AN APPROACH TO ENHANCE COMBINED DATA MINING

Neha[1], Dr. Rajeev Yadav[2]
[1] M.Tech (CSE) Student, [2] Assistant Professor
RPS College of Engineering and Technology, Balana Mohindergarh

Abstract: Data mining at big business level works on enormous measure of information, for example, government exchanges, banks, insurance agencies et cetera. Inevitably, these businesses produce complex data that might be distributed in nature. When mining is made on such data with a single-step, it produces business intelligence as a particular aspect [5]. It is challenging to dig for far reaching and instructive information in such complex information suited to genuine choice needs by utilizing the current techniques. The need of the hour is that the data mining methods used by enterprises must include the methods of multiple features, data sources and different approaches. The combination of all these approaches is called as combined mining. The combined mining can produce patterns that reflect all aspects of the enterprise. Thus the derived intelligence can be used to take business decisions that lead to profits. This sort of learning is known as significant information. The significant information is found through different information sources, various techniques and numerous components [2]. The intelligence thus obtained is dependable and reliable**.**
Keywords: Data Mining, Data Warehouse, Combined Data Mining, Multisource Combined mining, Multi-feature Combined Mining.

Introduction

Data mining that can be defined as the analysis step of the "Knowledge Discovery in Databases" process, or KDD. Data mining is the process of discovering information in large amount of data sets involving methods at the intersection of artificial intelligence, statistics, machine learning and database systems [7]. The main aim of data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining tools and techniques help to predict business trends those can occur in near future. Association rule mining is an important technique to discover hidden relationships among items in the transaction.

Mirroring this conceptualization of data mining, a few spectators consider data mining to be only one stage in a bigger procedure known as learning disclosure in databases (KDD). Different strides in the KDD procedure, in dynamic request, incorporate information cleaning, information coordination, information choice, information change, (data mining), design assessment, and learning introduction [12].

Data mining potential can be improved if the proper information has been gathered and put away in an information stockroom. An information distribution centre is a social database administration framework (RDMS) planned particularly to address the issues of exchange handling frameworks [15]. It can be approximately characterized as any unified information storehouse which can be questioned for business advantage. Information warehousing is another effective method making it conceivable to remove filed operational information and conquer irregularities between various heritage information groups. Difference is shown in Figure 1:

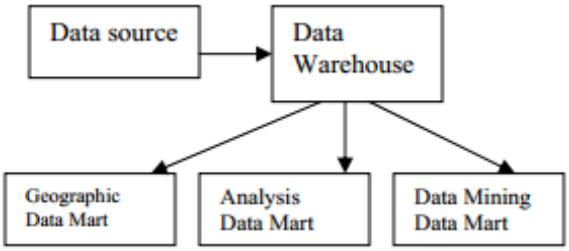


Figure 1: Data Mining and Data Warehouse Concept

The remaining paper is organized as follows. In Section 2, we conclude some present work related to data mining and its techniques. New approach for combined data mining is described in Section 3. Section 4, includes the evaluation results by using proposed technique. Lastly, in Section 5, we provide our conclusion and the future work that can be done to make the data mining techniques more efficient.

RELATED WORKS

An overview of distributed data mining methodologies, applications and frameworks given by Kargupta and Park (2002). The point is to bring up the crisscross between engineering of most off the information mining frameworks. They additionally brought up the need of information digging framework for appropriated applications. They additionally said and assert that such a mismatch of design may bring about key bottleneck in many conveyed applications and frameworks [4].

Karypis and Wang (2005) introduce another classifier, HARMONY, which is a case of direct digging for enlightening examples as HARMONY specifically mines the resultant arrangement of standards required for characterization. G. Dong and J. Li (1999) present another sort of examples i.e. developing examples (EPs), for finding information from databases. They characterize EPs as information thing sets whose bolster expands all the more altogether from one to other informational collection. They have utilized EPs to fabricate effective classifiers. W. Fan et al (2008) forms a model based inquiry tree, which parcels the information onto diverse hubs and at every hub, it specifically discover a discriminative example, which additionally separate its cases into more pure subsets [18].

J. Han et al (2006) proposed another approach called CrossMine, which fundamentally incorporates an arrangement of novel and intense techniques for multi-social order including 1) tuple ID engendering, 2) new definitions for predicates and choice tree hubs and 3) a particular testing strategy. They additionally proposed two exact and versatile strategies for multi-social order i.e. CrossMine-Rule and CrossMine-Tree. C. Zhang et al (2008) proposed a novel approach of joined examples to remove imperative, significant and affect situated data from a lot of affiliation standards. They additionally proposed meanings of joined examples and furthermore plan novel networks to gauge their intriguing quality and broke down the repetition in consolidated examples. Consolidated mining as a general approach is proposed by C. Zhang et al (2011) to mine the educational examples. They outline general structure, ideal models and fundamental procedures for different sorts of joined mining. They additionally create novel sorts of consolidated examples from their proposed systems. H. Yu, J. Yang and J. Han (2003) proposed another technique called as Clustering-Based SVM (CB-SVM), in which, they examine the entire informational index just once to have a SVM with tests that convey the factual data of the information by applying a progressive miniaturized scale grouping calculation. They additionally demonstrate that CB-SVM is likewise profoundly versatile for extensive informational indexes and furthermore creating high arrangement accuracy.Kargupta and Park (2002) give an outline of conveyed information mining calculations, frameworks and applications.
Consolidated mining as a general approach is proposed by C. Zhang et al (2011) to mine the educational examples. They outline general structure, ideal models and essential procedures for different sorts of consolidated mining. They additionally create novel sorts of joined examples from their proposed structures. H. Yu, J. Yang and J. Han (2003) proposed another technique called as Clustering-Based SVM (CB-SVM), in which, they check the entire informational index just once to have a SVM with tests that convey the measurable data of the information by applying a various leveled smaller scale grouping calculation. They additionally demonstrate that CB-SVM is likewise profoundly adaptable for huge informational collections and furthermore producing high arrangement precision

PROPOSED APPROACH

Our proposed work mainly aims at presenting a new comprehensive framework for combined mining. As a matter of fact, this thesis makes use of existing methods or techniques as part of the framework. Therefore it integrates multi source combined mining, multi-method combined mining, and multi-feature combined mining. Multiple features might include demographics of customer, behavior, business impacts and also transactional data. Multi method might include clustering, classification and so on. Multiple data sources do mean that the mining process takes data from multiple related data sources. The deliverables of the proposed framework combined patterns or combined association rules [7].

This work basically incorporates the idea of joined mining that depends on the current works. It talks about the system and different consolidated mining, for example, multi-source joined mining, multi-technique joined mining, and multi-highlight consolidated mining. It gives different systems to example collaboration and novel examples, for example, bunched designs, and consolidated examples and so on.

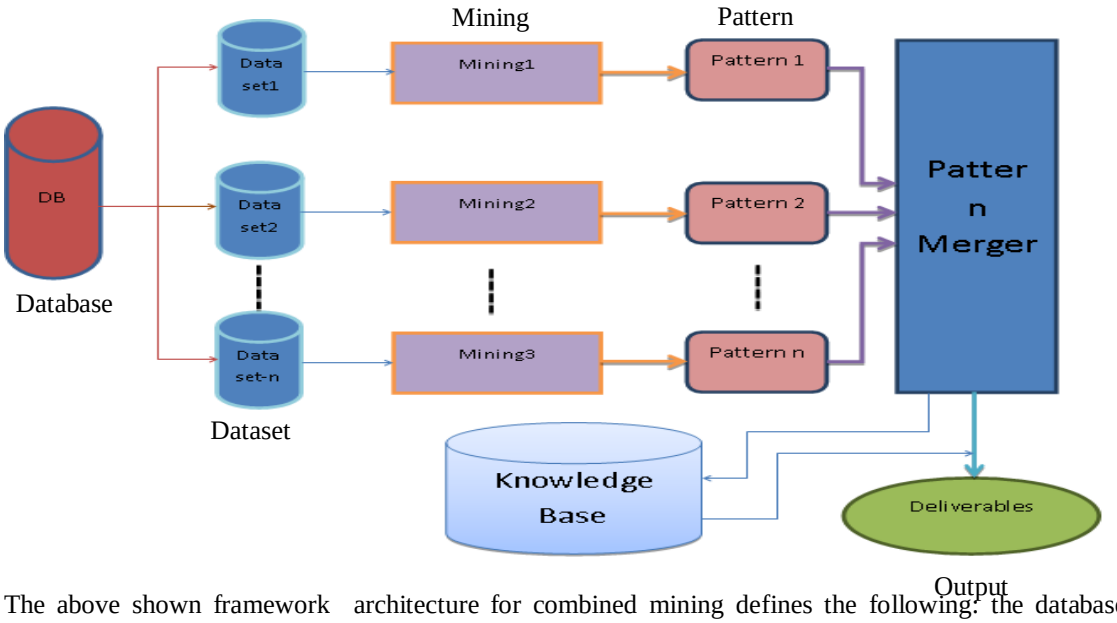
The general thoughts of combined/consolidated mining utilized as a part of our proposed system are as per the following:

- By including different heterogeneous components, consolidated examples are created which mirror various parts of concerns and attributes in organizations.
- By mining various information sources, consolidated examples are produced which mirror different parts of nature over the business lines.
- By applying various strategies in example mining, joined examples are produced which uncover a profound and exhaustive quintessence of information by exploiting distinctive techniques.

Rather than presenting a specific algorithm for mining a particular type of combined patterns, this work mainly focuses on abstracting several general and flexible frameworks from the architecture perspective, which can foster wide implications and particularly can be instantiated into many specific methods and algorithms to mine for various patterns in complex data [1].

Our proposed combined mining framework mainly combines the existing algorithms or methods to integrate multiple features, multiple data sources and multiple mining methods like classification. The architecture of the proposed framework is shown in Figure 2 which is common framework for all these combined mining approaches.

Figure 2: Framework for Combined Mining



The above shown framework architecture for combined mining defines the following: the database of larger amount of data is split into smaller datasets of purposed knowledgeable data for Knowledge Extraction by Combined Mining where mining is done on a multi thread basis. Any of Multi-source mining, Multi-feature mining and Multi-method mining techniques is applied to the datasets; all these approaches generate different types of set of patterns. Those patterns are then merged together using a component known as pattern merger. Pattern merger is mainly a program that combines all pattern sets obtained from mining approach. It uses the concept of Word Sense Disambiguity (WSD) which leads to a framework of Concept Based Pattern Mining (CBPM). After this the result is incorporated with the knowledge base for business intelligence to produce the deliverables which gives the informative knowledgeable patterns for business aspects.

A. ALGORITHMS

The main algorithm of this framework is as follows:
In this algorithm CAP = {p1, p2… pn} and BRR = {r1, r2… rn} where p1, p2… pn represents the patterns set and r1, r2… rn represents the business rules set.

Main Algorithm

INPUT : Target Data (data collected according to the problem defined) and required Threshold Values (Support Threshold, Confidence Threshold, lift Threshold).

OUTPUT: Combined Actionable Patterns (CAP) and Final Business Rules Report (BRR) Generated.
STEP 1: Identify a suitable data for initial mining exploration in the particular database of the enterprise.
STEP 2: Split the database into 'n' smaller datasets by importing those data.
STEP 3: Apply Multi-source, Multi-feature or Multi-method mining techniques on these datasets.
STEP 4: Generate the Combined Association Rules and patterns.
STEP 5: Merge the atomic pattern into combined pattern
For patterns 1 to n: -Design the Pattern merger Function to merge all relevant atomic patterns by involving Word Sense Disambiguity. Employ the method on the pattern set obtained. Generate the conceptual pattern.
STEP 6: Invoke Business Intelligence knowledge base on the final generated pattern.
STEP 7: Output the Deliverables.
This framework is general for multi-source, multi-feature and multi-method mining approaches. The generated datasets are mined by using these mining approaches, so it is necessary to describe the algorithms of multi-source mining, multi-feature-mining and multi-method mining.

Multi-source Algorithm

INPUT: Target the datasets.
OUTPUT: Informative Combined Patterns.
STEP1: Identify the informative datasets.
STEP2: Identify the required data from datasets for mining.
STEP3: Apply pattern mining and extract the atomic patterns from datasets.
STEP4: Merge the generated patterns by using the pattern merger.
STEP5: Enhance the actionable patterns for generating deliverables.
STEP6: Generate deliverables as output.

Multi-feature Algorithm

INPUT: Target the datasets.
OUTPUT: Informative Combined Patterns.
STEP1: From each datasets mine for the required atomic patterns.
STEP2: Merge these relevant atomic patterns generated in step-1as per pattern merging method.
STEP3: Make pairs of patterns from the resulting combined patterns from step-2.
STEP4: Add other related patterns to all the generated pair patterns for generating the cluster patterns.
STEP5: For those match designs on the off chance that one example is an augmentation of the other then yield it as a universal combine design.
STEP6:In a cluster/bunch design if there is an ordinal connection between the important neighboring example and the last examples comprises of extra data on top of its previous ones, yield them as incremental group designs/patterns.
STEP7: Generate deliverables as output.
Multi-method Algorithm

In many situations the patterns discovered by a particular method do not serve to user's perspective. Here one finds need of using more than one mining methods in order to discover more informative patterns. Multiple methods can be used in parallel, serially or in closed loop fashion.

Parallel Multi-method mining: In this approach different mining strategies are utilized as a part of parallel on various or same data indexes/sets.

- Initially different, independent data sources are mined using different data mining techniques, to find out respective atomic pattern sets.
- Then these patterns are merged together by merging method.

Serial Multi-method Mining: Here various data mining techniques are used one by one. Outcome of one technique is treated by another technique to discover in depth knowledge. This method works as follows.

- Initially data source is mined using a suitable method to obtain pattern set, say P1.
- After studying the initial pattern set, next suitable method is selected and P1 is again mined using it to discover next level pattern P2.
- Various techniques are applied according to domain knowledge and output requirement.

Closed loop Multi-method Mining: This approach takes into account the impact of one technique on other. Practically feedback from next method can be used to refine the results of previous method. Closed loop mining is carried out as follows.

- Initial pattern set P1 is formed by following the process of serial multi-method mining. At the end of this step some samples may not be identified. This is due to the limitations and conditions applied on various mining techniques.
- The patterns in P1 are checked for their validity. Some samples may not be valid to patterns. A separate data set is formed using such exceptional samples, say Dx. This dataset is again mined by multiple methods to discover new pattern P2.
- Process of step 2 repeated as many times required by miner to discover patterns P1, P2, ……, Pk.
- All patterns are then merged to form combined pattern.

EVALUATION

We have tested our framework method in randomly generated customer's data of store to generate combined patterns and business rules for a multipurpose retail store for good benefits and quick services. The cleaned sample data contains 100 customers with their demographic attributes. There are 82 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. Table 1 shows the demographic customer data and Table 2 shows shopping data.

Table 1: Demographic Customer Data

Sequence ID	Customer Type
S1	Man
S2	Women
S3	Boy
S4	Girl
S5	Kids

Table 2: Shopping based Data Mining Ordered and Unordered data

Sequence ID	Shopping Type	Customer Role	Shopping Cart
S1	Family	Father	Beans, Beer, Banana, Eggs, Soap, Medicine
S2	Family	Mother	Tomato, Perfume, Cleaner, Shampoo, Clothes
S3	Individual	Son	Book, Coke, Chips, Ice-cream, Pen, Trouser
S4	Individual	Daughter	Book, Pen, Pepsi, Nail Paint, Jeans, Chips, Lip
S5	Individual	Baby	Chocolates

After applying the algorithms we can find the combined association rules like as the following combined association which shows a single family shopping pattern. This is a complete package for a family so we put this in the category 'GP' (Great Purchase) and this leads to good marketing as it gives benefits to both the store owner and customers. On each GP customer gets 25% money back.

{S1= Father: Beer, iPhone, Medicine, Perfume;S2= Mother: Eggs, Apple, Shampoo, Clothes; S3= Son: Book, Bike; S4= Daughter: Nail Paint, Hair Color; S5= Baby: Chocolates —> gp = GP}
This can be represented as {S1, S2, S3, S4, S5 —> gp}

Finally, combined patterns can be transformed into operable business rules that may indicate direct actions for business decision making. We can generate the following business rule by extending the Business Rule specification.

DELIVERING BUSINESS RULES:
Customer Demographic- combination business rules…
For All customers;
Condition:
Satisfies = "A family member does not purchase the same item that other members have purchased at the same time."
Relates = "If father buys Beer, iPhone, Medicine and Perfume, Mother buys Eggs, Apple, Shampoo and Clothes, Son buys Book and Bike, Daughter buys Nail Paint and Hair Colour & Baby buys Chocolates."
Operation:
Alert = "The Family has Purchased the Items of category GP."
Action = "Give them surprise of 25% money back on purchasing items of category GP."
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 regular customers and monitoring the shopping characteristics of each customer.

CONCLUSION

Combined mining is a general approach for designing of real life data mining applications. It includes designing frameworks for multi-feature, multi-method, multi-source approaches. Sample frameworks are discussed in this paper. More such domain specific frameworks can be designed. Also combined mining involve designing of various representation in which patterns can be delivered.

Our proposed combined mining framework provides a paradigm shift from data-driven pattern mining to domain driven actionable pattern discovery. Using this framework various domain specific models for telecom fraud detection, trading evidence discovery, stock-market analysis, design of marketing policies can be built.

Additionally creating viable standards, combined pattern types, consolidated mining techniques, design combining strategies, and intriguing quality measures for taking care of vast and different wellsprings of information accessible in our industry ventures and and the proposed techniques are quite successful for mining the data from complex databases.

References

[1] Faraj A. El-Mouadib, Zakaria S. Zubi,Ahmed A. Almagrous, and Irdess S. El- Feghi"Generic Interactive Natural LanguageInterface to Databases (GINLIDB)",international journal of computers(2015).

[2] Siasardjahantighi F, Norouzifard M, Davarpanah S H, Shenassa M H. Using natural language processing in order to create SQL queries. In: IEEE International Conference on Computer and Communication Engineering (ICCCE); 13-15 May 2014; Kuala Lumpur, Malaysia: IEEE. pp. 600 - 604.

[3] Li H, Shi Y. A WordNet-based natural language interface to relational databases. In: IEEE 2nd International Conference on Computer and Automation Engineering (ICCAE); 26-28 Feb. 2010; Singapore: IEEE. pp. 514 – 518.

[4] Mrs. NeeluNihalani, Dr. Sanjay Silakari, Dr. Mahesh Motwani. "Natural language Interface for Database: A Brief review", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011 ISSN (Online): 1694-0814.

[5] Mrs. NeeluNihalani, Dr. Sanjay Silakari, Dr. Mahesh Motwani. "Natural language Interface to Database using Semantic Matching", International Journal of Computer Application, Vol. 31, no.11, Oct. 2011 ISSN: 0975 – 8887.

[6] A. Kaur, and P. Bhatiya, "Punjabi Language Interface to Database", M.tech thesis, Department of CSED, Thapar University, 2010.

[7] Arati K. Deshpande, and Prakash. R. Devale, "Natural Language Query Processing Using Probabilistic Context Free Grammar", International Journal of Advances in Engineering & Technology, May 2012.,ISSN:2231-1963.

[8] Owda M, Zuhair B, Crockett K. Conversation-Based Natural Language Interface to Relational Databases. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops; 5-12 Nov 2007; Silicon Valley: IEEE/WIC/ACM. pp. 363 – 367.

[9] VipinKumar.N et al, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 2 (4), 2011, 1706-1710.
[10] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", Scientific American, May 2013, Vol. 284,No. 5, pp. 34-43.
[11] W3C, Resource Description Framework, http://www.w3.org/RDF/, 2004.
[12] W3C, RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema/, 2004.
[13] W3C, Web Ontology Language, http://www.w3.org/2004/OWL/, 2004.
[14] SPARQL Query Language for RDF, W3C Working Draft, http://www.w3.org/TR/2006/WD-rdf-sparql-query-20061004/, 4 October 2006.
[15] SQL, XQUERY and SPAQL by Jim Melton.
[16] Michael Grobe : RDF, Jena, SparQL and the "Semantic Web".
[17] Querying the Semantic Web using a Relational Based SPARQL by Andrew Newman.