Review of Domain Driven Data Mining

K. Priya Karunakaran
Department of Computer Engineering
St. Francis Institute of Technology, Mumbai, India

Abstract- This paper presents a review of three papers based on Domain Driven Data Mining. In the first paper [1], Domain Driven Data Mining is proposed as a methodology and a collection of techniques targeting domain driven actionable knowledge delivery to drive Knowledge Discovery from Data (i.e. KDD) toward enhanced problem-solving infrastructure and capabilities in real business state of affairs. The second paper [2] emphasizes the development of methodologies, techniques, and tools for actionable knowledge discovery and delivery by incorporating relevantly ubiquitous intelligence surrounding data-mining-based problem solving. In the third paper [3] an application for intelligent credit scoring has been discussed using domain driven data mining techniques.

Keywords – Domain Driven Data Mining

I. INTRODUCTION

Knowledge Discovery from Data (KDD) is one of the most active areas in Information Technology [1]. Surveys of data mining for business applications has shown that there is a big gap between academic objectives and business goals, and between academic outputs and business expectations. Traditional data mining research mainly focuses on developing, demonstrating, and pushing the use of specific algorithms and models.

The process of data mining stops at pattern identification. Consequently, a widely seen fact is that 1) many algorithms have been designed of which very few are repeatable and executable in the real world, 2) often many patterns are mined but a major proportion of them are either commonsense or of no particular interest to business, and 3) end users generally cannot easily understand and take them over for business use. It is seen that the findings of KDD are not actionable, and lack soft power in solving real-world complex problems.

Domain-driven data mining (D3M) has been proposed to tackle the above issues, and promote the paradigm shift from "data-centered knowledge discovery" to "domain-driven, actionable knowledge delivery."

The rest of the paper is organized as follows. The shift from data driven data mining to domain driven data mining is discussed in II while the various challenges and prospects of domain driven data mining is presented in III and an application based on this technique is discussed in IV. Concluding remarks are given in section V.

II. DATA DRIVEN DATA MINING TO DOMAIN DRIVEN DATA MINING

A. Distinctive features of Data Driven Data Mining

The existing data mining methodology usually chains self-governing pattern discovery from data.

a) KDD (Knowledge Discovery from Data).
b) Elementary objectives of KDD are to discover knowledge that is of key concentration to genuine business requirements and user preferences but KDD is a presumed and preset process.
c) Targets the production of predefined and automatic algorithms and tools. As a consequence, the algorithms and tools developed have no potential to adapt to external environment constraints.
d) Lots of patterns are generated according to the problems but they are not enlightening and clear to business individuals.
e) Conventional KDD is a data centered and technically dominated course targeting automated hidden pattern mining.
f) The core objective of conventional data mining research is to let data verify research innovation, track the elevated performance of the algorithms and express novel algorithms.

B. Problems In Data Driven Data Mining

a) The mining process stops at discovering knowledge that is primarily of importance to academic or industrial individuals.
b) Existing data mining is principally data-centered and technically conquered, and stops at hidden pattern mining favoring technical concerns and expectation, while many other features surrounding business problems have not been thoroughly or exhaustively considered and balanced.
c) A large fraction of the identified patterns may be either commonsense or of no particular attention to business desires. Business grassroots are puzzled as to why and how they should be concerned regarding those conclusions.
d) Activities extracted or summarized through post investigation and post processing without in view of business concerns do not replicate the authentic expectations of business desires.

C. Key Elements of Domain Driven Data Mining

Domain driven data mining refers to the set of methodologies, frameworks, approaches, techniques, tools and systems that deliver for human, domain, organizational and social, and network and web factors in the environment, for the innovation and delivery of actionable knowledge. Actionable knowledge means business responsive and comprehensible, reflects user preferences and business needs, and can be effortlessly taken over by business individuals for decision-making and action-taking".

1. Restraint -Based framework

In human society, everyone is restrained either by communal regulations or by individual situations. Similarly, actionable knowledge only can be discovered in a restraint-based framework such as environmental authenticity, opportunities, and restraints in the mining procedure.

2. Incorporate Field Awareness

The incorporation of field awareness is subject to how it can be signifying and filled in to the knowledge discovery process. Ontology-based field awareness representation, transformation, and mapping between business and data mining systems is one of the proper approaches to form field awareness.

3. Collaboration Among Human beings and Mining Systems

Human involvement is embodied through the collaboration among humans (including users and business analysts, essentially domain experts) and data mining systems. This is accomplished through the complementation between human qualitative brainpower, such as field awareness and field supervision, and mining quantitative brainpower like computational potential.

4. Mining Exhaustively Patterns

Mining exhaustively patterns should think as how to get better both scientific and business interestingness in the previous restraint-based framework. Technically, it could be through enhancing or generating more effective interestingness Measures.

5. Improving Knowledge Actionability

The measurement of actionable patterns is to follow the actionablilty of a pattern. Both technical and business interestingness measures must be satisfied from both objective and subjective point of view.

6. Loop - clogged repetitive Improvement

Actionable knowledge discovery in a restraint based framework is probably to be a clogged rather than an open course of action. It includes repetitive feedback to varying phases such as sampling, assumption, feature selection, modeling, evaluation and interpretation in a human-involved approach.

7. Interactional and Concurrent Mining Supports

For interactional mining supports, clever agents and service-oriented computing are a number of high-quality technologies. They can support flexible, business-friendly, and user-oriented human-mining interaction through building facilities for user modeling; user knowledge achievement; domain knowledge modeling; personalized user services and recommendation; run-time Supports; and mediation and management of user roles, interaction, security, and cooperation.

Real-world data mining applications have projected critical desires for discovering actionable knowledge especially for real-users and industry needs. Actionable knowledge discovery is significant and also very challenging. It is listed as one of great challenges of KDD.

III. DOMAIN-DRIVEN DATA MINING: CHALLENGES AND PROSPECTS

Domain Driven Data mining (D3M) emphasizes the development of methodologies, techniques, and tools for actionable knowledge discovery and delivery by incorporating relevantly ubiquitous intelligence surrounding data-mining-based problem solving [2]. Ubiquitous intelligence consists of in-depth data intelligence, human intelligence, domain intelligence, network intelligence, and organizational/social intelligence. It is essential to synthesize such ubiquitous intelligence in actionable knowledge discovery and delivery. In recent years, researchers with strong industrial engagement have realized the need to shift from "data mining" to "knowledge discovery" [4] [5] [6]. This paper has presented an overview of D3M by systematically addressing a series of aspects, including challenges facing traditional data mining methodologies, the fundamental framework and relevant techniques for D3M referred to as Actionable Knowledge Discovery.

A. Actionable Knowledge Discovery: Macro Level Issues

1. Environment: Refer to any factors surrounding data mining models and systems, for instance, domain factors, constraints, expert groups, organizational factors, social factors, business processes, and workflows
2. Human role: To handle many complex problems, human-centered and human-mining-cooperated AKD is crucial.
3. Process: Real-world problem solving has to cater for dynamic and iterative involvement of environmental elements and domain experts along the way.
4. Infrastructure: The engagement of environmental elements and humans at runtime in a dynamic and interactive way requires an open system with closed loop interaction and feedback.
5. Dynamics: To deal with the dynamics in data distribution from training to testing and from one domain to another, in domain and organizational factors, in human cognition and knowledge, in the expectation of deliverables, and in business processes and systems.
6. Evaluation: Interestingness needs to be balanced between technical and business perspectives from both subjective and objective aspects.
7. Risk: Risk needs to be measured in terms of its presence and then magnitude, if any, in conducting an AKD project and system.
8. Policy: Data mining tasks often involve policy issues such as security, privacy, and trust existing not only in the data and environment, but also in the use and management of data mining findings in an organization's environment.
9. Delivery: Determining the right form of delivery and presentation of AKD models and findings so that end users can easily interpret, execute, utilize, and manage the resulting models and findings, and integrate them into business processes and production systems.

B. Actionable Knowledge Discovery: Micro Level Issues

On the micro-level, issues related to technical and engineering aspects supporting AKD need to be addressed. The following lists a few dimensions that address these concerns:

1. Architecture: AKD system architectures need to be effective and flexible for incorporating and consolidating specific environmental elements, AKD processes, evaluation systems, and final deliverables.
2. Process: Tools and facilities supporting the AKD process and workflow are necessary, from business understanding, data understanding, and human system interaction to result assessment, delivery, and execution of the deliverables.
3. Interaction: To cater for interaction with business people along the way of ADK process, appropriate user interfaces, user modeling, and servicing are required to support individuals and group interactions.
4. Adaptation: Data, environmental elements, and business expectations change all the time. AKD systems, models, and evaluation metrics are required to be adaptive for handling differences and changes in dynamic data distributions, cross domains, changing business situations, and user needs and expectations.
5. Actionability: What do we mean by "actionability?" How should we measure it? What is the trade-off between technical and business sides? Do subjective and objective perspectives matter? This requires essential metrics to be developed.
6. Deliverable: AKD deliverables are required to be easily interpretable, convertible into or presented in a business-oriented way such as business rules, and be linked to decision-making systems.

Table-1. Comparison of major aspects under the research of traditional data-driven and domain-driven data mining.

Aspects	**Data Driven**	**Domain Driven**
Rationale	Data tells a story	Data and ubiquitous intelligence disclose problem- solving solutions
Objective	Innovative and effective algorithm	Effective problem solving
Data	Abstract, synthetic and refined data	Real –life data and surrounding information
Process	One-off	Multi-step, iterative and interactive on demand
Mechanism	Automated	Human centered or human -mining-cooperated
Infrastructure	Closed pattern mining system	Closed-loop problem solving systems in open environment
Usability	Predefined models and processes	Ad-hoc, dynamic and customizable models and processes
Deliverable	Patterns	Business friendly decision support systems
Deployment	Solid validation	Well-founded artwork in problem solving
Evaluation	Technical metrics	Tradeoff between technical significance and business expectation

Table 1 shows a comparison of major aspects of Data Driven Data Mining and Domain Driven Data mining.

IV. DOMAIN-DRIVEN CLASSIFICATION FOR INTELLIGENT CREDIT SCORING

Credit scoring is often used to analyze a sample of past customers to differentiate present and future bankrupt and credit customers [3]. Extracting knowledge from the transaction records and the personal data of credit card holders has great profit potential for the banking industry. The challenge is to detect/predict bankrupts and to keep and recruit the profitable customers. However, grouping and targeting credit card customers by traditional data-driven mining often does not directly meet the needs of the banking industry, because data-driven mining automatically generates classification outputs that are imprecise, meaningless, and beyond users' control.

A. Problems

1. Achieving accurate credit scoring is a challenge due to a lack of domain knowledge of banking business.
2. Previous work used linear and logistic regression for credit scoring; regression cannot provide mathematically meaningful scores and may generate results beyond the user's control.
3. Traditional credit analysis, such as decision tree analysis, rule-based classification, Bayesian classifiers, nearest neighbor approaches, pattern recognition, abnormal detection, optimization, hybrid approach, neural networks, and Support Vector Machine (SVM), are not suitable since they are data-driven mining that cannot directly meet the needs of users.
4. The limitation of the data-driven mining approaches is that it is difficult to adjust the rules and parameters during the credit scoring process to meet the users' requirements.

B. Domain –Driven Score Based Approach

In this paper, the authors provide a novel domain-driven classification method that takes advantage of multiple criteria and multiple constraint level programming for intelligent credit scoring.

1. The method involves credit scoring to produce a set of customers' scores that allows the classification results actionable and controllable by human interaction during the scoring process.
2. Domain knowledge and experts' experience parameters are built into the criteria and constraint functions of mathematical programming and the human and machine conversation is employed to generate an efficient and precise solution.
3. Experiments based on various data sets validated the effectiveness and efficiency of the proposed methods.

C. Advantages Of Domain Driven Score Based Approach

1. Simple and meaningful: It outperforms traditional data-driven mining because the practitioners do not like black box approach for classification and prediction.
2. Measurable and Actionable: Credit card providers value Sensitivity—the ability to measure the true positive rate (that is, the proportion of bankruptcy accounts that are correctly identified-the formal definition will be given later) but not Accuracy.

IV.CONCLUSION

From the study of the three papers [1], [2] and [3], the authors have brought forward the importance of the need for a paradigm shift from the traditional Data Driven Data Mining approaches towards the Domain Driven Data Mining approach.

Serious efforts should be made to develop workable methodologies, techniques, and case studies to promote another round of booming research and development of date mining in real- world problem solving.

REFERENCES

[1] Mitu Kumari,"Data Driven Data Mining to Domain Driven Data Mining", Global Journal of Computer Science and Technology Volume 11 Issue 23 Version 1.0 December 2011

[2] Longbing Cao, "Domain-Driven Data Mining: Challenges and Prospects", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 6, JUNE 2010

[3] Jing He, Yanchun Zhang, Yong Shi, Guangyan Huang, "Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 6, JUNE 2010

[4] U. Fayyad and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad and P. Smyth, eds., pp. 1-34, 1996.

[5] M. Ankerst, "Report on the SIGKDD-2002 Panel the Perfect Data Mining Tool: Interactive or Automated?" ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 110-111, 2002.

[6] L. Cao and T. He, "Developing Actionable Trading Agents," Knowledge and Information Systems: An Int'l J., vol. 18, no. 2, pp. 183-198, 2009.