Domain-Driven In-Depth Pattern Discovery: A Practical Methodology [1]

Longbing Cao, Li Lin, Chengqi Zhang

Faculty of Information Technology, University of Technology Sydney, Australia
{lbcao, linli, chengqi}@it.uts.edu.au

Abstract. Traditional data mining is a data-driven trial-an-error process. Patterns discovered via predefined models in the above process, we call these patterns as *generic patterns*, are often not really interesting to constraint-based real business. In order to work out patterns really interesting and actionable to the real world, called *in-depth patterns*, pattern discovery is more likely to be a domain-driven human-machine-cooperated process. This paper proposes a practical data mining methodology named *domain-driven in-depth pattern discovery* (DDID-PD). Main ideas of the DDID-PD methodology are introduced. Guided by this methodology, we demonstrate some of our work in mining in-depth correlations in Australian Stock Exchange data. Real domain work has shown that our methodology is practical and potential for deeply analyzing patterns actionable to real business.

1 Introduction

Traditional data mining is a data-driven trial-and-error process [1] where data mining algorithms extract patterns from data via some predefined models. It targets fully automated mining process, algorithms and tools [1]. A data mining system is expected to be an automated tool without human involvement and the capability to adapt to external environment constraints.

However, data mining in the real world is highly constraint-based [2]. Real-world patterns interesting to business are often hidden in a large quantity of data with complex data structures and source distribution (*data constraints*). The real-world business process, problems and requirements are often tightly embedded in domain-specific information and expertise (*domain constraints*). Nonetheless, most of mined patterns would not be interesting or actionable to business even though they are sensible to research, or there exists interestingness conflicts between academia and business (*interestingness constraints*). Furthermore, the rules automatically discovered from domain-specific data often do not make sense to real business process or regulations, or they must be integrated with other business rules so that they can be deployed into real life (*rule constraints*).

[1] This project is sponsored by CMCRC (Capital Market CRC, www.cmcrc.com), AC3 data, and UTS Research Fund.

To deal with the above-mentioned constraints in the real world, it is essential to slough off superficial and captures the essential in data mining. Some real experience and lessons learned in artificial intelligence and pattern recognition [3], and integrated business intelligence for telecom industries [4] have taught us the involvement of domain knowledge and even domain experts can assist with filtering subtle concerns while capturing incisive issues and driving a practical design. Similarly, in order to effectively mine and deploy interesting patterns from the aforementioned constraint-based context, the involvement of domain knowledge and experts and the consideration of constraints are essential for knowledge discovery on a neatly definable domain problem. Combining these aspects together, a sleek DM methodology could be developed to find distilled core of a problem and build a deep domain model for advising the process of real-world data analysis and preparation, the selection of features, the design and fine-tuning of algorithms, and the evaluation and refinement of mining results in a more effective way. This leads to the *domain-driven in-depth pattern discovery* (DDID-PD) framework.

The key ideas of the DDID-PD framework include (i) dealing with constraint-based context, (ii) mining in-depth patterns, (iii) supporting human-machine-cooperated interactive knowledge discovery, and (iv) viewing data mining as a loop-closed iterative refinement process. Handling constraint-based context can improve the quality and effectiveness of data mining by extracting and transforming the domain-specific datasets in terms of guides taken from domain experts and their knowledge. In-depth pattern mining can discover more interesting and actionable patterns from domain-specific perspective. In this framework, data mining and domain experts complement each other on an in-depth granularity via an interactive interface. The involvement of domain experts and their knowledge can assist in developing highly effective domain-specific data mining techniques and reduce the complexity of the knowledge producing process in the real world. A system following the DDID-PD framework can embed effective supports for domain knowledge and experts' feedbacks, and refines the lifecycle of data mining in an iterative manner. Therefore, DDID-PD can benefit the real-world knowledge discovery in a more effective and efficient manner, and support the discovery of more interesting and actionable patterns from specific domains compared with current data-driven data mining methodology such as CRISP-DM [5].

Taking the real stock markets as an instance, this paper introduces some preliminary work of mining deep correlations in the markets through taking the DDID-PD methodology. These real-world in-depth analyses further show that the DDID-PD methodology is potential for discovering deep core domain model, and adapts to complex and dynamic business processes and requirements.

2 DDID-PD Framework

2.1 Fundamental concepts

In the DDID-PD framework, a collection of concepts are proposed in terms of practical requirements from the real world. These concepts bring either new ideas or deep thinking into the existing data mining framework, and enhance the efficiency and effectiveness of real-world data mining.

DEFINITION 1: Generic Pattern -- Referring to patterns automatically discovered by data mining models while taking little consideration of business requirements and interestingness.

For instance, in association rule mining, a large number of rules are often found while most of them might not make sense to business. These rules are called *generic patterns*. Another instance is trading strategy, for example, Moving Average (MA), discovered by financial experts. Taking MA as an instance, it actually represents a huge quantity of trading strategies. These strategies are generic rules since they are neither specifically developed for handling certain cases nor as effective as possible for daily trading decision. Obviously, generic patterns are interesting to data miners while not interesting enough to business for taking actions in the real world.

DEFINITION 2: In-depth Pattern – Referring to patterns which are highly interesting and actionable in business decision-making. These patterns are created through refining model or tuning parameters to optimize generic patterns; they may also be directly discovered from data set with sufficient consideration of business requirements and constraints.

In-depth patterns are not only interesting to data miners, but also to business decision-makers. In the afore-mentioned trading strategies, more actionable trading strategies can be found via model refinement or parameter tuning. We also call them *optimized strategies/rules*.

DEFINITION 3: Human-Machine Cooperation – The in-depth pattern discovery is conducted under the cooperation of business analysts and data analysts. Section 2.5 fills the human-machine cooperation concept into the data mining context.

DEFINITION 4: Domain-Driven Data Mining – In-depth pattern discovery is not only a data-driven trial-and-error process, rather is highly domain-dependent. It gets involved in domain expertise and constraints in a human-machine cooperation context. Domain-driven in-depth pattern discovery consists of the main ideas which will be discussed in the remaining of this section.

2.2 DDID-PD process model

The components of the DDID-PD framework are shown in Figure 1. The lifecycle of DDID-PD is as follows (the sequence is not rigid, some phase may be bypassed or moved back and forth in a real domain problem).

P1. Problem understanding and definition;
P2. Data understanding;
P3. Data preprocessing;
P4. Modeling;
P5. Results evaluation;
P6. Based on feedbacks and progress of the phases from P2 to P5, it is quite possible that each phase may be iteratively reviewed starting from P1 via the interaction with domain experts in a back-and-forth manner for the refinement of mining results;
P7. Results post-processing; or

P6': In-depth modeling on the mined results where applicable; then going to P7;
P8. Going back and reviewing phases from P1 on may be required;
P9. Deployment;
P10. Knowledge and report delivery.

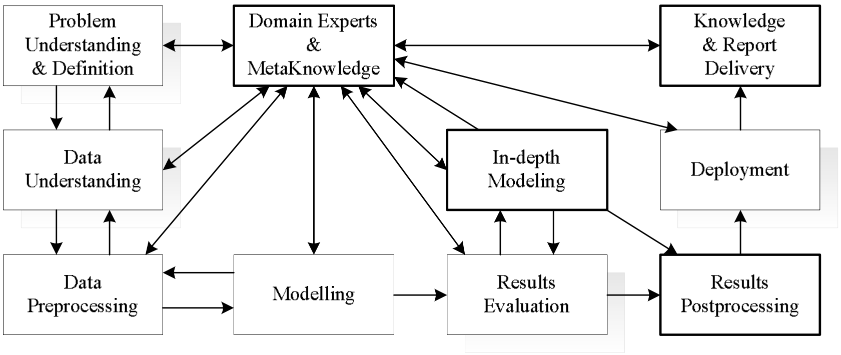

Fig. 1. DDID-PD process model

The DDID-PD process highlights four highly correlated ideas that are critical for the success of a data mining process in the real world. They are (i) *constraint-based context*, multiple types of constraints widely exist in the domain problem and its analysis objectives, (ii) *in-depth pattern mining*, another round of modeling on the first-round results may be necessary for mining patterns really interesting and actionable to business, (iii) *human-machine-cooperated interactive knowledge discovery*, the involvement of domain experts and their knowledge and the interaction between experts and mining system in the whole process are important for effective execution of the mining, (iv) *a loop-closed iterative refinement process*, patterns that can be deployed and adopted for smart business decision-making are the outcome of iterative refinement. The following sub-sections outline them individually.

2.3 Constraint-based context

In human society, every one is constrained by either social (environmental) regulations or personal situations. Similarly, advanced knowledge discovery and smart decision-making need to consider real-world aspects such as environmental reality, expectations and constraints in the whole process. More specifically, the following four kinds of constraints play important roles in building an effective and efficient data mining from requirements engineering to evaluation and refinement engineering. They consist of domain-specific, functional and environmental constraints, and form a constraint-based data mining context [2].

Data Constraints: this is related to data quantity, data structures, data distribution, data semantic complexity, etc.

Domain Constraints: it involves domain type, characteristics (eg privacy), business process and workflow, domain knowledge, human capability and role, qualitative and quantitative hypothesis and conditions, etc.

Interestingness Constraints: this is driven either by academic objectives or business goals, problem requirements and analytical goals, etc.

Rule Constraints: it gets involved in rule representation, rule interestingness to analytical goals, rule explanation, rule deployment in the integration with real-world business process and environment, etc.

All the above constraints must be to varying degree considered [2] in real-world data mining. They may get involved in the whole process of domain-specific data mining. In the development of data mining process and algorithms, we need to think of what they will bring to the improvement of traditional data mining, and what techniques or system supports can be used for utilizing, analyzing and avoiding these constraints. They must be closely connected to specific modeling methods, business environment, and analytical objectives in a systematic manner.

2.4 Mining in-depth patterns

Existing data mining methods, for instance association rule mining, often generate a huge number of patterns (or rules), but a majority of them either are redundant or do not reflect the true interestingness from business perspective. This has hindered the deployment and adoption of data mining in the real applications. Taking trading rules in finance as an instance, a trading rule, for instance Filter Rule, usually implies millions of individual rules. However, most of them are not actionable for a specific business environment. Therefore it is essential to further refine these rules so that more interesting and actionable rules can be discovered and recommended for more smart and effective decision-making. To overcome this obstacle in deploying data mining into the real world, we need to discover more interesting and actionable rules based on a domain-specific problem and its business requirements. This leads to in-depth mining.

In-depth mining refers to a further mining either on existing (mined) patterns/rules or in selected/refined datasets. Obviously, the involvement of domain knowledge and constraints are often necessary for conducting in-depth mining. More importantly, some appropriate in-depth mining techniques should be developed on the demand of a domain-specific problem. For instance, in Section 3, we illustrate some of our work in mining in-depth correlated patterns in real stock environment.

2.5 Human-machine-cooperated interactive knowledge discovery

Real-world data mining should be a human-machine-cooperated interactive knowledge discovery process rather than an autonomous system. Domain experts consist of the centre or an essential constituent of the data mining process via dynamic expert-model interaction. In fact, they and their knowledge play significant role in the whole data mining process such as business and data understanding, features selection, hypotheses proposal, model selection and learning, and evaluation and refinement of algorithms and resulting outcomes. For instance, domain experts can narrow down the selection of features and models, and create high quality hypotheses and efficient constraints based on their domain knowledge (especially their experience and imaginary thinking), which will effectively accelerate the mining process.

Instead of producing patterns or knowledge directly from data, the domain-driven data mining methodology allow domain experts and/or their knowledge to be the front or center of the mining process, and to interact with data and business via friendly interfaces and system supports to maximize the power of domain experts' knowledge and capability in complex problem solving. Domain experts and knowledge involvement is an essential constituent of the data integration, feature selection, hypotheses proposal, business modeling and learning, and evaluation, refinement and interpretation of algorithms and resulting outcomes. For instance, domain experts can incorporate their knowledge (especially their qualitative experience and imaginary thinking) into data and feature selection, model analysis and building via generating effective qualitative hypothesis and constraints on business data and problems. This point may also be called as *human-centered* [1, 7], *human-involved*, *supervised* or *guided* [6] data mining.

As discussed in the above, domain-driven in-depth data mining supports in-depth analysis with the assistance of domain knowledge. Furthermore, the mining is actually an interaction between domain-experts and mining system. To support the dynamic interaction, user-friendly human-machine interfaces are necessary. The interface needs to support domain expert-mining system dialogue, so that domain knowledge from domain experts can be online and instantly embedded into the mining system and knowledge base on demand, and refine tune the quality of final mined rules. This actually makes a data mining process and tool as highly interactive and dynamic rather than as fully automated as previously imagined. For this commitment, the knowledge base including expert systems, AI, PR and cognitive science needs to be involved. A good option is to build intelligent agents-based data mining platform [8, 9] to support user modeling, user interaction, and the like. This is also called *interactive mining* [10, 1].

2.6 Loop-closed iterative refinement

The data mining process and its system are closed rather than open, since it encloses iterative refinement and feedbacks of hypotheses, features, models, evaluation and explanations in a human-involved context. The real-world mining process is iterative because the evaluation and refinement of features, models and outcomes cannot be completed once rather is based on iterative feedbacks and interaction during the whole process.

The data mining process and its system are closed with iterative refinement and feedbacks of hypotheses, features, models, evaluation and explanations in the human-involved or -centered context. It iteratively evaluates and tunes features and models based on feedbacks from and the involvement of domain experts and their knowledge, and the interaction with the domain problem.

Specific data mining process needs to be designed for a particular problem. In the process, we may consider how to involve domain experts, their knowledge, feedbacks, fine-tuning work, evaluation and modification in an iterative and incremental manner.

To support the loop-closed iterative refinement, some appropriate human-computer interaction interface should be designed. Again, to this end intelligent agents [8] can play competitive role.

3 Mining in-depth correlations in real stock markets

Data mining in stock data is popular [11] but challenging in the real world because real stock markets are greatly complex. Taking the ASX as an instance, there are more than 1000 shares in this small market. The number of daily trades is up to 70,000. Data mining in this real data is highly expected by traders and is taken as a research priority for smart information use.

In the Data Mining Program (DMP) of Australian Capital Markets Cooperative Research Center (CMCRC) [12], we deploy the DDID-PD framework for mining in-depth correlations in stock data. Its main function consists of (i) high dimension reduction to generate a small quantity of data or rule representatives from huge data set or rule combinations, (ii) human-machine-cooperated interactive refinement to specify/refine correlation coefficients based on domain-specific knowledge and objectives, and (iii) in-depth pattern discovery to obtain the interesting correlations.

The correlation pattern mining in stock order stream targets patterns interesting and actionable to stock traders. In-depth correlation mining in stock market aims to finding correlations between stocks, searching correlated patterns from existing trading rules developed by financial experts in order to develop more actionable trading rules, and discovering correlated relations between trading rules and stocks.

In the following, we will present some results in utilizing the DDID-PD framework to mine correlated stocks, in-depth trading rules, and trading rule-stock correlations based on ASX stock data.

3.1 Mining correlated stocks

In real stock markets, for instance ASX, hundreds of stocks are traded by brokers and retailers every trading day. It is a common hypothesis that there may be some forms of correlations existing among stocks from the same or similar sectors, or belonging to a shared production chain. We have developed a set of correlation metrics for analyzing the correlations between stocks in a real market. The following lists basic idea of the correlated stock mining algorithm.

```
ALGORITHM: Mining Correlated Stocks
C1. Calculating the coefficient $\rho$ of two stock prices;
C2. Determining the scope of $\rho$ interesting to real trading through cooperation between miners and traders via considering other domain-specific aspects;
C3. Evaluating the correlation between stocks via some additional domain-specific elements;
C4. Recommending correlated stocks.
```

In order to testify the effectiveness of mined correlated stock pairs, we develop a Pairs Trading strategy and use it to trade in the historical market orderbook. Pairs trading involves the purchase of one security while simultaneously selling (or selling short) another security when a pair of highly correlated securities deviates from the normal relationship between them. Taking the ASX as instance, we targeted 32 stocks with quality data from Jan 1997 to Jun 2002. 13 stocks of them are found to be highly correlated. In all 78 pairs of combinations, 9 pairs are found to be actionable to real trading. For instance, it is found that there is a high correlation in the pair of CBA and GMF. Without considering the market impact, the return of the pair CBA-GMF is 40.51% in average on historical data from 1 Jan 1997 to 19 Jun 2002.

In this exercise, we found the following interesting points:

(1) Our real analytics also show that the correlated stocks actionable to traders cannot just be specified by the coefficient.

(2) Interestingly, the correlated stocks we mined in ASX market all come from different sectors, this means that correlated stocks are not necessary from the same industry as assumed by financial researchers.

(3) Our analytical results show that the profit of trading a correlated pair is greatly affected by the liquidity and the volatility of stocks. Therefore, an actionable (profitable) stock pair is based on correlation, liquidity and volatility of the parties.

3.2 Mining in-depth trading rules

In stock markets, since a long time ago, financial researchers have developed many trading rules to support traders' decision-making. These trading rules actually indicate possible patterns hidden in stock markets. For instance, the trading strategy MA actually indicates a correlated pattern between two features namely *short-run moving average* (*sr*) and *long-run moving average* (*lr*). The pattern MA (*sr*, *lr*, δ) is defined as follows (where δ is the fix band for the difference between *sr* and *lr*):

IF *sr* *(1-δ) >= *lr* THEN *Buy*
IF *sr* *(1+δ) <= *lr* THEN *Sell*

This pattern actually consists of a large number of rules (we call them *generic rules*) from finance perspective, for instance MA(2, 50) and MA(10, 50) represent two different MA rules. However, traders do not know which rule is actionable for assisting in their specific trading decision.

The in-depth pattern mining on existing trading rules aims to mining more actionable rules (called *optimized* or *in-depth rules*) which can better serve traders' objectives. To discover optimized rules from generic rules, Robust Genetic Algorithm [13] and a human-machine interaction interface are developed so that financial experts can dynamically and iteratively supervise and evaluate the training. Figure 1 (a) in [14] illustrates such an interface, where clients can supervise the construction of some features to narrow down the search space. Taking the MA as an instance, an in-depth rule MA(4, 19, 0.033) is found in the training data from 1 Jan 2000 to 31 Dec 2000 and testing set between 1 Jan 2001 and 31 Dec 2001. The number of trading signals generated by this rule is much more than other generic rules. Its Sharpe Ratio[2], as shown in Figure 1 (b) in [14] is greatly improved to positive scope compared with generic results. This demonstrates that the DDID-PD with the involvement of domain knowledge can lead to more interesting and actionable rules for trading support.

[2] It is taken as a benchmark for judging the performance of a trading rule.

3.3 Mining in-depth rule-stock correlations

It is assumed that some trading rules are suitable for a class of stocks, while others are more effective to guide the trading of other stocks in the market. This hypothesis actually indicates whether there are correlations between trading rules and stocks. If yes, and if we can discover the correlation, then it would be very helpful for guiding the real trading.

Based on this hypothesis, we develop algorithms to search the in-depth correlations between trading rules and stocks in real stock data. The basic ideas of the rule-stock correlation mining algorithms are as follows.

1) Mining in-depth rules for individual stock

For each ASX security, a set of in-depth rules are discovered for each class of trading rules by the algorithm described in Section 3.1. Furthermore, in-depth rules can be discovered from all classes of rules for all stocks respectively. As a result, a rule-stock set is found in which a trading rule is matched with one or multiple stocks.

2) Mining the highly correlated rule-stock pairs

In the above step, multiple in-depth rules from different rule classes may be found suitable for one stock. It is necessary to discover a highly correlated rule for a specific stock from the above resulting set. This leads to the most suitable rule for a stock, and forms a correlated rule-stock pair.

3) Refining and evaluating the rule-stock pairs

In order to find the interesting and actionable rule-stock pair, the assistance of domain experts and their suggestions are essential for the refinement and evaluation of pairs found in the above steps.

We analyzed the rule-stock correlations in ASX markets. Three classes of trading rules, they are MA, Filter Rule and Channel Breakout [6], and 26 ASX stocks have been chosen for the experiments. The data for training is from 1 Jan 2001 to 31 Jan 2001, and testing set is from 1 Feb 2001 to 28 Feb 2001. Five different investment plans are conducted on these rules and stocks. In order to organize the pairs, we rank them based on return, and generate 5% pair, 10% pair, and so forth from the whole pair population. The 5% pair means that the pairs are the top 5% based on the return.

The Figure 2 and 3 in [14] illustrates returns we have found for different investment plans on different pairs. These graphs are interesting to traders for them to make smart trading decision using these mined rule-stock pairs.

4 Conclusions and future work

In the real world, correlated patterns interesting to business are often hidden in domain-specific data and constraint-based context. This often leads to the scenario as too many rules are mined while few of them are truly interesting to business when using extant correlation mining techniques. Therefore, in-depth pattern discovery should be conducted on the domain-specific constraint-based context. To this end, we have developed the domain-driven in-depth pattern discovery (DDID-PD) framework to guide the real-world data mining. The DDID-PD framework has been outlined in this paper, which provides methodology for dealing with constraint-based context, mining in-depth patterns, supporting interactive mining in a loop-closed iterative refinement process.

The main phases and components of the DDID-PD framework (as shown in Figure 1) include almost all phases of the well-known industrial data mining methodology CRISP-DM. While there are three big differences from the CRISP-DM: (i) some new essential components highlighted by thick rims, such as *results postprocessing* and *in-depth modeling*, are taken into account in designing the lifecycle of the DDID-PD process, (ii) in the DDID-PD, the phases of CRISP-DM highlighted by shadow are enhanced via dynamic interaction with domain experts and the consideration of constraints and domain knowledge, (iii) the lifecycle of the DDID-PD is actually different from that of CRISP-DM. These differences are key to mine in-depth patterns in the real world.

Deploying the DDID-PD, we have analyzed in-depth correlations in stock markets. The experiments have shown that the mined correlations guided by the DDID-PD framework are interesting and actionable to real trading. The DDID-PD is potential in mining interesting real-world patterns in an effective and efficient manner.

Our further work has been on developing detailed process supports and interface design for DDID-PD for the real-world data mining.

References

[1] Panel members. The perfect data mining tool: Automated or interactive?, in Panel at ACM SIGKDD02', Edmonton, Canada, 2002.

[2] Ng, R., Lakshmanan, L., Han, J. & Pang, A. Exploratory mining and pruning optimizations of constrained association rules, in 'Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems', ACM Press, Seattle, Washington, pp. 13–24, 1998.

[3] L. Cao, R. Dai. "Human-Computer Cooperated Intelligent Information System Based on Multi-Agents", ACTA AUTOMATICA SINICA, 29(1):86-94, 2003, China (in English)

[4] L.B. Cao, et al. Ontology-Based Integration of Business Intelligence. Int. J. on Web Intelligence and Agent Systems, to appear.

[5] http://www.crisp-dm.org

[6] Ryan, S., Allan, T., Halbert, W., Data-snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Financial*, 1999. 54, (5):1647-1692.

[7] J. Han. *Towards Human-Centered, Constraint-Based, Multi-Dimensional Data Mining.* An invited talk at Univ. Minnesota, Minneapolis, Minnesota, Nov. 1999.

[8] C. Zhang, Z. Zhang, L. Cao. Agents and Data Mining: Mutual Enhancement by Integration, LNCS 3505, 2005

[9] F-Trade. http://datamining.it.uts.edu.au/f-trade

[10] Ankerst, M. (2001), Human involvement and interactivity of the next generation's data mining tools, in 'ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery', Santa Barbara, CA.

[11] B. Kovalerchuk and E. Vityaev. Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer Acad. Publ, 2000

[12] www.cmcrc.com

[13] L. Lin, et al. Genetic algorithms for robust optimization in financial applications. Proceedings of 4th IASTED conf. on Computational Intelligence Canada 2005.

[14] L. Lin, D. Luo, L. Liu. Mining Domain-Driven Correlations in Stock Markets. 18th Australian Joint Conf. on AI05.