JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Exploring Patterns Extraction Techniques and Methodologies from Multiple Data Sources: A Comprehensive Survey

[1]P. Sasikumar, [2]K.T.Meena Abarna

[1]Research Scholar, [2]Associate Professor
Department of Computer Science and Engineering
Annamalai University, Annamalai Nagar, India

Abstract: Progress in computer and communication technologies necessitates novel perspectives on distributed computing environments and the establishment of distributed data repositories for accommodating vast data volumes. In such contexts, the extraction of meaningful patterns from multiple data sources poses a formidable challenge within the data mining community. The domain of multi-database mining (MDM) arises as a propitious research area, evidenced by the considerable endeavors undertaken in recent epochs. Techniques for knowledge discovery from multiple data sources can be broadly classified into two categories, namely (1) mono-database mining and (2) local pattern analysis. The primary objective of this survey is to explicate the underlying principles of these methodologies, amalgamating research contributions, and elucidating their significance and constraints.

Index Terms **– Multi-database Mining, Patter Extraction, MDM, Mono-database Mining, Local Pattern Analyses**

I. INTRODUCTION

In recent times, there has been significant progress in communication technology, both over wired and wireless networks, leading to the emergence of diverse distributed applications. These distributed applications often involve data sources dispersed across various geographical locations to handle vast amounts of data. This setup enables organizations to employ multi-database applications to meet their operational requirements effectively. Consequently, many organizations find it necessary to extract valuable insights from their distributed multi-databases, which are spread across different branches, to support their decision-making processes.

Let us consider the case of a retail giant, Reliance India Ltd, which has initiated a remarkable retail revolution in India, rapidly expanding from zero stores to an impressive 1500 outlets within a mere six months. These outlets generate an enormous number of transactions on a daily basis. For such complex applications, the development of efficient data mining techniques to uncover meaningful patterns from multiple branches becomes critically important.

The realm of multi-database mining (MDM) garners considerable prominence owing to several key factors: (1) the escalating adoption of automated data collection instruments and the copious influx of data generated during organizational operations; (2) the dynamic nature of distributed repositories featuring diverse data sources and formats; (3) the imperative for organizations to analyze the contents and trends of branch databases; and (4) the exigency to bolster the efficacy of decision-making processes through the integration of high-quality knowledge gleaned from multi-databases.

The efficacy of Multi-Database Mining (MDM) applications is intricately tied to the diversity of data available across multiple databases. In real-world scenarios, data stored in various repositories often exhibit inconsistencies and conflicts. Bright et al. [1] elucidated pertinent data representation challenges within the multi-database environment, encompassing the following issues:

1. Name Differences: Databases may adopt disparate conventions for object naming, resulting in synonym and homonym complications. Synonyms denote instances where the same data item bears distinct names in different databases. To address this, the global system must discern semantic equivalence and map variant local names to a unified global name. Conversely, homonyms refer to instances where different data items share the same name in diverse databases, requiring the global system to identify semantic discrepancies and map common names to distinct global names.

2. Format Differences: Format disparities encompass variations in data types, domains, scales, precisions, and item combinations. For instance, a part number may be designated as an integer in one database but as an alphanumeric string in another. In some cases, data items are fragmented into separate components in one database while being recorded as a single quantity in another. Multi-database systems typically address format differences through the definition of transformation functions that facilitate conversions between local and global representations. These functions can range from straightforward numeric calculations, like converting square feet to acres, to more complex transformations necessitating conversion tables or algorithms. A critical concern in this domain is the potential complexity of local-to-global transformations, particularly when updates must be supported.

3. Structural Differences: Diverse local databases may exhibit dissimilar structures for the same object. A data item may possess a single value in one database but have multiple values in another. Similarly, an object might be represented as a single relation in one location and as multiple relations in another. Moreover, a single item could be construed as a data value

in one location, an attribute in another, and a relation in a third, leading to discrepancies in structure and content that require resolution.

4. Conflicting Data: The issue of conflicting data arises when two databases record the same data item but assign distinct values to it. This can be attributed to incomplete updates or system errors during data manipulation, necessitating mechanisms for conflict resolution and data cleansing.

In light of these multifaceted challenges, addressing data heterogeneity and ensuring data consistency are pivotal undertakings in the context of Multi-Database Mining applications.

The aforementioned concerns underscore the imperative of adopting appropriate methodologies for addressing the MDM problem, as decisions made at the global organizational level are profoundly influenced by the quality of knowledge synthesized from multiple databases. This survey is structured as follows: The Major Methods for Multi-Database Mining section expounds upon the two principal approaches for MDM, encompassing the definition, merits, and demerits of mono-mining and local pattern analysis strategies, complemented by schematic representations. Subsequently, in the Research Efforts Based on Mono-Database Mining and Research Efforts on Multi-Database Mining sections, endeavors grounded in mono-mining and local pattern analysis are subjected to comprehensive review and discussion, respectively. Finally, the Conclusion and Scope for Future Work section presents the summarizing insights and outlines potential avenues for subsequent research.

II. EXISTING TECHNIQUES FOR MULTI-DATABASE MINING

Over the recent decades, endeavors have been undertaken to augment the knowledge discovery process through the application of artificial intelligence techniques to databases, culminating in a captivating domain of research known as Knowledge Discovery from Multiple Databases (MDM). This paradigm encompasses the extraction of valuable patterns from multiple databases, often characterized by heterogeneity, with the goal of unveiling novel and significant insights [2]. Although diverse approaches exist for knowledge discovery from multiple data sources, they can be broadly categorized into two main classes: (1) mono-database mining and (2) local pattern analysis [3].

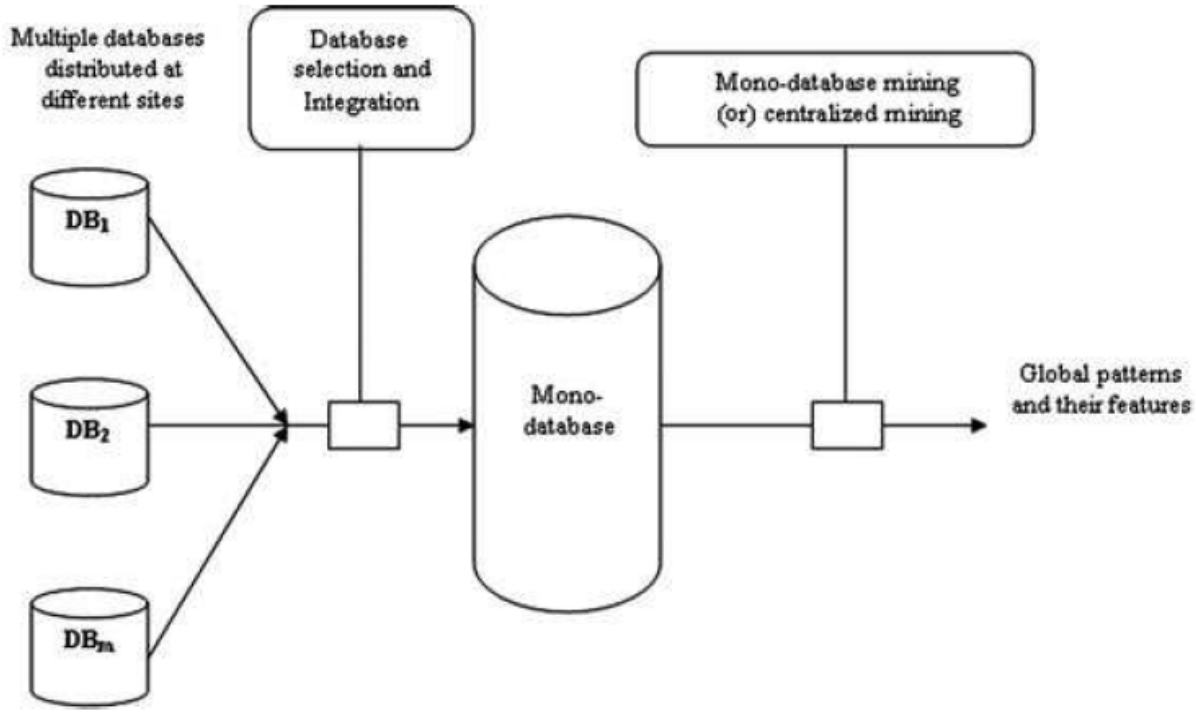

Figure-1: Mono-database Mining

2.1. Mono-database Mining

In the realm of mono-database mining, diverse data sources are consolidated into a centralized repository for the purpose of mining (as depicted in Figure 1). The principal objective of mono-database mining is to discern patterns that hold global significance among participating data sources. This entails the integration of data from various databases into a data warehouse, where mining is conducted to identify overarching patterns of interest. The primary technical challenge in this context resides in the communication cost incurred when dealing with distributed data sources, often proving exorbitant and, in some cases, impracticable to merge multiple data sources into a singular database [4].

Refining the preceding approach involves the judicious selection of pertinent data sources tailored to specific applications, subsequently integrating them to extract knowledge. Although this approach proves effective in reducing search costs for a given application, it exhibits a strong reliance on application-specific factors and necessitates multiple scans for each distinct application [5].

2.2. Limitations Mono-database Mining

The application of mono-database mining for extracting insights from multiple databases presents notable limitations, rendering it unsuitable for several critical reasons:

1. The conventional data warehouse architecture upon which mono-database mining relies is fundamentally unsuitable for distributed and ubiquitous data mining applications, particularly due to the varied formats of branch databases, demanding meticulous attention during the data preprocessing stage.

2. Processing the entire dataset on a single computer can be exceedingly time-consuming. While parallel computing and specialized software can alleviate this challenge, the substantial investments in associated hardware and software make it an impractical solution from a cost-benefit analysis perspective.

3. Collecting data from diverse branches for centralized processing poses a daunting task, especially considering the substantial volume of daily transactions handled by branch databases.

4. The privacy concern emerges as a significant issue in mono-database mining applications, especially when attempting to collaborate across multiple banks for fraud detection. Centralized collection of individual customer financial data from each bank compromises the privacy of bank customers, making this approach unfeasible.

5. Combining data from different databases into a single dataset risks erasing crucial information that reflects the unique characteristics of individual branches. Branch databases may carry different weights, and some branches might contribute significantly more to the overall company in terms of turnover, transactions, and other metrics.

In light of the aforementioned limitations, the conventional approach of mono-database mining proves inadequate, leading to the proposition of local pattern analysis as an alternative method for mining multiple data sources.

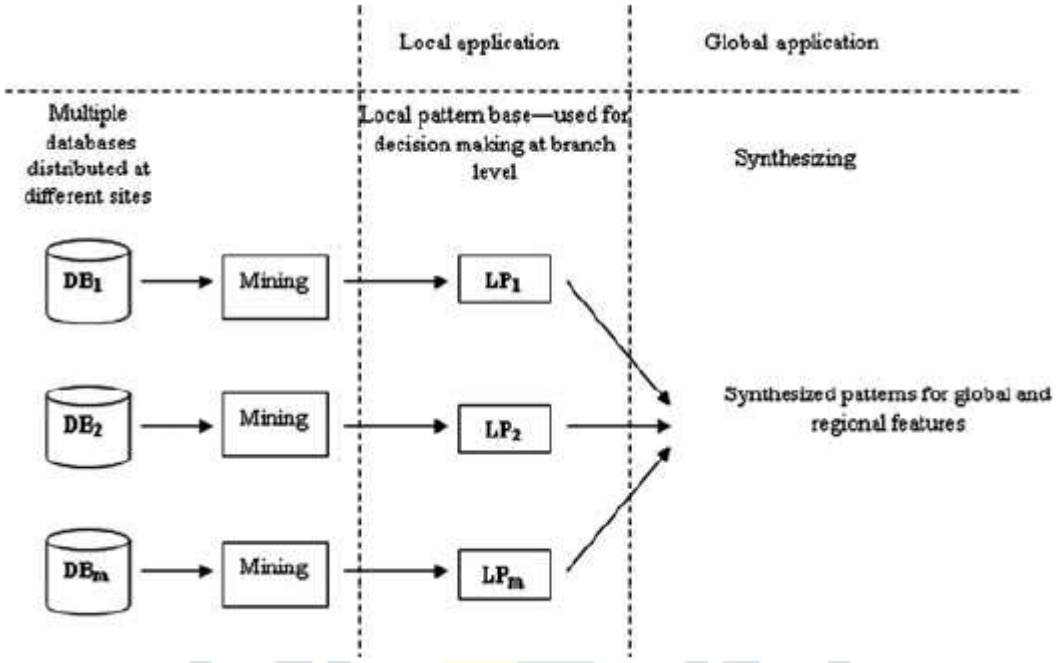

Figure-2: Local Pattern Analysis

2.3. Local Pattern Analysis

The primary objective of local pattern analysis lies in conducting data mining operations based on the specific type and availability of distributed resources, all while avoiding the need to transfer data to a central repository. It achieves this by extracting significant local patterns from individual data sources, forwarding these pattern bases, and minimizing data movement (see Figure 2). Consequently, a data mining application adopting the local pattern analysis strategy can learn models from distributed data without the exchange of raw data. In this context, a local pattern [7] may take the form of a frequent itemset, an association rule, a causal rule, a dependency, or other expressions that demonstrate the uniqueness of a branch site.

The process of Multi-Database Mining (MDM) using local pattern analysis can be defined as the synthesis of global patterns from the forwarded patterns generated by individual sites. This approach is particularly recommended for scenarios involving a large number of data sources, as it enhances scalability. The key focus here is on aggregating local mining results at multiple levels of abstraction, with the aim of promoting regional and global features.

2.3. Advantages Local Pattern Analysis

This approach presents several compelling benefits as follows:

1. It adopts an in-place strategy [8], thereby eliminating the need for data movement, which proves highly advantageous when dealing with vast volumes of data distributed across multiple sites.
2. It effectively captures the distinct characteristics of individual data sources and has the capability to identify special patterns that may hold greater significance than those found in the integrated and unified single database (mono-database).
3. The approach demonstrates low complexity, as it exclusively mines relevant individual data sources. Moreover, it offers a systematic approach for synthesizing forwarded patterns at multiple levels of abstraction, enabling the discovery of various types of patterns distributed within the data sources. For instance, global patterns (recognized by a majority of the sites), subglobal patterns (identified by some of the sites), and local patterns (acknowledged by few or single sites) [9].
4. It provides a dual-level decision-making mechanism:

(a) Global decisions: Central company decisions pertaining to global applications, formulated based on the synthesized patterns.

(b) Branch decisions: Local branches' decisions derived from features of the local patterns mined from their respective databases [3].

5. The primary aim of knowledge discovery from databases lies in uncovering interesting patterns from the perspective of the user. Although the user may not possess expertise in data mining, they are subject matter experts in the field being explored [10]. Consequently, the importance of any pattern relies on the user's interest. By adopting the local pattern analysis strategy, heads of local branches can employ diverse interestingness measures to evaluate local patterns within their respective databases, which may differ from the interestingness measure employed by the central head during global pattern synthesis. For example, branch 'B1' may use the measure of lift, while branch 'B2' may opt for the support [11, 12] measure when assessing the local rule A→B in their corresponding sites. Once synthesized, the rule A→B can be globally evaluated using another measure, such as correlation. This demonstrates how the strategy of local pattern analysis enables both local and central sites to adopt distinct interestingness metrics when evaluating patterns.

2.3. Limitations Local Pattern Analysis

While the utilization of the local pattern analysis strategy presents reasonably effective solutions for the Multi-Database Mining (MDM) problem, scholars have also observed certain drawbacks. Adhikari et al. [13] have critically assessed the matter and identified the frequency of data mining as a significant limitation of the local pattern analysis strategy when applied to the MDM problem. Unlike mono-mining, where the mining database occurs only once, the local pattern analysis strategy's frequency of mining is directly associated with the number of databases. Although the approach offers several advantages, this inherent dependence on database quantity can be considered a noteworthy constraint.

III. RESEARCH EFFORTS BASED ON MONO-DATABASE MINING

In the domain of database mining, an advanced knowledge discovery system known as INLEN [14] (Inference and Learning) has been developed at George Mason University. This system employs the AQ algorithm, which employs inductive inference over a set of training examples to learn decision rules. The resulting knowledge is presented in the form of IF-THEN rules, unveiling hidden characteristics and relationships within the database, thereby providing valuable insights to the user. However, the initial limitation of INLEN was its application restricted to small single databases.

To overcome this constraint, Ribeiro et al. [15] extended the INLEN approach to encompass multiple databases. They accomplished this by applying INLEN's methodology to individual databases and subsequently processing the acquired knowledge. This entailed modifying the AQ algorithm to handle primary and foreign key information from two data sources. In their approach, the databases should be co-located on the same machine. Wrobel [16] further extended the concept of foreign keys to include foreign links to account for nonkey attributes, considering that useful databases may exist in remote locations and contribute to the decision-making process. In response to such a scenario, Aronis et al. [17] introduced the WoRLD (Worldwide Relational Learning Daemon) system, utilizing an inductive rule-learning program that can glean knowledge from multiple databases distributed across the network. Their approach, called 'activation spreading,' computes the cardinal distribution of feature values in individual data sets and propagates this distribution across different sites.

Turinsky and Grossman [8] discuss two distinct strategies for mining multiple databases. The first strategy, an "in-place strategy," avoids moving large data sets over the Internet and instead builds local models that are combined at a central site. On the other hand, the "centralized strategy" involves moving all geographically distributed data to a central site and building a single model there. They also propose an intermediate strategy that optimizes data and model partitions to achieve a desired level of accuracy at a minimum cost.

Grossman et al. [18] introduced Papyrus, a distributed data mining system supporting various strategies based on data distribution, resource availability, and required accuracy, including moving data, models, results, or a mixture of these approaches. Prodromids et al. [19] adopted a metalearning strategy for mining multiple databases, which involves integrating multiple classifiers computed over different databases to form higher-level classifiers or a classification model. The metalearning process combines predictions from classifiers learned from data subsets through recursive learning of 'combiner' and 'arbiter' models in a bottom-up tree manner.

Kargupta and his colleagues [20, 21] proposed a collective data mining (CDM) framework for predictive data modeling in heterogeneous environments. This framework involves generating approximate orthonormal basis coefficients at each local site, selecting a representative sample of data sets from each site, and generating approximate basis coefficients for nonlinear cross terms. The local models are then combined, transformed into the user-described canonical representation, and output as the model. Various distributed data analysis algorithms have been developed based on this CDM framework, such as collective decision rule learning using Fourier analysis, collective hierarchical clustering, collective multivariate regression using wavelets, and collective principal component analysis. These algorithms facilitate distributed data analysis and modeling while ensuring the efficient exchange of a small sample of data compared to the entire dataset.

In the context of mining extensive databases, Savasere et al. [22] devised a partition algorithm to extract frequent itemsets from non-overlapping partitions of the database. Global candidate patterns were then generated from the union of these frequent itemsets. Subsequently, a second run on each partition yielded the frequency count of candidate patterns, which were aggregated to determine the global support count. If the support count surpassed the minimum threshold, the pattern was considered a global pattern, and global rules were generated accordingly. This approach offered an elegant solution for mining large centralized databases. However, its direct applicability to Multi-Database Mining (MDM) was limited. To adapt this method to MDM, each branch database was treated as a partition of the multi-database. Local frequent itemsets were forwarded to the center to create candidate global patterns, which were then transmitted back to the local sites for a second round of mining. The resulting patterns were again forwarded to the center for consolidation and evaluation of global rules. This scheme required two scans of local databases and three pattern transmissions across the network.

To discover novel and intriguing patterns concealed within data, Zhong et al. [23] introduced the concept of peculiarity-oriented mining in multiple databases. Peculiarity, denoting unique and unexpected relationships present in relatively few data instances, represented a fresh perspective on interesting patterns. The primary objective of mining peculiarity rules was to identify peculiar

data instances characterized by distinctiveness and rarity within the dataset. The authors proposed a peculiarity factor, which gauged whether an attribute value occurred in a relatively small number and significantly differed from other values by assessing the sum of the square root of the conceptual distance between them. The peculiarity factor allowed the selection of peculiar data based on predefined thresholds.

To address the challenges of handling multiple databases, Liu et al. [24] proposed a method for discovering relevant databases in the context of multi-database mining. They asserted that the first step in MDM involved identifying databases most pertinent to a specific application to enhance efficiency and accuracy. Their approach entailed constructing a cluster of multi-databases tailored for a given application, a process referred to as database selection. However, database selection had to be carried out multiple times to identify relevant databases for different real-world applications. Moreover, application-dependent techniques were inadequate when users aimed to mine multi-databases without a specific application in mind. In response to this requirement, Wu et al. [25] introduced an application-independent database classification strategy for MDM. They proposed a technique for clustering databases to facilitate mining multiple databases based on a relevance measure called similarity. This measure was derived from various metrics, including |class|, Goodness, and distance, designed to identify suitable clusters in multi-databases. Both approaches focused on efficient data preparation techniques for MDM.

While the aforementioned efforts have provided valuable insights and addressed essential aspects of mining multiple databases, they overlook numerous potentially useful patterns in local databases. In addition to the logistical challenges posed by transferring vast data over communication networks, the mono-database mining strategy disregards intriguing local patterns present at various sites. The subsequent section reviews research endeavors in local pattern analysis, which aims to overcome the limitations of mono-mining strategies.

IV. RESEARCH EFFORTS BASED ON MULTI-DATABASE MINING

Zhang et al. [26] have highlighted the disparities between mono-database mining and multi-database mining (MDM) by unveiling novel and significant patterns unique to MDM, eluding detection through mono-database mining. Their study emphasizes the two-tier decision-making process within business organizations possessing multiple branches: global decisions at the headquarters level and local decisions at the branch level. To this end, they classify patterns in multi-database systems as local patterns, high-vote patterns, exceptional patterns, and suggested patterns.

High-vote patterns gain robust support from the majority or all branches of an interstate organization, reflecting common features among the branch databases. Such patterns empower the head company to make decisions for the collective benefit of all branches. On the other hand, exceptional patterns exhibit higher support in some branches but lack support in others. These patterns enable the head company to tailor measures to local conditions and formulate specialized policies for the respective branches. Suggested patterns receive support from a subset of branches, which is smaller in comparison to the branches supporting high-vote patterns.

Considering that users are more likely to provide mined patterns instead of raw data, and given the abundance of forwarded local patterns from branch databases, a synthesizing model is crucial to derive global patterns from these local patterns. Wu and Zhang [27] propose a model for synthesizing high-frequency rules from multiple databases through a weighting approach, which draws inspiration from established methodologies in fields like probability and fuzzy set theory. To aggregate association rules from multiple databases, determining the weights of the data sources becomes imperative.

The weighting model advocated by Wu and Zhang [27] represents an initial endeavor in synthesizing global patterns from forwarded local patterns. They consider a rule as high-frequency if it garners support or votes from a substantial number of data sources, and its weight is proportional to the number of data sources supporting it. Consequently, the weight of each data source is calculated based on the number of high-frequency rules it supports. Higher weights are assigned to data sources supporting a larger number of high-frequency rules, while lower weights are allocated to those supporting fewer such rules. However, their model assumes similarity in data source sizes, which may not be practical when dealing with numerous data sources. In scenarios where data sources differ in size, complex operations like merging and splitting must be performed to equalize their sizes. Alternatively, data sources below a user-specified threshold may be disregarded if merging is infeasible due to data sharing concerns. As a result, certain data sources may not partake in the rule synthesis process. Although Wu and Zhang's model [27] aims to synthesize global association rules for the overall organization from the union of all data sources, it does not specifically target a comparative analysis of the synthesized results with the mono-mining results obtainable by the union of these data sources.

Nedunchezhian and Anbumani [28] address two key issues: data source selection and the selection of valid rules for synthesis. They calculate the data source weight based on two factors: (1) the number of high-frequency rules voted by the data source, and (2) the size of the data source. Utilizing these weights, they employ a data source selection threshold to identify candidate data sources for synthesizing high-frequency rules. To prune low-frequency rules at local sites, they present a procedure called support equalization, which equates the supports of data sources, thereby reducing the total number of rules forwarded to the central head.

Zhang et al. [29] advocate an approach for synthesizing global exceptional patterns in MDM applications. They propose an algorithm for identifying global exceptional patterns from multiple databases, mining each local database separately in a random order to derive these patterns. Kum et al. [30] develop a local mining approach for discovering sequential patterns in multiple databases. They introduce a novel algorithm to mine approximate sequential patterns, referred to as consensus patterns, from large sequence databases through two steps: sequence clustering based on similarity, followed by direct mining of consensus patterns from each cluster through multiple alignments. Adhikari and Rao [6] extend the local pattern analysis model and introduce the concept of heavy association rules. These rules possess synthesized global support exceeding a user-defined threshold, and they posit that heavy association rules can be more valuable than high-frequency association rules. They observe cases where heavy association rules may not be shared by all databases, defining a high-frequency rule as one shared by at least $n \times r1$ databases, and an exceptional rule as one shared by no more than $n \times r2$ databases, where 'n' denotes the number of databases, and r1 and r2 are user-defined thresholds.

Serial No.	Researchers	Issue-Focused	Contribution
1	Zhang et al.[26]	Local pattern analysis	Identification of new kinds of patterns in multi-database environments
2	Wu and Zhang[27]	Synthesizing model	Weighting model for synthesizing global patterns based on frequent rules voted by the data source
3	Nedunchezhian and Anbumani[28]	Database identification and setting up threshold values for synthesizing	Data source selection and support equalization for synthesizing global patterns.
4	Zhang et al.[29]	Discovering new kinds of patterns	Synthesizing procedure for globally exceptional patterns
5	Kum et al.[30]	Discovering new kinds of patterns	Algorithms for sequential pattern discovery
6	Adhikari and Rao[6]	Discovering new kinds of patterns	Notion of heavy association rules in synthesizing process on the basis of Wu and Zhang's model
7	Ramkumar and Srinivasan[2]	Synthesizing model	Transactions-population-based weighting model for synthesizing global patterns with a target of obtaining closer mono-mining result
8	Ramkumar and Srinivasan[9]	Discovering new kinds of patterns	Notion of Effective and nominal vote rate in rule synthesizing for pattern classification
9	Ramkumar and Srinivasan[31]	Optimization in synthesizing model	Notion of correction factor in rule synthesizing process for improved synthesized results
10	Adhikari et al.[32]	Database clustering	Mining global patterns in time- stamped databases
11	Adhikari et al.[34]	Grouping items	Model for mining selective items
12	He et al.[35]	Database clustering	Synthesizing model for databases of dissimilar in nature

Table-1: Analysis of Research Attempts in Local Pattern Analysis

The authors have presented an algorithm for the synthesis of heavy association rules from multiple data sources, exploring whether such rules exhibit high frequency or exceptional characteristics across various databases. Notably, this model offers an approximation of global patterns [7].

Ramkumar and Srinivasan [2] proposed a transactions-population-based weighting model to synthesize high-frequency rules from diverse data sources. Their approach assigns rule weight proportionally to the sum of weights contributed by supporting data sources, where the weight of a data source is determined by its transactional population. The aim of synthesizing global patterns from forwarded local patterns is to maintain similar levels of support and confidence as if all data sites were integrated and mono-mining was conducted. The authors disagreed with the notion that each branch within a large company should possess equal voting power for pattern synthesis, asserting that branches with higher business volumes should have a greater say in determining global policies based on global patterns.

Synthesizing models [2, 6, 27] have primarily focused on high-frequency rule synthesis, as these rules emerge as globally significant when data sources are integrated. While high-frequency rules are crucial for global decisions at the head branch of an interstate company, they tend to exclude regional patterns or rules. To address decision-making at regional levels, patterns exhibiting the individuality of regions or clusters of branches become essential and can be effectively explored through a multilevel perspective [9]. In response to this demand, Ramkumar and Srinivasan [9] extended their earlier work and proposed a framework for multilevel rule synthesis using two interesting rule evaluation measures: γ effective (effective vote rate) and γ nominal (nominal vote rate). These measures assist in synthesizing local patterns into global, subglobal, and local rules.

In the synthesizing process, rules or patterns that are weakly present in a site and fail to meet the minimum support threshold are excluded from the procedure. However, this does not imply that the rule is entirely absent, as it may still hold some significance in the site with a support value between 0 and the minimum support. To address this issue, Ramkumar and Srinivasan [31] introduced the concept of a correction factor in the rule synthesizing process, resulting in improved synthesized results. They suggested that domain experts choose a suitable correction factor based on their knowledge and estimates of the data distribution. In cases where detailed knowledge about data distribution is lacking, they recommended using a correction factor of 0.50.

Adhikari et al. [32] proposed a model for mining global patterns in multiple transactional time-stamped databases, emphasizing the importance of identifying variations in item sales over time. They introduced the notion of item stability, considering stable items as valuable for making strategic decisions. Based on the degree of stability, they devised an algorithm for clustering different databases. Zhang et al. [33], on the other hand, proposed a method to obtain local patterns from individual databases using customer lifetime values (CLVs) computed from customerid, customer expenditure, and the period of the customer lifecycle. Global patterns were then synthesized using the forwarded local patterns through a method called kernel estimation for mining global patterns (KEMGP).

Adhikari et al. [34] recognized the significance of association analysis for select items in multiple market databases and proposed a model for mining global patterns of select items. They also introduced a measure of overall association between two items in databases and designed an algorithm based on this measure for grouping frequent items in multiple databases.

In existing rule synthesizing methods, a common assumption is that relevant analyses have been conducted among databases, and the databases under consideration are highly relevant. However, this assumption is unrealistic as it implies that all stores have the

same type of business with identical metadata structures. He et al. [35] addressed this problem by proposing a synthesizing model suitable for databases containing different items that may not be relevant to each other. Their two-step clustering-based rule synthesizing framework involves clustering at the item level for databases with different items and further clustering for databases sharing similar items but different rules. This process leads to final clusters containing both similar items and similar rules, which are then used for weighted rule synthesis using the method proposed by Wu and Zhang.

Table-1 provides a summary of the salient features of research work based on the local pattern analysis strategy.

V. CONCLUSION AND SCOPE FOR THE FUTURE WORK

Research in Multi-Database Mining (MDM) gains paramount importance, becoming both imperative and formidable, as the proliferation of multi-databases continues to surge [1]. This paper undertakes a comprehensive survey of diverse research endeavors within this burgeoning field, with particular focus on the domains of mono-mining and local pattern analysis [2]. Notwithstanding the advancements made in local pattern analysis, several challenges persist, warranting further research efforts to address them adequately [3].

5.1. Incorporation of Quantitative Information

In the Context of Data Source Weight Allocation, considering two sites, S1 and S2, with respective transaction populations of 100 and 1000, a transactions-population-based weighting model assigns a weight of 10 times that of S1 for S2. However, such an approach may not be equitable if the turnover of S1 surpasses that of S2. Thus, relying solely on transactions-population for site weight allocation might lead to suboptimal decisions. To enhance decision-making, quantitative mining based on turnover quantity or cost of items sold can be employed. For instance, a frequent rule such as Wine→Salmon (support = 10%, confidence = 80%) may hold greater significance than another frequent rule like Bread→Milk (support = 30%, confidence = 80%), despite the former having a lower support value. This distinction arises because the items in the first rule typically yield higher profits per unit sale compared to those in the latter rule. Consequently, incorporating quantitative information in the allocation of site weights becomes essential, necessitating a synthesizing model based on multiple minimum supports for the corresponding quantitative data.

5.2. Weight Assignment for Transactions in Data Analysis

One of the prospective avenues for future research lies in the assignment of transaction weights. In real-world datasets, different transactions possess varying degrees of importance. For instance, in market basket analysis, each transaction is associated with a profit value, wherein transactions with a substantial number of items hold greater significance than those containing only a few items. This necessitates the adoption of distinct weightings for different transactions to adequately reflect their respective importance.

The process of assigning weights to transactions can be accomplished by leveraging factors such as recency, frequency, monetary value, and duration (RFMD) values, a well-established technique commonly used in market segmentation. The RFMD approach considers key parameters, namely, customers' recent purchase activities (recency), the frequency of their purchases (frequency), the monetary value of their transactions (monetary value), and the duration of their engagement with the sellers' website (duration).

By ascribing appropriate weights to each parameter, the weighted score for each transaction can be computed. This method of transaction weighting offers a viable solution to the predicament of considering all transactions in the rule mining process, allowing for the elimination of less relevant transactions and elevating the significance of extracted rules. As a result, this approach holds promise for enhancing the robustness and relevance of the rule synthesis process in data mining applications.

5.3. Comprehensive Global Classification Model for Rule Synthesis

Supervised learning represents a widely recognized data mining methodology employed for the purpose of classifying data records into predefined sets of class labels. In this context, employing classification techniques such as decision trees to mine the attributes of local data sources pertaining to a specific concept or class becomes an intriguing research avenue. The synthesis of these mined local features to construct a cohesive global classification model for mining multiple databases adds further significance to this area of investigation.

5.4. Negative Association Rules Synthesis

Prospective research endeavors in the realm of data mining encompass the exploration of negative association rule mining across multiple databases. A negative association rule elucidates the relationship between item sets, signifying occurrences of certain item sets characterized by the absence of others [2]. Alongside the well-established positive association rules denoted by 'A→B,' corresponding negative associations exist, including 'A→¬B,' '¬A→B,' and '¬A→¬B'. These negative association rules assume a pivotal role in decision-making processes. For instance, when dealing with diverse medical databases stemming from disparate areas, the Center for Disease Control might seek to identify factors that are relatively irrelevant or wholly irrelevant despite their frequent occurrence. Consequently, there exists a promising avenue for research in formulating an effective synthesizing model for negative association rules.

VI. REFERENCES

[1] Right MW, Hurson AR, Pakzad SH. A taxonomy and current issues in multidatabase systems. *IEEE Comput* 1992, 25: 50–60.

[2] Ramkumar T, Srinivasan R. Modified algorithms for synthesizing high-frequency rules from different data sources. *Knowl Inf Syst* 2008, 17:313–334.

[3] Zhang S, Wu X, Zhang C. Multi-database mining, *IEEE Comput Intell Bull* 2003, 2:5–13.

[4] Zhang S, Chen Q, Yang Q. Acquiring knowledge from inconsistent data sources through weighting. *Data Knowl Eng* 2010, 69:779–799.
[5] Liu H, Lu H, Yao J. Identifying relevant database for multidatabase mining. In: *Proceeding of the Sec-ond Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Melbourne, Australia; 1998, 210– 221.
[6] Adhikari A, Rao PR. Synthesizing heavy association rules from different real data sources. *Pattern Recognit Lett* 2008, 29:59–71.
[7] Zhang S, Zaki JM. Mining multiple data sources: lo- cal pattern analysis. *Data Min Knowl Discov* 2006, 12:121–125.
[8] Turinsky K, Grossman R. A framework for finding distributed data mining strategies that are intermedi- ate between centralized strategies and in-place strate-gies. In: *Workshop on Distributed and Parallel Knowl-edge Discovery at Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*. Boston, MA; 2000, 1–7.
[9] Ramkumar T, Srinivasan R. Multi-level synthesis of frequent rules from different data sources. *Int J Com- put Theory Eng* 2010, 2:195–204.
[10] Lenca P, Meyer P, Vaillant B, Lallich S. On selecting interestingness measures for association rules: user ori- ented description and multiple criteria decision aid. *Eur J Oper Res* 2008, 184:610–626.
[11] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: *Pro-ceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. Washington, D.C.; 1993, 207–216.
[12] Agrawal R, Srikant R. Fast algorithms for mining asso-ciation rules. In: *Proceedings of the Twentieth Interna-tional Conference on Very Large Databases(VLDB)*. Santiago de Chile, Chile; 1994, 478–499.
[13] Adhikari A, Jain CL, Ramana S. Analysing effect of database grouping on multi-database mining. *IEEE In- tell Inf Bull* 2011, 12:25–32.
[14] Michalski RS, Kerschberg L, Kaufman KA, Ribeiro JS. Mining for knowledge in databases: the INLEN archi-tecture, initial implementation and first results. *J Intell Inf Syst: Integr AI and Database Technol* 1992, 1:85–113.
[15] Ribeiro J, Kaufman K, Kerschberg L. Knowledge dis- covery from multiple databases. In: *Proceedings of the First International Conference on Knowledge Discov-ery and Data Mining (KDD-95)*. Montreal, Canada; 1995, 240–245.
[16] Wrobel S. An algorithm for multi-relational discovery of subgroups. In: *Proceedings of the First European Symposium on Principles of Data mining and Knowl- edge Discovery*. Trondheim, Norway; 1997, 78–87.
[17] Aronis J, Kolluri V, Provost F, Buchanan B. The WoRLD: knowledge discovery from multiple dis- tributed databases. In: *Proceedings of the Tenth In- ternational Florida AI Research Symposium*. Daytona Beach, FL; 1997, 337–341.
[18] Grossman RL, Bailey S, Ramu A, Malhi B, Turinsky A. The preliminary design of papyrus: a system for high performance, distributed data mining over clusters. *In:Advances in Distributed and Parallel Knowledge Dis-covery*. Menlo Park, CA: AAAI/MIT Press; 2000, 259–275.
[19] Prodromidis A, Chan P, Stolfo S. Meta-learning in dis- tributed data mining systems: issues and approaches. In: *Advances in Distributed and Parallel Knowledge Discovery*. Menlo Park, CA: AAAI/MIT Press; 2000.
[20] Kargupta H, Huang W, Sivakumar K, Johnson E. Dis- tributed clustering using collective principal compo- nent analysis. *Knowl Inf Syst* 2001, 3:422–448.
[21] Kargupta H, Huang W, Sivakumar K, Park B,Wang S. Collective principal component analysis from dis-tributed, heterogeneous data. In: *Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery*. Lyon, France; 2000, 452–457.
[22] Savasere A, Omiecinski E, Navathe S. An efficient algo-rithm for mining association rules in large databases. In: *Proceedings of the Twenty First International Con- ferences on Very Large Data Bases*. Zurich, Switzer- land; 1995, 432–444.
[23] Zhong N, Yao YY, Ohshima M. Peculiarity oriented multi-database mining. *IEEE Trans Knowledge Data Eng* 2003, 15:952–960.
[24] Liu H, Lu H, Yao J. Toward multi-database min- ing: identifying relevant databases. *IEEE Trans Knowl Data Eng* 2001, 13:541–553.
[25] Wu X, Zhang C, Zhang S. Database classification for multi-database mining. *Inf Syst* 2005, 30:71–88.
[26] Zhang S, Zhang C, Wu X. *Knowledge Discovery in Multiple Databases*. London: Springer-Verlag; 2004.
[27] Wu X, Zhang S. Synthesizing high-frequency rules from different data sources. *IEEE Trans Knowl Data Eng* 2003, 15:353–367.
[28] Nedunchezhian R, Anbumani K. Post mining– discovering valid rules from different sized data sources. *Int J Inf Technol* 2006, 3:47–53.
[29] Zhang C, Liu M, Nie W. Identifying global exceptional patterns in multidatabase mining. *IEEE Comput Intell Bull* 2004, 3:19–24.
[30] Kum HC, Chang JH, Wang W. Sequential pattern min-ing in multidatabases via multiple alignment. *Data Min Knowl Discov* 2006, 12:151–180.
[31] Ramkumar T, Srinivasan R. The effect of correction factor in synthesizing global rules in a multi-database mining scenario. *J Appl Comput Sci* 2009, 3:33– 38.
[32] Adhikari J, Rao PR, Adhikari A. Clustering items in different data sources induced by stability. *Int Arab J Inf Technol* 2009, 6:394–402.
[33] Zhang S, You X, Jin Z, Wu X. Mining globally in-teresting patterns from multiple databases using kernel estimation. *Expert Syst Appl* 2009, 36:10863–10869.
[34] Adhikari A, Rao PR, Pedrycz W. Study of select items in different data sources by grouping. *Knowledge Inf Syst* 2010, 27:23–43.

[35] He D, Wu X, Zhu X. Rule synthesizing from multiple related databases. In: *Proceedings of the Fourteenth Pacific-Asia Conference on Knowledge Discovery andData Mining*. Hyderabad, India; 2010, 201–213.
[36] Zhang S, Wu X. Fundamentals of association rules indata mining and knowledge discovery. *WIREs Data Min Knowl Discov* 2011, 1:97–116.