A Novel Hybrid Spatial Association Rule Mining Algorithm for Neuro Imaging

R. Parvathi
Senior Lecturer
Centre for Computer Applications
VLB Janakiammal College of Engg&
Technology, Coimbatore- 641 042

Dr. S. Palaniammal
Professor and Head
Department of Science & Humanities
VLB Janakiammal College of Engg&
Technology, Coimbatore- 641 042

ABSTRACT

Data mining methodologies have been developed for exploration and analysis of large quantities of data to discover meaningful patterns and rules. This paper presents a new approach that employs data mining, to find spatial association rules an effective method for discovering NeuroImaging. The propose system has been projected from the physical parameters which will be very helpful for and will make everything easier for the physicians in the diagnosis of Neuro imaging. Data of 492 patients are evaluated in the projected system. The results of the decision support system have completely matched with those of the physician's decisions.

General Terms

Data Mining, Spatial Association Rules, NeuroImaging

Keywords

Data Mining, Spatial Association Rules, NeuroImaging .

1. INTRODUCTION

In some works on spatial representation from the social scientist's perspective [1], socioeconomic phenomena have been conceptualized as *spatial objects*, without assuming any particular application such as marketing or resource allocation. Spatial objects in this sense are entities having both spatial location and spatially independent attribute characteristics [2]. Population data are among the potentially spatial socioeconomic data. They are most commonly related to geographic locations by reference to areal spatial objects such as census zones, electoral constituencies, local government areas, or regular grid squares.

Statistical spatial analysis has been the most common approach to the analysis of georeferenced data [3,4]. Being a well-studied area, it supplies a large number of algorithms including various optimization techniques. It handles numerical data very well and usually comes up with realistic models of spatial phenomena. The statistical approach to spatial analysis shows some weaknesses in dealing with dependency in spatial data. Most methods are exploratory and when applied to patially correlated data some of them are of unknown reliability having been developed initially, like so many areas in statistics, for situations where observations are independent [5]. This contrasts with the nature of spatial data where spatial objects are influenced by their neighboring objects as pointed out by [6]. In recent times, alternative approaches to spatial analysis have been emerged. In particular, the extension of data mining methods and techniques to spatial databases has been attempted to allow *the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases* [7]. Extracted knowledge can take on various forms according to the spatial data mining task at hand (discrimination, characterisation, clustering, classification, etc.). We are concerned with the task of mining *spatial association rules*, namely the detection of associations between spatial objects [8]. In this paper we propose the application of logic-based methods and techniques to the discovery of spatial association rules. In particular, we resort to Inductive Logic Programming (ILP) which is a form of inductive learning derived from the theory of computational logic [9]. Computational logic relies on an augmented expressive power which allows us to represent spatial relations and symbolic background knowledge (such as spatial hierarchies, spatial constraints and rules for spatial qualitative reasoning) in a very elegant and natural way. Thus, it enables applications which can not be tackled by traditional statistical techniques in spatial data analysis. The technique being proposed has been implemented in the ILP system SPADA (**Spatial Pattern Discovery Algorithm**) [10].

Several data mining techniques (*Table 1*) were developed and applied to discover new and interesting pattern relationships from large spatial data [11,12,13].

From the application side, the aim of the project is the development of an integrated interactive Internet-enabled spatial data mining system to enhance decision-making based on spatio-temporally referenced data and publish geographical data mining services on the WWW. The application to Medical data from Coimbatore - one of the Metropolitan Districts of Tamil Nadu.

The paper is organized as follows. Section 2 will introduce the task of mining spatial association rules, by discussing some issues raised by the application of computational logic to spatial data mining and the solutions adopted by the system SPADA. In Section 3 we report preliminary results on the application of SPADA to Coimbatore District Medical data. Conclusions and future work will be drawn in Section 4.

2. SPATIAL ASSOCIATION RULE MINING

A spatial association rule is of the form $X => Y (c\ \%)$, where X is called antecedent and Y consequent of the rule. The antecedent contains a set of predicates from the exploring database, the consequent only represents one predicate, which is not yet included in the antecedent. The rule itself then reflects an existing relationship between predicates in antecedent and consequent. A measure of a rule's strength is confidence *(c%)*, which indicates that *c* percent of the items satisfying the antecedent also satisfies the consequent. In a large database many association relationships may exist but some may occur rarely or may not hold in most cases. To focus our study to the patterns which are relatively strong, i.e., which occur frequently and hold in most cases, the concepts of minimum support and minimum confidence are introduced. spatial association rules represents object predicate relationships containing spatial predicates. For example, the following rules are spatial association rules.

Nonspatial consequent with spatial antecedent(s)

is_a(x,house) ^ close_to(x,beach) □ Is_expensive(x) (90%)

Spatial consequent with non-spatial /spatial antecedent(s).

Is_a(x,gas_station) □ close_to(x,highway) (75%)

Various kinds of spatial predicates can be involved in spatial association rules .

2.1 The logical framework

The problem of mining spatial association rules can be formally stated as follows:

Given

_ a spatial database (SDB),

_ a set of reference objects *S*,

_ some task-relevant geographic layers *Rk*, 1_ *k*_ *m*, together with spatial hierarchies defined on them, _ two thresholds for each level *l* in the spatial hierarchies, *minsup*[*l*] and *minconf*[*l*] *Find* strong multiple-level spatial association rules.

The basic idea in our ILP approach is that a *spatial database* can be boiled down to a DDB once that reference objects and task-relevant objects, their aspatial properties and the spatial relationships among them have been extracted according to a predefined semantics (*feature extraction*). As for topological relations, we have adopted the 9-intersection model [14]. Thus our approach requires that spatial data are transformed into ground facts of a logical language for relational databases. In particular, we resort to Datalog [15] whose expressive power allows us to specify background knowledge (BK) such as spatial hierarchies, spatial constraints and rules for spatial qualitative reasoning. Formal details follow. From now on, we denote the DDB at hand *D*(*S*) to mean that it is obtained by adding spatial relations extracted from SDB as concerns the set of reference objects *S* to the previously supplied *BK*. The tuples in *D*(*S*) can be grouped into distinct subsets: Each group, uniquely identified by the corresponding reference object *s* _*S*, is called *spatial observation* and denoted *O*[*s*]. It is given

$$O[s] = O[s|s] \cup \{O[r_i|s] \mid \exists \text{ tuple } \theta \in D(S)\text{: } \theta(s, r_i)\}_{1\leq i\leq n}$$

where $O[s|s]$ contains spatial relations between s and some task-relevant object $r_i \in R_k$ and each $O[r_i|s]$ contains spatial relations between r_i and some $s' \in S$.

Example 1 Suppose the mining task is to discover associations relating large towns (*S*) with water bodies (*R*1), roads (*R*2) and province boundaries (*R*3) in the Province of Bari, Italy. We are also given a BK including the spatial hierarchies of interest (see Figure 1 for a graphical representation of the layer of roads).

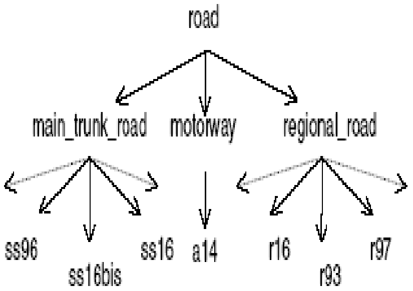


Figure 1 : Is-a assertions are deducted by means of rules from BK. Here, the *is-a* relationship is overloaded, namely it may stand for *kind-of* as well as for *instance_of* depending on the context. Spatial relations between objects in *S* and objects in any of *R*1, *R*2 and *R*3, are extracted by means of spatial computation and transformed into facts of the kind <spatial relation>(RefObj, TaskRelevantObj) to be added to *D*(*S*). Spatial observations are portions of *D*(*S*), each concerning a reference object [16].

2.2 Main procedure of SPADA

Procedure mineMultipleLevelAssociations (*A*, *keyAtom*, *maxLevel*, *maxDepth*)

FP: set of frequent patterns

IP: set of infrequent patterns

SR: set of strong rules

l: level in task-relevant spatial hierarchies (<= *maxLevel*)

k: depth in the pattern space (<= *maxDepth*)

```
begin
FP<-ø; SR<-ø; l<- 1;
        foreach level l do
        IP(l)<-ø; k<-1;FP(l,k)<-{keyAtom}
        while k>maxDepth and FP(l, k)≠φ do
                k<- k +1;
                [FP(l,k),IP(l)]<-
        generateFrequentPatterns(FP(l, k-1), IP(l), A);
                SR(l, k)<-generateStrongRules(FP(l, k));
                FP(l)<-FP(l) FP(l, k); SR(l) <-SR(l) SR(l, k)
        endwhile
        FP<-FP FP(l); SR<-SR  SR(l); l<-l +1
endforeeach
return[FP,RS]
```

2.3 FAST (Finding Association from Sampled Transactions) TRIM Algorithm

Given a specified minimum support p and confidence c, the FAST algorithm [17] for data

reduction proceeds as follows:

1. Obtain a simple random sample S from D.

2. Compute f(A; S) for each 1-itemset A 2 I1(S).

3. Using the supports computed in Step 2, obtain the final small sample S0 from S.

4. Run a standard association-rule mining algorithm against S0 — with minimum

support p and confidence c—to obtain the final set of association rules.

Steps 1, 2, and 4 are straightforward. The drawing of a sample in Step 1 can be performed with a computational cost of O(|S|) and a memory cost of O(|S|). The computational cost of Step 2 is at most O(Tmax ·|S|), where Tmax denotes the maximal transaction length. From a computational point of view, because the cost of Step 2 is relatively low, the sample S can be relatively large, thereby helping to ensure that the estimated supports are accurate. Step 4 computes the frequent itemsets using a standard association rule mining algorithm such as Apriori [18,19]. The crux of the algorithm is Step 3. Two approaches (trimming and growing) for computing the final small sample S0 from S are given in [20].

In this chapter, we discuss only the trimming method, which removes the “outlier” transactions from the sample S to obtain S0. In this context an outlier is defined as a transaction whose removal from the sample maximally reduces (or minimally increases) the difference between the supports of the 1-itemsets in the sample and the corresponding supports in the database D. Since the supports of the 1-itemsets in D are unknown, we estimate them by the corresponding supports in S as computed in Step 2

Algorithm of FAST TRIM

```
Obtain a simple random sample S from D;
Compute f(A;S) for each item A in S;
Set $S_0$ = S
While (|$S_0$| > n)
{
divide $S_0$ into disjoint groups of min(k,| $S_0$|)
transactions each;
for each group G
{
Compute f(A;$S_0$) for each item A in $S_0$
Set $S_0$ = $S_0$ – { t* } where
    Dist($S_0$ – {t*}, S) = min $_{t \varepsilon G}$ Dist($S_0$ – {t},S);
}
}
run a standard association-rule algorithm against $S_0$ to obtain the final set of association rules
```

Stopping Criteria

As formulated so far, the trimming procedure stops when the sample size reaches a specified value *n*. Note that after the desired final sample size is reached, additional trimming may further reduce the processing time without decreasing the accuracy much.

2.4 FAST GROW Algorithm

```
Obtain a simple random sample S from D;
Compute f(A,S) for each A ε $x_1$ (S);
Set I =0;  $S_0$(i) = Φ,minDist =  ∞  and minStage = -1;
While (| $S_0$ (i) < n) {
Divide S-$S_0$(i) into disjoint groups of min( | S -  $S_0$(i)|,k)
Transaction each;
For each group G {
Set $S_0$(i) = $S_0$(i) ∪ {t*} where Dist( $S_0$(i) ∪ {t*},S ) = min $_{t \varepsilon G}$
Dist( $S_0$(i) ∪ {t},S )
}
Compute f(A;$S_0$(i)) for each A ε x1(S)
If (Dist( $S_0$(i), S)< MinDist(
{
set minDist = dis( $S_0$(i), S) and minStage = i; }
set $S_0$(i+1) = $S_0$(i);
}}
run a standard association-rule algorithm against $S_0$(minStage) to obtain final set of association rules.
```

First consider a version of fast-grow with a specified final sample size of *n* transactions. Like fast-trim, the fastgrow algorithm has an input parameter *k 2 f* 1*;* 2*; : : : ; jSj g* and proceeds in stages. Initially, *S*0 is empty. At each stage, fast-grow increments *S*0 by adding representative transactions.

In order to identify representative transactions, the transactions in *S ¡S*0 are divided into disjoint groups, with each group having min(*jS ¡ S*0*j; k*) transactions. For each group, the algorithm selects a transaction *t*¤ that minimizes the function Dist(*S*0 *[ftg; S*) over all transactions in the group and adds *t*¤ to *S*0. The algorithm proceeds until *jS*0*j = n*. As with the fast-trim algorithm, *k* is chosen to trade off speed and accuracy: the larger the *k* value, the higher the accuracy, but the slower the speed. To quantify the complexity in terms of the number of Dist() evaluations required, let *m* be, as before, the number of transactions in the initial sample *S*.

3. PROPOSED METHOD(HYBRID SPATIAL ASSOCIATION RULE MINING ALGORITHM)

The most common and main tests that help the physicians decide on a diagnosis are biochemistry tests which are mostly successful in diagnosing.. The construction of the improved decision support system is showed in Figure 2. The functions of system are constituted of these following steps:

Figure 3 . Mining with Hybrid Spatial Association Rule Mining Algorithm

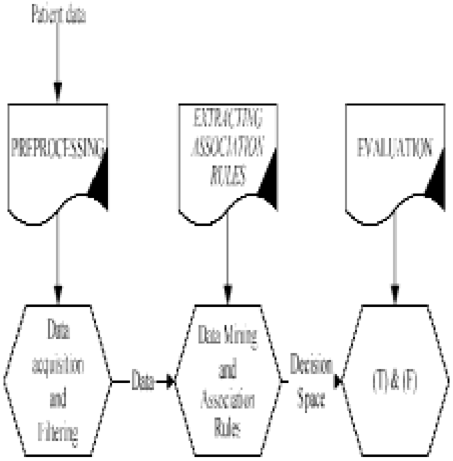

Patient data
PREPROCESSING
EXTRACTING ASSOCIATION RULES
EVALUATION
Data acquisition and Filtering
Data
Data Mining and Association Rules
Decision Space
(T) & (F)

Figure 2. Steps of Proposed Work

Step – 1: Pre-processing

Collecting the data is the step in which the ones those are proper for the goal are choosen. The necessary data were obtained by collecting the scan biochemistry test results of the patients applied for the internal medicine with the direction of the physicians from various locations near by coimbatore.

Step – 2: Feature Extraction and Classification

This is the most important step of association rule based decision support system which was developed for processing the data in the collected temporal database. The data collected are being processed in this stage using SPADA.

In this paper, medical scan prescription are collected from the various hospital and forwarded to scan centre where the transactions of the same type of scans are classified. Based on threshold value, support is calculated for the transactions are shown in figure 3.

Cluster Support = (clusters_area(antecedent)) => cluster_area(consequent) / area(s)

Cluster Confidence = cluster_area(x=>y)/cluster_area(x)

The threshold value is used to classify the minimum support and maximum support transaction.

Step – 3: Evaluation

In this step, rules which are chosen support and confidence values evaluated using fast grow and fast trim algorithm. Fast algorithm mainly used for analyzing the data set by strong confidence and week confidence based on their support values.

As an enhancement an algorithm "**FAST ALGORITHM**" is being used where it takes as subsets and gives their maximum support and the minimum support. The Fast trim is used to remove all the combinations that has the support level zero. Fast Grown and Fast Trimmed algorithms are performed till the process is completed and the confidence level is calculated using the formula:

$$\text{Confidence Level} = \frac{Support(A \cup B)}{Support(A) or Support(B)} * 100$$

Eq(1)

strong confidence is maximum confidence and maximum support values and week confidence is one which has maximum confidence and minimum support values are represented in Figure 4.

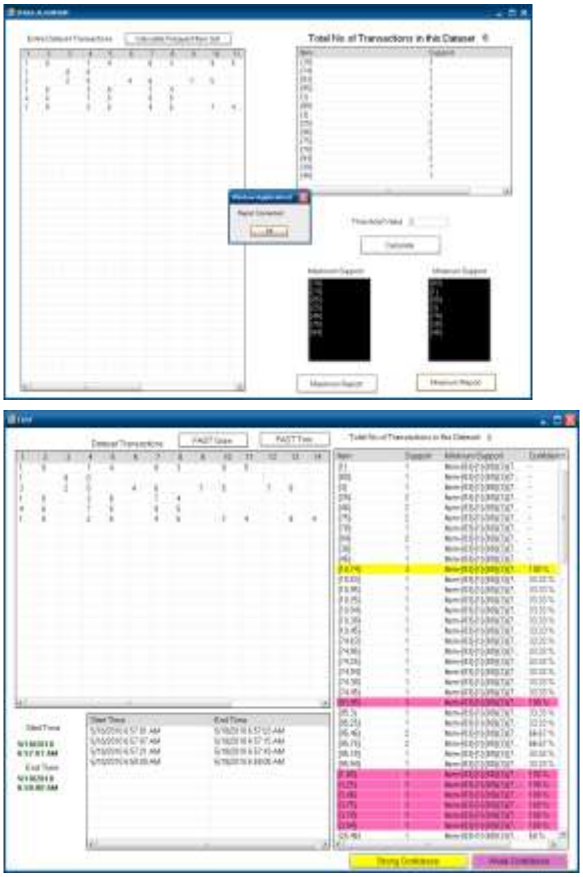


Figure 4. Calculation of Strong and Week Confidence Level

4. RESULT AND DISCUSSION

The increase in the knowledge obtained under the light of biomedical searches, the processes of reading, simplifying, classifying these findings and making a decision are being so complex day by day. Understanding the major risk factors of a diesease is an important factor for clinicians in prevention strategy. The attending physician plays an important role providing information to reduce those risk factors. It is up to the physician whether to warn patients at risk about the major causes of a particular disease and the degree of risk that they are facing. These processes are automized after the data mining took a part in this field. It has considerably helped the medical experts and made it easier to prepare a guide. Also, a lot of findings hidden between the data mining and the data stacks obtained in these kinds of fields have turned to be useful information in this way.

In this paper we dealt with spatial association rule. We restricted ourselves to the "classic" association rule problem, that is the generation of all association rules that medical data with respect to minimal thresholds for support and confidence. The advised system is based on association rules on which so many clever diagnosis systems are constituted. There, the association rules, which take an important part in data mining for the feature extracting and classification stage, have been used. For this disease, decision tree technique, a kind of data mining techniques, has been applied [38]. On the other hand, since association rules find results with percentage rates, it's better than decision tree technique for those diseases which can't be separated by definite rules.

The focus of this career development plan is to build an education and research program that will focus on the discovery of patterns and relations between anatomy (structure) and function through the effective and efficient analysis of large repositories of medical images and other clinical data. Medical centers almost everywhere today are facing an interesting challenge in analyzing the huge volumes of image and associated clinical data collected daily as part of several ongoing studies. By focusing on the regions of interest (ROIs), the approach uses novel techniques to extract their most discriminative features and uses them in classification and similarity searches. New representations of the information content of medical images are also provided. Moreover, spatial data mining tools are developed to efficiently discover associations between image data and non-image (functional) data. The approaches have applicability to medical images from a wide range of modalities (e.g., CT, MRI, fMRI, angiography, confocal microscopy, etc). Information about the function of structures related to various medical conditions is extracted from clinical assessment

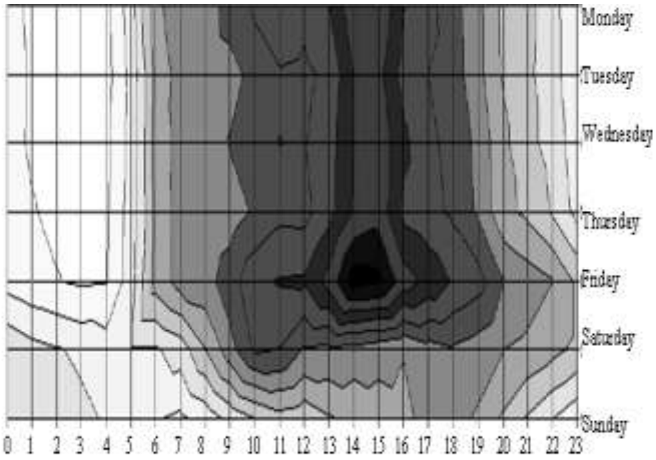
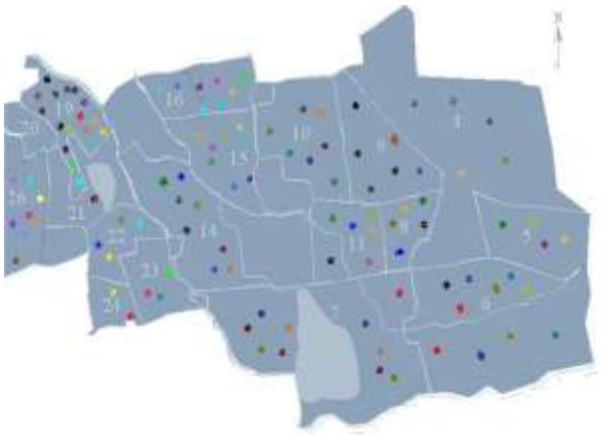


Figure 5: The total number of Scans by hours and days of the week.

The first step in the data mining process was data visualization, performed to enable better understanding of the data. Selected visualizations are presented below. Figure 5 is used to analyze neuroimaging peaks in terms of the total number of scan in certain time periods. As indicated in Figure 5, the number of scans during the week has an obvious peak on Friday, which is probably due to weekly migrations. The daily peak of the number of any type of scan is between 2pm and 3pm when people start leaving work. It is interesting to notice, that there in no morning peak, when people go to work.



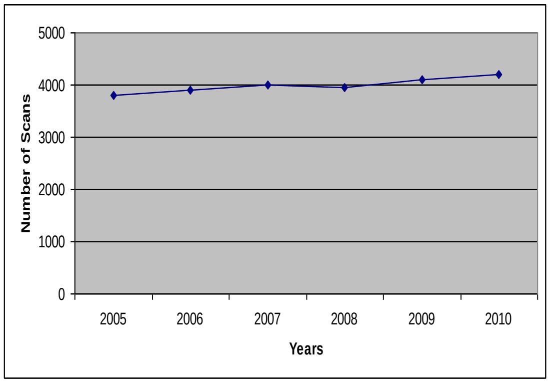


Figure 6: The total number of scans in the years 2005-2010.

Figures 6 show that the total number of scan details has not changed much during the period of five years. A growing number of scan can be understood as a consequence of more unknown diseases.

Table 3: Sample of Transaction on Friday(13/08/2010) carried on Ease zone of Coimbatore

Region Code	Transaction ID
1	13 14 15 16 17 18 19 20
2	14 15 16 17 18 19 20 21 22 23
3	6 8 9 10 11
4	6 8 9 10 11 12 19 20 21 22 23 24 25 26 27 28
5	1 2 3 4 5 6 7 12 13 14 15 16 17 19 20 21 22 23 24 25 26 27 28

6	1 2 3 4 5 6 7 12 13 14 15 16 17 19 20 21 22 23 24 25 26 27 28 29 30 32 33
7	8 9 10 11 19 20 21 22 23 24 25 26 27 28 29 30 32 33 36 37
8	8 9 10 11 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
9	1 2 3 4 5 6 7 9 10 11 12 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
10	1 2 3 4 5 6 7 31 32 33 34 35 36 37 38 39 40 41 46

Figure 9 : Transaction on the East Zone of Coimbatore

Figure 8 represents the city map of Coimbatore and Figure 9 is representing the different zones of the city along with different transactions performed on particular date.

5. CONCLUSION

Conversely, nowadays immense collections of spatio-temporal data are gathered without any previous hypothesis, and less as parts of a structured experiment. Moreover, it is impossible that trained researchers examine all possible interesting patterns in such huge amounts of data. We require an intelligent assistant to process the data and to autonomously (or at least with very little guidance) analyze data. The presented methods and algorithms that increase even more the capacity of GIS to become intelligent pattern spotters beyond just as repositories to store, manipulate and retrieve data. Note that, this direction complements traditional statistical analyses on geo-referenced data. The intent is to design and deploy autonomous model in suggesting hypothesis in the knowledge discovery process. Urbanized algorithms that will enable this exploratory capability in computers within the context of spatial data and medical data.

This helps us to seek inference with the threshold value of the count as eight, there are twelve categories out of 45 categories of scan are more major and occur in one out of four zones. Therefore the result expose that the more number of scan centers in the above mentioned categories can be deployed. In addition, the association rules mining and GIS technique helps to reveal the spatial pattern of scan and zones in the Coimbatore city of Tamil Nadu

6. REFERENCES

[1] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the Twentieth VLDB Conference, Santiago: Cile (1994).

[2] Bogorny, V., Kuijpers, B., Alvares, L.O.(2007c). Reducing Non-Interesting Spatial Association Rules in Geographic Databases using Background Knowledge: a Summary of Results. IJGIS International Journal of Geographical Information Science, Taylor and Francis

[3] Ester, M.; Kriegel, H.P.; and Sander, J. 1999. Knowledge Discovery in Spatial Databases. In Burgard, W.; Christaller, T.; Cremers, A.B. (Eds.): *KI-99: Advances in Artificial Intelligence*, LNCS 1701, Springer-Verlag, 61-74,.

[4] Gatrell, M.R. 1991. Concepts of space and geographical data. In Maguire D.J., Goodchild M.F., Rhind D.W. (eds): Geographical Information Systems: principles and application. Harlow, Longman/New York, John Wiley & Sons Inc. Vol. 1: 119-134.

[5] Han, J and Fu, Y., 1999, "Mining Multiple-Level Association Rules from Large Databases", *IEEE Transactions on Knowledge and Data Engineering*, 11(5), September 1999.

[6] Koperski, K.; Adhikary, J. and Han, J. 1996. Spatial Data Mining: Progress and Challenges. In *Proceedings Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada

[7] Lavrac, N.; and Dzeroski, S. 1994. *Inductive Logic Programming: Techniques and Applications*. Chichester, UK: Ellis Horwood

[8] Lee SM, Park RW. Basic concepts and principles of data mining in clinical practice. J Korean Soc Med Inform 2009; 15: 175-189

[9] Malerba, D.; Esposito, F.; and Lisi, F.A. 2001. A Logical Framework for Frequent Pattern Discovery in Spatial Data. To appear in *Proc. 14th International FLAIRS Conference (special track on spatiotemporal reasoning)*, Key West, Florida, USA, May 2001

[10] Miller H J and Han J (2001), "Geographic data mining and knowledge discovery: an Overview". In Miller, H.J. and Han, J. (eds) *Geographic data mining and knowledge discovery*. London, New York: Taylor & Francis, 3-32.

Ms. R.Parvathi received B.Sc Degree in Physics in 1992, MCA Degree in Computer Applications in 1996. She is Currently working as a Sr.Lecurer in the Department of MCA, VLB Janakiammal College of Engineering and Technology, Coimbatore She is currently perusing Ph.D. Her research interest includes Spatial Data Mining. She has published technical papers in International conferences and journals.

Dr S.Palaniammal is working as Professor and Head, Department of Science and Humanities,VLB Janakiammal College of Engineering and Technology,Coimbatore,Tamil Nadu. She has 25 years of teaching experienceat the Under graduate and Post graduate Levels for the Engineering students.She is Board of studies Member and Doctoral Committee member of various universities and Advisory Committee Member for several National/ International Conferences.She has organised many Seminars/Conferences/Workshop. 10 Research Scholars are pursuing their Ph.D under her guidance.Her area of interest includes DataMining, Queuing Theory and Computer Networks. She has published more than 32 papers in the National and International Journals and Conferences. She has authored 7 Books in Mathematics for various branches of Engineering Students.