International Journal of Computer Theory and Engineering, Vol. 2, No. 2 April, 2010
1793-8201

Multi-Level Synthesis of Frequent Rules from Different Data-Sources

Thirunavukkarasu Ramkumar[1], Rengaramanujam Srinivasan[2]

***Abstract*—An interstate business organization may have a large number of transactions in each of its branches operating at different locations. In such situations, for making effective head-quarter decisions, multi-database mining using local pattern analysis has been considered as an efficient strategy. During this process, individual data sources are mined and discovered patterns are forwarded to the head branch. To reap meaningful patterns from the large number of forwarded patterns, a comprehensive synthesizing process is a necessity. Earlier we had proposed a synthesizing model to synthesize global rules from high-frequent rules, in which weights are based on the transaction - population of sites. In this paper, we are proposing a synthesizing model for multi-level synthesis of local patterns on the basis of two, rule-selection measures - namely effective and nominal vote rates. Using these rule selection measures, synthesized patterns are classified into groupings of global, sub-global and local rules. With this, not only high-frequent rules but also frequent rules are taken care of and synthesized into appropriate set of sub-global rules. Examples and experimental results presented clearly establish the validity of the proposed model in meeting the requirements of multi-level rule synthesizing strategy.**

***Index terms* - Multi-Databases, Multi-level rule synthesis, Local pattern analysis, Effective vote rate, Nominal vote rate**

I. Introduction

Advances in information and communication technologies demand various new perceptions in industrial growth and business activities. In practice, the required data for corporate decisions are spread over multiple branches, which are physically located in different regions. In such situations, traditional data mining techniques based on data warehousing have serious limitations like high investment on software, hardware and difficulties in finding the individuality of local patterns, apart from the necessity of moving huge quantity of data to the central data warehouse. The problem could be overcome by using Multi-Database mining based on *local* pattern analysis. *Multi-database mining can be defined as the process of mining data from multiple data sources, which may be heterogeneous, and finding novel and useful patterns of significance.* Patterns can be discovered using various data mining functionalities in which association analysis receives significant attention among KDD community. In multi-database mining using local instance analysis, patterns are mined from individual data sources by various data mining tasks such as association analysis and the mined patterns are forwarded to the head quarter.

Since the number of rules forwarded from individual data sources is high, a synthesizing process is necessary to find out the interesting ones. We have advocated a weighting model [5] for synthesizing high frequent rules from different data sources. In our weighting model, rule-weight is assigned on the basis of cumulative score of transaction - populations of data sources where the rule is present and in turn, weight of any data source is calculated on the basis of its own transaction population. Synthesized global support and confidence values of a rule is obtained by using the data source weights, local support and confidence values where the rules are present. In this paper, we present a multi-level synthesizing strategy for synthesizing local patterns by introducing two interesting rule evaluation measures, namely *effective* and *nominal* vote rates. $\gamma_{effective}$ is defined as the *effective* vote rate, which is the cumulative percentage of votes received from different data sources for a given rule on the basis of the transaction population of respective data sources.$\gamma_{nominal}$ is defined as the nominal vote rate, which is the cumulative percentage of vote received from different data sources for a given rule on the basis of *equal* votes for sites. Using these rule selection measures, local patterns are synthesized in multi-level perspectives. *Multi-Level synthesis of frequent rules can be defined as the process of synthesizing frequent rules from different data sources at multiple levels of abstraction using effective and nominal vote rates to form global, sub-global and local rules.* Thus the process of multi-level synthesis is comprehensive, since all rules are accounted for in one way or other. Global rules are meant for head branch of an interstate company to make decisions; predictably sub-global rules may help regional lead branches while local rules will be helpful to the respective local branches in making effective decisions.

The rest of the paper is organized as follows: Related research works in Multi-database mining are discussed in section 2; Section 3 reveals the proposed approach for multi-Level rule synthesis with an example; experimental results are furnished in the section 4; section 5 summarizes the conclusions.

II. Related Research Work

Multi-Database mining has been recognized as an important research topic among KDD community. Zhang et al. [10] in their paper, discuss the limitations of traditional techniques in mining multiple databases. They have developed a new method for mining multi databases; based on the method they have divided the patterns in multi databases into local patterns, high-vote patterns, exceptional patterns and suggested patterns. Liu, Lu and Yao [4] focus on the identification of databases that are most relevant to an application. They have proposed a relevance measure to identify relevant databases for mining with an objective to

[1]Thirunavukkarasu Ramkumar is with the Faculty of Computer Applications, AVC College of Engineering, Tamilnadu, India. (email: ramooad@yahoo.com)

[2]Rengaramanujam Srinivasan is with the School of Computer and Information sciences, BSA Crescent University, Tamilnadu, India. (email: drrsrs@yahoo.com)

find patterns or regularities with in certain attributes. Their work can be considered as first step towards multi-database mining. Xindong Wu [6] etal. deal with the problem of classifying multiple databases by providing application independent relevance measure. They have proposed an algorithm for the best classification of relevant databases. Zhong et al. [8] have proposed a way for mining peculiarity patterns from multiple statistical and transactional databases. They have developed a framework for peculiarity-oriented mining in multiple data sources.

Wu and Zhang [7] have advocated a weighting model for synthesizing high-frequency association rules or patterns from different data sources on the basis of data source weights. They calculate data source weight based on the number of high-frequency rules supported by it and they calculate rule-weight as being proportional to the number of data sources supporting the rule. They have also presented a synthesizing model for clustering of association rules that are extracted from different *unknown* data sources. As an extension of the above work Ramkumar and Srinivasan [5] proposed a transaction- population based weighting model for synthesizing high-frequency rules from different data sources. According to them, rule weight is proportional to the sum of the weights of the data sources supporting the rule. The weight of any data source is calculated based on the population of data sources – i.e., by the number of transactions in the database. Their goal in synthesizing local patterns to obtain global pattern is that, the support and confidence of synthesized pattern should be very nearly same if all data sites were integrated and mono-mining has been done. Chengi et al. [9] developed a model for identifying exceptional patterns from multiple databases, which could be considered as a post processing work after mining multiple relevant databases. Kum et al. [3] have presented a novel algorithm, *Approxmap, to* mine approximate sequential patterns called consensus patterns from large sequence databases in two steps. Firstly, sequences are clustered by similarity. Then consensus patterns are mined directly from each cluster through multiple alignments.

III. Proposal for Multi-Level Synthesis of Rules

Most of the current synthesizing models focus on synthesizing high-frequency rules only [5,7] since they emerge as global rules when all the data sources are integrated. High-frequency rules are truly valid in making global decisions by the head branch of any interstate company. However in such a synthesis, regional patterns or rules get eliminated. For making decisions, say at regional levels, patterns that show the individuality of regions or cluster or groups of branches become important. They can be explored by synthesizing them in a multi-level perspective alone. Hence it is necessary to adopt a strategy for synthesizing local patterns from individual data sources at multiple levels of abstraction. Weighting method is a commonly used approach for rule synthesizing. In our multi-level rule-synthesizing model, weight of the data source is calculated on the basis of their transaction-population. We have proved in our previous paper [5] that transaction population based weighting model is a necessity in order to ensure that synthesized results tally with mono-mining results. The Multi-Level synthesizing problem can be formulated as follows: "There are m sites: $S_1, S_2,\ldots,S_m$. Each site or data source support a set of patterns or rules called local patterns or local rules which are of the form A→B {supp,conf}. It is proposed to get synthesized rules from different data sources at multiple levels of abstraction using effective and nominal vote rate-measures. It is assumed that transaction-population of data sources are known. Our goal in the synthesis is that synthesized results must be close to mono-mining values that could be obtained, if the data in corresponding databases are put together and then mining has been done"

The Multi-level rule synthesis process consists of the following steps: i) obtaining the weights of data-sources based on their respective transaction population ii) calculating effective and nominal vote rate-measures; $\gamma_{effective}$, $\gamma_{nominal}$ - of patterns or rules presented by the data sources. iii) classifying the rules into candidate global rules, candidate sub-global rules and local rules on the basis of these two measures. iv) calculating synthesized support and confidence of rules to form global, sub-global and local rules based on the rule groupings. The following expressions formally define our proposed approach.

A. Normalized site weights

Let $S_1,S_2,\ldots,S_m$ be the set of relevant databases taking part in the synthesizing process. $w'_1, w'_2,\ldots,w'_m$ are the weights corresponding to the transaction population of the data sources. In general, $W'j$ is the un-normalized site weight of site j, based on transaction population. Normalized weight of a site is the ratio between transaction population in respective data-sources and total transaction population of all participating sites.

$$\text{Normalized weight of Site j} = W_j = \frac{W'_j}{\sum_{j=1}^{m} W'_j} \quad -- \quad (1)$$

B. Effective and Nominal Vote rates

Effective vote rate is represented as $\gamma_{effective}$, which is the percentage of votes received from different data sources for a given rule on the basis of the transaction population of corresponding data sources. Nominal vote rate is defined as $\gamma_{nominal}$, which is the percentage of vote received from different date sources for a given rule on the basis of equality of votes for sites. For any rule R_i, $\gamma_{effective}$, $\gamma_{nominal}$ are given by:

$$\gamma_{effective}(R_i) = \sum_{j=1}^{m} \delta(i,j) * W_j \quad -- \quad (2)$$

$$\gamma_{nominal}(R_i) = \frac{1}{m} * \sum_{j=1}^{m} \delta(i,j) \quad -- \quad (3)$$

$\delta(i, j) = 1$ if R_i is present in site j otherwise

$$\delta(i, j) = 0 \quad -- \quad (4)$$

C. Rule Selection

Let S' be set of all rules of the form A→ B {supp,conf}.

$$S' = S_1' \cup S_2',\ldots,\cup S_m' \quad -- \quad (5)$$

Clearly S' is the union of set of rules presented by all the

sites put together. The number of rules so generated may have a huge value. We would like to reduce the synthesizing effort by segregating rules in $S^{|}$ into three groups; candidate global rules, candidate sub-global rules and local rules. Candidate global rules are those rules that have a potential of becoming global rules and are selected by using γeffective, the effective vote –rate as a gating measure. Thus candidate global rules have $\gamma_{effective}$ >= min.$\gamma_{effective}$. This raises the question what should be $\gamma_{effective}$ threshold limit. We suggest 0.50 as a lower level limit of min.$\gamma_{effective}$. With this limit, some of the candidate global rules may not be confirmed as global rules since they fail to pass the tests of min.support and min.confidence values. If we adopt two-thirds majority approach and keep min.$\gamma_{effective}$ = 0.67, most of the candidate global rules are likely to be conferred global status.

Candidate sub-global rules are those, which fail to satisfy min. $\gamma_{effective}$ but satisfy min. $\gamma_{nominal}$ value. The min. $\gamma_{nominal}$ value is chosen by the enterprise based up on the minimum number of sites required to form a cluster and sub-global rule. Clearly it depends on such parameters like total number of participating sites, their geographical distribution etc. Further during processing, such of the those candidate global rules, which fail to satisfy min. support and min. confidence limits, will also be added to candidate sub global rule kitty.

The rest are local rules that exist only at a single site or at very few sites insufficient to form a cluster.

D. Local support and confidence of a rule in a site

Let R_i be the rule in the form of *antecedent → consequent.* Local support and local confidence of a rule are the support and confidence values obtained in the respective sites. Local support of a rule R_i at site j is represented as $Supp_j(R_i)$. Local confidence of a rule R_i at site j is represented as $Conf_j(R_i)$. Support for antecedent of Rule R_i at site j is represented as $Supp\_ante_j(R_i)$. To find the interesting rules,

Support (*antecedent → consequent*) ≥ *min. Support*
Confidence (*antecedent → consequent*) = Support (*antecedent → consequent*) / Support (*antecedent*) ≥ *min. Confidence*

Where *min.Support* and *min.Confid*ence are user specified thresholds. [1,2]

E. Expressions for synthesized global support and confidence values

The global support of a rule R_i can be defined as the support obtained for a rule R_i from all the sites put together. In our synthesizing model, global support of R_i is calculated as:

$$Supp_G(R_i) = \sum_{j=1}^{m} W_j * Supp_j(R_i) \quad -- \quad (6)$$

Similarly, the global confidence of a rule R_i can be defined as the confidence obtained for the rule R_i from all the sites. Global support of an antecedent of a rule R_i can be defined as the support of antecedent of R_i in all the sites put together. To find global support of antecedent of R_i,

$$Supp\_ante_G(R_i) = \sum_{j=1}^{m} W_j * \frac{Supp_j(R_i)}{Conf_j(R_i)} \quad -- \quad (7)$$

Expression 7 is valid in the case of a global rule, which is supported by all sites; however it has to be amended to take care of the situation when a global rule is not being supported at some sites. Suppose we have a rule R_i, A → B that is not supported by site j which means support for R_i < min.supp at site j. Does that mean $Supp\_ante_j(R_i)$ < min.supp? Clearly the answer is in the negative. In order to ensure that our synthesized results tally exactly with mono-mining results, we have to find support for antecedent of R_i (support for A) at site S_j. There are two ways to capture this information. The site S_j may have some other rule R_k which has the same antecedent as R_i..

If antecedent (R_i) = antecedent (R_k) at site S_j then

$$Supp\_ante_j(R_i) ? = Supp\_ante_j(R_k) = \frac{Supp_j(R_k)}{Conf_j(R_k)} \quad -- \quad (8)$$

In case there is no such rule R_k with the same antecedent we can get lower level meta data corresponding to lower level item sets also forwarded from respective sites along with the rules. The required information can be captured from them. To find synthesized global confidence, we evaluate each support value separately and then calculate the ratio.

$$Conf_G(R_i) = \frac{Supp_G(R_i)}{Supp\_ante_G(R_i)} \quad -- \quad (9)$$

F. Expressions for synthesized sub-global support and confidence values

The expressions for sub-global support and confidence values are given below. They are similar to global values but the summations are restricted to sub grouped sites only.

$$W_{SG}(R_i) = \sum_{\forall jinclass-list} W_j \quad -- \quad (10)$$

$$Supp_{SG}(R_i) = \frac{1}{W_{SG}} \sum_{\forall jinclass-list} W_j * Supp_j(R_i) \quad -- \quad (11)$$

$$Supp\_ante_{SG}(R_i) = \frac{1}{W_{SG}} \sum_{\forall jinclass-list} W_j * \frac{Supp_j(R_i)}{Conf_j(R_i)} \quad -- \quad (12)$$

$$Conf_{SG}(R_i) = \frac{Supp_{SG}(R_i)}{Supp\_ante_{SG}(R_i)} \quad -- \quad (13)$$

Note : The above expressions are applicable to the groups as given below. (i) For SGR-II, the summation will be corresponding to all the sites j where R_i is present.(ii) For SGR-III, Sub-global confidence corrections are to be made similar to global rules.

(SGR - I) :- Grouping under standard label

In this classification, features of the patterns are captured using existing, standard groupings. Groupings are labeled on the basis of the general characteristics of the patterns. For example, groupings are labeled based upon region (west, east, north, south) or based upon type (metropolis, suburb) etc. Patterns, which exactly match with the existing classifier label constitute sub-global rule set under SGR-I.

(SGR-II):- Classification to be labeled by the domain expert

Patterns that do not comply with general characteristics of the existing standard groupings are synthesized and put into SGR–II group. Suitable class labels are given to the new groupings by the domain expert.

(SGR - III) :-Sub-global rules under standard classifiers with reduced support

In this grouping, sub-global rules, which do not exactly match with the corresponding class label lists, are synthesized on the basis of the effective vote rate of the classifiers. Rules whose effective vote rate meets the minimum effective vote rate of the classifier will form the candidate sub-global rule set. After synthesizing, candidate sub-global rules whose synthesized support and confidence values meets the *min.support and min.confidence* thresholds are put in SGR-III. These set of rules will be found in SGR-II also; the support at SGR- III grouping will be lower as compared to support at SGR-II. It is for the domain expert to decide which grouping is ultimately to be adopted. To completely capture the full significance of any synthesized pattern, we recommend representing any synthesized pattern by a sixtuple (*Pattern, effective vote rate, nominal vote rate, classifier label, synthesized support, synthesized confidence*). The entire procedure has been given as a flowchart in Fig.1. The figure is self-explanatory. We illustrate the multi-level rule synthesis procedure with two examples. Example 1 demonstrates rule selection procedure while the complete multi-level rule synthesis process is presented in Example 2.

Example 1*:* Let us consider five sites; S1, S2, S3, S4 and S5 having transaction populations of 10000, 5000, 4000, 3000 and 2000 respectively. Rule R1 is supported by S1, S2, S3, S4 and S5.Rule R2 is supported by S2, S3 and S4. Rule R3 is supported S3, S4 and S5.Rule R4 is supported S2, S3 and S4.Rule R5 is supported by S1, S2 and S3 and Rule R6 is supported by S4 alone. Effective and nominal vote rates of the rules are calculated as follows:

Rule R1

Effective vote rate $= \gamma_{effective}(R1) = (10{,}000 + 5{,}000 + 4{,}000 + 3{,}000 + 2{,}000)/24{,}000 = 1$

Nominal vote rate $= \gamma_{nominal}(R1) = 5/5 = 1$

Rule R2

$\gamma_{effective}(R2) = (5{,}000 + 4{,}000 + 3{,}000)/24{,}000 = 0.50$

$\gamma_{nominal}(R2) = 3/5 = 0.60$

Rule R3

$\gamma_{effective}(R3) = (4{,}000 + 3{,}000 + 2{,}000)/24{,}000 = 0.37$

$\gamma_{nominal}(R3) = 3/5 = 0.60$

Rule R4

$\gamma_{effective}(R4) = (5{,}000 + 4{,}000 + 3{,}000)/24{,}000 = 0.50$

$\gamma_{nominal}(R4) = 3/5 = 0.60$

Rule R5

$\gamma_{effective}(R5) = (10{,}000 + 5{,}000 + 4{,}000)/24{,}000 = 0.79$

$\gamma_{nominal}(R5) = 3/5 = 0.60$

Rule R6

$\gamma_{effective}(R6) = 3{,}000/24{,}000 = 0.12$

$\gamma_{nominal}(R6) = 1/5 = 0.20$

Let $\min.\gamma_{effective} = 0.50$ and $\min.\gamma_{nominal} = 0.30$, Focusing our attention only to $\min.\gamma_{nominal}$ and $\min.\gamma_{effective}$ thresholds, the three groupings are $\gamma_{effective} >= 0.50$ Grouping 1 : Candidate Global Rules: R1,R2,R4,R5 $\gamma_{effective} < 0.50$ & $\gamma_{nominal} >= 0.30$ Grouping 2 : Candidate Sub-Global Rules: R3 and Grouping 3: Local Rules: R6.

Example 2: To explain the multi-level synthesis process on the basis of the proposed approach, local rules, their support and confidence values corresponding to 10 sites are given in Table 1. The minimum support and confidence thresholds are 0.20 and 0.50 and minimum effective and minimum nominal vote rates are chosen as 0.50 and 0.20 respectively.

The effective and nominal vote rates for each of the rules is calculated as in Example1 and rules are classified into the following groups:

Group 1: Candidate Global Rules: **R1**: A→ B, **R2**: A→ C

Group 2: Candidate Sub- Global Rules: **R3**: AB → D, **R4**: CD→E, **R5**: AC→F, **R6**: AD → G, **R9**: CF → J, **R10**: AB → G

Group 3: Local Rules: **R7**: G → H, **R8**: B → H

The calculation of synthesized support and confidence values for few of the above rules are illustrated below.

Rule R2 : A→ C

Rule R2 is supported by the sites S1, S2, S3, S4, S5, S6, S7 and S8.

$Supp_G(R2) = 0.16 * 0.30 + 0.12 * 0.35 + 0.12 * 0.26 + 0.08 * 0.30 + 0.08 * 0.22 + 0.04 * 0.28 + 0.16 * 0.35$

$= +0.04 * 0.30$

$= 0.2420$

R2 is not supported by the sites S9 & S10.However antecedent of (R2) = antecedent of (R1) = A; Hence $Supp\_ante_G(R2)$ is calculated as

$Supp\_ante_G(R2) = 0.16 * (0.30/0.45) + 0.12 * (0.35/0.76) + 0.12 * (0.26/0.43) + 0.08 * (0.30/0.51) + 0.08 * (0.22/0.44) + 0.04 * (0.28/0.73) + 0.16 * (0.35/0.72) + 0.04 * (0.30/0.50) + 0.12 * (0.45/0.75) + 0.08 * (0.35/0.70) = 0.5503$

$Conf_G(R2) = 0.2420 / 0.5503 = 0.4397$

Rule R3: AB → D

Rule R3 is supported by the south region: S1, S2 & S3 and emerges as sub-global rule for the region south (SGR-I)

$Supp_{SG}(R3) = 0.4 * 0.40 + 0.3 * 0.35 + 0.3 * 0.45$

$= 0.40$

$Supp\_ante_{SG}(R3) = 0.4 * (0.40/0.70) + 0.3 * (0.35/0.50) + 0.3 * (0.45/0.60) = 0.6635$

$Conf_{SG}(R3) = 0.40 / 0.6635 = 0.6028$

Rule R10: AB → G

Rule R10 is supported by sites S1 and S2 of south region. It does not exactly match with c (j, k) of south region class label. Because two of the three sites of south region only support R10, Rule R10 will be synthesized in two ways.

(I) SGR-II

$Supp_{SG}(R10) = 0.5714 * 0.40 + 0.4285 * 0.50 = 0.4428$

$Supp\_ante_{SG}(R10) = 0.5714 * (0.40/0.70) + 0.4285 * (0.50/0.71) = 0.6282$

$Conf_{SG}(R10) = 0.4428 / 0.6282 = 0.7048.$

(II) SGR - III

$\gamma_{effective}$ of the south region classifier = $\gamma_{eff.\ South} = 35{,}000 / 50{,}000 = 0.7 \geq$ min. effective

$Supp_{SG}$ (R10) = 0.4 * 0.40 + 0.3 * 0.50 = 0.31 $Conf_{SG}$ (R10) = 0.31 / 0.6648 = 0.4663.
$Supp\_ante_{SG}$(R10) = * (0.40/0.70) + 0.3 * (0.50/0.71)
+ 0.3 * (0.45 / 0.60) = 0.6648

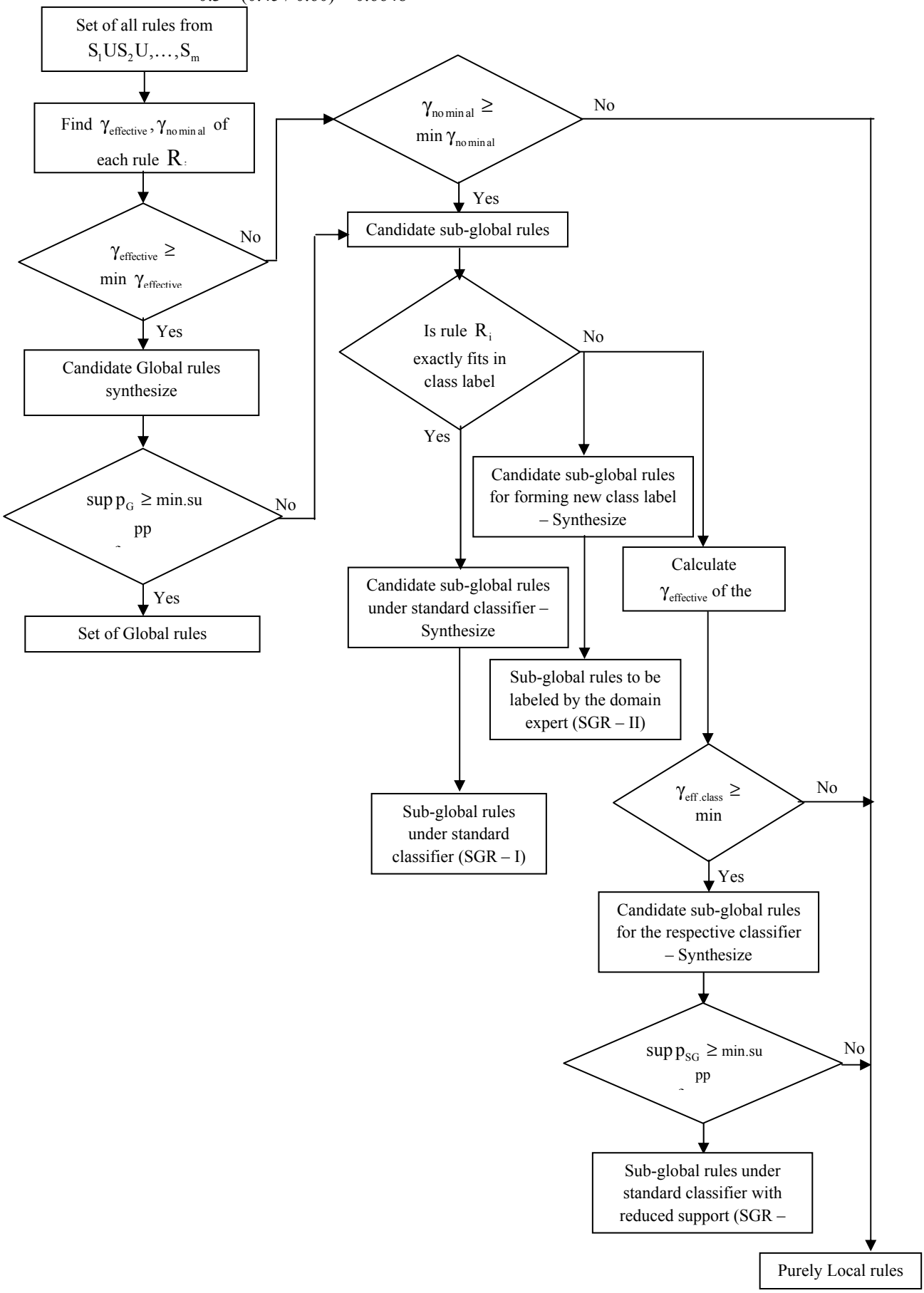

Fig. 1 - The Multi-Level rule synthesis model

Table 1 Data for Example1

Region: South						
Rule / Pattern	Site S1 (20,000)		Site S2 (15,000)		Site S3 (15,000)	
	Sup	Conf	Sup	Conf	Sup	Conf
A→ B	0.40	0.60	0.30	0.65	0.45	0.75
A→C	0.30	0.45	0.35	0.76	0.26	0.43
AB→ D	0.40	0.70	0.35	0.50	0.45	0.60
AC→F	0.40	0.70				
AD → G	0.35	0.65				
B → H					0.28	0.58
AB→G	0.40	0.70	0.50	0.71		
Region: North						
Rule / Pattern	Site S4 (10,000)		Site S5 (10,000)		Site S6 (5,000)	
	Sup	Conf	Sup	Conf	Sup	Conf
A→ B	0.35	0.60	0.30	0.60	0.25	0.65
A→ C	0.30	0.51	0.22	0.44	0.28	0.73
AC→ F			0.25	0.50		
AD → G	0.40	0.55				
G → H			0.40	0.70		
Region: West						
Rule / Pattern	Site S7 (20,000)		Site S8 (5,000)			
	Sup	Conf	Sup	Conf		
A→ B	0.28	0.58	0.34	0.56		
A→ C	0.35	0.72	0.30	0.50		
CD → E	0.45	0.80	0.30	0.65		
AC → F	0.35	0.55				
Region: East						
Rule / Pattern	Site S9 (15000)		Site S10 (10,000)			
	Sup	Conf	Sup	Conf		
A→ B	0.45	0.75	0.35	0.70		
CD → E	0.40	0.55	0.30	0.60		
AC → F			0.25	0.55		
CF →J	0.26	0.56	0.40	0.60		

Therefore, R10 will be included in both Sub-global rule categories SGR-II and SGR-III. In such cases, synthesized support in SGR-III will be lower than SGR-II. Appropriate classification will be determined by the domain expert. Analysis of synthesized support and confidence values of all rules are presented in Table 2.

Tallying with Mono-mining Result values:

Our goal in multi level synthesizing process is that the synthesized support and confidence values must tally with mono mining results obtainable if all the site data in the case of global rules and the relevant site-data-in the case of sub global rules are put together and then mining had been done. By adopting transaction population based site weights and rule weights we have ensured that exact matching will occur in most of the cases. For example in the synthesized results presented in Table 2, values corresponding to rules R1, R3, R4, R5, R6, R7, R8, R9 and R10 (under SGR II) are values that match exactly with mono mining results. Only results corresponding to rule R2 and rule R10 (under SGR III) classification requires further attention. Let us consider rule R2 first and consider global support values. The synthesized support value = 0.2420 where as the values obtained from mono – mining is 0.2652. The deviation occurs because R2 is a global rule supported by 8 sites only and is not supported by sites S9 and S10. If the support at sites S9 and S10 were to be zero at each of these two sites then mono-mining and synthesized results would tally exactly. However actual support at S9 and S10 for R2 is 0.14 and 0.08 (< min. supp = 0.20). When mono mining is done these two sites contribute 0.14 and 0.08, while in synthesis process their contributions are assumed to be zero. We can handle the situations in two ways. In all such cases the synthesized support values will be marginally lower then mono-mining support values. We can accept synthesized result as a conservative estimate and ignore the deviation. Otherwise we can assume that support for rule Ri at sites which has failed to pick up min.support is some value between 0 and h* min.supp value where h is a correction factor such that $0 \leq h \leq 1$. Clearly 0.50 is a convenient figure for h. With this correction, the synthesized mono-support = 0.2620. Similar correction can be applied to global confidence values. The results are available in Table 3. We find with corrections, synthesized values tally closely with mono-mining results.

IV. EXPERIMENTS

We have conducted several studies to test the efficiency of the proposed method. The results of one of the studies are presented here. To carry out the multi-level synthesis process, nine artificial transactional databases namely S1, S2, S3, S4, S5, S6, S7, S8 and S9 have been generated using a database generator. Their populations are 5000, 7000, 8000, 15000, 20000, 25000, 20000, 40000 and 20000 respectively. The number of items = 26. The average length of transactions in the data items is approximately 12.23. The distribution of data items in the respective data set is not uniform. Using standard apriori algorithm, frequent patterns are mined from the individual data sources with minimum thresholds for support and confidence of 0.20 and 0.40. Mono-mining has also been done with the union of all data sources, ie with 1,60,000 records, while keeping the same minimum support and confidence thresholds. To reap the global, sub-global and local patterns, we have kept effective and nominal gating measures as 0.50 and 0.20 respectively.

A. Synthesizing global patterns

During the rule synthesis process, 45 patterns are selected as candidate global rules since they meet the stipulated minimum effective vote rate of 0.50.Out of these, 15 patterns are supported by all the nine sites. Their synthesized support and confidence values along with mono-mining values is shown in Table 4. From table 4, it can be observed that synthesized values exactly tally with mono-mining results. Among the remaining 30 candidate global patterns, 9 patterns fail to get the global status since their synthesized support and confidence are below threshold values. By the proposed approach, these failed candidate global patterns are reverted as candidate sub-global patterns in the second level of synthesis process. The remaining 21 patterns are not supported by all the nine sites. From Table 5 (a), we find, out of 21 patterns the synthesized results exactly tally with mono-mining values for 15 patterns.

This is rather fortuitous since in these cases, where rules are not supported by certain sites, actual support at the

Table 2 Analysis of results

Rule / Pattern	Voting Measures		$Supp_G$	$Conf_G$	Classifier Label	Remarks
	γ_{eff}	γ_{nom}				
A → B	1	1	0.3564	0.6440	Global rule	Supported by all the sites
A → C	0.80	0.80	0.2420	0.4397	Global rule	Supported by eight sites
AB → D	0.40	0.30	0.4000	0.6028	Sub global rule for south region	Sub-global rule under standard classifier
CD → E	0.40	0.40	East 0.3600 West 0.4200 Over all 0.3900	East 0.5657 West 0.7744 Over all 0.6618	Sub global rule for east and west region	Sub –global rule under standard classifier
AC → F	0.48	0.40	0.3332	0.5934	Sub global rule	To be labeled by the domain expert
AD → G	0.24	0.20	0.3666	0.6096	Sub global rule	To be labeled by the domain expert
G → H	0.08	0.10	0.4000	0.7000	Local rule	Purely local rules
B → H	0.12	0.10	0.2800	0.5800	Local rule	Purely local rules
CF → J	0.20	0.20	0.3160	0.5796	Sub global rule for east region	Sub-global rule under standard classifier
AB → G	0.28	0.20	0.4428	0.7048	Sub global rule	To be labeled by the domain expert
AB → G	0.28	0.20	0.3100	0.4663	Sub global rule	Sub global rule under standard classifier with reduced support

Table 3 Comparison with mono mining results

Rule / Pattern	Classifier Label	$Supp_G$		Mono-Supp	$Conf_G$		Mono- Conf
		Without Correction	With Correction		Without Correction	With Correction	
A → C	Global Rule	0.2420	0.2620	0.2652	0.4397	0.4761	0.4819
AB → G	Sub Global Rule-SGRIII	0.3100	0.3400	0.3196	0.4663	0.5115	0.4807

sites happened to be very nearly zero. Table 5 (b) presents results with correction factor h = 0.25 and Table 6 presents error analysis. We find that, with correction there is an increase in error rates for support while there is a reduction in error rates for confidence. Actual value of the correction factor has to be decided based upon the intuitive knowledge about data distribution. Possibly for the present study we could have even dispensed with the correction factor.

B.Synthesizing Sub-global Patterns

In the previous section, we had focused attention on synthesizing high-frequent global rules. Let us now direct our attention to candidate sub-global rules. For these we have proposed three groupings SGR-I, SGR-II and SGR-III. SGR-I grouping corresponds to already existing classification. The existing multi-level classification of sites for the current study is presented in Fig.2.Thus there are three classes viz. region, zone and location. Each class is again having corresponding sub-classes.

Table 7 represents the patterns which are exactly fitting with standard groupings and their synthesized values exactly matches with mono- mining results. In Table 7 pattern id 158 to 167 belongs to continental classifier label. Pattern 49 and 50 constitutes north zone where as pattern 104 and 105

Table 4 Synthesized global patterns supported by all the sites

P_ id	Pattern	$Supp_G$	Mono-Supp	$Conf_G$	Mono-Conf
5	F→C	0.3351	0.3351	0.6624	0.6624
6	C→E	0.3684	0.3684	0.6339	0.6339
7	C→D	0.5383	0.5383	0.9262	0.9262
21	F→B	0.3424	0.3424	0.6767	0.6767
22	B→E	0.3772	0.3772	0.6578	0.6578
23	F→E	0.5060	0.5060	1.0000	1.0000
24	B→D	0.3556	0.3556	0.6201	0.6201
25	B→C	0.3746	0.3746	0.6532	0.6532
37	F→A	0.3424	0.3424	0.6767	0.6767
38	A→E	0.3772	0.3772	0.6310	0.6310
39	F→D	0.3361	0.3361	0.6642	0.6642
40	A→D	0.3556	0.3556	0.5949	0.5949
42	E→D	0.3696	0.3696	0.6344	0.6344
43	C→A	0.3746	0.3746	0.6446	0.6446
44	B→A	0.5734	0.5734	1.0000	1.0000

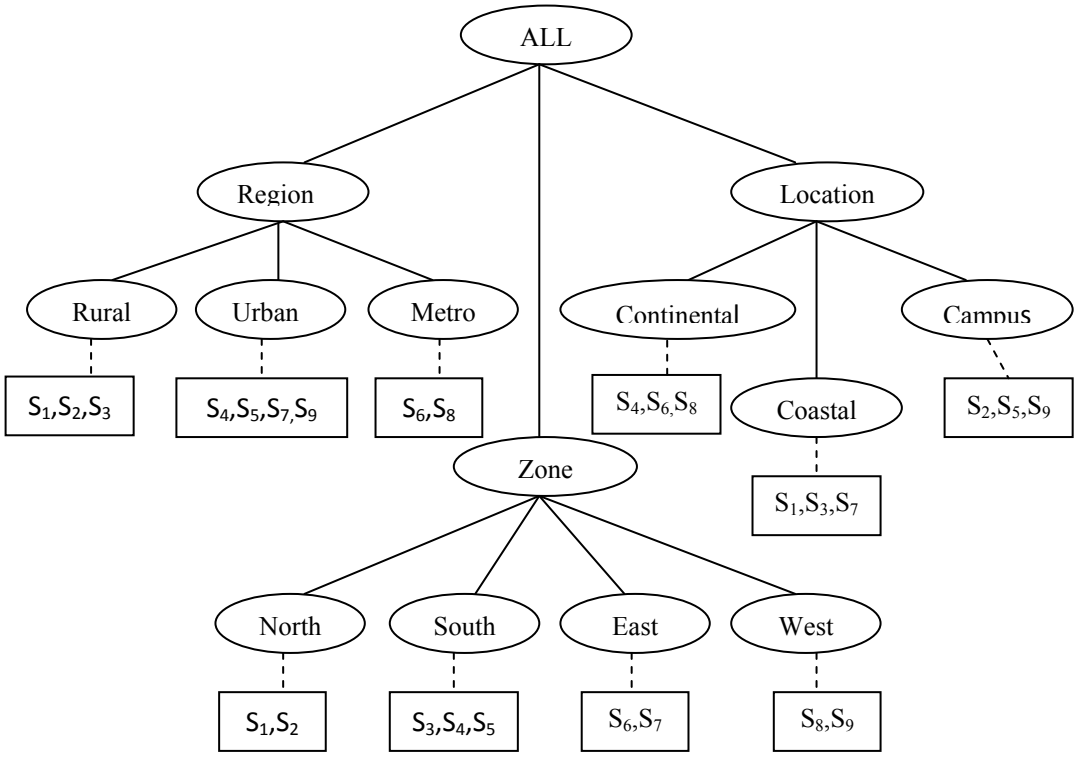

Fig.2 The Multi-Level hierarchy of sites

Table 5 (a) Synthesized global patterns – with out correction

P_ id	Pattern	$Supp_G$	Mono-Supp	| Error |	$Conf_G$	Mono-Conf	| Error |
1	C→G	0.3106	0.3106	0.0000	0.5344	0.5344	0.0000
18	H→B	0.2373	0.2373	0.0000	0.7012	0.7012	0.0000
19	B→G	0.3126	0.3126	0.0000	0.5452	0.5452	0.0000
26	H→F	0.2326	0.2326	0.0000	0.6872	0.6872	0.0000
27	H→E	0.2326	0.2326	0.0000	0.6872	0.6872	0.0000
28	H→D	0.2357	0.2357	0.0000	0.6965	0.6965	0.0000
34	H→A	0.2373	0.2373	0.0000	0.7012	0.7012	0.0000
35	A→G	0.3126	0.3126	0.0000	0.5230	0.5230	0.0000
41	H→G	0.3384	0.3384	0.0000	1.0000	1.0000	0.0000
45	F→G	0.3074	0.3074	0.0000	0.6075	0.6075	0.0000
46	E→G	0.3074	0.3074	0.0000	0.5277	0.5277	0.0000
47	D→G	0.3120	0.3120	0.0000	0.5773	0.5773	0.0000
48	BC→D	0.2969	0.3546	0.0577	0.7926	0.9464	0.1538
51	AF→E	0.2879	0.3425	0.0546	0.8407	1.0000	0.1593
55	BH→G	0.2373	0.2373	0.0000	1.0000	1.0000	0.0000
61	GF→E	0.3074	0.3074	0.0000	1.0000	1.0000	0.0000
63	AC→D	0.2539	0.3546	0.1007	0.6777	1.0000	0.3223
69	BF→E	0.2879	0.3425	0.0546	0.8407	1.0000	0.1593
70	AH→G	0.2373	0.2373	0.0000	1.0000	1.0000	0.0000
71	DF→E	0.2419	0.3361	0.0942	0.7199	1.0000	0.2801
74	CE→D	0.2640	0.3684	0.1044	0.7166	1.0000	0.2834

Table 5 (b) Synthesized global patterns – with correction

P_ id	Pattern	$Supp_G$	Mono-Supp	| Error |	$Conf_G$	Mono-Conf	| Error |
1	C→G	0.3193	0.3106	0.0087	0.5495	0.5344	0.0151
18	H→B	0.2482	0.2373	0.0109	0.7105	0.7012	0.0093
19	B→G	0.3214	0.3126	0.0088	0.5605	0.5452	0.0153
26	H→F	0.2435	0.2326	0.0109	0.6970	0.6872	0.0098
27	H→E	0.2435	0.2326	0.0109	0.6970	0.6872	0.0098
28	H→D	0.2466	0.2357	0.0109	0.7060	0.6965	0.0095
34	H→A	0.2482	0.2373	0.0109	0.7105	0.7012	0.0093
35	A→G	0.3214	0.3126	0.0088	0.5376	0.5230	0.0146
41	H→G	0.3493	0.3384	0.0109	1.0000	1.0000	0.0000
45	F→G	0.3162	0.3074	0.0088	0.6248	0.6075	0.0173
46	E→G	0.3162	0.3074	0.0088	0.5427	0.5277	0.0150

47	D→G	0.3207	0.3120	0.0087	0.5935	0.5773	0.0162
48	BC→D	0.3047	0.3546	0.0499	0.8135	0.9464	0.1329
51	AF→E	0.2963	0.3425	0.0462	0.8653	1.0000	0.1347
55	BH→G	0.2482	0.2373	0.0109	1.0000	1.0000	0.0000
61	GF→E	0.3162	0.3074	0.0088	1.0000	1.0000	0.0000
63	AC→D	0.2679	0.3546	0.0867	0.7153	1.0000	0.2847
69	BF→E	0.2963	0.3425	0.0462	0.8653	1.0000	0.1347
70	AH→G	0.2482	0.2373	0.0109	1.0000	1.0000	0.0000
71	DF→E	0.2566	0.3361	0.0795	0.7636	1.0000	0.2364
74	CE→D	0.2781	0.3684	0.0903	0.7548	1.0000	0.2452

Table 6 Error - Analysis

Parameter	Without Correction	With Correction
$Supp_G$ – Mean Error	0.0222	0.0261
$Supp_G$ - RMS Error	0.0432	0.0379
$Conf_G$ - Mean Error	0.0647	0.0624
$Conf_G$ - RMS Error	0.1267	0.1098

Table 7 Synthesized sub-global patterns under SGR-I

P_ id	Pattern	$Supp_{SG}$	Mono-Supp	$Conf_{SG}$	Mono-Conf
158	I→C	0.3190	0.3190	0.6724	0.6724
159	F→I	0.3100	0.3100	0.6016	0.6016
160	I→B	0.3110	0.3110	0.6555	0.6555
161	I→E	0.3100	0.3100	0.6535	0.6535
162	I→H	0.3096	0.3096	0.6898	0.6898
163	A→I	0.3110	0.3110	0.6099	0.6099
164	I→D	0.3190	0.3190	0.6724	0.6724
165	G→I	0.3096	0.3096	0.6005	0.6005
167	HI→G	0.3096	0.3096	1.0000	1.0000
49	BG→A	0.3447	0.3447	1.0000	1.0000
50	AG→B	0.3447	0.3447	1.0000	1.0000
104	FD→C	0.3205	0.3205	1.0000	1.0000
105	FC→D	0.3205	0.3205	1.0000	1.0000
192	O→C	0.3451	0.3451	0.6289	0.6289
193	H→O	0.3401	0.3401	0.8143	0.8143
194	O→G	0.3401	0.3401	0.6197	0.6197
195	O→F	0.3514	0.3514	0.6403	0.6403
196	O→E	0.3514	0.3514	0.6403	0.6403
197	O→B	0.3640	0.3640	0.6633	0.6633
198	O→A	0.3640	0.3640	0.6633	0.6633
199	O→D	0.3451	0.3451	0.6289	0.6289
204	OH→G	0.3401	0.3401	1.0000	1.0000
205	OG→H	0.3401	0.3401	1.0000	1.0000
206	DB→A	0.3579	0.3579	1.0000	1.0000
207	DA→B	0.3579	0.3579	1.0000	1.0000
208	CF→G	0.2550	0.2550	0.7640	0.7640

comes under campus classification. The remaining patterns 192 to 208 exactly fit with east zone classifier.

Table 8 presents rules under SGR-II; these rules do not fit exactly with any of the already existing class labels. The domain expert has to supply a suitable labels based upon the background knowledge. For example rule R29 is supported by S1, S4, S5, S6 and S7 and does not fit under any existing class/subclass label. Possibly on the basis of background knowledge, the domain expert can supply a label say, "Old_ Settlements". Table 9 presents rules under SGR-III.All the 19 rules will be found in table 8 also. However the values for support at SGR-III will be lower than the corresponding values of SGR-II. For example, if we consider rule R81, C→U, the synthesized support values are 0.3099 and 0.2324 under SGR-II and SGR-III respectively. Rule 81 is supported only by S2, S3 of rural region but is not supported by S1.The domain expert has to decide whether to put the rule under "rural" bracket or given a new name for the cluster formed by the two sites S2 and S3.

Table 8 Synthesized sub-global Patterns under SGR-II

P_ id	Pattern	$Supp_{SG}$	$Conf_{SG}$	Supporting Sites
29	H→C	0.2809	0.6711	S1,S4,S5,S6, S7
54	CF→E	0.3483	1.0000	S1,S4,S6, S8
116	FA→G	0.2594	0.7581	S2,S4,S5,S6, S7
133	J→I	0.2444	1.0000	S3,S4,S6,S8
…	…	…	…	…
68	FK→G	0.2930	0.7269	S1,S9
81	C→U	0.3099	0.5413	S2,S3
…	…	…	…	…
128	C→M	0.3453	0.5956	S3,S5
129	B→M	0.3641	0.6174	S3,S5
130	M→A	0.3641	0.6663	S3,S5
142	MD→C	0.3453	1.0000	S3,S5
143	MC→D	0.3453	1.0000	S3,S5

Table 9 Synthesized sub-global Patterns under SGR-III

Pattern	$Supp_{SG}$	Mono-Supp	$Conf_{SG}$	Mono-Conf	Class Label
C→U	0.2324	0.2324	0.4199	0.4199	Rural
B→U	0.2374	0.2374	0.4390	0.4390	Rural
U→E	0.2401	0.2401	0.6428	0.6428	Rural
A→U	0.2374	0.2374	0.4208	0.4208	Rural
D→U	0.2324	0.2324	0.4375	0.4375	Rural
V→D	0.2061	0.2061	0.6442	0.6442	Campus
V→C	0.2061	0.2061	0.6442	0.6442	Campus
V→F	0.2156	0.2156	0.6740	0.6740	Campus
V→B	0.2280	0.2280	0.7127	0.7127	Campus
V→E	0.2156	0.2156	0.6740	0.6740	Campus
V→A	0.2280	0.2280	0.7127	0.7127	Campus
F→M	0.2266	0.2963	0.4520	0.5910	South
M→E	0.2266	0.2963	0.4899	0.6406	South
D→M	0.2249	0.2932	0.4209	0.5486	South
C→M	0.2249	0.2932	0.3886	0.5069	South

B→M	0.2371	0.3064	0.4106	0.5305	South
M→A	0.2371	0.3064	0.5126	0.6624	South
MD→C	0.2249	0.2932	0.7673	1.0000	South
MC→D	0.2249	0.2932	0.7673	1.0000	South

C. *Identifying Local Patterns*

Patterns which are not satisfying the minimum nominal vote measure are labeled as local patterns. By their local occurrence, they are exhibiting the individuality of sites. In all 150 patterns were found in this category. A few of them are shown in Table 10. The Summary of synthesized patterns is accounted in Table 11.

Table 10 Local patterns

P_ id	Pattern	Local Support	Local Confidence	Supporting Site
8	R→K	0.2416	0.5615	S1
9	R→F	0.2420	0.5625	S1
…….	……..	………..	………..	……
77	DF→K	0.2008	0.6952	S1
83	L→U	0.3748	0.8489	S2
84	L→C	0.2870	0.6499	S2
…….	……..	………..	………..	……
131	N→C	0.2832	0.6472	S3
132	N→F	0.2436	0.7291	S3
…….	……..	………..	………..	……
219	CH→O	0.2421	0.8909	S7
229	HO→A	0.2587	0.7125	S7
230	AH→I	0.2682	0.8263	S8
…….	……..	………..	………..	……
239	FH→I	0.2657	0.8260	S8
240	D→P	0.3646	0.7418	S9
…….	……..	………..	………..	……
273	PX→V	0.2916	1.0000	S9

Table 11 Summary of synthesized patterns

Pattern Nature	Total No
Candidate global patterns	45
Synthesized global patterns supported by all the sites	15
Synthesized global patterns not supported by all the sites	21
Candidate sub-global patterns	69
Synthesized sub-global patterns under standard classifier (SGR-I)	26
Synthesized sub-global patterns to be labeled by the domain expert(SGR-II)	43
Synthesized sub-global patterns under standard classifier (SGR-III)	19
Local patterns	150

V. Conclusions

A novel approach for synthesizing local rules from different data sources at multiple levels of abstraction into global rules, sub-global rules, while segregating local rules has been proposed. The emerging sub-global rules may fit in with existing class definitions or may be labeled suitably by a domain expert. The proposed weighted model is sufficiently comprehensive and general and thus extends our earlier work. The synthesized values tally with mono-mining results.

References

[1] Agrawal.R and Srikant.R (1994) Fast algorithms for mining association rules. In: Proceedings of 20th International Conference on Very Large Databases, pp.487-499.

[2] Agrawal.R, Imielinski.T, Swami. A (1993) Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp.207-216.

[3] Kum H C, Pai J, Wan W, Duncan D SIAM international conference on Data mining (2003)

[4] Liu H, Lu H, Yao J (2001) Toward multi database mining: Identifying relevant databases. IEEE Transactions on Knowledge and Data Engineering 13(4) : 541- 553.

[5] Ramkumar T, Srinivasan R (2008) Modified algorithm for synthesizing high-frequency rules from different data sources. Knowledge and Information System 17(3) :313-334

[6] Wu X, Zhang C, Zhang S (2005) Database classification for multi-database mining: Information System 30(1) : 71-88.

[7] Wu X and Zhang S (2003) Synthesizing high-frequency rules from different data sources. IEEE Transactions on Knowledge and Data Engineering 15(2): 353-367.

[8] Zhang N, Yao.Y.Y, Ohshima M (2003) Peculiarity oriented Multi-Database Mining. IEEE Transactions on Knowledge and Data Engineering 15(4): 952- 960

[9] Zhang C, Liu M, Nie W, et al. (2004) Identifying global and exceptional patterns in multi-database mining. IEEE Computational Intelligence Bulletin 3(1): 19-24.

[10] Zhang S, Wu X, Zhang C(2003) Multi-Database Mining. IEEE computational Intelligence Bulletin 2(1) : 5-13.

[11] Zhang S, Zhang C, Wu X (2004) Knowledge Discovery in Multiple Databases. Springer.