Volume 3, No. 3, May-June 2012

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Combined Data Mining for Discovering Knowledge

Mrs. Rupali A. Mahajan
M.E. Student
Department of Computer Sci. & Engg., P.R.M.I.T & R,
Badnera-Amravati, India.
rupalimahajan@gmail.com

Prof. S.R. Gupta
Assistant Professor
Department of Computer Sci. & Engg., P.R.M.I.T & R,
Badnera-Amravati, India.
srg_99@rediffmail.com

Abstract: Mining Complex Knowledge from Complex Data has been recognized as one of the most challenging problems in data mining. In many real world scenarios, the data is not extracted from single data source but from distributed and heterogeneous ones. The discovered knowledge is expected comprehensive so that it can better fit in business environment and enhance its actionability in decision-making. To this end, concept of combined mining is useful to extract actionable knowledge from complex data.

Keywords: component; formatting; style; styling; insert (Minimum 5 to 8 key words)

I. INTRODUCTION

The amount of data being generated and stored is growing exponentially so it is very important to gather data from different data sources, store and maintain the data, generate information and then generate knowledge. Complex data such as multiple large heterogeneous data sources, user preferences, and business impact. In such situations, a single method or one-step mining is often limited in discovering informative knowledge. It would also be very time and space consuming. So we proposes combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. We summarize general frameworks, paradigms, and basic processes for multifeature combined mining, multisource combined mining, and multimethod combined mining.

Data mining [5] is the process of discovering meaningful new correlations, patterns and trends through vast amounts of data using statistical and mathematical techniques. Over recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. It is also useful for extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data.

Patterns identified by traditional method involve homogenous features from single source of data and it is not informative in business decision making. Whereas combined data mining extracting actionable patterns from complex data. Knowledge reflecting full business settings is more business friendly, comprehensive, and informative for business decision makers to accept the results and to take operable actions accordingly. With the accumulation of ubiquitous enterprise data, there is an increasing need to mine for such informative knowledge in complex data.

The general ideas of combined mining are as follows.

By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.

By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.

By applying multiple methods in pattern mining, combined patterns are generated which disclose a deep and comprehensive essence of data by taking advantage of different methods.

A. Concept of Combined Mining:

a. Basic Concept:

For a given business problem (Ψ), we suppose that the key entities associated with it in discovering interesting knowledge for business decision support are as follows: data set D collecting all data relevant to a business problem, feature set F including all features for data mining, method set R consisting of all data mining methods that can be used on the data D, interestingness set I composed of all measures from all methods R, impact set T referring to business impacts or outcomes such as fraud or no fraud, and pattern set P.

They are described as follows.

a) ***Data set D:*** D = {Dk; k = 1, . . ., K} consists of all K sub data sets relevant to the underlying business problem, and Xk is the set of all items in the data set Dk∀ k ≠ j, Xk ∩ Xj = φ.

b) ***Feature set F:*** F = {Fk; k = 1, . . . , K} refers to all features used for pattern mining on K subdata sets, where Fk is the feature set corresponding to the data set Dk.

c) ***Method set R:*** R = {Rl; l = 1, . . . , L}, where Rl is a data mining method set deployed on the data set Dk involving the feature set Fk.

d) ***Interestingness set I:*** I = {Im,l;m = 1, . . . , M; l = 1, . . . , L}, where Im,l is an interestingness metric set corresponding to a particular data mining method Rl, which is associated with m interestingness metrics.

e) ***Impact set T :*** T = {Tj ; j = 1, . . . , J} consists of the categorized business impacts associated with certain

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html

patterns; in some cases, impacts can be categorized into impact (T) and nonimpact (¯ T), for instance, fraud or nonfraud. If a pattern is associated with an impact (T), represented by X→ T, then we call it an impact - oriented pattern. Similarly, if a pattern is mainly relevant to nonimpact, indicated by X → ¯ T, we call it a nonimpact oriented pattern.

Combined mining is a process defined as follows.

f. Definition 1 (Combined Mining): Combined mining is a two-to-multistep data mining procedure, consisting of the following:

a) Mining atomic patterns *Pn,m,l* as described in (1).

b) Merging atomic pattern sets into combined pattern set *P k* = *Gk*(*Pn,m,l*) for each data set *Dk* by pattern merging method *Gk* ,3) If multiple data sets are involved, combined patterns identified in specific data sets are then further merged into the combined pattern set *P* = *G*(*Pk*).

c) The combination of multiple features (*F*): The combined pattern set *P* involves multiple features, namely, *P* = *{Fk|Fk* ⊂ *F,Fk* ∈ *Dk,Fj+k* ∈ *Dj+k*; *j*, *k* _= 0*}*,e.g., features of customer demographics and behavior.

d) The combination of multiple methods (*R*): The patterns in the combined set reflect the results mined by multiple data mining methods, namely, *P* = *{P k* |*Rk* →*Pk}*, for instance, association mining and classification.

b. Basic Paradigms:

Some basic paradigms of combined mining involve combined pattern types, structures formed by atomic patterns, and relationships and time frames among atomic patterns. From the pattern type perspective, combined patterns can be classified into nonimpact-oriented combined patterns (NICPs) and impact-oriented combined patterns (ICPs), depending on whether a pattern is associated with a certain target item or business impact. For a NICP, its item sets are associated with each other under certain interestingness metrics.

$$Pn: Rl(X1 \wedge \cdot, \cdot \cdot \wedge Xi) \rightarrow Im \quad (3)$$

$$P: = G(P1 \wedge \cdot, \cdot \cdot \wedge Pn) \rightarrow I \quad (4)$$

An ICP is associated with either a target itemset or resulting impact (*Tj* ; *Tj* ⊂ *T*, where T is the target or impact set).

$$Pn : \{Rl(X1 \wedge \cdot, \cdot \cdot \wedge Xi) \rightarrow Im\} \rightarrow T1 \quad (5)$$

$$P: = G(P1, \cdot \cdot \cdot, Pn) \quad (6)$$

The number of the constituent atomic patterns in a combined pattern can vary.

There are two kinds of general structures.

a) Pair patterns: P ::= G(P1, P2), where two atomic patterns P1 and P2 are correlated to each other in terms of pattern merging method G into a pair. From such patterns, contrast and emerging patterns [5] can be further identified.

b) Cluster patterns: P ::= G(P1, . . . , Pn)(n > 2), where more than two patterns are correlated to each other in terms of pattern merging method G into a cluster. A group of patterns, such as combined association clusters [13], can be further discovered.

From the ***time frame*** perspective, patterns may be correlated in terms of different temporal relationships, for instance, as follows:

a) Independent relation: as illustrated by {P1 : P2} in which P1 and P2 occur independently from the time perspective;

b) Concurrent relation: as illustrated by {P1| | P2} in which P1 and P2 occur concurrently;

c) Sequential relation: as illustrated by {P1; P2} in which P2 happens after the occurrence of P1;

d) Hybrid relation: as illustrated by *{P*1 ⊗ *P*2 · · · ⊗*Pn;*⊗ ∈ *{*:,| |, ; *}*—there are more than two patterns existing in P in which some of them happen concurrently (| |) or independently (:) while others occur sequentially (;).

II. MULTIFEATURE COMBINED MINING

Multifeature combined mining approach considers features from multiple data sets during the direct generation of more informative patterns. In multifeature combined pattern (MFCP) mining, a combined pattern is composed of heterogeneous features of different data types, such as binary, categorical, ordinal, and numerical, or of different data categories, such as customer demographics, transactions, and time series.

a. ***Definition (MFCPs):*** Assuming that *Fk* denotes the set of features in data set *Dk*∀*i* ≠ *j*, *Fk,i* ∩*Fk,j* = ∅, MFCP *P* is in the form of

Pk : *Rl*(*F*1, · · · ,*Fk*) and *P* := *GF* (*Pk*)

Where ∃*i*, *j*, *i* ≠*j*, *Fi* ≠ ∅, *Fj* ≠ ∅, and *GF* is the merging method for the feature combination.

III. MULTIMETHOD COMBINED MINING

Multimethod combined mining is another approach to discover more informative knowledge in complex data. The focus of multimethod combined mining is on combining multiple data mining algorithms as needed in order to generate more informative knowledge. In fact, the combination of multiple data mining methods has been recognized as an essential and effective strategy in dealing with complex applications. In dealing with complex real-world applications, the general process of multimethod combined mining is as follows.

a. First, based on the domain knowledge, business understanding, data analysis, and goal definition, a user determines which methods should be used in the framework.

b. Second, the patterns discovered by each method are combined with the patterns by the other methods in terms of merging method *G*. In reality, the merger could be through either *serial* or *parallel* combined mining.

a) Definition (Multimethod Combined Mining): Assuming that there are *l* data mining methods

Rl (*l* = 1, . . . , *L*), their respective interestingness metrics are in the set *Im* (*m* = 1, . . . , *M*). The features available for

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html

mining the data set are denoted by *F*, and *multimethod combined mining* is in the form of

$Pl : Rl(F) \rightarrow Im,l$

$P := GM(Pl)$

Where *GM* is the merging method integrating the patterns identified by multiple methods.

Following advantages of combined mining in discovering informative knowledge in complex data, compared to a single use of existing methods. ([1], [11],[12], and [13])

a. Flexible frameworks for combining multifeatures, multisources, and multimethods covering various needs in mining complex data, which are customizable for specific cases. With combined mining, the advantage of specific algorithms can be well taken in handling particular tasks.
b. Effective in discovering patterns with constituents from multiple heterogeneous sources and a large scale of real life data, which can provide patterns reflecting a full picture rather than a single line of business.
c. Novel combined patterns can be produced which cannot be identified by directly applying existing methods.

IV. CONLUSION

There is an increasing need to mine for patterns consisting of multiple aspects of the aforementioned information so as to reflect comprehensive business scenarios and present patterns that can inform decision-making actions. Thus presented a comprehensive and general approach named combined mining for discovering informative knowledge in complex data. . It consists of frameworks for handling multi feature, multisource, and multi method related issues.

V. REFERENCES

[1]. Longbing Cao, Huaifeng Zhang, Yanchang Zhao, " Combined Mining: Discovering Informative Knowledge in Complex Data " IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 41, NO. 3, JUNE 2011

[2]. L. Cao, Y. Zhao, and C. Zhang, "Mining impact-targeted activity patterns in imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 20, no. 8, pp. 1053–1066, Aug. 2008.

[3]. L. Cao, Y. Zhao, H. Zhang, D. Luo, and C. Zhang, "Flexible frameworks for actionable knowledge discovery," IEEE Trans. Knowl. Data Eng., vol. 22, no. 9, pp. 1299–1312, Sep. 2010.

[4]. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in Proc. ICDE, 2007,pp. 716–725.

[5]. S. Dzeroski, "Multirelational data mining: An introduction," ACM SIGKDD Explor. Newslett., vol. 5, no. 1, pp. 1–16, Jul. 2003.

[6]. G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in Proc. KDD, 1999, pp. 43–52.

[7]. W. Fan, K. Zhang, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure, "Direct mining of discriminative and essential graphical and itemset features via model-based search tree," in Proc. KDD, 2008, pp. 230–238.

[8]. K. K. R. Hewawasam, K. Premaratne, and M.-L. Shyu, "Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections," IEEE Trans. Syst., Man, Cybern.B, Cybern., vol. 37, no. 6, pp. 1446–1459, Dec. 2007.

[9]. Jorge, "Hierarchical clustering for thematic browsing and summarization of large sets of association rules," in Proc. SDM, 2004, pp. 178–187.

[10]. N. Lesh, M. J. Zaki, and M. Ogihara, "Mining features for sequence classification," in Proc. KDD, 1999, pp. 342–346.

[11]. H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Combined association rule mining," in Proc. PAKDD, 2008, pp. 1069–1074.

[12]. Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang, "Mining for combined association rules on multiple datasets," in Proc. DDDM, 2007, pp. 18–23.

[13]. Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Combined pattern mining: From learned rules to actionable knowledge," in Proc. AI, 2008, pp. 393–403.

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html