Actionable Knowledge discovery using MSCAM

V.Vijay[1], M.Satyanarayana [2]
[1]*M.Tech Final Year Student Department of C.S.E .,* Swarnandhra Collage of Engg&tech
[2]Asst professor CSE Dept, Swarnandhra Collage of Engg&tech

ABSTRACT

Actionable Association Rule mining (AAR) is a closed optimization problem solving process from problem definition, model design to actionable pattern discovery, and is designed to deliver apt business rules that can be integrated with business processes and technical aspects. To support such processes, we correspondingly propose, formalize and illustrate a generic AAR model design: Multisource Combined-Mining-based AAR (MSCM-AAR). In this paper, we present a view of actionable association rule (AAR) from the technical and decision-making perspectives. A real-life case study of MSCM-based AAR is demonstrated to extract debt prevention patterns from social security data. Substantial experiments show that the proposed model design are sufficiently general, flexible and practical to tackle many complex problems and applications by extracting actionable deliverables for instant decision making.

KEYWORDS: Data mining, domain-driven data mining (D^3M), actionable association rule, business decision making.

I. INTRODUCTION

In the last decade, data mining, or KDD (knowledge discovery in database) has become an active research and development area in information technology fields. In particular, data mining is gaining rapid development in various aspects such as the data mined, the knowledge discovered, the techniques developed, and the applications involved.

A typical feature of traditional data mining is that KDD is presumed as an automated process. It targets the production of automatic algorithms and tools. As a result, algorithms and tools developed have no capability to adapt to external environment constraints. Millions of patterns and algorithms are published in academia but unfortunately very few of them have been transferred into real business.

Generally, association rules are used in conjunction with transaction (or basket) type data, but their use is not limited to this domain. When dealing with customer demographic data, for example, the database schema defines a fixed set of attributes for each record, and each customer is represented by one record. Each record contains a value for each attribute, i.e., (attribute=value). By replacing the sets of items in the traditional definition of association rules with conjunctions of (attribute=value) equalities, we can generalize the above definition of association rules to include this type of non-transactional data. For example, (age=40) □ (salary=$50,000) =□ (own. home=yes).

In general, data mining (or KDD) algorithms and tools only focus on the discovery of patterns satisfying expected technical significance. The identified patterns are then handed over to business people for further employment. Surveys of data mining for business applications following the above paradigm in various domains [5] have shown that business people cannot effectively take over and interpret the identified patterns for business use. This may result from several aspects of challenges besides the dynamic environment enclosing constraints [3]. 1) There are often many patterns mined but they are not informative and transparent to business people who do not know which are truly interesting and operable for their businesses. 2) A large proportion of the identified patterns may be either commonsense or of no particular interest to business needs. Business people feel confused by why and how they should care about those findings. 3) Further, business people often do not know, and are also not informed, how to interpret them and what straight-forward actions can be taken on them to support business decision-making and operation.

The above issues inform us that there is a large gap [22], [8], [7], [6] between academic deliverables and business expectations, as well as between data miners and business analysts. Therefore, it is critical to develop effective methodologies and techniques to narrow down and bridge the gap. Clearly, there is a need to develop general, effective, and practical methodologies for actionable knowledge discovery (AAR).

One essential way is to develop effective approaches for discovering patterns that not only are of technical significance [24], but also satisfy business expectations [7], and further indicate the possible actions that can be explicitly taken by business people [1], [4]. Therefore, we need to discover actionable knowledge that is much more than simply satisfying predefined technical interestingness thresholds. Such actionable knowledge is expected to be delivered in operable forms for transparent business interpretation and action taking.

It has been increasingly recognized that traditional data mining is facing crucial problems in satisfying user preferences and business needs. For example, research work has been reported on developing actionable interestingness [7], [1] and subjective interestingness such as profit mining [37] to extract more interesting patterns, and on enhancing the interpretation of

findings through explanation [27]. However, the nature of the existing work on actionable interestingness development is mainly technical-significance-oriented, e.g., by developing alternative and subjective metrics. The critical problem to a great extent comes from the oversimplification of complex domain factors surrounding business problems, the universal focus on algorithm innovation and improvement, and the little attention taken of enhancing KDD system infrastructure to tackle organizational and social complexities in real-world applications.

Fundamental work on AAR is therefore necessary to cater for critical elements in real-world applications such as environment, expert knowledge, and operability. This is related to, but much beyond, algorithm innovation and performance improvement. To this end, AAR must cater for domain knowledge [28] and environmental factors, balance technical significance and business expectations from both objective and subjective perspectives [7], and support automatically converting patterns into deliverables in business- friendly and operable forms such as actions or rules. It is expected that the AAR deliverables will be business friendly enough for business people to interpret, validate, and action, and that they can be seamlessly embedded into business processes and systems. If that is the case, data mining has good potential to lead to productivity gain, smarter operation, and decision making in business intelligence. Such efforts actually aim at the KDD paradigm shift from traditionally technical interestingness-oriented and data centered hidden pattern mining toward business-use-oriented and domain-driven actionable knowledge discovery [6].

Relevant preliminary work on AAR mainly addresses specific algorithms and tools for the filtration, summarization, and post processing [29]of learned rules. There is a need to develop general AAR frameworks that can cater for critical elements in the real world and can also be instantiated into various approaches for different domain problems. To the best of our knowledge, very limited research work has been reported in this regard.

This paper features the definition and development of several general AAR frameworks from the system viewpoint, which follow the methodology of Domain-Driven Data Mining (DDDM, or D3M for short) [6], [7], [8], [4], [5]. Our focus is on introducing their concepts, principles, and processes that are new, effective to AAR, flexible, and practical. Such frameworks are necessary and useful for implementing real-world data mining processes and systems, but are often ignored in the current KDD research.

The main contributions of this work are:

1. Stating the AAR problem from system and micro economy perspectives to define fundamental concepts of actionability and actionable patterns,
2. Defining knowledge actionability by highlighting both technical significance and business expectations that need to be considered, balanced, and/or aggregated in AAR,
3. Demonstrating the effectiveness and flexibility of the proposed framework in tackling real-life AAR.

Table 1

Notations	Explanations
AAR	Actionable Association Rule
MSCM- AAR	Multi-source + combined mining-based AAR
P	$P = \{p_1, \ldots, p_u\}$ is a pattern set
$\widetilde{P}$	$\widetilde{P} = \{\tilde{p_1}, \tilde{p_2}, \cdots\}$ is an actionable pattern set
$\widetilde{R}$	$\widetilde{R} = \{\tilde{r_1}, \tilde{r_2}, \cdots\}$ is a business rule set

MSCM-AAR: Handles AAR in either multiple data sources or large quantities of data. One of the data sets is selected for mining initial patterns. Some learned patterns are then selected to guide feature construction and pattern mining on the next data set(s). The iterative mining stops when all data sets are mined, and the corresponding patterns are then merged/summarized into actionable deliverables.

II. RELATED WORK

Actionable knowledge discovery is critical in promoting and releasing the productivity of data mining and knowledge discovery for smart business operations and decision making. Both SIGKDD and ICDM panelists pointed it out as one of the great challenges in developing the next generation KDD methodologies and systems [2], [10]. In recent years, some relevant work has been emerging.

The term "actionability" measures the ability of a pattern to suggest a user to take some concrete actions to his/her advantage in the real world. It mainly measures the ability to suggest business decision-making actions. Existing efforts in the development of effective interestingness metrics are basically on developing and refining objective technical interestingness metrics ($t_o()$) [21], [25]. They aim to capture the complexities of pattern structure and statistical significance. Other work appreciates subjective technical measures ($t_s()$) [29], [21], [34], which also recognize to what extent a pattern is of interest to particular user preferences. For example, probability- based belief is used to describe user confidence of unexpected rules [21]. There is very limited research on developing business-oriented interestingness, for instance, profit mining [37].

The main limitations for the existing work on interestingness development lie in a number of aspects. Most work is on developing alternative interest measures focusing on technical interestingness only [20]. Emerging research on general business-oriented interestingness is isolated from technical significance. A question to be asked is "what makes interesting patterns actionable in the real world?" For that, knowledge actionability needs to pay equal attention to both technical and business-oriented interestingness from both objective and subjective perspectives [7].

With regard to AAR approach, the existing work mainly focuses on developing post analysis techniques to filter/prune rules [27], reduce redundancy [26], and summarize learned rules [27], as well as on matching against expected patterns by similarity/difference [28]. In post analysis, a recent highlight is to extract actions from learned rules [26]. A typical effort on learning action rules

is to split attributes into “hard/soft” [26]or “stable/flexible” [25]to extract actions that may improve the loyalty or profitability of patients. Other work is on action hierarchy [1]. Some other approaches include a combination of two or more methods, for instance, class association rules (or associative classifier) that build classifiers on association rules (A ! C) [23]. In [23], external databases are input into characterizing the item sets. In [22], clustering is used to reduce the number of learned association rules. Some other work is on the transformation from data mining to knowledge discovery [11], and developing a general KDD framework to fit more factors into the KDD process [27].

III. ACTIONABLE ASSOCIATION RULE DISCOVERY

Discovery of actionable patterns

We use action trees for the discovery of actionable patterns using the following steps.

Building an action tree. First, an action tree must be built (and maintained later on) for a given application. This can be done using techniques described in the previous section.

Assigning data mining queries. Second, data mining queries defining actionable patterns for the specific actions should be assigned to the corresponding nodes of the tree. For example, a possible data mining query assigned to the node *"Based on customer demographics"* of the tree in Figure 1 could be the query (2). Additional examples of data mining queries expressed in pattern template language similar to (Klemettinen *et al.* 1994) are: Query "Find what kinds of product categories sell well on different days of week" (assigned to the action *"Based on season"): DayOfWeek * Category*+ (0.4, 0.01) **(3)** Query "Find 'cross-selling' categories, that is, find categories of products that are selling together" (assigned to the action *"Determining how to AARang products in the store"):*

Executing data mining queries. Given an attributed action tree, the pattern discovery process consists of the traversal of the whole action tree (say, using depth-first search) and execution of all the data mining queries. The discovered actionable patterns are written to the files associated with data mining queries.

Discovery optimization

The method for discovering actionable patterns described in the previous section does not give an answer to the question: *when* or *how often* to reexecute data mining queries that are assigned to the nodes of an action tree to obtain up-to-date patterns. The straightforward approach, which would be to reexecute all data mining queries whenever data changes in the database, is too computationally expensive in general. This is especially true for big applications with large action trees and many data mining queries. In the remainder of this section we present two optimization techniques and explain when they can be used in practice.

Partial tree traversal. The natural optimization of the action tree traversal technique is a *partial* traversal of an action tree. In this case, only the nodes of the tree selected bv the user are traversed and only those data mining "queries that are assigned to these nodes are executed. Nodes can be selected as individual nodes or as belonging to the user specified subtree. The partial tree traversal approach can be used for applications in which there is no need to keep patterns up-to-date all the time. Therefore, data mining queries can be executed "on demand". That is, whenever there is a need to consider some specific action, only then data mining queries assigned to that action must be reexecuted to supply the user with the latest patterns to help make decisions.

MSCM-AAR:

Enterprise applications often involve multiple-subsystems based and heterogeneous data sources that cannot be integrated, or are too costly to do so. Another common situation is that the data volume is so large that it is too costly to scan the whole data set. Mining such complex and large volumes of data challenges existing data mining approaches. To this end, we propose a Multisource + combined-mining-based AAR framework. Fig. 1 shows the idea of MSCM-AAR.

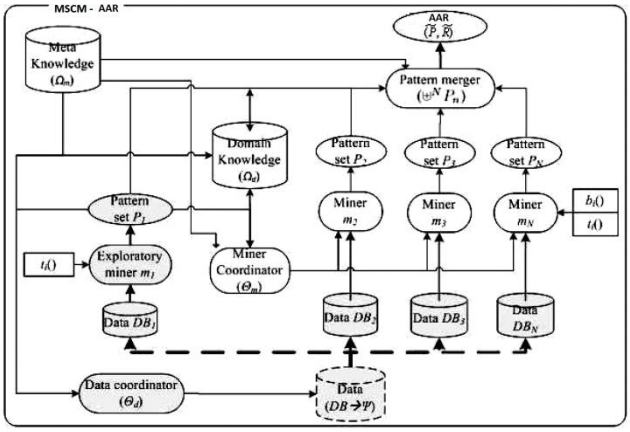

Fig. 1. Multisource + combined-mining-based AAR.

MSCM-AAR discovers actionable knowledge either in multiple data sets or data subsets (DB_1, . . ., DBn) through partition. First, based on domain knowledge, business understanding, and goal definition, one of the data sets or certain partial data (say DBn) is selected for mining exploration (m_1). Second, the exploration results are used to guide either data partition or data set management through a data coordinator agent Θ_{db} (coordinating data partition and/or data set/feature selection in terms of iterative mining processes, see more from AMII-SIG[1] regarding agents in data mining), and to design strategies for managing and conducting parallel pattern mining on each data set or subset and/or combined mining [9] on relevant remaining data sets. The deployment of method m_n, which could be either in parallel or combined, is determined by data/business understanding and objectives. Third, after the mining of all data sets, patterns Pn identified from individual data sets are merged $(\uplus^N P)$ and extracted into final deliverables $(\widetilde{P}, \widetilde{R})$.

MSCM-AAR can be expressed as follows:

$$MSCM - AKD:$$
$$\underbrace{DB_n[DB \xrightarrow{\otimes} DB_n]^{e,t_{i,n}()[u_{i,n}()],m_n,\Omega_m}\{P_n\}}_{N}$$
$$^{e,b_{i,n}(),\uplus^N P_n,\Omega_d,\Omega_m}\longrightarrow \widetilde{P}, \widetilde{R},$$

Where $t_{i,n}$ and bi,n are technical and business interestingness of model m_n on data set/subset n, and $[i_{i,n}()]$ indicates the alternative checking of unified interestingness as in UI-AAR, $\uplus^N P_n$ is the merger function, $\otimes$ indicates the data partition if the source data needs to be split.

Algorithm

Multi-Source + Combined Mining Based AAR (MSCM-AAR)
INPUT: target data sets DB, business problem ψ, and thresholds ($t_{o,0}$, $t_{s,0}$, $b_{o,0}$ and $b_{s,0}$)
OUTPUT: actionable patterns $\widetilde{P}$ and business rules $\widetilde{R}$
Step 1: Identify or partition whole source data into N data sets DBn (n = 1,…, N);
Step 2: Data Set-n mining: Extracting general patterns P_n on data set/subset DB_n;
FOR l = n to (N)
Develop modeling method m_n with technical interestingness $t_{i,n}()$ (i.e., $t_o()$, $t_b()$) or unified $i_{i,n}()$
Employ method m_n on the environment e and data DB_n engaging meta-knowledge Ω_m;
Extract the general pattern set P_n;
ENDFOR
Step 3: Pattern merger: Extracting actionable patterns $\widetilde{P}$;
FOR l = n to N
Design the pattern merger functions $\uplus^N P_n$ to merge all patterns into $\widetilde{P}$ by involving domain and meta knowledge Ω_d and Ω_m, and business interestingness $b_i()$;
Employ the method $\uplus P_n$ on the pattern set P_n;
Extract the actionable pattern set $\widetilde{P}$;
ENDFOR
Step 4: Converting patterns $\widetilde{P}$ to business rules $\widetilde{R}$.

The MSCM-AAR framework can also be instantiated into a number of mutations. For instance, for a large volume of data, MSCD-AAR can be instantiated into data partition + unsupervised + supervised-based AAR by integrating data partition into combined mining. An example is as follows: First, the whole data set is partitioned into several data subsets based on the data/business understanding and domain knowledge jointly by data miners and domain experts, say data sets 1 and 2. Second, an unsupervised learning method is used to mine one of the preference data sets, say data set 1. Some of the mined results are then used to design new variables for processing the other data set. Supervised learning is further conducted on data set 2 to generate actionable patterns by checking both technical and business interestingness. Finally, the individual patterns mined from both data subsets are combined into deliverables.

IV. RESULTS & DISCUSSION

We test the MSCM-AAR method in randomly generated medical data. The cleaned sample data contains 55,800 patients with their demographic attributes. There are 711 traditional associations mined. Combined associations cannot be discovered by traditional association rule techniques.

Compared with the single associations from respective data sets, the combined associations and combined association clusters are much more workable than single rules presented in the traditional way. They contain much richer information from multiple aspects rather than from a single one, or a collection of separated single rules. For instance, the following combined association shows that patients aged 65 or more, whose arrangement method is of “smoking” plus “regular,” then they have more chances of getting cancer i.e can be classified into class “C” (high risk of life). Obviously, this pattern combines heterogeneous information regarding the specific group of the patient’s demographic method

$$\{x = age : 65+, y = \text{smoking \& repeated} + \text{cancer} \rightarrow c = C\}.$$

Finally, combined patterns can be transformed into operable business rules that may indicate direct actions for business decision making. For instance, for the above combined association, it actually connects key business elements with segmented patient characteristics, and we can generate the following business rule by extending the Business Rule specification:

DELIVERING BUSINESS RULES:
Patient Demographic- combination business rules
For All patients i (i 2 I is the number of valid patients)
Condition:
satisfies S/he is a patient aged 65 or plus;
relates
S/he is under arrangement of “smoking” and “regularly”, and
S/he is also having “cancer”;
Operation:
Alert = “S/he has ‘High’ risk of life in short timeframe.”
Action = “Try to avoid the smoking habit or take medical advises.’”
End-All

The converted business rules are deliverables presented to business people. They are convenient and it is easy for clients to embed them into their routine business processes and operational systems for filtering patients and monitoring the cancer patients. Our clients feel more comfortable in understanding, interpreting, and actioning these business rules than those patterns directly mined in the data. Therefore, combined patterns are more business friendly and indicate much more straightforward decision-making actions to be taken by business analysts in the business world, while this cannot be achieved by patterns identified by traditional methods.

In addition, the use of combined mining leads to combined patterns consisting of attributes from different business units or by partitioning into organized segments. Through attribute segmentation or merger, it is

manageable to differentiate attribute impact on business objectives, and thus, extract more and more informative patterns and more operable decision-making actions.

IV. CONCLUSION

This paper has formally defined the AAR concepts, processes, actionability of patterns, and operable deliverables. With such components, we have proposed MSCM AAR framework capable of handling various business problems and applications. These framework support closed-optimization-based problem solving from a business problem/environment definition, to actionable pattern discovery, and to operable business rule conversion. Deliverables extracted in this way are not only of technical significance but also are capable of smoothly integrating into business processes.

Substantial experiments in significant data mining applications such as financial data mining and mining social security data have shown that the proposed framework have the potential to handle the limitations in existing methodologies and approaches. This framework is sufficiently general, flexible, and workable to be instantiated into various approaches for tackling complex data and business applications.

Following the D3M theory, there are many issues to be studied, for instance, defining operable business rules by involving ontological techniques for representing both syntactic and semantic components.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD '97), pp. 111-114, 1997.

[2] 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.

[3] J.F. Boulicaut and B. Jeudy, "Constraint-Based Data Mining," The Data Mining and Knowledge Discovery Handbook, pp. 399-416, Springer, 2005.

[4] L. Cao, "Domain-Driven Actionable Knowledge Discovery," IEEE Intelligent Systems, vol. 22, no. 4, pp. 78-89, July/Aug. 2007.

[5] L. Cao, P. Yu, C. Zhang, and H. Zhang, Data Mining for Business Applications. Springer, 2008.

[6] L. Cao, P. Yu, C. Zhang, and Y. Zhao, Domain Driven Data Mining Springer, 2009.

[7] L. Cao and C. Zhang, "Knowledge Actionability: Satisfying Technical and Business Interestingness," Int'l J. Business Intelligence and Data Mining, vol. 2, no. 4, pp. 496-514, 2007.

[8] L. Cao and C. Zhang, "The Evolution of KDD: Towards Domain-Driven Data Mining," Int'l J. Pattern Recognition and Artificial Intelligence, vol. 21, no. 4, pp. 677-692, 2007.

[9] L. Cao, H. Zhang, Y. Zhao, and C. Zhang, "Combined Mining: Discovering More Informative Knowledge in e-Government Services," technical report, Univ. of Technology Sydney, 2008.

[10] U. Fayyad, G. Shapiro, and R. Uthurusamy, "Summary from the KDD-03 Panel—Data mining: The Next 10 Years," ACM SIGKDD Explorations Newsletter, vol. 5, no. 2, pp. 191-196, 2003.

[11] 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, AAAI Press/MIT Press, 1996.

[12]A. Freitas, "On Objective Measures of Rule Surprisingness," Proc. Second European Symp. Principles of Data Mining and Knowledge Discovery (PKDD '98), pp. 1-9, 1998.

[13]O.G. Ali and W. Wallace, "Bridging the Gap between Business Objectives and Parameters of Data Mining Algorithms," Decision Support Systems, vol. 21, pp. 3-15, 1997.

[14]H. Kargupta, B. Park, D. Hershbereger, and E. Johnson, "Collective Data Mining: A New Perspective toward Distributed Data Mining," Advances in Distributed Data Mining, H. Kargupta and P. Chan, eds., AAAI/MIT Press, 1999.

[15]R. Hilderman and H. Hamilton, "Applying Objective Interestingness Measures in Data Mining Systems," Proc. Symp. Principles of Data Mining and Knowledge Discovery (PKDD), pp. 432-439, 2000.

[16]B. Lent, A.N. Swami, and J. Widom, "Clustering Association Rules," Proc. 13th Int'l Conf. Data Eng., pp. 220-231, 1997.

[17]B. Liu, W. Hsu, and Y. Ma, "Pruning and Summarizing the Discovered Associations," Proc. ACM SIGKDD, 1999.

[18]B. Liu and W. Hsu, "Post-Analysis of Learned Rules," Proc. Nat'l Conf. Artificial Intelligence/Innovative Applications of Artificial Intelligence Conf. (AAAI/IAAI), 1996.

[19]B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing Subjective Interestingness of Association Rules," IEEE Intelligent Systems, vol. 15, no. 5, pp. 47-55, Sept./Oct. 2000.

[20]E. Omiecinski, "Alternative Interest Measures for Mining Associations," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 1, pp. 57- 69, Jan./Feb. 2003.

[21]B. Padmanabhan and A. Tuzhilin, "A Belief-Driven Method for Discovering Unexpected Patterns," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 94-100, 1998.

[22]B. Park and H. Kargupta, "Distributed Data Mining: Algorithms and Systems, Applications," Data Mining Handbook, pp. 341-358, 2002.

[23]A. Silberschatz and A. Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 275-281, 1995.

[24]P. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns," Proc. ACM SIGKDD, pp. 32-41, 2002.

[25]A. Tzacheva and Z. Ras, "Action Rules Mining," Int'l J. Intelligent Systems, vol. 20, no. 7, pp. 719-736, 2005.

[26]Q. Yang, J. Yin, C. Ling, and R. Pan, "Extracting Actionable Knowledge from Decision Trees," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 43-56, Jan. 2007.

[27]Y. Yao and Y. Zhao, "Explanation-Oriented Data Mining," Encyclopedia of Data Warehousing and Mining, J. Wang, ed., pp. 492-497, 2005.

[28]S. Yoon, L. Henschen, E. Park, and S. Makki, "Using Domain Knowledge in Knowledge Discovery," Proc. Eighth Int'l Conf. Information and Knowledge Management, pp. 243-250, 1999.

[29]Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction, Y. Zhao, C. Zhang, and L. Cao, eds. IGI Press, 2008.