin, 2000). Conventional SQL and reporting tools cannot handle this. This requires running complex algorithms. A large number of data mining algorithms (Wu et al. 2008) and tools are on the market to process data, find relationships and patterns in raw data and

Related Content

Brave New Digital Tools for Action Research in Education: A Beginner's Guide
Reinhard Bauer, Klaus Himpsl-Gutermann, Martin Sankofi, Petra Szucsichand Ruth Petz (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 469-490).*
www.irma-international.org/chapter/brave-new-digital-tools-for-action-research-in-education/252039

Healthcare Information Security in the Cyber World
Brian S. Coatsand Subrata Acharya (2019). *Machine Learning and Cognitive Science Applications in Cyber Security (pp. 1-56).*
www.irma-international.org/chapter/healthcare-information-security-in-the-cyber-world/227575

Big Data and Its Visualization With Fog Computing
Richard S. Segalland Gao Niu (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 341-377).*
www.irma-international.org/chapter/big-data-and-its-visualization-with-fog-computing/252034

The Diagnosis of Dengue Disease: An Evaluation of Three Machine Learning Approaches
Shalini Gambhir, Sanjay Kumar Malikand Yugal Kumar (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 1076-1095).*
www.irma-international.org/chapter/the-diagnosis-of-dengue-disease/252072

Designs of Mixed Method Research
Amir Manzoor (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 95-121).*
www.irma-international.org/chapter/designs-of-mixed-method-research/252022