CHAPTER-1

INTRODUCTION

1.1 BACKGROUND

The proliferation of spatial information during last several years has attracted the researchers and policy makers to explore the possibility of extracting useful patterns from large spatial datasets. The various combinational approaches of spatial data mining [SDM] fulfill real needs of many such applications. The overall objective of spatial data mining is to take advantage of the potentially rich growing geographically referenced data. For many applications the geographical attribute that is present in the spatial database is considered as an important aspect. There are various applicative areas for which some framework of SDM can be modeled to extract novel and useful pattern from spatial database. Spatial analysis of epidemic risk is one of the promising and challenging areas where the appropriate use of SDM technique is required to be applied.

Spatial data mining is a non trivial process to extract knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. For spatial data mining it is required to integrate data mining with spatial database technologies. It can be used to understand spatial data, relationships between spatial and non-spatial data, discovery of spatial relationships, constructing spatial
knowledge bases, reorganizing spatial databases, and optimizing spatial queries. It has a wide application in geographic information systems (GIS), vector and raster image database, geo-marketing, remote sensing, image database exploration, environmental studies, and many other areas where spatial data are used. However, in comparison to extracting patterns from traditional numeric and characterized data, it is much difficult to extract interesting and useful patterns from spatial databases because of the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Spatial epidemiology database is the volume of data which could be used for the mining purpose to extract useful pattern and rule for prediction.

Elliott and Wartenberg [1] described spatial epidemiology as “Spatial epidemiology is the description and analysis of the geographic, or spatial, variations in disease with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors”. The spread of infectious disease is closely associated with the concept of spatial or spatio-temporal proximity. The individuals who are linked in a spatial and temporal sense are at a high risk of getting infected[2]. Thus the knowledge of spatial and temporal variations of disease and characterizing its spatial structures is essential for the epidemiologist to understand better the population’s interaction with its environment [3]. The history of spatial epidemiology dates back to 1800s, when maps of disease rates in different countries began to emerge to characterize the
spread and possible causes of outbreak of infectious disease such as yellow fever and cholera [4]. Proximity to environmental risk factors is therefore important. Spatial epidemiology analysis comprises of wide range of methods. Now it is a big challenge to determine which one to use [5].

The recent global health report shows that our population is quite vulnerable to the infectious disease. Government and the planners have extensively recognized the need to improve the well being of population. They need a system of exactly knowing the root cause of the problem and to undertake prompt preventive and control measures. The infectious diseases are complex to control and prevent, leading to questions on how best to combat them through novel and creative solutions.

In the early 1980s HIV/AIDS was first recognized as a disease. Since then it has started showing its presence throughout the world. According to WHO annual report, it is now the leading cause of death in Africa and is responsible for one in five deaths. Globally it is the fourth most common cause of death. There is a considerable variation in the pattern of epidemic spread within countries, locally, and between countries. Consequently there is also variation in the impact of the resulting illness and pre mature deaths. The burden of this epidemic falls on the world’s poorer countries and on poorer communities in some richer countries.
One of the main facilitating conditions of HIV transmission is migration[88]. Study shows that a proper understanding of linkages between migration and HIV risk factors is crucial in controlling further spread of the disease. Vulnerability to this disease is often greatest when people find themselves living and working in conditions of poverty, social instability, and powerlessness, which apply to many migrants. When we consider migrants for work it is the men who first migrate then followed by link migrants like spouse and other family members. The most vulnerable chunk of migrants who are considered to be the bridge population for the spread of HIV/AIDS are the temporary laborers and truckers. They are the men having to leave behind families and their social groups and redefine their identities. They have to abstain or look for other alternatives to satisfy their sexual needs. Although HIV/AIDS and migration do not have a linear, cause-effect link but they are linked laterally[6,7]. The table below shows the percentage distribution of migration status among male HIV-positive cases and HIV-negative controls.
With respect to geographical variation and effect of different behavioral characteristics, the epidemic HIV/AIDS in India has been analyzed [8]. In the study it has been found that in the distribution of HIV/AIDS is heterogeneous with respect to the vast geographical stretch of the country, and also with respect to the variation in human development index (HDI), gender development index (GDI), occupational structure and socio-cultural lifestyle [9,10].

Figure 1.1: Migration versus HIV status
1.2 ISSUES

The main reason behind the spread of the disease is the higher level of urbanization which triggers related migration. The potential reason of rural to urban movement is because of the lack of economic opportunities. The census data of 2001 reveals that all of India’s million-plus cities have over one third of their population made up of migrants [11]. In India higher percentage of men show high-risk behavior and the infection is more common among men than among women. The study shows that out of the estimated numbers of adults living with HIV, 62 percent are male. It is from the married men the virus is being transferred from sex-workers to their wives.

The Census of India defines migrants as a person who has moved from one politically defined area to another similar area. For the purpose of HIV programming in the country the revised migrant operation guideline defines migrants as;

- People (Both male and female) who move from their place of origin in rural areas (source) to a town or city (destination) – irrespective of district/state/country
- Return to their place of origin at least once in 6-12 months
• Move frequently between districts for work purpose
• Move directly between the places (or) via the transit locations
• Move either alone or with their partners
• Those returned to places of origin (at source areas)
• Female spouses of migrants (at source areas)

Unlike the core group interventions, the migrant intervention program is required to be evolved. It is also required that identification of high migration and identification of locations/states within high out-migration areas where interventions could be initiated. The National Aids Control Program (NACP IV) of NACO provides an opportunity to further strengthen strategies by enhancing the evidence and designing interventions tailored to the dynamics of migrant populations including the typologies that influences their vulnerability. Within this research work I attempted to draw together as much information as possible on the spatial cause and effect of HIV epidemiology.

Discovery of implicit and useful knowledge in spatial database of epidemiology like HIV/AIDS is a highly demanding field because very large amount of spatial data has been collected and it is witnessed that this disease is showing location wise different characteristic in terms of its prevalence and growth. The spread of infectious disease is closely associated with the concepts of spatial and spatio-temporal proximity, as individuals who are linked in a spatial
and temporal sense are at a high risk of getting infected. Proximity to environmental risk factors is therefore important. Thus acknowledging the spatial and temporal variations of disease and characterizing its spatial structure is essential for the epidemiologist to understand better the population’s interactions with its environment. Therefore, it is highly required that some spatial data mining framework is designed which is able to take into account the spatial and non spatial factors of the spread of this disease. The general objective of this work is to formulate a predictive learning model of data mining based on the spatial database consisting of spatial and non spatial attributes related to some epidemiology.

The study of the HIV/AIDS based reports prompted us to undertake a deep research work of mining the spatial database of HIV epidemiology. Here my effort is to formulate a framework of spatial data mining which will consider the geographical (spatial), demographic, socio-economic factors. It will not only identify the pattern and root cause of the disease but also will predict the situation of the epidemiology like HIV/AIDS in particular and other epidemiology in general. To carry out the spatial data mining task a spatial data base, consisting of spatial and non-spatial information, is required to be created. Further some efficient algorithm is required to be formulated to find out spatial and non spatial predicates for classification of spatial database objects. The method of finding out the predicate for classification and preparation of decision tree should be
supported with the learning approach of ANN to propose an efficient learning and predictive model.

We know that there is heterogeneous spatial distribution of HIV/AIDS epidemiology in India. Now in India there is an interesting report of a recent study conducted by population council of India with support of the United Nations Development Program (UNDP) and in collaboration with the National AIDS Control Organization (NACO). As per the report the recent estimates suggest a decline in adult HIV prevalence at national level by 50% over the past decade. At the same time, the report highlight the gradual rate of increase in the number of new infections over the past couple of years in certain states which are considered to be of low prevalence. At this crossroad, increasing HIV infection in states with high out-migration rate provided the impetus for this study. It is therefore inferred that the contribution of the phenomena of migration to the spread of the HIV epidemic is considerable.

The present work of spatial data mining examines the linkages between migration and HIV transmission along with understanding the socio economic determinants of HIV transmission in men and women, with a particular focus on male out-migration. The main objective of this research work is to develop an intelligent computational model which is able to take into account the important socio economic factors and develop a model which is able to forecast the
prevalence, growth or declining trend of the epidemic at various geographical locations. The knowledge given by the model can be used to perform spatial prediction that could help the policy makers to plan and monitor the impact of HIV prevention and care intervention program.
CHAPTER-2

LITERATURE REVIEW AND OBJECTIVE

2.1 INTRODUCTION

A huge volumes of data is being generated with the growing production of maps which exceed people's capacity to analyze them. Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. Generally speaking, geography and related spatial sciences have moved from a data-poor era to a data-rich era [12]. Therefore the knowledge discovery methods like data mining can appropriately be applied to spatial data. This recent concept is an extension of the conventional data mining tasks and is applied to alphanumerical and spatial data. However in spatial data mining the spatial analysis takes into account spatial relations between objects. During the last decade, due to the widespread applications of GPS technology, web-based spatial data sharing and mapping, high-resolution remote sensing, and
location-based services, more and more research domains have created or gained access to high-quality geographic data to incorporate spatial information and analysis in various studies, such as social analysis [13] and business applications [14].

For spatial data analysis and spatial data mining in database the two main teams have contributed a lot. The first one, DB Research Lab, developed GeoMiner [15], which is an extension of DBMiner and the second one devised as structure-of-neighborhood graph [16], on which some algorithms are based.

Important contribution has also been done by them on clustering methods which is based on association rules (based upon an efficient spatial join), classification (extension of ID3 and DBLearn), a hierarchical partitioning (extension of DBSCAN with a R*Tree), characterization and spatial trends. The clustering algorithm STING (University of California) uses a hierarchical grid [17] to perform optimization. It is worth mentioning here the work on Datawarehouse dedicated to spatial data (University of Laval) [18].

In spatial data mining (SDM) process we try to find out the spatial relationship, extract knowledge and any other properties that are not explicitly stored in the database. SDM is used to find implicit relations, regularities between spatial and/or non-spatial data.
Spatial data mining is becoming quite popular as there are many applicative areas such as ecology, transportation, epidemiology etc, where the concept is being utilized. There are many diversified approaches and algorithms that have been proposed in the literature to extract knowledge from such data. However the main challenge to apply any of the algorithm is how to do or rather automate the data preparation task, which consumes around 60-80% of the effort and time required for knowledge discovery in geographic databases. In spatial databases data are stored in different relations and it is required to joined them spatially for finding out novel and useful pattern of data.

The important property of SDM is that it interacts in space. A geographical database contains spatio-temporal data. In such databases the properties concerning a particular place are linked and explained in terms of the properties of its neighborhood. Thus in the analysis process the importance of spatial relationships is important. The temporal aspects for spatial data are also important but are rarely taken into account.

Since the conventional data mining methods [19] are not suited for spatial data as they neither support location data nor the implicit relationships between spatial objects. Therefore, it is needed to develop some new methods of spatial data mining to find spatial relationships. Calculating the spatial relationship is a
time consuming task in spatial database because the data is generated by encoding geometric location. Because of this complexity the global performances will suffer.

However, it is highly desired that the existing methods of data mining are extended and incorporated into spatial data mining methods. Spatial and GIS methods are crucial for spatial join, data access and graphical map display. The conventional data mining methods are able to generate knowledge about alphanumerical properties only.

### 2.2 SPATIAL DATA MINING TASKS

Spatial data mining tasks are the extension of conventional data mining tasks where spatial data and criteria are combined. The following tasks are performed in spatial data mining:

(i)Associations and dependencies identified to characterize data
(ii)Spatial and non-spatial data are summarized.
(iii)Classification rules are discovered
(iv)Clusters of similar objects are formed, and
(v)Deviations detection is done after looking for general trends.
To carry out these tasks different methods are used. Some of these methods are derived from statistics and others from the field of machine learning. Following is a summary of different SDM tasks:

<table>
<thead>
<tr>
<th>SI No</th>
<th>SDM Task</th>
<th>Statistic Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>Summarization</td>
<td>• Global Autocorelation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Density Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Smooth and Contrast analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Factorial analysis</td>
</tr>
<tr>
<td>ii</td>
<td>Generalization</td>
<td>• Concept hierarchy</td>
</tr>
<tr>
<td>iii</td>
<td>Classification / Class</td>
<td>• Spatial classification</td>
</tr>
<tr>
<td></td>
<td>identification</td>
<td></td>
</tr>
<tr>
<td>iv</td>
<td>Clustering</td>
<td>• Point Pattern analysis</td>
</tr>
<tr>
<td>v</td>
<td>Dependencies</td>
<td>• Local autocorrelation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Correspondence analysis</td>
</tr>
<tr>
<td>vi</td>
<td>Trends and deviations</td>
<td>• Krigging / Trend rule</td>
</tr>
</tbody>
</table>

Table 2.1: Different Spatial Data Mining Tasks
i. STATISTICAL DATA SUMMARIZATION

It is the most common approach in which we apply elementary statistics, such as the calculation of mean, average, variance, histogram and pie chart formation etc. For Neighborhood dependency measuring task some new methods have been developed, such as local variance, local covariance, and spatial autocorrelation such as Geary and Moran indices[15,20].

To establish spatial relationship between the objects we use the concept of contiguity matrix. In this we represent any spatial relationships, such as adjacency, a distance gap, and so on.

Contrary to the autocorrelation measure, density analysis forms part of Exploratory Spatial Data Analysis (ESDA), which does not require any knowledge about data. In this the non-spatial properties are ignored. In Geographic data analysis both the alphanumerical property data (called attributes) and spatial data are taken into consideration. This requires two things: using multidimensional data to analyze multiple attributes, and integrating spatial data with attributes in the analysis process.
ii. **GENERALISATION**

In this method abstract level of non-spatial attributes are raised and detail of geometric description are reduced by merging adjacent objects. It is in fact derived from attribute-oriented induction concept as described in [15]. The concept hierarchy is the important representation which can be spatial (like the hierarchy of administrative boundaries) or the thematic map of non-spatial data [21].

**Generalizations are of two types:**

1. Non-spatial dominant generalization: In this the thematic hierarchy is used first and then adjacent objects are merged.

2. Spatial dominant generalization: This generalization is based on a spatial hierarchy. After this the aggregation or generalization of non-spatial values for each generalized spatial value is formed.

iii. **CLASSIFICATION**

Classification is about grouping data items into classes (categories) according to their properties (attribute values). Classification is also called supervised classification, as opposed to the unsupervised classification (clustering). “Supervised” classification needs a training dataset to train (or configure) the classification model, a validation dataset to validate (or optimize)
the configuration, and a test dataset to evaluate the performance of the trained model.

Classification methods include, for example, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbor methods and case-based reasoning (CBR).

Spatial regression or prediction models form a special group of regression analysis that considers the independent and/or dependent variable of nearby neighbors in predicting the dependent variable at a specific location, such as the spatial autoregressive models (SAR) [22]. However, spatial regression methods such as SAR often involve the manipulation of an n by n spatial weight matrix, which is computationally intensive if n is large. Therefore, more recent research efforts have sought to develop approaches to find approximate solutions for SAR so that it can process very large data sets [23].

Classification tasks are called supervised classification which provides a logical description that partitions the database with some logical rule. The result of the classification rules is that it forms a decision tree. Each root of the decision tree contains a criterion on an attribute. When we consider spatial database this criterion could be a spatial predicate. Because spatial objects are dependent on
neighborhood, a rule involving the non-spatial properties of an object should be extended to neighborhood properties.

iv. CLUSTERING

Cluster analysis is mainly used for data analysis. It organizes a set of data items into groups known as clusters. The items in the same cluster are similar to each other and different from those in other clusters[24,25,26]. Many different clustering methods have been developed in various research fields such as statistics, pattern recognition, data mining, machine learning, and spatial analysis.

The methods of clustering are broadly classified into two groups. They are partitioning clustering and hierarchical clustering. K-means and self-organizing map (SOM) [27] are the partitioning clustering methods. They divide a set of data items into a number of non-overlapping clusters. A data item is assigned to the “closest” cluster based on a proximity or dissimilarity measure. Hierarchical clustering, on the other hand, organizes data items into a hierarchy with a sequence of nested partitions or groupings [25]. Commonly-used hierarchical clustering methods include the Ward’s method [28], single-linkage clustering, average-linkage clustering, and complete-linkage clustering [24,25].

To consider spatial information in clustering, three types of clustering analysis have been studied, including spatial clustering (i.e., clustering of spatial
points), regionalization (i.e., clustering with geographic contiguity constraints), and point pattern analysis (i.e., hot spot detection with spatial scan statistics). For the first type, spatial clustering, the similarity between data points or clusters is defined with spatial properties (such as locations and distances). Spatial clustering methods can be partitioning or hierarchical, density-based, or grid-based. Increasingly statistics for the detection of spatial clusters are available for non-Euclidean spaces, particularly network spaces [29,30,31].

There are many algorithms that have been proposed for performing clustering. They are CLARANS [32], DBSCAN [33] or STING [17]. They are mainly focused on cost optimization. There is a method, GDBSCAN[34], that has recently been proposed which is specifically applicable to spatial data. This method is applicable to any spatial shape of spatial data.

v. SPATIAL DATA DEPENDENCIES

It is the local autocorrelation method to work with spatial data which reflects how data are related. Spatial association rule, which works on spatial data dependence, have been adapted to spatial data.
Spatial auto-correlation is concerned with the assessment of the degree of spatial dependence. This is done by using the concept of spatial weight matrix \[35,36\]. It is equivalent to a residual test in regression analysis. This makes it possible to measure the difference between the actual spatial distribution of variable values and a random one.

Another approach is association rule which is well known in data mining and is applied to market analysis by looking for items that are frequently associated in a commercial transaction \[19\].

### 2.3 IMPORTANT METHODS

Generalization based method \[21,32,37\] is the widely used tuple-oriented technique in machine learning \[38\]. The method is often combined with generalization \[39\]. Since this approach does not handle inconsistent and noise data very well it cannot be used for large spatial database. The algorithms are also exponential in the number of examples. It requires the existence of background knowledge in the form of concept hierarchies. Now as per the requirement of the analysis the two kinds of concept hierarchies, spatial and non-spatial, are given by experts. Following (Figure:2) could be a concept hierarchy of epidemiology study.
As we move upward in the concept hierarchy the information becomes more and more general. Similar concept hierarchy can be formed for spatial data for example hierarchy related to region, state, district, village etc. W. Lu and J. Han[15] described two generalization based algorithms one \textit{spatial-datadominant} and another \textit{non-spatial-datadominant} generalization. In the first approach, until the spatial \textit{generalization threshold} is reached, the generalization of the spatial objects continues i.e. the no of region is not bigger than a threshold value. When the spatial oriented induction process is complete, non-spatial data are retrieved and analyzed for each of the spatial object using the attribute oriented induction technique.
In the second approach the algorithm performs attribute oriented induction, generalizing them to a higher concept level and for that it considers the non-spatial attributes. Here based on a threshold value the generalization approach determines whether to continue or stop the generalization process. Finally the neighboring area with the same generalized attributes are merged together based on the spatial function of adjacency (adjacent_to). For example the adjacent area having no of malaria epidemiology count, both in male and female, more than 5% of population are merged together forming a high-prevalence cluster of malaria epidemiology. Similarly low-prevalence and no-prevalence clusters can be identified.

Thus the concept hierarchy is generated automatically in the above described generalization based algorithms. However there are cases when in such hierarchy is not present a priori. At the same time the mined characteristic rules is going to be much dependent on the given concept hierarchy(ies). Such concept hierarchy is given by the experts and therefore the above mentioned approaches falls under the category of supervised classification or supervised knowledge discovery methods.

On the other hand we have some unsupervised techniques of knowledge discovery methods. The method of clustering is one such approach. The conventional clustering algorithms like PAM (Partition Around Medoids) or CLARA (Clustering LARge Applications) [40], are not appropriate from
computational complexity point of view. The difference between these two algorithms is that CLARA algorithm is based on sampling. CLARA can deal with large data set than PAM. Both PAM and CLARA were developed by Kaufman and Rousseeuw [40].

Then CLARANS was developed for cluster analysis and it outperformed the previous two algorithms. This algorithm was proposed by Ng and Han [32] which tries to mix both PAM and CLARA by searching only the subset of data set and it does not confirm itself to any sample at any given time. Experimentally it has been shown that CLARANS is more efficient than PAM and CLARA. Its every iteration computational complexity is linearly proportional to number of objects [33]. Some of the drawback of CLARANS has been pointed out by Ester, Kriegel, and Xu[33]. It assumes that the objects to be clustered are stored in main memory. For a large database it is not possible and hence a disk based method would be required. This method has been shorted out by integrating CLARANS with efficient spatial access methods, like R*-tree. But the construction of R*-tree is time consuming. Zhang, Ramakrishnan and Livny[41] resented another method BIRCH (Balanced Iterative Reducing and Clustering) for clustering of large set of points. The method is incremental one with possibility of adjustment of memory requirements to the size of memory that is available. It uses the concept called *Clustering Feature and CF tree.*
To minimize the number of costly spatial computation the two-step spatial computation technique [42] for optimization during the search for association was introduced. Spatial association rule is a rule that associate one or more spatial object with other spatial objects. Agarwal, Imielinski and Swami [43] introduced the concept of association rules in the study of mining large transaction database. Later Koperski and Han[44] extended this concept to spatial database. In order to discover the useful rule the concept of minimum support and minimum confidence are used. A strong rule is a rule having large support and large confidence.

2.4 OBJECTIVE

Discovery of implicit and useful knowledge in spatial database of epidemiology like HIV/AIDS is a highly demanding field because very large amount of spatial data has been collected and it is witnessed that this disease is showing location wise different characteristic in terms of its prevalence and growth. The spread of infectious disease is closely associated with the concepts of spatial and spatio-temporal proximity, as individuals who are linked in a spatial and temporal sense are at a high risk of getting infected. Proximity to environmental risk factors is therefore important. Thus acknowledging the spatial and temporal variations of disease and characterizing its spatial structure is essential for the epidemiologist to understand better the population’s interactions with its environment. Therefore it is highly required that some spatial data mining
framework is designed which is able to take into account the spatial and non spatial factors of the spread of this disease.

The general objective of this work is to formulate a predictive learning model of data mining based on the spatial database consisting of spatial and non spatial attributes related to some epidemiology.

The study of the HIV/AIDS based reports prompted us to undertake a deep research work of mining the spatial database of HIV epidemiology. Here my effort is to formulate a framework of spatial data mining which will consider the geographical (spatial), demographic, socio-economic factors. It will not only identify the pattern and root cause of the disease but also will predict the situation of the epidemiology like HIV/AIDS in particular and other epidemiology in general.

To carry out the spatial data mining task a spatial data base, consisting of spatial and non-spatial information, is required to be created. Further some efficient algorithm is required to be formulated to find out spatial and non spatial predicates for classification of spatial database objects. The method of finding out the predicate for classification and preparation of decision tree should be supported with the learning approach of ANN to propose an efficient learning and predictive model.
CHAPTER-3

METHODOLOGY

3.1 INTRODUCTION

Knowledge discovery from large spatial database or spatial data mining is the process of discovering spatial relation, extracting implicit knowledge or finding out other patterns from spatial database [45]. There has been a lot of research on data mining which set a foundation and provide some unique methods for exploring the concept of spatial data mining. Discovering useful data pattern from spatial database discloses interesting relationship among spatial and non-spatial data in large spatial database.

A spatial database can stores any kind of space related data, such as maps, topological and distance information and preprocessed remote sensing data or medical imaging data etc. For accessing they often require spatial reasoning, geometric computation, and spatial knowledge representation techniques and are accessed by spatial data access methods. Mining of such kind of spatial database requires integration of database technologies and data mining techniques.

Mining spatial databases is much more difficult than finding out the corresponding patterns from traditional numeric and characterized data due to the
complexity of spatial data types, spatial relationships, and spatial autocorrelation [1]. Spatial data mining has got some specialty that distinguishes it from conventional data mining and those are in the form of:

I. Data input
II. Statistical foundation
III. Computational process

I. DATA INPUT

The data input process of spatial database and the data mining process is relatively more complex than traditional data mining approach. It is because, in addition to non spatial data, in data input we consider extended objects like point, line and polygon. In this process the non spatial attributes are used to characterize the non-spatial feature of the spatial object. The non-spatial attribute of a spatial object are like name, population, area, length etc. and the spatial attribute of spatial objects are shape, elevation, latitude, longitude etc. The spatial objects and their spatial attributes are used to define the spatial location and spatial extent [46]. The relationships among spatial objects are implicit whereas relationships among non-spatial objects are explicit.
II. STATISTICAL FOUNDATION

To represents observations among random variables statistical models[47] are often used. These models can then be used for description, estimation and prediction based on probability theory. It is a fact that for data mining the conventional statistical approach is widely used, thus we can think that for spatial data mining also the same statistical approach is sufficient enough. However, as stated earlier, the method has got one limitation that it assumes statistical independence among spatial distributed data, which is not true because spatial objects are always interrelated. This assumption violates the Tobler’s [48] first law of geography which says that “Everything is related to everything else, but nearby things are more related than distant things”. Thus we can say that, of the nearby spatial objects, the values of attributes tend to systematically affect each other. In spatial statistics the concept of spatial autocorrelation is devoted to the analysis of spatial data where researchers have created, adapted, and applied statistical techniques to spatial data.

III. COMPUTATIONAL PROCESS

Many conventional data mining algorithms have been applied to spatial and non-spatial data in the form of spatial clustering, spatial association rule, spatial classification etc. The modern learning approach with genetic and neural network strategies and that to considering the fuzzy characteristics of spatial and non-
spatial data are yet to get established. In spatial database correlation based queries are computationally expensive thus the spatial indexing approach proposed by Zhang [46] uses spatial autocorrelation to facilitate correlation based queries.

In spatial data analysis various kinds of spatial predicates are involved. They represents topological relationship between spatial objects such as disjoints, intersects, equal, inside/outside, adjacent_to, and covers/covered_by etc. They can also represent spatial orientation or ordering such as north, east, left, right etc., or some distance information such as close_to, far_away etc.

3.2 HYPOTHESIS

Present work is based on the assumption that every spatial object stored in relational database is having spatial relationship with other object in terms of proximity characterized by distance parameter and dissimilarity in their attributes. We consider every Indian state as a separate spatial object. We find various computed and compiled statistical data that are available in various government reports. The data could be related to health, agriculture, economy etc. Now the question that automatically arises is there any spatial cause and effect relationship between the data so available.

The present work of spatial data mining tries to establish the cause and effect relationship between the spatial objects and the epidemiology related statistical data available in public domain.
Specifically the work focuses on the data related to the different pattern of HIV epidemiology in different states and the factor of distance measure related to the migration distance of the migrants. As migration by workers is found to be the major cause of HIV spread in the remote areas of India, this factor is required to be better analyzed with an objective of finding out more meaningful measure and its relation with the spread of HIV in different parts of India.

The assumptions of the work are enlisted below:

1. States of the country are considered as spatial objects.
2. The spatial and non spatial attributes are stored in a spatial database.
3. Such spatial database will help creating the appropriate spatial measure.
4. Distance measure of the worker migrants with respect to the major metropolitan destinations are considered here as a major spatial attribute.
5. The non spatial attributes such as HDI (Human Development Index) and GDI (Gender Development Index) are taken as the highly correlated socio economic factors that triggers in-migration or out-migration of the states.
6. The spatial temporal data will establish the trend of HIV prevalence either increasing or declining.
7. The HDI, GDI, migration distance measure and migration rate of the states are the major determinants of different HIV prevalence pattern of the state. This research work has tried to establish relationship between these antecedents and infection rate as consequent.
3.3 METHODOLOGY AND FRAMEWORK FOR THE PROPOSED RESEARCH

The present work proposes a model which covers various aspects of spatial data mining and gives a new framework of intelligent spatial data mining to study the spatial distribution of epidemiology (Appendix 3.1). The figure below (Figure: 3.1) is a diagrammatic representation of a spatial analysis framework:

![Diagram of Spatial Data Mining Framework]

Figure3.1 : Framework of spatial data mining task

1. Feature Data
2. Attribute Data
Following are the four active groups of the framework:

**A. DATA**

Data is the basic need of epidemiological analysis which is conducted for description of spatial patterns, identification of disease cluster, and explanation or prediction of disease risk [5]. Geographic data system includes georeferenced feature data and attributes, be they point and area. These data are obtained by taking field survey, remotely sensed imagery or use of existing data generated either by government organizations or those closely linked to government such as cadastral, meteorological or national census statistics and health organizations.

In the present work the state boundary geographical data has been taken from NSDI (National Spatial Database Infrastructure), Govt. of India, and other epidemiology and socio-economic data has been taken from NACO, Ministry of Statistics & Program Implementation, Human development research paper 2009/13, and India labour market report 200. The development of spatial database and computation of spatial measure has been covered in Chapter 4.
B. VISUALIZATION AND EXPLORATION

It covers technique that focus solely on examining the spatial dimension of the data. Visualization tools are used resulting in maps that describe spatial patterns and which are useful for both stimulating more complex analysis and for communicating the results of such analysis. Statistical methods are used in spatial data exploration to determine whether observed patterns are random in space. However there is some overlap between visualization and exploration, since meaningful visual presentation will require the use of quantitative analytical methods.

For the present work the MapViewer tool of Oracle has been used for visualization. In the present work the MapViewer tool of Oracle has been used for visualization. The visualization of the spatial computational output has been demonstrated in Chapter 4, Chapter 7 and in Chapter 8.

C. MODELING

Modeling introduces the concept of cause-effect relationships using both spatial and non-spatial data sources to explain or predict spatial patterns. It involves the activity of deciding the methodology, applying the method to work
with real life data and validation of the model with test data. The proposed methodology of this work can be classified as follows:

1. Proposing a model of spatial data mining which utilizes the potential of spatial database, spatial statistic methods and modern computational techniques which are capable of describing non-linear relationship among spatial and non-spatial attribute values of spatial database objects (database tuples). The activities that would be design the following framework:

   a) Spatial database

   b) Thematic layer of non spatial and spatial data

   c) Computational model of rule based classification with spatial predicate.

   d) Validation of the model with Spatial predicate based neural network- BPN model and fuzzy NN-BPN model.

2. Applying the model in the study of spatial distribution of epidemiology and proposing a predictive model which would highlight the spatial and non-spatial characteristics responsible in the spread and growth of the epidemiology. This work would help planner to undertake corrective measures timely and accurately.

3. For training, validation and testing the non spatial temporal data from year 2002 to 2011 has been used.
For analyzing spatial data the statistical spatial analysis has been the most common approach. It handles very efficiently the numerical data which comes from the realistic model of spatial phenomena. It is worth mentioning here that our conventional assumption about statistical data assumes statistical independence among the distributed data. In this work this assumption is not relevant because the spatial objects are influenced by their neighboring objects. At the same time the statistical approach cannot model non linear rules very well. In case of incomplete and inconsistent data the conventional statistical methods do not work well. Another problem that is related to statistical spatial analysis is the expensive computation of the result. To supplement the work the soft computing techniques and the spatial database potential has nicely utilized.

Data mining and knowledge discovery is an interactive process that involves multiple step, including data selection, cleaning, preprocessing, and transformation; incorporating the prior knowledge; analysis with computational algorithms and/or visible approaches, interpretation and evaluation of the results; formulation or modification of hypotheses and theories; adjustment of data and analysis method; evaluation of result again and so on[50]. Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, whether computational, statistical visual, or some combination of them [50].

36
Development of spatial database and computation of spatial measure is described in chapter 4. The further work of doing rule based spatial classification in spatial database for getting classification rule has been done which is described in Chapter 5. In order to examine the non linear behavior of the data model a neural network-BPN model and a fuzzy NN-BPN model is proposed to be developed to work with the spatial database. To carry out the work a PL/SQL based program has been developed. This task is described in Chapter 6 and Chapter 7 respectively. The detail description about intelligence in spatial data mining task is described in Appendix 3.2
CHAPTER-4

DEVELOPMENT OF SPATIAL DATABASE AND COMPUTATION OF SPATIAL MEASURE

4.1 INTRODUCTION

Large amount of spatial data is being collected in various applications like remote sensing, computer cartography, geographical information system (GIS), environmental assessment and planning, etc. Now the real challenge is to discover interesting, implicit, and previously unknown knowledge from this large data set. This is what the objective is of spatial data mining. The real work is to extend the scope of data mining from relational and transactional database to spatial database and apply it in the study of spatial distribution of epidemiology.

In India the geographical disparities in the levels of economic development are very profound. Such disparities have contributed a lot to influence migratory patterns from impoverished rural areas to prospective urban destinations [51]. In addition to socio-geographical disparity and the urban-rural divide, there are consistent differences in indicators of socioeconomic development between the
states of India. Such regional disparities in the level of literacy are also very significant.

The various reports published by government agencies shows that it is a geographically heterogeneous distribution of HIV/AIDS epidemiology (Appendix 4.1). In India the states shows altogether different cultures, language, status if livelihood etc. There are some vital indicators published in the government reports that shows the social and gender disparities in different states. Some of the highly correlated indicators are human development index (HDI), gender development index (GDI), rate of migration, difference in the literacy rate of male and female etc.

The main objective of the present work is to develop an intelligent model which is able to take into account the important socio economic and spatial factors and forecast the prevalence, growth or declining trend of the epidemic like HIV/AIDS at various geographical locations [52,53]. The knowledge given by the model could help the policy makers to plan and monitor the impact of HIV prevention and care intervention program.

The work related to statistics [40,54,55,56,57], machine learning[38,39,58] and database systems [59,60] laid the foundation of knowledge discovery from database. Then after, with respect to spatial database, the study related to
computational geometry [61], spatial data structure[62,63,64] and spatial reasoning [65,66] paved the way for the study of spatial data mining. The detail description of spatial database and the assumption of this research work is described in Appendix 4.2

4.2 SPATIAL AND NON SPATIAL DATASET

The spatial database consists of spatial objects and non-spatial description. The non-spatial description of the spatial object can be stored in the traditional relational database. There are two different properties of spatial data and they are geometric and topological. The geometric properties could be spatial location, area, perimeter etc. whereas topological properties can be adjacencies, inclusion etc. The figure below (Figure:4.1) describes how the non-spatial and spatial attribute values about the states of India are mapped in a database.

Figure 4.1: Spatial and non-spatial attribute of Indian States.
4.3 SPATIAL DATA PREPARATION

Spatial and non spatial data is the basic need of epidemiological analysis which is conducted for description of spatial patterns, identification of disease cluster, and explanation or prediction of disease risk. Geographic data system includes geo referenced feature data and attributes, be they point and area. These data are obtained by taking field survey, remotely sensed imagery or use of existing data generated either by government organizations or those closely linked to government such as cadastral, meteorological or national census statistics and health organizations.

The explosive growth of spatial database is in fact posing challenge to interpret it. This creates an urgent need for new technology and tools that support the human in transforming the data into useful information and knowledge. Spatial Database Management System (SDBMS) is the database systems for the management of spatial data [67]. Spatial Data Mining (SDM) is the process to find the implicit regularities, rules or patterns hidden in such large spatial database [16,19,21,68].

There are many solutions that have been proposed in the literature for spatial data mining, but we can see that only a few of them focuses on data
preparation aspects. Therefore a carefully designed spatial database would reduce the burden of data preparation task for data mining. Rather it will help in providing relevant and filtered data to the mining algorithm and will keep on maintaining spatial relation.

Spatial database [SDB] stores spatial features, which are the real world entities, located in specific region. Spatial features (e.g. UP, Rajasthan) belongs to a feature type (e.g. state) and have both spatial (geographic coordinates, x, y) and non – spatial (e.g. name, population etc) attributes. Most of the conventional data mining algorithms runs on separate spatial data file specifically designed for spatial data set. However if we are able to integrate this spatial data with a database management system then it will have many advantages such as inconsistency and redundancy that would be removed. The potential of various indexing methods of a DBMS system would help running various database queries efficiently. All this will speed up the steps of data mining.

When we analyze public health data the most important factors that are to be considered are disease count, proportions or rate. However these counts and rates are not continuous outcome as it is generally seen in linear regression. Modeling spatial relation that arises in spatially referenced data is commonly done by incorporating spatial dependence into the covariance structure either explicitly or implicitly via an auto regressive model. The influence between two location
based events depends on factors such as topology, distance between the events and direction. For example a new industrial plant may pollute the neighboring localities which will mainly depend upon distance of the locality and direction of the wind. Topology is a branch of mathematics which studies the relationships of spatial objects that do not change due to elastic deformation of underlying space. Topological relationships such as adjacent, connect, inside etc. plays an important role to spatial data analysis. Similarly the topological relation of migrant state & destination states are of quite significance.

**4.3.1 CONFIGURING SPATIAL AND NON SPATIAL DATABASE**

Here we have created the database of thirty two Indian states and union territories (combined state of UP & Uttarakhand, Bihar & Jharkhand, MP & Chattisgarh) with their spatial (scaled state boundary coordinates) information and non spatial temporal data like epidemiology data and socio economic data. The scaled boundary coordinates has been made available from NSDI. The example layout of one state boundary is shown in table below (table 4.1):
<table>
<thead>
<tr>
<th>scode</th>
<th>sname</th>
<th>GEOM(SDO_GTYPE, SDO_SRID, SDO_POINT(X, Y, Z), SDO_ELEM_INFO, SDO_ORDINATES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andhra Pradesh</td>
<td>SDO_GEOMETRY(3, NULL, NULL, SDO_ELEM_INFO_ARRAY(1, 3, 1), SDO_ORDINATE_ARRAY(82.00063, 17.95354, 82.11718, 18.02457, 82.24023, 17.99232, 82.28539, 18.02516, 82.36536, 18.24984, 82.33528, 18.39703, 82.39066, 18.45479, 82.52105, 18.44178, 82.57897, 18.21425, 82.64119, 18.2096, 82.80577, 18.33431, 82.83324, 18.4054, 82.92, 468, 18.33981, 83.05514, 18.33667, 83.02056, 18.5853, 83.14518, 18.73362, 83.254, 39, 18.72796, -----))</td>
</tr>
<tr>
<td>Query 1</td>
<td>CREATE TABLE INDIA_ST (</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCODE NUMBER(2),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STATE VARCHAR2(35),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GEOM MDSYS.SDO_GEOMETRY);</td>
<td></td>
</tr>
<tr>
<td>Query 2</td>
<td>DELETE FROM USER_SDO_GEOM_METAD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TABLE_NAME = 'INDIA_ST' AND</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COLUMN_NAME = 'GEOM' ;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>INSERT INTO USER_SDO_GEOM_METAD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(TABLE_NAME, COLUMN_NAME, DIMINFO,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SRID) VALUES ('INDIA_ST', 'GEOM',</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDSYS.SDO_DIM_ARRAY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(MDSYS.SDO_DIM_ELEMENT('X',</td>
<td></td>
</tr>
<tr>
<td></td>
<td>68.358590000, 97.462120000, 0.000000050),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDSYS.SDO_DIM_ELEMENT('Y',  6.733651000,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.034730000, 0.000000050) ;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NULL); COMMIT;</td>
<td></td>
</tr>
<tr>
<td>Query 3</td>
<td>CREATE INDEX india_st_idx</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ON india_st(GEOM) INDEXTYPE IS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDSYS.SPATIAL_INDEX;</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Spatial queries for creating and populating spatial database in Oracle.

**Query 1**: Creates the table india_st which contains non spatial columns SCODE and NAME and one spatial column GEOM.
**Query 2**: Populate the USER_SDO_GEOM_METADATA view to reflect the dimensional information for the areas. Spatial users are responsible for populating these views. For each spatial column, we must insert an appropriate row into the USER_SDO_GEOM_METADATA view. This Spatial database ensures that the ALL_SDO_GEOM_METADATA view is also updated to reflect the rows that we insert into USER_SDO_GEOM_METADATA.

**Query 3**: After the data is loaded in the INDIA_ST table, we create the spatial index with this command. The statement creates R-tree index.

Thus our india_st table contains non spatial fields SCODE (state code) and STATE (state name) and one spatial column GEOM (coordinate of state boundary). The field SCODE will serve the purpose of key column to join the india_st table with other state related non spatial tables which contains information regarding various epidemiology and socio-economic factors. The whole database schema related to my work is described below (Figure 4.2). Putting the non spatial data related to a spatial object into a spatial database will provide us the opportunity to consider various topological relations.
Figure 4.2: Database Schema
4.3.2 WORKING WITH SPATIAL DATABASE

The spatial relationships among the spatial data are initially implicit. They require spatial join operations (i.e. matching data collections according to spatial criterion). The problem is that the spatial joins are fastidious and time consuming. Here our effort would be to make the relationships explicit. It is also our idea to bring back them to semantic properties as well as other relational references. The method consists broadly of two stages:

1. Data Cleaning: It is done by elimination attribute (information) that is useless for analysis. The combination of projection and combination of the relational database model makes this happen.

2. Retrieving spatial relationship: This is done by spatial join using spatial join index. In this schema the India_st table contains the GEOM(Geometry) attribute that represents the state coordinate system. The value of this attribute is used for spatial join. Based on this value I computed a spatial matrix which is termed as ‘Distance Measure’ of each state. These spatial attribute values along with the various thematic values of non-spatial attributes are used together for further analysis.
**Spatial Queries:** The spatial database queries constitute a way of analyzing spatial data which can use spatial relationship predicates. For example, we can query the database for the states having more than 5% of HIV infected peoples where the population is more than 50,000. However, such analysis are much confirmatory than exploratory in nature. Another disadvantage of spatial database query is the lack of advance statistic computation and spatial statistic models [69] such as Moran and Geary indices. However this approach can be used in filtering phase of the knowledge discovery process i.e. while preparing the dataset on which the analysis will focus.

The database objects has been examined and three different class labels has been identified which are as follows:

i. States with growing trend of HIV infection.

ii. States with decreasing trend of HIV infection

iii. States with stable trend of HIV infection.

For this the year wise HIV infection data value (in percentage) stored in database has been queried. The state wise HIV prevalence temporal data set data is shown in table below (table 4.3).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDHRA PRADESH</td>
<td>1.16</td>
<td>1.13</td>
<td>1.1</td>
<td>1.08</td>
<td>1.05</td>
</tr>
<tr>
<td>ARUNACHAL PRADESH</td>
<td>0.18</td>
<td>0.18</td>
<td>0.09</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>ASSAM</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>BIHAR</td>
<td>0.1</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>GOA</td>
<td>1.01</td>
<td>0.92</td>
<td>0.84</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>GUJARAT</td>
<td>0.54</td>
<td>0.51</td>
<td>0.48</td>
<td>0.45</td>
<td>0.43</td>
</tr>
<tr>
<td>HARYANA</td>
<td>0.5</td>
<td>0.32</td>
<td>0.21</td>
<td>0.14</td>
<td>0.1</td>
</tr>
<tr>
<td>HP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>J&amp;K</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>KARNATAKA</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>KERALA</td>
<td>0.59</td>
<td>0.39</td>
<td>0.25</td>
<td>0.17</td>
<td>0.13</td>
</tr>
<tr>
<td>MADHYA PRADESH</td>
<td>0.17</td>
<td>0.15</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>MAHARASHTRA</td>
<td>1.08</td>
<td>0.98</td>
<td>0.89</td>
<td>0.8</td>
<td>0.74</td>
</tr>
<tr>
<td>MANIPUR</td>
<td>2.42</td>
<td>2.2</td>
<td>2.01</td>
<td>1.83</td>
<td>1.67</td>
</tr>
<tr>
<td>MEGHALAYA</td>
<td>0.19</td>
<td>0.14</td>
<td>0.1</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>MIZORAM</td>
<td>1.13</td>
<td>0.99</td>
<td>0.91</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>NAGALAND</td>
<td>2</td>
<td>1.83</td>
<td>1.62</td>
<td>1.45</td>
<td>1.26</td>
</tr>
<tr>
<td>ORISSA</td>
<td>0.06</td>
<td>0.08</td>
<td>0.11</td>
<td>0.15</td>
<td>0.22</td>
</tr>
<tr>
<td>PUNJAB</td>
<td>0.18</td>
<td>0.16</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>RAJASTHAN</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
<td>0.12</td>
<td>0.17</td>
</tr>
<tr>
<td>SIKKIM</td>
<td>0.24</td>
<td>0.17</td>
<td>0.13</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>TAMIL NADU</td>
<td>0.93</td>
<td>0.73</td>
<td>0.59</td>
<td>0.47</td>
<td>0.39</td>
</tr>
<tr>
<td>TRIPURA</td>
<td>0.41</td>
<td>0.29</td>
<td>0.21</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>UTTAR PRADESH</td>
<td>0.14</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>WEST BENGAL</td>
<td>0.1</td>
<td>0.13</td>
<td>0.16</td>
<td>0.21</td>
<td>0.3</td>
</tr>
<tr>
<td>A&amp;N</td>
<td>0.81</td>
<td>0.66</td>
<td>0.54</td>
<td>0.44</td>
<td>0.37</td>
</tr>
<tr>
<td>CHANDIGARH</td>
<td>0.45</td>
<td>0.43</td>
<td>0.38</td>
<td>0.35</td>
<td>0.34</td>
</tr>
<tr>
<td>D&amp;NH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAMAN AND DIU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DELHI</td>
<td>0.35</td>
<td>0.32</td>
<td>0.3</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>LAKSHADWEEP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PONDICHERRY</td>
<td>0.4</td>
<td>0.43</td>
<td>0.47</td>
<td>0.5</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 4.3: HIV Prevalence data 2002 to 2006
Now from 2002 to 2006 we see that the HIV infection difference (INFDIFF) is of three types. Some states are showing decline in infection, some are showing growth in its prevalence whereas some are showing stable state over this period. Figure 4.3 below shows the state code wise three states of HIV prevalence.

Figure 4.3: State wise HIV infection patterns from 2002 to 2006
Now the following set of queries divides the states into three categories hence three class labels gets created:

Table 4.4: Queries for creating class labels
<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>INFDIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>MAHARASHTRA</td>
<td>-0.19</td>
</tr>
<tr>
<td>14</td>
<td>MANIPUR</td>
<td>-0.41</td>
</tr>
<tr>
<td>15</td>
<td>MEGHALAYA</td>
<td>-0.09</td>
</tr>
<tr>
<td>16</td>
<td>MIZORAM</td>
<td>-0.22</td>
</tr>
<tr>
<td>17</td>
<td>NAGALAND</td>
<td>-0.38</td>
</tr>
<tr>
<td>19</td>
<td>PUNJAB</td>
<td>-0.04</td>
</tr>
<tr>
<td>21</td>
<td>SIKKIM</td>
<td>-0.11</td>
</tr>
<tr>
<td>22</td>
<td>TAMIL NADU</td>
<td>-0.34</td>
</tr>
<tr>
<td>23</td>
<td>TRIPURA</td>
<td>-0.2</td>
</tr>
<tr>
<td>29</td>
<td>ANDAMAN AND NICOBAR ISLANDS</td>
<td>-0.27</td>
</tr>
<tr>
<td>1</td>
<td>ANDHRA PRADESH</td>
<td>-0.06</td>
</tr>
<tr>
<td>2</td>
<td>ARUNACHAL PRADESH</td>
<td>-0.09</td>
</tr>
<tr>
<td>3</td>
<td>ASSAM</td>
<td>-0.04</td>
</tr>
<tr>
<td>30</td>
<td>CHANDIGARH</td>
<td>-0.07</td>
</tr>
<tr>
<td>33</td>
<td>DELHI</td>
<td>-0.05</td>
</tr>
<tr>
<td>5</td>
<td>GOA</td>
<td>-0.17</td>
</tr>
<tr>
<td>6</td>
<td>GUJARAT</td>
<td>-0.06</td>
</tr>
<tr>
<td>7</td>
<td>HARYANA</td>
<td>-0.29</td>
</tr>
<tr>
<td>10</td>
<td>KARNATAKA</td>
<td>-0.03</td>
</tr>
<tr>
<td>11</td>
<td>KERALA</td>
<td>-0.34</td>
</tr>
<tr>
<td>12</td>
<td>MADHYA PRADESH</td>
<td>-0.04</td>
</tr>
<tr>
<td>24</td>
<td>UTTAR PRADESH</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Table 4.5: Output of query 1 i.e. epidemiology decreasing class
Table 4.6: Output of query 2 i.e. epidemiology increasing category

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>INFDIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>ORISSA</td>
<td>0.05</td>
</tr>
<tr>
<td>35</td>
<td>PONDICHERRY</td>
<td>0.07</td>
</tr>
<tr>
<td>20</td>
<td>RAJASTHAN</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>BIHAR</td>
<td>0.02</td>
</tr>
<tr>
<td>25</td>
<td>WEST BENGAL</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 4.7: Output of query 3 i.e. epidemiology stable category

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>INFDIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIMACHAL</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PRADESH</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>JAMMU AND KASHMIR</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>KASHMIR</td>
<td>0</td>
</tr>
</tbody>
</table>
The data related to the rate of migration from all the states from 2002 to 2006 is as follows:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDHRA PRADESH</td>
<td>-0.07</td>
<td>-0.09</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>ARUNACHAL PRADESH</td>
<td>6.5</td>
<td>6.7</td>
<td>6.9</td>
<td>7.1</td>
<td>7.2</td>
</tr>
<tr>
<td>ASSAM</td>
<td>-1.1</td>
<td>-0.9</td>
<td>-0.8</td>
<td>-0.71</td>
<td>-0.7</td>
</tr>
<tr>
<td>BIHAR</td>
<td>-3.2</td>
<td>-3.1</td>
<td>-2.9</td>
<td>-2.8</td>
<td>-2.7</td>
</tr>
<tr>
<td>GOA</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>GUJARAT</td>
<td>1.1</td>
<td>1.3</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
</tr>
<tr>
<td>HARYANA</td>
<td>5</td>
<td>3.2</td>
<td>4.7</td>
<td>4</td>
<td>4.1</td>
</tr>
<tr>
<td>HIMACHAL PRADESH</td>
<td>0.8</td>
<td>1.1</td>
<td>0.94</td>
<td>1.2</td>
<td>1</td>
</tr>
<tr>
<td>JAMMU AND KASHMIR</td>
<td>-0.3</td>
<td>-0.5</td>
<td>-0.54</td>
<td>-0.48</td>
<td>-0.4</td>
</tr>
<tr>
<td>KARNATAKA</td>
<td>1.1</td>
<td>0.7</td>
<td>0.58</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>KERALA</td>
<td>-1.2</td>
<td>-0.99</td>
<td>-1.1</td>
<td>-0.81</td>
<td>-0.6</td>
</tr>
<tr>
<td>MADHYA PRADESH</td>
<td>-0.2</td>
<td>-0.4</td>
<td>-0.3</td>
<td>-0.13</td>
<td>-0.1</td>
</tr>
<tr>
<td>MAHARASHTRA</td>
<td>4</td>
<td>3.5</td>
<td>5</td>
<td>4.3</td>
<td>3</td>
</tr>
<tr>
<td>MANIPUR</td>
<td>-1.8</td>
<td>-2.1</td>
<td>-1.9</td>
<td>-1.6</td>
<td>-1.4</td>
</tr>
<tr>
<td>MEGHALAYA</td>
<td>0.9</td>
<td>0.78</td>
<td>0.81</td>
<td>0.79</td>
<td>0.8</td>
</tr>
<tr>
<td>MIZORAM</td>
<td>-0.1</td>
<td>-0.17</td>
<td>-0.21</td>
<td>-0.18</td>
<td>-0.1</td>
</tr>
<tr>
<td>NAGALAND</td>
<td>-1.7</td>
<td>-0.98</td>
<td>-1.21</td>
<td>-1.2</td>
<td>-1.4</td>
</tr>
<tr>
<td>ORISSA</td>
<td>-1.2</td>
<td>-1.6</td>
<td>-0.98</td>
<td>-1.1</td>
<td>-0.7</td>
</tr>
<tr>
<td>PUNJAB</td>
<td>2.5</td>
<td>3.4</td>
<td>2.9</td>
<td>1.98</td>
<td>1.7</td>
</tr>
<tr>
<td>RAJASTHAN</td>
<td>-1.8</td>
<td>-1.76</td>
<td>-0.93</td>
<td>-0.81</td>
<td>-0.6</td>
</tr>
<tr>
<td>SIKKIM</td>
<td>6.1</td>
<td>5.89</td>
<td>4.67</td>
<td>5.1</td>
<td>5.9</td>
</tr>
<tr>
<td>TAMIL NADU</td>
<td>-0.82</td>
<td>-0.76</td>
<td>-0.83</td>
<td>-1.1</td>
<td>-0.7</td>
</tr>
<tr>
<td>TRIPURA</td>
<td>0.9</td>
<td>0.91</td>
<td>1.1</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>UTTAR PRADESH</td>
<td>-2.1</td>
<td>-1.9</td>
<td>-2.8</td>
<td>-2.7</td>
<td>-2</td>
</tr>
<tr>
<td>WEST BENGAL</td>
<td>0.6</td>
<td>0.71</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>ANDAMAN AND NICOBAR</td>
<td>6.1</td>
<td>7.1</td>
<td>6.9</td>
<td>5.89</td>
<td>7.9</td>
</tr>
<tr>
<td>CHANDIGARH</td>
<td>18</td>
<td>17</td>
<td>20.3</td>
<td>21.2</td>
<td>21.4</td>
</tr>
<tr>
<td>DADRA AND NAGAR HAVELI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32.6</td>
</tr>
<tr>
<td>DAMAN AND DIU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44.1</td>
</tr>
<tr>
<td>DELHI</td>
<td>19.2</td>
<td>17.9</td>
<td>18.1</td>
<td>18.6</td>
<td>18.7</td>
</tr>
<tr>
<td>LAKSHADWEEP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.4</td>
</tr>
<tr>
<td>PONDICHERRY</td>
<td>7.9</td>
<td>8.2</td>
<td>7.9</td>
<td>8.6</td>
<td>8.8</td>
</tr>
</tbody>
</table>

Table 4.8: state wise migration rate temporal data
The Human Development Index of the states during 2002 to 2006 is as follows:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDHRA PRADESH</td>
<td>0.494</td>
<td>0.519</td>
<td>0.526</td>
<td>0.574</td>
<td>0.585</td>
</tr>
<tr>
<td>ARUNACHAL PRADESH</td>
<td>0.5992</td>
<td>0.612</td>
<td>0.623</td>
<td>0.637</td>
<td>0.647</td>
</tr>
<tr>
<td>ASSAM</td>
<td>0.498</td>
<td>0.514</td>
<td>0.527</td>
<td>0.584</td>
<td>0.595</td>
</tr>
<tr>
<td>BIHAR</td>
<td>0.492</td>
<td>0.498</td>
<td>0.501</td>
<td>0.502</td>
<td>0.507</td>
</tr>
<tr>
<td>GOA</td>
<td>0.758</td>
<td>0.711</td>
<td>0.721</td>
<td>0.761</td>
<td>0.764</td>
</tr>
<tr>
<td>GUJARAT</td>
<td>0.612</td>
<td>0.628</td>
<td>0.63</td>
<td>0.631</td>
<td>0.634</td>
</tr>
<tr>
<td>HARYANA</td>
<td>0.598</td>
<td>0.611</td>
<td>0.624</td>
<td>0.639</td>
<td>0.643</td>
</tr>
<tr>
<td>HIMACHAL PRADESH</td>
<td>0.612</td>
<td>0.628</td>
<td>0.649</td>
<td>0.659</td>
<td>0.667</td>
</tr>
<tr>
<td>JAMMU AND KASHMIR</td>
<td>0.51</td>
<td>0.52</td>
<td>0.532</td>
<td>0.547</td>
<td>0.59</td>
</tr>
<tr>
<td>KARNATAKA</td>
<td>0.611</td>
<td>0.628</td>
<td>0.621</td>
<td>0.6005</td>
<td>0.622</td>
</tr>
<tr>
<td>KERALA</td>
<td>0.712</td>
<td>0.724</td>
<td>0.731</td>
<td>0.758</td>
<td>0.764</td>
</tr>
<tr>
<td>MADHYA PRADESH</td>
<td>0.511</td>
<td>0.521</td>
<td>0.527</td>
<td>0.528</td>
<td>0.529</td>
</tr>
<tr>
<td>MAHARASHTRA</td>
<td>0.654</td>
<td>0.661</td>
<td>0.625</td>
<td>0.678</td>
<td>0.689</td>
</tr>
<tr>
<td>MANIPUR</td>
<td>0.701</td>
<td>0.698</td>
<td>0.699</td>
<td>0.701</td>
<td>0.702</td>
</tr>
<tr>
<td>MEGHALAYA</td>
<td>0.601</td>
<td>0.615</td>
<td>0.621</td>
<td>0.626</td>
<td>0.629</td>
</tr>
<tr>
<td>MIZORAM</td>
<td>0.651</td>
<td>0.661</td>
<td>0.671</td>
<td>0.682</td>
<td>0.688</td>
</tr>
<tr>
<td>NAGALAND</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>ORISSA</td>
<td>0.521</td>
<td>0.529</td>
<td>0.532</td>
<td>0.535</td>
<td>0.537</td>
</tr>
<tr>
<td>PUNJAB</td>
<td>0.642</td>
<td>0.653</td>
<td>0.661</td>
<td>0.665</td>
<td>0.668</td>
</tr>
<tr>
<td>RAJASTHAN</td>
<td>0.522</td>
<td>0.529</td>
<td>0.534</td>
<td>0.539</td>
<td>0.541</td>
</tr>
<tr>
<td>SIKKIM</td>
<td>0.641</td>
<td>0.649</td>
<td>0.655</td>
<td>0.661</td>
<td>0.665</td>
</tr>
<tr>
<td>TAMIL NADU</td>
<td>0.63</td>
<td>0.645</td>
<td>0.656</td>
<td>0.661</td>
<td>0.666</td>
</tr>
<tr>
<td>TRIPURA</td>
<td>0.611</td>
<td>0.627</td>
<td>0.642</td>
<td>0.657</td>
<td>0.663</td>
</tr>
<tr>
<td>UTTAR PRADESH</td>
<td>0.515</td>
<td>0.519</td>
<td>0.522</td>
<td>0.527</td>
<td>0.528</td>
</tr>
<tr>
<td>WEST BENGAL</td>
<td>0.621</td>
<td>0.629</td>
<td>0.637</td>
<td>0.641</td>
<td>0.642</td>
</tr>
<tr>
<td>ANDAMAN AND NICOBAR</td>
<td>0.701</td>
<td>0.704</td>
<td>0.706</td>
<td>0.707</td>
<td>0.708</td>
</tr>
<tr>
<td>CHANDIGARH</td>
<td>0.752</td>
<td>0.769</td>
<td>0.772</td>
<td>0.782</td>
<td>0.784</td>
</tr>
<tr>
<td>DADRA AND NAGAR HAVELI</td>
<td>0.642</td>
<td>0.651</td>
<td>0.666</td>
<td>0.672</td>
<td>0.677</td>
</tr>
<tr>
<td>DAMAN AND DIU</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>DELHI</td>
<td>0.51</td>
<td>0.67</td>
<td>0.71</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>LAKSHADweep</td>
<td>0.615</td>
<td>0.644</td>
<td>0.677</td>
<td>0.682</td>
<td>0.697</td>
</tr>
<tr>
<td>PONDICHERRY</td>
<td>0.712</td>
<td>0.716</td>
<td>0.721</td>
<td>0.725</td>
<td>0.729</td>
</tr>
</tbody>
</table>

Table 4.9: State wise HDI temporal data
The Gender Development Index of the states during 2002 to 2006 is as follows:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDHRA PRADESH</td>
<td>0.522</td>
<td>0.537</td>
<td>0.545</td>
<td>0.562</td>
<td>0.574</td>
</tr>
<tr>
<td>ARUNACHAL PRADESH</td>
<td>0.631</td>
<td>0.635</td>
<td>0.639</td>
<td>0.641</td>
<td>0.642</td>
</tr>
<tr>
<td>ASSAM</td>
<td>0.552</td>
<td>0.567</td>
<td>0.578</td>
<td>0.581</td>
<td>0.585</td>
</tr>
<tr>
<td>BIHAR</td>
<td>0.437</td>
<td>0.444</td>
<td>0.456</td>
<td>0.475</td>
<td>0.479</td>
</tr>
<tr>
<td>GOA</td>
<td>0.721</td>
<td>0.732</td>
<td>0.739</td>
<td>0.745</td>
<td>0.747</td>
</tr>
<tr>
<td>GUJARAT</td>
<td>0.612</td>
<td>0.617</td>
<td>0.622</td>
<td>0.623</td>
<td>0.624</td>
</tr>
<tr>
<td>HARYANA</td>
<td>0.612</td>
<td>0.619</td>
<td>0.627</td>
<td>0.629</td>
<td>0.632</td>
</tr>
<tr>
<td>HIMACHAL PRADESH</td>
<td>0.651</td>
<td>0.656</td>
<td>0.661</td>
<td>0.663</td>
<td>0.664</td>
</tr>
<tr>
<td>JAMMU AND KASHMIR</td>
<td>0.515</td>
<td>0.535</td>
<td>0.545</td>
<td>0.565</td>
<td>0.568</td>
</tr>
<tr>
<td>KARNATAKA</td>
<td>0.601</td>
<td>0.605</td>
<td>0.608</td>
<td>0.609</td>
<td>0.611</td>
</tr>
<tr>
<td>KERALA</td>
<td>0.721</td>
<td>0.731</td>
<td>0.735</td>
<td>0.741</td>
<td>0.745</td>
</tr>
<tr>
<td>MADHYA PRADESH</td>
<td>0.502</td>
<td>0.509</td>
<td>0.512</td>
<td>0.514</td>
<td>0.516</td>
</tr>
<tr>
<td>MAHARASHTRA</td>
<td>0.641</td>
<td>0.652</td>
<td>0.663</td>
<td>0.671</td>
<td>0.677</td>
</tr>
<tr>
<td>MANIPUR</td>
<td>0.665</td>
<td>0.672</td>
<td>0.682</td>
<td>0.691</td>
<td>0.699</td>
</tr>
<tr>
<td>MEGHALAYA</td>
<td>0.605</td>
<td>0.615</td>
<td>0.619</td>
<td>0.621</td>
<td>0.624</td>
</tr>
<tr>
<td>MIZORAM</td>
<td>0.661</td>
<td>0.669</td>
<td>0.672</td>
<td>0.682</td>
<td>0.687</td>
</tr>
<tr>
<td>NAGALAND</td>
<td>0.665</td>
<td>0.679</td>
<td>0.682</td>
<td>0.691</td>
<td>0.697</td>
</tr>
<tr>
<td>ORISSA</td>
<td>0.505</td>
<td>0.512</td>
<td>0.521</td>
<td>0.522</td>
<td>0.524</td>
</tr>
<tr>
<td>PUNJAB</td>
<td>0.621</td>
<td>0.635</td>
<td>0.649</td>
<td>0.659</td>
<td>0.663</td>
</tr>
<tr>
<td>RAJASTHAN</td>
<td>0.512</td>
<td>0.522</td>
<td>0.523</td>
<td>0.524</td>
<td>0.526</td>
</tr>
<tr>
<td>SIKKIM</td>
<td>0.651</td>
<td>0.653</td>
<td>0.657</td>
<td>0.658</td>
<td>0.659</td>
</tr>
<tr>
<td>TAMIL NADU</td>
<td>0.622</td>
<td>0.629</td>
<td>0.637</td>
<td>0.649</td>
<td>0.655</td>
</tr>
<tr>
<td>TRIPURA</td>
<td>0.615</td>
<td>0.619</td>
<td>0.621</td>
<td>0.624</td>
<td>0.626</td>
</tr>
<tr>
<td>UTTAR PRADESH</td>
<td>0.501</td>
<td>0.502</td>
<td>0.505</td>
<td>0.507</td>
<td>0.509</td>
</tr>
<tr>
<td>WEST BENGAL</td>
<td>0.605</td>
<td>0.612</td>
<td>0.617</td>
<td>0.621</td>
<td>0.622</td>
</tr>
<tr>
<td>ANDAMAN AND NICOBAR</td>
<td>0.671</td>
<td>0.682</td>
<td>0.687</td>
<td>0.691</td>
<td>0.692</td>
</tr>
<tr>
<td>CHANDIGARH</td>
<td>0.751</td>
<td>0.757</td>
<td>0.759</td>
<td>0.761</td>
<td>0.763</td>
</tr>
<tr>
<td>DADRA AND NAGAR HAVELI</td>
<td>0.612</td>
<td>0.633</td>
<td>0.655</td>
<td>0.667</td>
<td>0.673</td>
</tr>
<tr>
<td>DAMAN AND DIU</td>
<td>0.652</td>
<td>0.663</td>
<td>0.669</td>
<td>0.671</td>
<td>0.677</td>
</tr>
<tr>
<td>DELHI</td>
<td>0.688</td>
<td>0.691</td>
<td>0.695</td>
<td>0.699</td>
<td>0.701</td>
</tr>
<tr>
<td>LAKSHADWEEP</td>
<td>0.612</td>
<td>0.619</td>
<td>0.625</td>
<td>0.629</td>
<td>0.635</td>
</tr>
<tr>
<td>PONDICHERRY</td>
<td>0.701</td>
<td>0.703</td>
<td>0.704</td>
<td>0.705</td>
<td>0.706</td>
</tr>
</tbody>
</table>

Table 4.10: State wise GDI temporal data
The map viewer tool of Oracle has been used to get the output of the vector data related to state boundary. The query output, in MapViewer, of the entire state boundary data is as follows:

Figure 4.4 : India State Map in MapViewer
4.3.3 SPATIAL WEIGHT MATRIX

As per our hypothesis, which is based on the study of NACO, the following states (Table 4.11) are the major destination states for the migrants for getting employment.

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>DELHI</td>
</tr>
<tr>
<td>6</td>
<td>GUJARAT</td>
</tr>
<tr>
<td>22</td>
<td>TAMIL NADU</td>
</tr>
<tr>
<td>13</td>
<td>MAHARASHTRA</td>
</tr>
<tr>
<td>25</td>
<td>WEST BENGAL</td>
</tr>
</tbody>
</table>

Table 4.11: Major destination states for migrants

With appropriate spatial query those states are visualized in MapViewer as follows (Figure: 4.4).
With the help of the state boundary geometry and spatial query we calculate the centroid of each states which becomes the geographical points having the non spatial attributes like HDI, GDI, MGRRATE (Migration Rate). This concept diagram of centroid can be represented in the following example diagram. Here dots as a black circle are the centroid of the migrant states and the dots in red circle are the states as a destination states.

Figure 4.6: Centroid of states Concept Diagram
The computed centroid that is visualized in MapViewer for some states as follows:

![Figure 4.7 Visualization of state centroid](image)

The computation mechanism of centroid is as follows:

*Calculating the centroid of a polygon or the mean center of a set of points.*
For calculating the centroid of the state SDO_CENTROID function is used which returns a point geometry that is the centroid of a polygon.

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}
\]
The centroid so computed of the example figure (Figure 4.7) is represented in the following table (Table 4.12).

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>sum</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td>Centroid/MC</td>
<td><strong>5.2</strong></td>
<td><strong>4.4</strong></td>
</tr>
</tbody>
</table>

Table 4.12: Centroid of example polygon of Figure 4.7
For calculating the centroid of each state and distance between the two state centroids, the spatial query used is as follows:

<table>
<thead>
<tr>
<th>Query</th>
<th>Query Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query 1</td>
<td>SELECT c.SCODE, SDO_GEOM.SDO_CENTROID(c.geom, m.diminfo) FROM india_st c, user_sdo_geom_metadata m WHERE m.table_name = 'INDIA_ST' AND m.column_name = 'GEOM' AND c.SCODE = 4</td>
</tr>
<tr>
<td>Query 2</td>
<td>SELECT SDO_GEOM.SDO_DISTANCE(c.geom, m.diminfo, d.geom, m.diminfo) FROM india_st c, india_st d, user_sdo_geom_metadata m WHERE m.table_name = 'INDIA_ST' AND m.column_name = 'GEOM' AND c.STATE = 'BIHAR' AND d.state = 'DELHI';</td>
</tr>
</tbody>
</table>

Table 4.13: Spatial query for calculating centroid of states and distance between two centroids.

Query 1: With the help of this query the coordinate of the state centroid is computed which is stored in database.

Query 2: With the help of this query the distance between the two centroid points are computed.
With the help of spatial query the centroid so computed for all the states are in the database table. The data is as follows:

<table>
<thead>
<tr>
<th>STATE</th>
<th>LATI(X)</th>
<th>LONGI(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDHRA PRADESH</td>
<td>79.6</td>
<td>16.5</td>
</tr>
<tr>
<td>ARUNACHAL PRADESH</td>
<td>94.7</td>
<td>28.1</td>
</tr>
<tr>
<td>ASSAM</td>
<td>92.9</td>
<td>26.3</td>
</tr>
<tr>
<td>BIHAR</td>
<td>85.6</td>
<td>24.7</td>
</tr>
<tr>
<td>GOA</td>
<td>74</td>
<td>15.3</td>
</tr>
<tr>
<td>GUJARAT</td>
<td>71.6</td>
<td>22.7</td>
</tr>
<tr>
<td>HARYANA</td>
<td>76.3</td>
<td>29.16</td>
</tr>
<tr>
<td>HIMACHAL PRADESH</td>
<td>77.2</td>
<td>31.8</td>
</tr>
<tr>
<td>JAMMU AND KASHMIR</td>
<td>76.39</td>
<td>34.5</td>
</tr>
<tr>
<td>KARNATAKA</td>
<td>76.14</td>
<td>14.68</td>
</tr>
<tr>
<td>KERALA</td>
<td>76.37</td>
<td>10.39</td>
</tr>
<tr>
<td>MADHYA PRADESH</td>
<td>79.4</td>
<td>22.8</td>
</tr>
<tr>
<td>MAHARASHTRA</td>
<td>76.14</td>
<td>19.42</td>
</tr>
<tr>
<td>MANIPUR</td>
<td>94.01</td>
<td>24.77</td>
</tr>
<tr>
<td>MEGHALAYA</td>
<td>91.36</td>
<td>25.54</td>
</tr>
<tr>
<td>MIZORAM</td>
<td>92.76</td>
<td>23.31</td>
</tr>
<tr>
<td>NAGALAND</td>
<td>94.57</td>
<td>26.12</td>
</tr>
<tr>
<td>ORISSA</td>
<td>84.41</td>
<td>20.45</td>
</tr>
<tr>
<td>PUNJAB</td>
<td>75.42</td>
<td>30.82</td>
</tr>
<tr>
<td>RAJASTHAN</td>
<td>73.78</td>
<td>26.56</td>
</tr>
<tr>
<td>SIKKIM</td>
<td>88.5</td>
<td>27.55</td>
</tr>
<tr>
<td>TAMIL NADU</td>
<td>78.38</td>
<td>10.94</td>
</tr>
<tr>
<td>TRIPURA</td>
<td>91.73</td>
<td>23.74</td>
</tr>
<tr>
<td>UTTAR PRADESH</td>
<td>80.3</td>
<td>27.44</td>
</tr>
<tr>
<td>WEST BENGAL</td>
<td>87.95</td>
<td>23.83</td>
</tr>
<tr>
<td>ANDAMAN AND NICOBAR</td>
<td>93.09</td>
<td>11.34</td>
</tr>
<tr>
<td>CHANDIGARH</td>
<td>76.77</td>
<td>30.71</td>
</tr>
<tr>
<td>DAMDRA AND NAGAR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAVELI</td>
<td>73.1</td>
<td>20.19</td>
</tr>
<tr>
<td>DAMAN AND DIU</td>
<td>71.97</td>
<td>20.55</td>
</tr>
<tr>
<td>DELHI</td>
<td>77.08</td>
<td>28.6</td>
</tr>
<tr>
<td>LAKSHADWEEN</td>
<td>71.8</td>
<td>12.18</td>
</tr>
<tr>
<td>PONDICHERRY</td>
<td>79.68</td>
<td>12.26</td>
</tr>
</tbody>
</table>

Table 4.14: Centroid of each state
The centroid and the distance between two centroids can be represented using visualization tool as follows:

Figure 4.9: Distance between the two states
Now based on the centroid of each states, its distance from the destination state centroid is measured and the reciprocal of the average distance is the weight of the centroid. For example some of the distance and computed weighted matrix are given below. The detail program to compute the weight matrix is present in Annexure 4.1

\[(w_{ij} = 1/d_{ij})\]

<table>
<thead>
<tr>
<th></th>
<th>J&amp;K</th>
<th>Rajasthan</th>
<th>Pondicherry</th>
<th>Bihar</th>
<th>Orissa</th>
<th>WB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>0.267</td>
<td>3.1812</td>
<td>0.789</td>
<td>0.1515</td>
<td>0.11199</td>
<td>0.101</td>
</tr>
<tr>
<td>Gujarat</td>
<td>0.1222</td>
<td>0</td>
<td>0.1142</td>
<td>0.1139</td>
<td>0.1313</td>
<td>0.0888</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>0.01</td>
<td>0.975</td>
<td>0.366</td>
<td>0.2862</td>
<td>0.952</td>
<td>0.1837</td>
</tr>
<tr>
<td>Kolkata</td>
<td>0.011</td>
<td>0.1199</td>
<td>0.1444</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TN</td>
<td>0.053</td>
<td>0.0936</td>
<td>0</td>
<td>0.1026</td>
<td>0.2223</td>
<td>0.093</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>0.4632</strong></td>
<td><strong>4.3697</strong></td>
<td><strong>1.4136</strong></td>
<td><strong>0.6542</strong></td>
<td><strong>1.41759</strong></td>
<td><strong>0.4665</strong></td>
</tr>
<tr>
<td><strong>Mig. Weight ((\rho))</strong></td>
<td><strong>0.09264</strong></td>
<td><strong>0.87394</strong></td>
<td><strong>0.28272</strong></td>
<td><strong>0.13084</strong></td>
<td><strong>0.283518</strong></td>
<td><strong>0.0933</strong></td>
</tr>
</tbody>
</table>

Table 4.15: Example Centroid Weight Matrix

The following program code is used to compute the total centroid weights of all the states with respect to the destination states mentioned in Table 4.11 is calculated with the help of the program code presented in Annexure 4.1
The weights matrix of each state so calculated, with respect to the destination states, are as follows:

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>DIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ANDHRA PRADESH</td>
<td>0.307972</td>
</tr>
<tr>
<td>2</td>
<td>ARUNACHAL PRADESH</td>
<td>0.077276</td>
</tr>
<tr>
<td>3</td>
<td>ASSAM</td>
<td>0.091999</td>
</tr>
<tr>
<td>4</td>
<td>BIHAR</td>
<td>0.174782</td>
</tr>
<tr>
<td>5</td>
<td>GOA</td>
<td>0.141927</td>
</tr>
<tr>
<td>6</td>
<td>GUJARAT</td>
<td>0.197306</td>
</tr>
<tr>
<td>7</td>
<td>HARYANA</td>
<td>0.145671</td>
</tr>
<tr>
<td>8</td>
<td>HIMACHAL PRADESH</td>
<td>0.112826</td>
</tr>
<tr>
<td>9</td>
<td>JAMMU AND KASHMIR</td>
<td>0.09768</td>
</tr>
<tr>
<td>10</td>
<td>KARNATAKA</td>
<td>0.216789</td>
</tr>
<tr>
<td>11</td>
<td>KERALA</td>
<td>0.121868</td>
</tr>
<tr>
<td>12</td>
<td>MADHYA PRADESH</td>
<td>0.643827</td>
</tr>
<tr>
<td>13</td>
<td>MAHARASHTRA</td>
<td>0.300542</td>
</tr>
<tr>
<td>14</td>
<td>MANIPUR</td>
<td>0.073594</td>
</tr>
<tr>
<td>15</td>
<td>MEGHALAYA</td>
<td>0.092406</td>
</tr>
<tr>
<td>16</td>
<td>MIZORAM</td>
<td>0.07926</td>
</tr>
<tr>
<td>17</td>
<td>NAGALAND</td>
<td>0.070785</td>
</tr>
<tr>
<td>18</td>
<td>ORISSA</td>
<td>0.226372</td>
</tr>
<tr>
<td>19</td>
<td>PUNJAB</td>
<td>0.117762</td>
</tr>
<tr>
<td>20</td>
<td>RAJASTHAN</td>
<td>0.245616</td>
</tr>
<tr>
<td>21</td>
<td>SIKKIM</td>
<td>0.099627</td>
</tr>
<tr>
<td>22</td>
<td>TAMIL NADU</td>
<td>0.128321</td>
</tr>
<tr>
<td>23</td>
<td>TRIPURA</td>
<td>0.085207</td>
</tr>
<tr>
<td>24</td>
<td>UTTAR PRADESH</td>
<td>0.255116</td>
</tr>
<tr>
<td>25</td>
<td>WEST BENGAL</td>
<td>0.133868</td>
</tr>
<tr>
<td>29</td>
<td>A&amp;N</td>
<td>0.065049</td>
</tr>
<tr>
<td>30</td>
<td>CHANDIGARH</td>
<td>0.106143</td>
</tr>
<tr>
<td>31</td>
<td>DADRA AND NAGAR</td>
<td>0.164551</td>
</tr>
<tr>
<td>32</td>
<td>DAMAN AND DIU</td>
<td>0.161594</td>
</tr>
<tr>
<td>33</td>
<td>DELHI</td>
<td>0.134103</td>
</tr>
<tr>
<td>34</td>
<td>LAKSHADWEEP</td>
<td>0.099053</td>
</tr>
<tr>
<td>35</td>
<td>PONDICHERY</td>
<td>0.160933</td>
</tr>
</tbody>
</table>

Table 4.16: Spatial weight matrix of states
4.4 NON SPATIAL AND SPATIAL DATA ASSOCIATION

Related to each spatial object i.e. state, we have several non-spatial data that represents the significant attribute values of the object.

A systematic and logical study is required to be carried out to identify the root cause of the spread of the HIV disease in many new parts of our country. As stated earlier the available evidence suggests that migration could be fuelling the spread of HIV epidemic in high out migration states such as Uttar Pradesh, Bihar, Rajasthan, Orissa, Madhya Pradesh and Gujarat. The sentinel surveillance data (2008-09) has shown an increase in HIV prevalence in these states. Studies have also shown that migrants per se are not at risk but it is the conditions and the environment that puts them at risk of acquiring HIV infections. Data also has demonstrated that HIV infection in couples (sero-concordant or sero-discordant) was significantly more likely among those couples where a man is a migrant and those couples where man is a migrant as well as mobile, relative to those couples where men were neither migrant nor mobile.

In India the progress of socio economic development is not uniform. There are many variability of inter state development and so there are many indicators responsible for the diversity in development. There is wide disparity in socio-economic development among the states of India. The factors, which are found out
to be more important for the overall development process, relates to basic need like availability of food, education, minimum purchasing power, and facilities like health care infrastructure, safe drinking water etc.

There are several social indicators which triggers the phenomena of migration. The illiterate, poor and labor class people across India who travels in search of employment to various metropolitan cities where they get employment and live their life relatively in a comfortable condition. So the economic and social backwardness of the state triggers the out migration process. So the economic and social indicators of the states should be taken into consideration.

![Diagram](image.png)

Figure 4.10: Relation between social factors, Migration and growth of HIV

The statistics shows that among the various factors that characterizes a state as under developed or poor state are low income, poor health facility, poor education facility, imbalanced gender development etc. So the statistical figure released by *Ministry of Statistics and Program Implementation* - “Selected Socio
Economic Statistics, India 2011”, gives as major socio economic indicators to study the prevailing condition of a state. Similarly the figure related to migration rate released jointly by Population Council & NACO gives us the opportunity to find out strong correlation between HDI,GDI and Migration rate. The net out migration is the difference between in migration and out migration.

4.5 THE SOCIO ECONOMIC INDICATORS

Among the various socio economic indicators, the indicators like Human Development Index (HDI) and Gender Development Index (GDI) are more meaningful and are the combination of various sub factors.

HDI: Human Development Index was released for the first time by the UNDP for 30 countries of the world in 1990. The UNDP has defined human development as the process of enlarging people’s choices. It defines that income is one of those choices but it does not cover the totality of human life. Health, a good physical environment, education, and freedom of action and expression are also very important. The Human Development Index (HDI) is in fact is a combination of various indicators like national income, life expectancy and educational attainment to give a composite measure of human progress. Government published HDI for each state. So for my study I have taken HDI as a composite indicator for economic and social status of a state.
Then for considering the inequality of states from male and female I have taken GDI indicator of each state. The GDI is defined as below.

**GDI:** This index also measures the achievements measuring the same dimensions but additionally it captures inequalities in achievement between women and men. It is simply the HDI adjusted downward for gender inequality. The greater the gender disparity in basic human development, the lower a country’s GDI compared with its HDI.

The time series data of the following time period has been used for training, validation and testing purpose of the model.

Training Data Set: from 2002 to 2005

Validate data: 2005 to 2008

Test Data Set: 2008 to 2011

The non spatial data used for the model is as follows:
<table>
<thead>
<tr>
<th>SCODE</th>
<th>HDI</th>
<th>GDI</th>
<th>MGRRATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.585</td>
<td>0.574</td>
<td>-0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.647</td>
<td>0.642</td>
<td>7.2</td>
</tr>
<tr>
<td>3</td>
<td>0.595</td>
<td>0.585</td>
<td>-0.7</td>
</tr>
<tr>
<td>4</td>
<td>0.507</td>
<td>0.479</td>
<td>-2.7</td>
</tr>
<tr>
<td>5</td>
<td>0.764</td>
<td>0.747</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>0.634</td>
<td>0.624</td>
<td>1.7</td>
</tr>
<tr>
<td>7</td>
<td>0.643</td>
<td>0.632</td>
<td>4.1</td>
</tr>
<tr>
<td>8</td>
<td>0.667</td>
<td>0.664</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.59</td>
<td>0.568</td>
<td>-0.4</td>
</tr>
<tr>
<td>10</td>
<td>0.622</td>
<td>0.611</td>
<td>0.3</td>
</tr>
<tr>
<td>11</td>
<td>0.764</td>
<td>0.745</td>
<td>-0.6</td>
</tr>
<tr>
<td>12</td>
<td>0.529</td>
<td>0.516</td>
<td>-0.1</td>
</tr>
<tr>
<td>13</td>
<td>0.689</td>
<td>0.677</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>0.702</td>
<td>0.699</td>
<td>-1.4</td>
</tr>
<tr>
<td>15</td>
<td>0.629</td>
<td>0.624</td>
<td>0.8</td>
</tr>
<tr>
<td>16</td>
<td>0.688</td>
<td>0.687</td>
<td>-0.1</td>
</tr>
<tr>
<td>17</td>
<td>0.7</td>
<td>0.697</td>
<td>-1.4</td>
</tr>
<tr>
<td>18</td>
<td>0.537</td>
<td>0.524</td>
<td>-0.7</td>
</tr>
<tr>
<td>19</td>
<td>0.668</td>
<td>0.663</td>
<td>1.7</td>
</tr>
<tr>
<td>20</td>
<td>0.541</td>
<td>0.526</td>
<td>-0.6</td>
</tr>
<tr>
<td>21</td>
<td>0.665</td>
<td>0.659</td>
<td>5.9</td>
</tr>
<tr>
<td>22</td>
<td>0.666</td>
<td>0.655</td>
<td>-0.7</td>
</tr>
<tr>
<td>23</td>
<td>0.663</td>
<td>0.626</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0.528</td>
<td>0.509</td>
<td>-2</td>
</tr>
<tr>
<td>25</td>
<td>0.642</td>
<td>0.622</td>
<td>0.4</td>
</tr>
<tr>
<td>29</td>
<td>0.708</td>
<td>0.692</td>
<td>7.9</td>
</tr>
<tr>
<td>30</td>
<td>0.784</td>
<td>0.763</td>
<td>21.4</td>
</tr>
<tr>
<td>31</td>
<td>0.677</td>
<td>0.673</td>
<td>32.6</td>
</tr>
<tr>
<td>32</td>
<td>0.7</td>
<td>0.677</td>
<td>44.1</td>
</tr>
<tr>
<td>33</td>
<td>0.74</td>
<td>0.701</td>
<td>18.7</td>
</tr>
<tr>
<td>34</td>
<td>0.697</td>
<td>0.635</td>
<td>6.4</td>
</tr>
<tr>
<td>35</td>
<td>0.729</td>
<td>0.706</td>
<td>8.8</td>
</tr>
</tbody>
</table>

Table 4.17: Data of HDI, GDI and Migration rate
Now we combine the column DIST of table 4.15 and the socio-economic indicators of table 4.16 of the respective state and get a comprehensive data for our study. Based on the spatial and non-spatial predicates we need to formulate some classification module which is able to classify the states into three classes labels i.e. states with increasing rate of HIV infection, states with stable rate of HIV infection and states with decreasing rate of HIV infection.

For the present work the data under consideration and its thematic map is as below:

Figure 4.11: Thematic map of spatial and non-spatial data
The combined data for the study is as follows:

<table>
<thead>
<tr>
<th>SCODE</th>
<th>HDI</th>
<th>GDI</th>
<th>MGRATE</th>
<th>DIST</th>
<th>INFDIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.585</td>
<td>0.574</td>
<td>-0.3</td>
<td>0.307972</td>
<td>-0.06</td>
</tr>
<tr>
<td>2</td>
<td>0.647</td>
<td>0.642</td>
<td>7.2</td>
<td>0.077276</td>
<td>-0.09</td>
</tr>
<tr>
<td>3</td>
<td>0.595</td>
<td>0.585</td>
<td>-0.7</td>
<td>0.091999</td>
<td>-0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.507</td>
<td>0.479</td>
<td>-2.7</td>
<td>0.174782</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.764</td>
<td>0.747</td>
<td>8</td>
<td>0.141927</td>
<td>-0.17</td>
</tr>
<tr>
<td>6</td>
<td>0.634</td>
<td>0.624</td>
<td>1.7</td>
<td>0.197306</td>
<td>-0.06</td>
</tr>
<tr>
<td>7</td>
<td>0.643</td>
<td>0.632</td>
<td>4.1</td>
<td>0.145671</td>
<td>-0.29</td>
</tr>
<tr>
<td>8</td>
<td>0.667</td>
<td>0.664</td>
<td>1</td>
<td>0.112826</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.59</td>
<td>0.568</td>
<td>-0.4</td>
<td>0.09768</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.622</td>
<td>0.611</td>
<td>0.3</td>
<td>0.216789</td>
<td>-0.03</td>
</tr>
<tr>
<td>11</td>
<td>0.764</td>
<td>0.745</td>
<td>-0.6</td>
<td>0.121868</td>
<td>-0.34</td>
</tr>
<tr>
<td>12</td>
<td>0.529</td>
<td>0.516</td>
<td>-0.1</td>
<td>0.643827</td>
<td>-0.04</td>
</tr>
<tr>
<td>13</td>
<td>0.689</td>
<td>0.677</td>
<td>3</td>
<td>0.300542</td>
<td>-0.19</td>
</tr>
<tr>
<td>14</td>
<td>0.702</td>
<td>0.699</td>
<td>-1.4</td>
<td>0.073594</td>
<td>-0.41</td>
</tr>
<tr>
<td>15</td>
<td>0.629</td>
<td>0.624</td>
<td>0.8</td>
<td>0.092406</td>
<td>-0.09</td>
</tr>
<tr>
<td>16</td>
<td>0.688</td>
<td>0.687</td>
<td>-0.1</td>
<td>0.07926</td>
<td>-0.22</td>
</tr>
<tr>
<td>17</td>
<td>0.7</td>
<td>0.697</td>
<td>-1.4</td>
<td>0.070785</td>
<td>-0.38</td>
</tr>
<tr>
<td>18</td>
<td>0.537</td>
<td>0.524</td>
<td>-0.7</td>
<td>0.226372</td>
<td>0.05</td>
</tr>
<tr>
<td>19</td>
<td>0.668</td>
<td>0.663</td>
<td>1.7</td>
<td>0.117762</td>
<td>-0.04</td>
</tr>
<tr>
<td>20</td>
<td>0.541</td>
<td>0.526</td>
<td>-0.6</td>
<td>0.245616</td>
<td>0.04</td>
</tr>
<tr>
<td>21</td>
<td>0.665</td>
<td>0.659</td>
<td>5.9</td>
<td>0.099627</td>
<td>-0.11</td>
</tr>
<tr>
<td>22</td>
<td>0.666</td>
<td>0.655</td>
<td>-0.7</td>
<td>0.128321</td>
<td>-0.34</td>
</tr>
<tr>
<td>23</td>
<td>0.663</td>
<td>0.626</td>
<td>1</td>
<td>0.085207</td>
<td>-0.2</td>
</tr>
<tr>
<td>24</td>
<td>0.528</td>
<td>0.509</td>
<td>-2</td>
<td>0.255116</td>
<td>-0.02</td>
</tr>
<tr>
<td>25</td>
<td>0.642</td>
<td>0.622</td>
<td>0.4</td>
<td>0.133868</td>
<td>0.06</td>
</tr>
<tr>
<td>29</td>
<td>0.708</td>
<td>0.692</td>
<td>7.9</td>
<td>0.065049</td>
<td>-0.27</td>
</tr>
<tr>
<td>30</td>
<td>0.784</td>
<td>0.763</td>
<td>21.4</td>
<td>0.106143</td>
<td>-0.07</td>
</tr>
<tr>
<td>31</td>
<td>0.677</td>
<td>0.673</td>
<td>32.6</td>
<td>0.164551</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>0.7</td>
<td>0.677</td>
<td>44.1</td>
<td>0.161594</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>0.74</td>
<td>0.701</td>
<td>18.7</td>
<td>0.134103</td>
<td>-0.05</td>
</tr>
<tr>
<td>34</td>
<td>0.697</td>
<td>0.635</td>
<td>6.4</td>
<td>0.099053</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>0.729</td>
<td>0.706</td>
<td>8.8</td>
<td>0.160933</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 4.18: Actual data for model formation
Based on the actual data (Table 4.18) the table below (Table 4.19) is prepared which is a coded table. The scheme of coding the table is as follows:

For the data value of the HDI and GDI column if it is less than their mean values then it is labeled as -1 otherwise it is +1. Similarly for migration rate (MGRRATE) of the states if out-migration is more than in-migration then it is coded as -1 and if out-migration is less than in-migration it is +1.

The value of DIST column has been categorized in three categories. If the value is in between 10% plus or minus value from the mean value then the value is coded as ‘Moderate’. Now the value more than this range value is labeled as ‘Far’ and the value less than this range value is labeled as ‘Near’.
<table>
<thead>
<tr>
<th>SCODE</th>
<th>HDI</th>
<th>GDI</th>
<th>MGRRATE</th>
<th>DIST</th>
<th>INFDIFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Moderate</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>Moderate</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Near</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>15</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>18</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Far</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>20</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Far</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>24</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>Far</td>
<td>-1</td>
</tr>
<tr>
<td>25</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Moderate</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Moderate</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Near</td>
<td>-1</td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>Near</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Moderate</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.19: Coded data from table 4.17
4.6 SPATIAL AUTO CORRELATION AND REGRESSION ANALYSIS

However it is important to make distinction between spatial data mining and spatial data analysis. The spatial data analysis covers a broad spectrum of techniques that deals with both the spatial and non spatial characteristics of spatial objects whereas spatial data mining techniques are often derived from spatial statistics, spatial analysis, machine learning, and data base and are customized to analyzed massive data sets [70]. The pattern of information that is expected to be the output of data mining could be summary of statistics or simple rule.

Therefore in data mining process we use a set of techniques to generate hypothesis. After that we do validation and verification via standard statistical tools[70].

The first law of Geography says that “All things are related but nearby things are more related.” The measures of spatial auto correlation are K-function, Moran’s I and Variogram etc.

The approach of spatial auto correlation is used to measure and analyze the degree of spatial dependency i.e. correlation among observations in a geographic space. A
weight matrix is used in spatial auto correlation measure. We have already computed a weight matrix. This matrix is based on distance measure.

<table>
<thead>
<tr>
<th>Name</th>
<th>Model</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Linear Regression</td>
<td>$y = x\beta + \varepsilon$</td>
<td>Low</td>
</tr>
<tr>
<td>Spatial Auto-Regression</td>
<td>$y = \rho Wy + x\beta + \varepsilon$</td>
<td>High</td>
</tr>
</tbody>
</table>

During classification process when we consider spatial dependency and model spatial dependency, it improves the classification accuracy [71] . Spatial dependency is also known as spatial context. Spatial Autoregressive Model (SAR) is an important approach of modeling spatial dependency. In this model, the dependent variables i.e. the spatial dependencies of the error term, are directly modeled in the linear regression equation.

$$y = aWy + x\beta + e$$

The classification accuracy of this linear regression equation is higher than the classical linear regression.

Here $W = \text{Neighborhood relationship contiguity matrix}$ and

$a = \text{strength of the spatial dependencies between the elements of the dependent variable.}$
When some of the statistical analysis performed to test the correlation between non spatial attributes and with dependent variable infection difference, we get the following results. The result of the correlation analysis is as follows:

1. Residual case order plot of all the non spatial data of spatial database

Figure 4.12: Residual case order plot
2. Partial Linear Regression

Figure 4.13: Partial linear regression
Residual case order plot for states showing infection increasing trend

Figure 4.14: Residual case order plot for states showing infection increasing trend
4. PLS Regression

\[ y = p_1 x^2 + p_2 x + p_3 \]

Coefficients:
\[ p_1 = -4.7407 \]
\[ p_2 = 30.918 \]
\[ p_3 = 51.181 \]

Norm of residuals = 4.4692

Figure 4.15: PLS Regression
5 Residual case order plot for states showing decreasing infection trend

Figure 4.16: Residual case order plot for states showing decreasing infection trend
6. PLS Regression of HIV Decreasing pattern data

Figure 4.17: PLS regression of HIV Decreasing

Statistics:

\[ y = p_1 z^3 + p_2 z^2 + p_3 z + p_4 \]

where \( z \) is centered and scaled:

\[ z = (x - \mu) / \sigma \]

\( \mu = 2.5 \)

\( \sigma = 1.291 \)

Coefficients:

\( p_1 = -0.29935 \)
\( p_2 = -1.6528 \)
\( p_3 = 3.1698 \)
\( p_4 = 12.216 \)

Norm of residuals = 1.7764e-015
Since the data under study are correlated predictors the Partial least-squares (PLS) regression technique is used here. This technique constructs new predictor variables, known as components, leading to a parsimonious model with reliable predictive power. As a result we are able to construct a reliable predictive model.

4.7 DISCUSSION

As mentioned earlier the process of creation of appropriate database structure for performing the data mining task is very important. Simultaneously selection of appropriate and significantly correlated data is equally important. The work described in this chapter illustrates how the spatial database has been created using the scaled geographical vector data related to state boundaries. In the same spatial database we have incorporated the non spatial data like epidemiology and socio economic parameters influencing the different behavior of the spread of epidemiology. We have computed the spatial distance measure related to each state location. Now with the methods of statistical computation, the highly correlated attributes has been identified. But we can see that the regression model, based on those attributes, are not sufficient enough to incorporate the non linear behavior of the attributes. Therefore, a rule based spatial classification technique has been evolved which is discussed in chapter 5.
5.1 INTRODUCTION

Classification of objects based on spatial relation enables researches to explore interesting relation between spatial and non spatial data. Here a model has been proposed which is a spatial database oriented interesting and efficient method for the classification of objects based on spatial relation. The proposed method enables classification of spatial objects based on aggregate values of non-spatial attributes for different state wise epidemiology and socio-economic values. It also takes into account spatial relation between objects (states) on the map which may be represented in the form of predicates.

The goal of spatial classification is to find rules that divide the set of classified objects \( O_c \) into number of groups, where objects in each group belongs mostly to a single class. In spatial database we have the data of following categories:
• Non-spatial attributes of objects: HDI, GDI, Migration Rate of states, number of HIV infected people in state.

• Spatial attribute: Distance matrix of states (from centroid) i.e. source state to destination state of migrants.

• Spatial predicates like distance_more_than_thresold_value (X, State_Centroid)

• Spatial function like driving_average_distance (X,State Centroid)

Such categories of information may be used to extract values both for:

• Class label attribute i.e. attribute dividing data into classes.

• Predicting attribute i.e. attributes on whose values decision tree is branched

Here we want to build a classification rule for classifying object $O_i$ based on two types of information:

1. Thematic layers of non spatial attributes.

2. Description of objects based on distance metrics of the objects.

A combined thematic map has been presented in figure 4.11. All the related data has been given in table 4.18. The value of table 4.18 represents the non-spatial and spatial class data. The function used for these predicates are as follows:
The predicates are classified based on the classification function mentioned in the table 5.1 above. The objects i.e. states are characterized by distance from migrant destination locations. The spatial function \textit{driving_distance} is used to describe classified objects. The whole classified objects are presented in chapter-4 (Table 4.18). The process of finding spatial predicate and functions may be time consuming. To accelerate this process some rough computations are performed first and then fine computations are done only for the promising patterns. Based on the rule triggered by the record the proposed rule-based-classifier classifies a test record.

In the first step we can find coarse description for only a sample of objects. For example we can use \textit{driving_distance} predicate which imply that distance of object from the destination object is more, less or within the specific distance threshold.

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Function</th>
<th>Class value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI</td>
<td>above_average</td>
<td>Yes/No</td>
</tr>
<tr>
<td>GDI</td>
<td>above_average</td>
<td>Yes/No</td>
</tr>
<tr>
<td>MGRRATE</td>
<td>net_migration</td>
<td>Negative/Positive</td>
</tr>
<tr>
<td>DIST</td>
<td>driving_dist(x,Migration-Destination)</td>
<td>Far/Moderate/Near</td>
</tr>
<tr>
<td>INFDIFF</td>
<td>growth_rate (x, infdiff)</td>
<td>Increasing/Decreasing/Stable</td>
</tr>
</tbody>
</table>

Table 5.1: Classification predicate functions
Then some learning method is used for the extraction of the relevant predicates. We need to extract a set of rules that identifies key relationships between the attributes of a data set and the class label. The classification approach used here uses distance measure to identify spatial relationship of objects. Based on the spatial relationship, such as Euclidean distance, network distance, adjacent and so on, these objects will be classified to get rules that satisfy the class labels obtained by data filtering condition of spatial query language. Then method like threshold value can be applied to find collocation pattern.

This work generalizes the concept of collocation pattern to extended spatial objects. That includes boundary of the regions to get its centroid and associate its non spatial attributes. Thus this work provided a more general algorithm by introducing the notion of buffer, which is a zone of specified distances from the fixed spatial object. This is the major contribution. Therefore it makes it applicable to spatio-temporal collocation pattern mining.
5.2 PROPOSED ALGORITHM

In the algorithm we first take an empty decision tree say R. The function \textit{rule\_learn} is used to extract the best rule for class y that covers the current set of training records. During rule extraction, all training records of class y are considered to be positive examples, while those that belong to other classes are considered to be negative example. A rule is then taken into consideration if it covers most of the positive examples and none of the negative examples. The algorithm (Algorithm 5.1) then proceeds to generate rules for the class.

\begin{algorithm}
\caption{Rule generation algorithm}
\begin{algorithmic}
\STATE 1: Let we have \( T \) as the training records.
\STATE 2: Let we have \( A \) as the set of attribute-value pairs, \( \{(A_j, v_j)\} \).
\STATE 3: Let \( C_0 \) be an ordered set of classes \( \{c_1, c_2, \ldots, c_k\} \).
\STATE 4: for each class \( c \in C_0 \) do
\STATE 5: \quad while stopping condition is not met do
\STATE 6: \quad \quad \quad \quad r \leftarrow \text{learn\_rule} (T, A, c).
\STATE 7: \quad \quad \quad \quad \text{Remove training records from } T \text{ that are covered by } r.
\STATE 8: \quad \quad \quad \quad \text{Add } r \text{ at the bottom of the rule set: } R \rightarrow R \bigcup \{r\}.
\STATE 9: \quad \quad \textbf{end while}
\STATE 10: \quad \textbf{end for}
\STATE 11: \text{Add the default rule, } \{\} \rightarrow c_k, \text{ at the bottom of the rule list } R.
\end{algorithmic}
\end{algorithm}

Illustration of algorithm:

The approach of \textit{learn\_rule} is to extract the classification rules that cover many of the positive examples and minimum no of negative examples in the
training set. The algorithm follows general-to-specific rule growing strategy. Initially rule \( r: \{ \} \rightarrow y \) is created, where the left hand side is an empty set and the right hand side contains the target class. This first rule is of poor quality because it covers all the examples in the training set. Subsequently new conjuncts are added to improve the rule’s quality. Out of the given attribute the first conjunct is select which is having bigger likelihood ratio.

After generating a rule all the positive and negative examples covered by the rule are eliminated. The rule is then added in the rule set as long as it does not violate the stopping condition. The violating condition is based on the minimum description length principle. The stopping condition used by this algorithm is that the error rate of the rule on the validation set must not exceed 40%.

**Algorithm:**

**Procedure learn_rule** (T,Predicate,class)
FOR \( j := 1 \) TO max_predicate DO
R\{\} ← predicate
FOR sample 1 TO max_sample
Nearest_hit := FIND_NEAREST_HIT(sample_i)
Nearest_miss := FIND_NEAREST_MISS(SAMPLE_i);
Predicate_weight:= information gain
IF predicate_weight > threshold
THEN Predicate_relevant:=TRUE
   R\{\} ← + Predicate
ELSE Predicate_relevant:=FALSE;

Algorithm 5.2: Learning rule algorithm called by algorithm 5.1
In this algorithm for every object in the sample distance measure are found where one measure belongs to the same class as the object (nearest hit) and the other measure belongs to the class different than this class. Subsequently the weights of the predicates are calculated. The predicates with weights larger than the predefined threshold are used for classification. The value for the threshold can be set based on statistical theory. Since it is a multi class problem the classes are ordered according to their frequencies. In the first iteration, the objects that belong to the least frequent class are labeled as positive examples, while those that belong to other classes are labeled as negative examples. The algorithm employs a general-to-specific strategy to grow a rule and on every addition of conjugate the biggest likelihood ratio is computed for selecting the rule having biggest likelihood ratio. In between the rule the information gain measure is also computed to choose the best conjunct to be added into the rule antecedent. The algorithm stops adding conjuncts when the rule starts covering negative examples.

I have implemented the above algorithm in relational database using PL/SQL program.(The program code of the classification model is attached in Annexure 5.1).
The database layout to implement the program is as follows:

![Diagram]

Figure 5.1: Rule processing diagram

The rule so generated with high likelihood ratio and high information gain is stored in one of the three rule class tables. The table R1 contains the rule related to the states showing growth rate of infection, table R2 contains the rule related to the states showing declining rate of infection and table R3 contains the rules related to the states showing stable pattern of infection.
For this work I have taken data of year 2002 to 2005 for rule generation. The statistics and the method of computation of likelihood ratio and information gain is mentioned ahead.

5.2.1 BIGGER LIKELIHOOD RATIO COMPUTATION OF THE RULE

Based on the given data in table 4.18 & table 4.19 we can see that there are three classes of infection prevalence in the states. Since we are considering 32 states so we have 32 objects in the database. Now out of the given INFDIFF (Infection difference) columns we have the following three classes based on infection prevalence:

<table>
<thead>
<tr>
<th>Infection status from year 2002 to 2005</th>
<th>INFDIFF &lt; 0 Negative Prevalence (NP) (Disease Decreasing)</th>
<th>INFDIFF &gt; 0 Positive Prevalence (PP) (Disease Increasing)</th>
<th>INFDIFF = 0 Stable Prevalence (SP) (Disease Stable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Prevalence (Frequency)</td>
<td>22</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.2: Different infection prevalence class

Now an evaluation metric is needed to determine which conjunct should be added (or removed) during the rule-growing process. In the training data set of 32 objects
we have 22, 5 and 5 examples of negative, positive and stable prevalence states respectively. Now let us assume the following two candidate rules:

**Rule r₁**: If HDI >0 then INFDIFF would be<0;

If HDI is more than the average value then the state will have decreasing trend of HIV prevalence

**Rule r₂**: If HDI > then INFDIFF would be >0;

If HDI is more than the average value then the state will have increasing trend of HIV prevalence

Now from the given data set the rule:

r₁ fetches 13 NP, 1 PP and 4 SP objects

r₂ fetches 9 NP, 4 PP and 1 SP objects

For rule r₁:

Expected frequency of NP class= 13 x (22/32) =8.9375

Expected frequency of PP class= 1 x (5/32) =0.15625

Expected frequency of SP class= 4 x (5/32) = 0.625

Therefore likelihood ratio of r₂:

\[ R(r₁)= 2 \times [13 \times \log_2 (13/8.9375) + 1 \times \log_2 (1/0.15625) + 4 \times \log_2 (4/0.625)] \]
= 40.8342

For rule \( r_2 \):

Expected frequency of NP class = \( 9 \times \frac{22}{32} = 6.1875 \)

Expected frequency of PP class = \( 4 \times \frac{5}{32} = 2.75 \)

Expected frequency of SP class = \( 1 \times \frac{5}{32} = 0.6875 \)

Therefore likelihood ratio of \( r_1 \):

\[
R(r_2) = 2 \times [9 \times \log_2 (9/26.1875) + 4 \times \log_2 (4/2.75) + 1 \times \log_2 (1/0.6875)]
\]

\[
= 15.1334
\]

Because \( R(r_1) > R(r_2) \), so Rule \( r_1 \) is better than rule \( r_2 \). thus rule \( r_2 \) is pruned.

Now on adding a new conjunct, say GDI, the new rule \( r_n : HDI \land GDI \) covers say \( NP_1 \) examples of NP class, \( PP_1 \) examples of PP class and \( SP_1 \) examples of SP class. The information gain is computed using the following relation:

\[
\text{Information gain} = NP_1 \times (\log_2 (NP_1/(NP_1+PP_1+SP_1)) - \log_2 (NP_0/(NP_0+PP_0+SP_0)))
\]

The predicate/attribute with the highest information gain is used to partition the dataset.
Now the new rule that is formed based on the addition of the new conjunct GDI into rule $r_1$ gives rise to a new rule $r_3$ and $r_4$. This can be described as follows:

**Rule $r_3$:** If $HDI^GDI >0$ then $INFDIFF$ would be$<0$;

**Rule $r_4$:** If $HDI^GDI >0$ then $INFDIFF$ would be$>0$;

When we calculate the information gain for rule $r_3$ and $r_4$ we see that it is 53.24 for $r_3$ and 21.22 for $r_4$, so rule $r_3$ is better than rule $r_4$. Thus the rule $r_4$ is pruned.

This way our algorithm keeps on adding the conjugate and gradually we get the better rule. The best rule based on non spatial predicate i.e. the predicates HDI,GDI and MGRRATE it is derived as follows:

**Rule $r_5$:** $HDI >0 ^ GDI >0 ^ MGRRATE >0$ results into $INFDIFF<0$

The rule states that if HDI and GDI is more than average value and at the same time Migration rate is positive i.e. Out migration population is less than in migration population then infection difference rate is negative i.e. state has shown decline in the HIV infection rate.

Similarly , the another rule also gets high information gain which is as follows:

**Rule $r_6$:** $HDI >0 ^ GDI >0 ^ MGRRATE<0$ results into $INFDIFF>0$
When algorithm calculates the information gain for this rule it comes 59.45. The rule states that if HDI and GDI is more than the average value but the Migration rate is negative i.e. Out-migration population is more than in-migration population then infection difference rate is positive i.e. state shows the increase in the HIV infection rate.

Now we consider the spatial predicate DIST. As stated earlier this is the distance of states from highly urbanized destination states (described in Chapter 4). When this predicate is added to this rule then the rule having highest information gain 62.43 is as follows:

\[ \text{Rule } r_7: \text{HDI} > 0 \land \text{GDI} > 0 \land \text{MGRRATE} < 0 \land \text{DIST='FAR'} \Rightarrow \text{INFDIFF} > 0 \]

The rule states that when HDI and GDI is more than average, Migration rate is negative (i.e. out migration is more than in-migration) and for migrants migration distance is of highest category ‘Far’ then it is more chance that the HIV infection will have increasing growth rate. So the impact of spatial predicate DIST is a significant predicate.
So the rules having bigger likelihood ratio is summarized as follows:

<table>
<thead>
<tr>
<th>No</th>
<th>Rule</th>
<th>Bigger likelihood ratio &amp; Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rule $r_2$: If $\text{HDI} &gt; 0$ then $\text{INFDIFF would be}&lt;0$;</td>
<td>40.8342</td>
</tr>
<tr>
<td>2</td>
<td>Rule $r_3$: If $\text{HDI}^\land \text{GDI} &gt; 0$ then $\text{INFDIFF would be}&lt;0$;</td>
<td>53.24</td>
</tr>
<tr>
<td>3</td>
<td>Rule $r_6$: $\text{HDI} &gt; 0 \land \text{GDI} &gt; 0 \land \text{MGRRATE}&lt;0$ results into $\text{INFDIFF}&gt;0$</td>
<td>59.45</td>
</tr>
<tr>
<td>4</td>
<td>Rule $r_7$: $\text{HDI} &gt; 0 \land \text{GDI} &gt; 0 \land \text{MGRRATE}&lt;0 \land \text{DIST}='\text{FAR}'$ results into $\text{INFDIFF}&gt;0$</td>
<td>62.43</td>
</tr>
</tbody>
</table>

Table 5.3: Rules and their likelihood ratio.

### 5.2.2 SPATIAL PREDICATE AND RULE VALIDATION

The result shows that when the spatial predicate is associated in the rule with the non-spatial predicates the overall information gain increases. The composite rules, having bigger likelihood ratio and high information gain, are summarized below (Table 5.4) with their percentage of accuracy when validated with validation dataset.
The data for validation is taken from the period of 2005 to 2008 (Table 5.6). Randomly the data of 10 states has been taken for the validation of the model. Here the non spatial data is changing but the spatial data (DIST) is not changing.

Table 5.4 : Rule Validation Result

<table>
<thead>
<tr>
<th>Rule No</th>
<th>Rule</th>
<th>Validation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\text{HDI} &gt; 0 \land \text{GDI} &gt; 0 \land \text{MGRRATE} &lt; 0 \land \text{DIST} = \text{NEAR}$ results into $\text{INFDIFF} &lt; 0$</td>
<td>86%</td>
</tr>
<tr>
<td>2</td>
<td>$\text{HDI} &gt; 0 \land \text{GDI} &gt; 0 \land \text{MGRRATE} &gt; 0 \land \text{DIST} = \text{NEAR}$ results into $\text{INFDIFF} &lt; 0$</td>
<td>89%</td>
</tr>
<tr>
<td>3</td>
<td>$\text{HDI} &lt; 0 \land \text{GDI} &lt; 0 \land \text{MGRRATE} &lt; 0 \land \text{DIST} = \text{FAR}$ results into $\text{INFDIFF} &gt; 0$</td>
<td>94%</td>
</tr>
<tr>
<td>4</td>
<td>$\text{HDI} &lt; 0 \land \text{GDI} &lt; 0 \land \text{MGRRATE} &lt; 0 \land \text{DIST} = \text{NEAR}$ results into $\text{INFDIFF} &lt; 0$</td>
<td>93%</td>
</tr>
</tbody>
</table>

Table 5.5: Data of 10 states 2005 to 2008 for model validation
5.3 SPATIAL AUTOCORELATION

The extent to which neighboring values are correlated was measured using Global Moran’s Index. A significant assessment under randomization procedure was run in MATLAB to determine the significance of the computed Moran’s Index. There is positive and spatial autocorrelation for HIV incidence in the Bihar, Orissa (Moran’s I = 0.138, p = 0.045), as shown in Table 5.3 below.

<table>
<thead>
<tr>
<th>Moran’s I</th>
<th>p-value</th>
<th>exception</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.138</td>
<td>0.045</td>
<td>-0.015</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 5.6: Moran’s I Value

The mathematical model of computing Moran’s I value is as follows:

\[
I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]
5.4 DISCUSSION

Conducted experiments showed the importance of relevance analysis which resulted in better quality of classification rule. It is important to mention here that the time to build decision tree is shorter when relevance analysis is performed on the sample of the objects and only relevant attributes are computed for all objects in comparison to the time when all predicates were used for the construction of the decision tree.

I have integrated the classification algorithm with the spatial query engine. For example when the distance of state is of “Near” category from the identified destination states then among the set of predicates the aggregate value of the relevant predicate is computed.

The proposed algorithm is a direct method which enables classification of spatial objects and generate rule based on aggregate value of non spatial attributes of neighboring regions. It takes into account the spatial relation between objects on the map represented in the form of predicates. The algorithm first performs less
costly, approximate spatial computations to obtain a sample of approximate spatial predicates. Refined computations are performed only for the set of promising patterns that are capable of producing smaller and more accurate decision rules. In this work the rules so obtained are validated and are summarized in table 5.4.

As discussed in Chapter 4, a model is required to properly map the non-linear nature and behavior of the attribute values under consideration. An indirect method to extract the classification rule and validate the rule generated by direct method is required. I further undertook the task of validating the rules so generated here by neural network based learning models. The challenge is to implement the NN-BPN model in the spatial database. I implemented this in our spatial database using PL/SQL program. All these tasks and its justification are discussed in Chapter 6.
CHAPTER-6

NEURAL NETWORK-BPN BASED SPATIAL CLASSIFICATION MODEL

6.1 INTRODUCTION

The rule based spatial classification performed in chapter 5 is useful. However looking into the complexity of the data model and considering the computational efficiency issues it requires a great amount of expert knowledge of the non-linear nature of the model and domain knowledge of the application area. Artificial neural network technology offers an alternative to the classification problem. With the learning and adaptive capability of a neural network algorithm, empirical relations between cause and effect of multisource of data can be established. Empirical relations themselves can be used for constructing a basis for expert knowledge. They can also be examined for constructing expert knowledge basis. The neural network model is useful in situations in which the structure of a problem has been poorly recognized, input data are incomplete and approximated results are satisfactory. Neural network makes it possible to model any complex non-linear functions.
Neural network can be used either as a standalone tool for spatial data classification or a complementary tool for knowledge acquisition when expert knowledge is ambiguous and structurally unclear.

The process of classification of database objects based on spatial and non-spatial data involves use of discriminant variables called features, and a list of classes or patterns. Discriminant analysis is employed to partition the feature space and associate each partitioned portion to a specific class or pattern.

The main challenge in modeling an NN model is to choose the right transformation function, the ideal size of the network and calibrate the weights to increase classification accuracy.

Since neural network allow non linear relations and complex interactions among predictor variables and thus it score over parametric methods. In the present work the relation between the data is very much non-linear in nature so a neural network method of classification gives a better answer to the classification problem.
When addressing the issue related to appropriate topology of the network it has been established that MLP (Multi Layer Perceptron) is better choice than RBF (Radial Base Function)[72].

6.2 DESIGN ISSUE

The following design issues are important to be considered before we train a neural network to learn a classification task.

1. The number of nodes in the input layer should be determined. To each numerical or binary input variable a separate input node is assigned.

2. The number of nodes in the output layer should be established.

3. An appropriate network topology (i.e. the number of hidden layers and hidden nodes, and feed forward or recurrent network architecture) is required to be selected. As we know that the target function representation depends on the weights of the link, the number of hidden nodes and hidden layers, biases in the nodes, and type of activation function. Finding the right topology is not an easy task.

4. The weights and biases are initialized and initially random assignment is done.

5. If there is missing value or error in the training example, it should be replaced or removed with appropriate value.
6.3 PROPOSED MODEL

The most critical step in the use of a neural network is how to setup the net. This involves determining the number of layers, the number of nodes on the hidden layer, the weights and thresholds associated with each node. There can be more than one hidden layer, but one is usually sufficient for characterizing any complicated pattern. At the same time the difficult learning tasks can sometimes be simplified by increasing the number of internal layers. The number of nodes in the input layer would be equal to the number of input patterns whereas the number of nodes on the output layer is the number of classes. The number of nodes in the hidden layer is determined empirically.

Indirectly the neural network uses the domain knowledge of the expert. The classification result carried out earlier can be used to train the network such that the expert knowledge is implicitly encoded in the net through weights and threshold. This is in fact a supervised learning model.

The net adjusts the weights and thresholds in the net according to the pattern presented using a back-propagation algorithm. An iterative gradient algorithm is designed to minimize the mean square error between the actual output
and predetermined (or desired) output values. After all the input data and the output values are presented to the net, a new iteration is initiated and the iteration process is terminated until either the number of iterations is reached or the mean square error is below a specified small value.

With any of the termination scheme the training results are preserved through writing the weights and thresholds into two different files. Training can be continued by loading back the weight and threshold files. During the training, the convergence rate is controlled by two parameters: a gain factor and a momentum coefficient.

In the present work neural network has been implemented in relational database using self written PL/SQL program and a stored procedure. The program records all the adjusted weights right from the beginning to the end of the iteration into two separate tables. For that a table named ‘Training’ has been created for the training data and two different tables ‘V’ and ‘W’ for the weights. The ‘V’ table is for the weights between input layer to hidden layer and the ‘W’ table is for the weights between hidden layer to output layer. The first record of both the tables is the assumed weights between the layers. The program fetches the first records from the table and passes it to the stored procedure for computing the output, using feed forward algorithm, and comparing it with desired output. On finding out the difference between the desired output and computed output the same
procedure, using back propagation algorithm, adjusts the weights and makes entry into the two tables as a next record. The so adjusted weights between hidden layer to output layer is recorded into ‘W’ table and the adjusted weights between input layer to hidden layer is recorded into ‘V’ table.

Figure 6.1: Program layout of the proposed model
The following steps assumes sigmoid logistic nonlinearity:

Algorithm: NN-BPN in Relational Database
Input Data: Training data in ‘Training’ table
Initial weight assumption: Table ‘V’ for input to hidden & table ‘W’ for hidden to output.
Step 1: Read one training record from ‘Training’ table
Step 2: Read initial weights from table ‘V’ and ‘W’.
Step 3: for each hidden or output layer unit j{
    \( I_j = \sum_i W_{ij} O_i + \Theta_j \);
    \( O_j = \frac{1}{1 + e^{-I_j}} \);
    //compute the net input of unit j with respect to the previous layer, i
Step 4: Back propagate the errors: for each unit j in the output layer
    \( Err_j = O_j (1 - O_j) (T_j - O_j) \);
Step 5: Compute the error for each unit j in the hidden layers, from the last to the first hidden layer
    \( Err_j = O_j (1 - O_j) \sum_k Err_k W_{jk} \);
Step 6: Compute the error with respect to the next higher layer, k
    for each weight \( W_{ij} \) in network{ \( \Delta W_{ij} = (l) Err_j O_i \);
Step 7: weight Update
    \( W_{ij} = W_{ij} + \Delta W_{ij} \); //weight update
    for each bias \( \Theta_j \) in the network{ \( \Delta \Theta_j = (l) Err_j \); //bias increment
    \( \Theta_j = \Theta_j + \Delta \Theta_j \); //bias update
Step 8: Insert weights into tables and go to step 1.

Algorithm 6.1: NN-BPN implementation with RDBMS

After the training process, the set of final weights and thresholds for each nodes for the input to hidden to output layer are stored in the tables. Now for each set of the input features, a feed-forward calculation can then be used to obtain the final output values. The back-propagation training algorithm is an iterative gradient designed to minimize the mean square error between the actual output of multi layer feed forward perceptron and the desired output.
The advantage of storing the adjusted weight values for all the iterations into the tables is that we can further analyze the weight correction pattern of all the individual nodes of the layer. This can further be used to decide the appropriate number of layers, appropriate number of nodes in the layer and appropriate number of network connections. The detail program implementation in PL/SQL is mentioned in Annexure 6.1 and in Annexure 6.2.

6.4 DATA USED FOR TRAINING, VALIDATION AND TESTING

The Artificial Neural Network [ANN] model used in this work is the multi layer perceptron (MLP). For machine learning, the model requires a desired output which correctly maps input to output. For training, validation and testing of the model the relevant spatial and non spatial data of various have been collected from the following sources:

1. National AIDS Control Organization [NACO]

2. Planning Commission

3. UNDP Human Development Report 2009

The attributes that are significantly correlated have been identified with various statistical methods and are summarized in Chapter 4 (table 4.18 & 4.19).

Data of the following time period has been used for the model:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>Data from year 2002 to 2005</td>
</tr>
<tr>
<td>Validation data</td>
<td>Data from year 2005 to 2008</td>
</tr>
<tr>
<td>Testing data</td>
<td>Data from year 2008 to 2011</td>
</tr>
</tbody>
</table>

Table 6.1: Data period for model formation

6.5 IMPLEMENTATION OF ANN-BPN MODEL ON SPATIAL RDBMS DATA USING SQL-PL/SQL

It is indeed a challenging task to develop a computational procedure as a database object which is able to retrieve the database resident spatial and non spatial data and performs the NN based model training task inside the RDBMS environment and stores the adjusted weights of each step as a record in database. Since the spatial (vector) and non spatial data related to my work is stored in
RDBMS database I designed the database schema for computing the learning model of ANN. The schema of the work is as follows:

After some statistical computations which identify the highly correlated predicates among the various non spatial parameters, the prepared data is taken into a separate table as a training data set. For performing basic statistical task I have used MATLAB tool. After data has been prepared and stored into training
dataset table, the following PL/SQL program and stored procedures perform the network training task.

**Process 1:** It is a PL/SQL code that recursively invokes the stored procedure (Process 2) by passing the training data record one-by-one. The number of call (i.e. for number of epoch) to process 2 can be specified in process 1 by the user. For one epoch the number of call to process 2 is equal to the number of training data record.

**Process 2:** The procedure collects the training data, fetches the first record of the pre assumed weight records from the table 2 which contains the weights for input layer to hidden layer perceptrons. The relation between the input of the input layer and output of the input layer is linear. Sigmoidal function is used as an activation function for the hidden layer. Similarly table 3 contains the weight for the hidden layer to output layer. The first record of this table is the assumed weight. For the output layer also the activation function used is sigmoidal. Now on finding the difference between the computed output of the output layer and the expected out of the training data set (table 1) the back propagation process computes the adjusted weights both for the hidden-to-output layer and input-to-hidden layer. The process records the new adjusted weights into table 1 and table 2. As stated earlier the adjusted weights for hidden layer-to-output layer is recorded into table 3 and for input layer-to-output layer is recorded in table 2. It is worth mentioning
here that for each training records and for every subsequent computation the mean square error (MSE) is also recorded into table 3. This recorded MSE helps to analyze the MSE convergence rate to minimum.

6.5.1 TRAINING DATA

The class label (INFDIFF) here which shows the infection prevalence pattern is again of three categories. For training the model the census and NACO data used is of the period from 2002 to 2005. The negative value of the class label shows the decreasing prevalence of the disease, the positive value shows the increasing prevalence and the zero value shows the stable prevalence over the period 2002 to 2005.

Class Label 1: INFDIFF < 0; Infection Decreasing

Class Label 2: INFDIFF = 0; Infection Stable

Class Label 3: INFDIFF > 0; Infection Increasing

Now in order to find out the appropriate model for this problem the following topology has been tested and compared:
6.5.2 MULTI LAYER PERCEPTION WITH MULTI VALUE INPUT AND MULTIPLE OUTPUT: MODEL I

Multi layer perception with multi value input and multiple output model is described as follows:

![Multi Layer Perception Scheme](image)

*Figure 6.3: Multi Layer Perception scheme - Back Propagation Artificial Neural Network Architecture*
The topological information of this model is as follows:

- **Inputs are:** I1: HDI, I2: GDI, I3: MGRRATE, I4: DIST
- **Output Y:** HIV prevalence [INFDIFF]
- **Parameters of ANN:**
  - Nodes in input nodes: 4
  - No. of hidden layer: 1
  - No. of output layer: 1
  - Nodes in hidden layer: 3
  - Nodes in output layer: 3
  - Weight assigned in input to hidden layer: 12 weights
  - Weight assigned in each hidden to output layer: 3 weights
  - Learning rate: 0.6
  - Transfer Function: sigmoid

Table 6.2: Topology information of BPB-ANN Model-1

The input to the sigmoid function can have any value between plus and minus infinity. Then the sigmoid transfer function squashes the output into the range 0 to 1.

In this model the class label of the output is defined as below:

<table>
<thead>
<tr>
<th>Label Value</th>
<th>0.05</th>
<th>0.15</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label:</td>
<td>Infection</td>
<td>Infection</td>
<td>Infection</td>
</tr>
<tr>
<td></td>
<td>Decreasing</td>
<td>Increasing</td>
<td>Stable</td>
</tr>
</tbody>
</table>

Table 6.3: Class label values
Against these class label values the convergence of error square value (MSE) in hundred epoch is as follows:

For the states having increasing pattern of HIV prevalence the MSE (Mean Square Error) value has been minimized after 100 Epoch. The epoch to error graph for this class of data is as follows:

Figure 6.4: Convergence of MSE for states showing increasing sign of HIV
For the states having decreasing pattern of HIV prevalence the MSE (Mean Square Error) value has been minimized after 100 Epoch. The epoch-to-error graph for this class of data is as follows:

![MSE for HIV Decreasing Pattern States](image)

Figure 6.5: Convergence of MSE for states showing decreasing sign of HIV
Similarly for the states having stable pattern of HIV prevalence the MSE (Mean Square Error) value has been minimized after 100 Epoch. The epoch-to-error graph for this class of data is as follows:

![Graph showing convergence of MSE for states showing stable sign of HIV](image)

**Figure 6.6**: Convergence of MSE for states showing stable sign of HIV
The weight value and the Mean Square Error value after 100 epoch is recorded as follows:

<table>
<thead>
<tr>
<th>PERCEPTRON</th>
<th>INCREASING</th>
<th>DECREASING</th>
<th>STABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1W1</td>
<td>1.2362641</td>
<td>1.2735195</td>
<td>1.275757</td>
</tr>
<tr>
<td>H2W1</td>
<td>0.7822157</td>
<td>0.5077289</td>
<td>0.148866</td>
</tr>
<tr>
<td>H3W1</td>
<td>1.5597478</td>
<td>1.5997078</td>
<td>1.561459</td>
</tr>
<tr>
<td>I1W1</td>
<td>0.2863029</td>
<td>0.293445</td>
<td>0.231533</td>
</tr>
<tr>
<td>I1W2</td>
<td>0.4334321</td>
<td>0.3942083</td>
<td>0.365164</td>
</tr>
<tr>
<td>I1W3</td>
<td>0.8483722</td>
<td>0.848171</td>
<td>0.761855</td>
</tr>
<tr>
<td>I2W1</td>
<td>-0.0202087</td>
<td>-0.0098179</td>
<td>-0.07195</td>
</tr>
<tr>
<td>I2W2</td>
<td>0.2323129</td>
<td>0.1942755</td>
<td>0.166101</td>
</tr>
<tr>
<td>I2W3</td>
<td>-0.060393</td>
<td>-0.0559849</td>
<td>-0.14236</td>
</tr>
<tr>
<td>I3W1</td>
<td>0.1661802</td>
<td>0.3641367</td>
<td>0.468257</td>
</tr>
<tr>
<td>I3W2</td>
<td>-0.3229289</td>
<td>-0.3480084</td>
<td>-0.34788</td>
</tr>
<tr>
<td>I3W3</td>
<td>0.1286066</td>
<td>0.3889961</td>
<td>0.547294</td>
</tr>
<tr>
<td>I4W1</td>
<td>0.2652057</td>
<td>0.2537308</td>
<td>0.221849</td>
</tr>
<tr>
<td>I4W2</td>
<td>-0.0886318</td>
<td>-0.1027286</td>
<td>-0.10574</td>
</tr>
<tr>
<td>I4W3</td>
<td>0.487744</td>
<td>0.46877</td>
<td>0.426947</td>
</tr>
<tr>
<td>ERR_SQR</td>
<td>0.00664</td>
<td>0.01053</td>
<td>0.00347</td>
</tr>
</tbody>
</table>

Table 6.4: Adjusted weight value and MSE
6.5.3 MULTI LAYER PERCEPTION WITH MULTI VALUE INPUT AND SINGLE OUTPUT: Model II

Multi layer perception with multi value input and single output model is described as follows:

**Inputs are:** I1: HDI, I2: GDI, I3: MGRRATE, I4: DIST

**Output Y:** HIV prevalence [INFDIFF]

**Parameters of ANN:**
- Nodes in input nodes: 4
- No. of hidden layer: 1
- No. of output layer: 1
- Nodes in hidden layer: 3
- Nodes in output layer: 1
- Weight assigned in input to hidden layer: 12 weights
- Weight assigned in each hidden to output layer: 3 weights
- Learning rate: 0.6

Table 6.5: Topology information of BPN-ANN Model-2

![Diagram](image-url)
In this approach each class of data is separated at the database label and the different class data is used to train the model separately. In our problem we have identified three class labels. The class labels are the different pattern of HIV prevalence. The labels of the classes are computed as the mean value of the infection difference.

Error Square Pattern for Increasing Prevalence State:

\[ y = p_1z + p_2 \]

where \( z \) is centered and scaled:

\[ z = \frac{(x - \mu)}{\sigma} \]

\[ \mu = 51 \]
\[ \sigma = 29.3 \]

Coefficients:
\[ p_1 = -0.024517 \]
\[ p_2 = 0.023541 \]

Norm of residuals = 0.23402

Figure 6.8: MSE Plot for states showing increasing infection prevalence
Residual Plot:

Figure 6.9: Residual Plot for states showing increasing infection prevalence

Error Squire Pattern for Decreasing Prevalence State:

Figure 6.10: MSE Plot for states showing increasing infection prevalence
Residual Plot:

\[ y = p_1 z + p_2 \]
where \( z \) is centered and scaled:
\[ z = (x - \mu)/\sigma \]
\( \mu = 51 \)
\( \sigma = 29.3 \)

Coefficients:
\( p_1 = -0.036463 \)
\( p_2 = 0.036683 \)
Norm of residuals = 0.40319

Figure 6.11: Residual Plot for states showing increasing infection prevalence
Error Squire Pattern for Stable Prevalence State:

\[ y = p_1 z + p_2 \]

where \( z \) is centered and scaled:

\[ z = \frac{x - \mu}{\sigma} \]

\[ \mu = 51 \]
\[ \sigma = 29.3 \]

Coefficients:

\[ p_1 = -0.032819 \]
\[ p_2 = 0.027753 \]

Norm of residuals = 0.37368

Figure 6.12: MSE plot for states showing stable infection prevalence.
The final adjusted weight values of the three individual modules for the states having increasing pattern, decreasing pattern and stable pattern of HIV prevalence is as follows:

<table>
<thead>
<tr>
<th>PERCEPTRON</th>
<th>INCREASING</th>
<th>DECREASING</th>
<th>STABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1W1</td>
<td>1.2362641</td>
<td>1.2735195</td>
<td>1.275757</td>
</tr>
<tr>
<td>H2W1</td>
<td>0.7822157</td>
<td>0.5077289</td>
<td>0.148866</td>
</tr>
<tr>
<td>H3W1</td>
<td>1.5597478</td>
<td>1.5997078</td>
<td>1.561459</td>
</tr>
<tr>
<td>I1W1</td>
<td>0.2863029</td>
<td>0.293445</td>
<td>0.231533</td>
</tr>
<tr>
<td>I1W2</td>
<td>0.4334321</td>
<td>0.3942083</td>
<td>0.365164</td>
</tr>
<tr>
<td>I1W3</td>
<td>0.8483722</td>
<td>0.848171</td>
<td>0.761855</td>
</tr>
<tr>
<td>I2W1</td>
<td>-0.0202087</td>
<td>-0.0098179</td>
<td>-0.07195</td>
</tr>
<tr>
<td>I2W2</td>
<td>0.2323129</td>
<td>0.1942755</td>
<td>0.166101</td>
</tr>
<tr>
<td>I2W3</td>
<td>-0.060393</td>
<td>-0.0559849</td>
<td>-0.14236</td>
</tr>
<tr>
<td>I3W1</td>
<td>0.1661802</td>
<td>0.3641367</td>
<td>0.468257</td>
</tr>
<tr>
<td>I3W2</td>
<td>-0.3229289</td>
<td>-0.3480084</td>
<td>-0.34788</td>
</tr>
<tr>
<td>I3W3</td>
<td>0.1286066</td>
<td>0.3889961</td>
<td>0.547294</td>
</tr>
<tr>
<td>I4W1</td>
<td>0.2652057</td>
<td>0.2537308</td>
<td>0.221849</td>
</tr>
<tr>
<td>I4W2</td>
<td>-0.0886318</td>
<td>-0.1027286</td>
<td>-0.10574</td>
</tr>
<tr>
<td>I4W3</td>
<td>0.487744</td>
<td>0.46877</td>
<td>0.426947</td>
</tr>
<tr>
<td>ERR_SQR</td>
<td>0.00664</td>
<td>0.01053</td>
<td>0.00347</td>
</tr>
</tbody>
</table>

Table 6.6: Adjusted weight value and MSE
6.6 RESULT ANALYSIS

I performed validation and testing of the two modules (Model 1 & Model 2). Based on the result this analysis gives us the answer of the question that what kind of model is appropriate for our this kind of spatial data analysis.

Here again the data for validation (Table 5.6) is taken for the period of 2005 to 2008. When this data is executed with our NN-BPN based computational model the output is validated with the actual HIV prevalence information published by NACO.

For further testing the model the data from 2008 to 2011 has been taken and used in the model. Here also I have taken 15 random state data and tried to see whether the output pattern matched with the actual report presented by NACO. The test data so taken is as follows:
Now when this data is executed with our NN-BPN based computational model the output that we get is presented in the following table.

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>HDI</th>
<th>GDI</th>
<th>MGRRATE</th>
<th>DIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ARUNACHAL PRADESH</td>
<td>0.648</td>
<td>0.638</td>
<td>7.21</td>
<td>0.077276</td>
</tr>
<tr>
<td>18</td>
<td>ORISSA</td>
<td>0.548</td>
<td>0.525</td>
<td>-0.79</td>
<td>0.226372</td>
</tr>
<tr>
<td>14</td>
<td>MANIPUR</td>
<td>0.703</td>
<td>0.698</td>
<td>-1.38</td>
<td>0.073594</td>
</tr>
<tr>
<td>19</td>
<td>PUNJAB</td>
<td>0.664</td>
<td>0.661</td>
<td>1.69</td>
<td>0.1177616</td>
</tr>
<tr>
<td>24</td>
<td>UTTAR PRADESH</td>
<td>0.526</td>
<td>0.507</td>
<td>-1.9</td>
<td>0.2551165</td>
</tr>
<tr>
<td>7</td>
<td>HARYANA</td>
<td>0.644</td>
<td>0.635</td>
<td>4.2</td>
<td>0.1456709</td>
</tr>
<tr>
<td>12</td>
<td>MADHYA PRADESH</td>
<td>0.527</td>
<td>0.518</td>
<td>-0.11</td>
<td>0.6438269</td>
</tr>
<tr>
<td>11</td>
<td>KERALA</td>
<td>0.766</td>
<td>0.747</td>
<td>-0.61</td>
<td>0.1218676</td>
</tr>
<tr>
<td>5</td>
<td>GOA</td>
<td>0.765</td>
<td>0.741</td>
<td>8.2</td>
<td>0.1419272</td>
</tr>
<tr>
<td>30</td>
<td>CHANDIGARH</td>
<td>0.786</td>
<td>0.766</td>
<td>21.2</td>
<td>0.1061433</td>
</tr>
<tr>
<td>8</td>
<td>HIMACHAL PRADESH</td>
<td>0.67</td>
<td>0.666</td>
<td>1.2</td>
<td>0.1128265</td>
</tr>
<tr>
<td>33</td>
<td>DELHI</td>
<td>0.76</td>
<td>0.704</td>
<td>18.9</td>
<td>0.1341025</td>
</tr>
<tr>
<td>34</td>
<td>LAKSHADWEEP</td>
<td>0.699</td>
<td>0.638</td>
<td>6.7</td>
<td>0.0990527</td>
</tr>
<tr>
<td>35</td>
<td>PONDICHERRY</td>
<td>0.731</td>
<td>0.709</td>
<td>9</td>
<td>0.1609334</td>
</tr>
<tr>
<td>16</td>
<td>MIZORAM</td>
<td>0.69</td>
<td>0.689</td>
<td>-0.11</td>
<td>0.0792603</td>
</tr>
</tbody>
</table>

Table 6.7: Test Data

Now when this data is executed with our NN-BPN based computational model the output that we get is presented in the following table.

<table>
<thead>
<tr>
<th>SCODE</th>
<th>STATE</th>
<th>OUTPUT</th>
<th>STATECLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ARUNACHAL PRADESH</td>
<td>0.51</td>
<td>DECREASING</td>
</tr>
<tr>
<td>18</td>
<td>ORISSA</td>
<td>0.148</td>
<td>INCREASING</td>
</tr>
<tr>
<td>14</td>
<td>MANIPUR</td>
<td>0.5</td>
<td>DECREASING</td>
</tr>
<tr>
<td>19</td>
<td>PUNJAB</td>
<td>0.52</td>
<td>DECREASING</td>
</tr>
<tr>
<td>24</td>
<td>UTTAR PRADESH</td>
<td>0.6</td>
<td>DECREASING</td>
</tr>
<tr>
<td>7</td>
<td>HARYANA</td>
<td>0.59</td>
<td>DECREASING</td>
</tr>
<tr>
<td>12</td>
<td>MADHYA PRADESH</td>
<td>0.58</td>
<td>DECREASING</td>
</tr>
<tr>
<td>11</td>
<td>KERALA</td>
<td>0.57</td>
<td>DECREASING</td>
</tr>
<tr>
<td>5</td>
<td>GOA</td>
<td>0.51</td>
<td>DECREASING</td>
</tr>
<tr>
<td>30</td>
<td>CHANDIGARH</td>
<td>0.52</td>
<td>DECREASING</td>
</tr>
<tr>
<td>8</td>
<td>HIMACHAL PRADESH</td>
<td>0.091</td>
<td>STABLE</td>
</tr>
<tr>
<td>33</td>
<td>DELHI</td>
<td>0.58</td>
<td>DECREASING</td>
</tr>
<tr>
<td>34</td>
<td>LAKSHADWEEP</td>
<td>0.092</td>
<td>STABLE</td>
</tr>
<tr>
<td>35</td>
<td>PONDICHERRY</td>
<td>0.151</td>
<td>INCREASING</td>
</tr>
<tr>
<td>16</td>
<td>MIZORAM</td>
<td>0.59</td>
<td>DECREASING</td>
</tr>
</tbody>
</table>

Table 6.8 : Output of test data
The percentage of rule accuracy of both the models is summarized below:

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation Data Accuracy</th>
<th>Test Data Accuracy</th>
<th>No of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>93%</td>
<td>92%</td>
<td>100</td>
</tr>
<tr>
<td>Model II</td>
<td>96%</td>
<td>95%</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 6.9: Model accuracy result

The table 6.9 shows that Model-II gives us more accuracy when we perform validation and testing on the data. Secondly the computational load is again less in terms of number of iteration.

**6.7 DISCUSSION**

As noted earlier, classification and prediction is a fundamental data-mining task and ANNs are among the commonly used classification methods. However, conventional ANN does not consider the spatial dependence and associations between neighboring objects. We seek to address this issue in developing an ANN for space–time prediction. The ANN approach incorporates feedback from previous iterations of the model inputs and outputs. Such feedbacks model makes it a good candidate for modeling time-series data. Here we propose that a target prediction can be improved by not only incorporating the value of the target at the
The main advantage of neural network over conventional system is their ability to perform non linear input output mapping, generalization, adaptivity and fault tolerance. Here I tried to find out an appropriate neural network and optimal solution. The modeling of neural network for spatial database is quite unique as mostly the spatial data are temporal in nature and voluminous. So the spatial DBMS system is required to keep those data in relational database environment.

The comparative study of the neural network model shows that this model is having some limitations. Secondly the incomplete and fuzziness in the data available further prompts us to makes the model appropriate enough. This can be done if we consider fuzziness in the data available and propose a fuzzy NN-BPN model to work with spatial database. This task has further been discussed in Chapter 7
CHAPTER-7

FUZZY-BPN-NN MODEL OF SPATIAL CLASSIFICATION

7.1 INTRODUCTION

ANN can model complex non-linear relationships. It is appropriately suitable for classification phenomenon into predetermined classes. But the output precision is often limited to least square error, the training time required is quite large, the training data is chosen over entire range where the variables are expected to change.

Fuzzy logic system addresses the imprecision of inputs and outputs defined by fuzzy sets and allow greater flexibility in formulating detail system description[73]. The integrated neuro-fuzzy system have turned out to be useful in:

1. Accomplishing mathematical relationships among many variables in a complex dynamic process.
2. Performing mapping with some degree of imprecision.
3. Controlling non-linear systems to an extent not possible with conventional linear control system.

The direct fuzzification of conventional neural networks is to extend connection weights and/or inputs and/or fuzzy desired outputs (or targets) to fuzzy numbers. This various extensions are summarized below in Table 6.10

<table>
<thead>
<tr>
<th>Fuzzy NN</th>
<th>Weight</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>Crisp</td>
<td>Fuzzy</td>
<td>Crisp</td>
</tr>
<tr>
<td>Type 2</td>
<td>Crisp</td>
<td>Fuzzy</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>Type 3</td>
<td>Fuzzy</td>
<td>Fuzzy</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>Type 4</td>
<td>Fuzzy</td>
<td>Crisp</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>Type 5</td>
<td>Crisp</td>
<td>Crisp</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>Type 6</td>
<td>Fuzzy</td>
<td>Crisp</td>
<td>Crisp</td>
</tr>
<tr>
<td>Type 7</td>
<td>Fuzzy</td>
<td>Fuzzy</td>
<td>Crisp</td>
</tr>
</tbody>
</table>

Table 7.1 : Different Fuzzy- NN-BPN

Fuzzy neural networks of Type 1 are used in classification problem of a fuzzy input vector to a crisp class. The networks of Type 2, 3 and 4 are used to implement fuzzy IF-THEN rules. However, the last three types in Table 6.10 are unrealistic.

In our problem of extracting useful classification rule out of spatial database, I have examined the potential of Fuzzy-BP Type-1 hybrid architecture, which maps fuzzy inputs to crisp outputs. This model is the makeup of LR-type
fuzzy numbers. The triangular type of LR-type fuzzy numbers have been used for simplification of architecture and reduction of computational load.

![Symmetric triangular LR-Fuzzy Number](image)

**Figure 7.1: Symmetric triangular LR-Fuzzy Number**

### 7.2 FUZZY DATA SET AND MODEL

A fuzzy set A is called triangular fuzzy number with peak (or center) m, left width(a) \( a > 0 \) and right width(b) \( b > 0 \). The choice of L and R functions is specific to the problem. For the data I have chosen a symmetric triangular LR number. The left and right shift from the centre is taken as 10% of the peak value. Based on the given “Training” data I further created a “Fuzzy_Training” table and its layout is as follows:
Model III:

The topology information


Output Y : HIV prevalence [INFDIFF]

Parameters of ANN:
Nodes in input nodes: 4x3=12
No. of hidden layer: 1
No. of output layer: 1
Nodes in hidden layer: 3
Nodes in output layer: 1
Weight assigned in input to hidden layer: 12 weights
Weight assigned in each hidden to output layer: 3 weights
Learning rate: 0.6

Table 7.3: Topology information of Fuzzy BPB-ANN Model-3
The algorithm of the fuzzy NN-BPN model is as follows:

**Algorithm: Fuzzy NN-BPN in Relational Database**

Input Data: Training data in ‘Training’ table
Initial weight assumption: Table ‘V’ for input to hidden & table ‘W’ for hidden to output.

Step 1: Read one training record from ‘Fuzzy_Training’ table
Step 2: Read initial weights from table ‘V’ and ‘W’.
Step 3: for each hidden or output layer unit j{
    \[ I_j = \sum_i W_{ij} O_i + \Theta_j; \]
    \[ O_j = \frac{1}{1+e^{-(I_j)} }; \]
    \[
    //compute the net input of unit j with respect to the previous layer, i
    \]
Step 4: Back propagate the errors: for each unit j in the output layer
    \[ \text{Err}_j = O_j(1-O_j)(T_j-O_j); \]
Step 5: Compute the error for each unit j in the hidden layers, from the last to the first hidden layer
    \[ \text{Err}_j = O_j(1-O_j) \sum_k \text{Err}_k W_{jk}; \]
Step 6: Compute the error with respect to the next higher layer, k
    for each weight \( W_{ij} \) in network
    \[ \Delta W_{ij} = (l) \text{Err}_j O_i; \]
Step 7: weight Update
    \[ W_{ij} = W_{ij} + \Delta W_{ij}; \]
    //weight update
    for each bias \( \Theta_j \) in the network
    \[ \Delta \Theta_j = (l) \text{Err}_j; \]
    //bias increment
    \[ \Theta_j = \Theta_j + \Delta \Theta_j; \]
    //bias update
Step 8: Insert weights into tables and go to step 1.

Algorithm 7.1: Fuzzy NN-BPN implementation with RDBMS
The detail program written in PL/SQL for this model is present in Annexure 7.1 and in Annexure 7.2.

7.3 RESULT ANALYSIS

Here the input value to the input layer is taken as a fuzzy data where as the weights and outputs are the crisp values. From the mean Square Error (MSE) culve which shows the error in output value from the computed value, it is clear that with a very less number of iterations the MSE value has declined sharply and the model has established itself to an accurate state. For all the three classes of our HIV prevalence data separate training modules are trained which requires less iteration (forty here) to get trained. The iteration vs MSE graph of this module is as follows (Figure 6.15):

![Figure 7.3: MSE convergence of three models of fuzzy NN-BPN](image)

Figure 7.3: MSE convergence of three models of fuzzy NN-BPN
It can be seen that with 50 iterations the MSE of the three models has become uniform. Now after validating and testing of this model for the validation and test data of Table 5.6 & Table 6.7 respectively, we find the results as follows:

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation Data Accuracy</th>
<th>Test Data Accuracy</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model III</td>
<td>97%</td>
<td>96%</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 7.4: Model-3 accuracy result

It can be seen that the accuracy of the validation and test data results has gone up to 97% and 96% respectively.

With the help of spatial query language the testing output of the model is presented in a graphical way into the map of India as follows:

Figure 7.4: Different pattern of HIV spread in India (MapViewer)
CHAPTER-8
COMMENT, CONCLUSION AND FUTURE SCOPE

8.1 COMMENT

The framework of spatial database and mining algorithm when applied to our spatial database consisting of vector data of the states of India, its socio-economic and HIV epidemiology data, it reveals and confirms the high significant association between different patterns of HIV prevalence and spatial distance of the state locations and its other non spatial significant socio-economic parameters. The specific objectives and research questions in this study are addressed below.

Objective 1 - To build a spatial database in the existing RDBMS environment with the vector data of Indian states and potential non-spatial parameters which are influencing the spread of HIV in India.

- For doing the spatial data mining task what vector or raster data related to
the geographical location of India is to use.

- How to design a database where the spatial/ geographical information related to different state location and its non spatial parameters stored.

- Computing the centroid point of the states which represents the mean geographical location of the states.

- Finding out the state spatial predicate value.

**Objective 2** – To find out the different class labels related to the prevalence of the HIV disease in states and then building an efficient algorithm which is capable of finding out rules based on potentially significant spatial and non spatial predicates influencing the different pattern of spread of this disease.

- With the help of the DBMS potential finding out the different class labels of the HIV prevalence pattern.

- Preparing categorical data first and then formulating an algorithm based on two-step method to find out the rule, based on the potentially significant spatial and non spatial predicate values influencing the different pattern of spread of this disease.

- Developing an ANN model which takes into account the actual data of the different significant predicates to test and validate the rule extracted by the algorithm.
Objective 3- To map out the different growth and decline pattern of HIV prevalence in different states.

- With the help of MapViewer the output of the prediction can be visualized

8.2 CONCLUSION

We have seen that Spatial Data Mining is an important field of research with wide applicative areas. Although the field is quite young, a number of algorithms and techniques have been proposed to discover various kinds of knowledge from spatial database. I have surveyed the existing methods of spatial data mining which led us to design an appropriate framework for applying the spatial data mining into the study of heterogeneous pattern of HIV epidemiology prevalence in different states of India. The variety of yet unexplored topics and problems makes knowledge discovery in spatial database an attractive and challenging research field. The major contribution of the study can be summarized as follows:

i. Modeling a spatial database in existing relational database environment.

ii. Associating the non spatial and spatial predicates to further study the characteristic of spatial epidemiology.
An algorithm and program in PL/SQL has been modeled for rule-based spatial classification.

The NN-BPN and fuzzy NN-BPN model has been proposed to work with spatial database for training, validation and testing purpose. This model can be used for prediction by the policy makers to undertake planning an intervention mechanism.

8.3 FUTURE SCOPE

One extension of the current work in spatial data mining toward spatio-temporal database is to study data deviation and evolution rule. We can find spatial characteristic evolution rules which summarizes the general characteristics of the changing data. For example during the mining process one can discover properties of the regions with average growth of certain disease over 1% per year. The rule discriminates the properties of objects in the target class from those in the contrasting classes. For example we can make a comparison of the states where prevalence of disease increased last year with the states where health index has been improved.

Such rule can be used in health planning and monitoring process where one would like to find out how certain features are deviating from norm or how they are evolving over time.
8.4 LIMITATION

Lack of accurate spatial, non spatial data and presence of noise in data are the major deterrent for the study in this area. Some of the limitations of the present study are stated below so that sufficient precaution are exercised while applying the proposed methodology.

i. Artificial neural network is a powerful algorithm for classification. However the accuracy of forecasting depends not only on the forecasting technique but also on the accuracy of training data available for the model.

ii. Rule based classification utilizes only very limited expert knowledge. The expert system, therefore, is a problem solving system which supports expert knowledge in a computer based system. The model developed is a domain specific rather then generalized.
REFERENCES


[74] J. Han, and Y. Fu. Exploration of the power of Attribute-Oriented Induction in Data Mining. In[16]


PROGRAM CODE TO COMPUTE WEIGHT MATRIX OF STATE CENTROID (DIST)

CREATE OR REPLACE PROCEDURE DISTANCE(MSCODE
INDIA_ST.SCODE%TYPE) AS
MS METRO.SCODE%TYPE;
MDIST NUMBER;
MSUM NUMBER:=0;
CURSOR CMETRO IS
    SELECT SCODE FROM METRO;
BEGIN
OPEN CMETRO;
LOOP
    FETCH CMETRO INTO MS;
    EXIT WHEN CMETRO%NOTFOUND;
    SELECT SDO_GEOM.SDO_DISTANCE(c.geom, m.diminfo, d.geom,
        m.diminfo) INTO MDIST
        FROM india_st c, india_st d, user_sdo_geom_metadata m
        WHERE (m.table_name = 'INDIA_ST' AND m.column_name = 'GEOM') AND
        (c.SCODE = MSCODE AND d.SCODE = MS);
    MSUM:=(MSUM+MDIST);
END LOOP;
UPDATE DISTANCE_WT SET DIST = 1/(MSUM/5) WHERE SCODE=MSCODE;
CLOSE CMETRO;
END;
ANNEXTURE 5.1

--PROCEDURE FOR COMPUTING BIGGER LIKELYHOOD RATIO---
create or replace function learn_rule(
R1NP  NUMBER,
R1PP  NUMBER,
R1SP  NUMBER,
R2NP  NUMBER,
R2PP  NUMBER,
R2SP  NUMBER,
) RETURN NUMBER IS

LR1  NUMBER;
LR2  NUMBER;

EFNP  NUMBER
EFPP  NUMBER
EFSP  NUMBER;

BEGIN
EFNP:=R1NP*(22/32);
EFPP:=R1PP*(5/32);
EFSP:=R1SP*(5/32);
LR1:=2*[R1NP*[LOG(R1NP/EFNP)/LOG(2)]+R1PP*[LOG(R1PP/EFPP)/LOG(2)]+R1SP*[LOG(R1SP/EFSP)/LOG(2)];

EFNP:=R2NP*(22/32);
EFPP:=R2PP*(5/32);
EFSP:=R2SP*(5/32);
LR2:=2*[R2NP*[LOG(R2NP/EFNP)/LOG(2)]+R2PP*[LOG(R2PP/EFPP)/LOG(2)]+R2SP*[LOG(R2SP/EFSP)/LOG(2)];

IF LR1>LR2
    RETURN LR1;
ELSE IF
    RETURN LR2;
create or replace function I_GAIN(
NP1 NUMBER,
PP1 NUMBER,
SP1 NUMBER,

NP0 NUMBER,
PP0 NUMBER,
SP0 NUMBER,

) RETURN NUMBER IS

IG NUMBER;

BEGIN
IG:= NP1*[(LOG(NP1/(NP1+PP1+SP1))/LOG(2))-
(LOG(NP0/(NP0+PP0+SP0))/LOG(2))];

RETURN IG;
END;
---Program to invoke the stored procedure of NN-BPN---
Declare
n number(2);
m integer(2):=1;
x integer(2);

vi1 xtraining.i1%TYPE;
vi2 xtraining.i2%TYPE;
vi3 xtraining.i3%TYPE;
vi4 xtraining.i4%TYPE;
voutput xtraining.output%TYPE;

begin

dbms_output.put_line('Training of BPN is in progress.....');
select count(*) into x from xtraining;
for n in 1..100loop
    select i1,i2,i3,i4,output into vi1,vi2,vi3,vi4,voutput from xtraining where data=m;
    xcalculation_4(n,vi1,vi2,vi3,vi4,voutput);
    m:=m+1;
    if(m>x) then
        m:=1;
    end if;
end loop;
dbms_output.put_line('Training of BPN is completed.....');
end;
---Stored Procedure of NN-BPN---
create or replace procedure xcalculation_4(
m xtraining.data%type,
pi1 xtraining.i1%TYPE,
pi2 xtraining.i2%TYPE,
pi3 xtraining.i3%TYPE,
pi4 xtraining.i4%TYPE,
po xtraining.output%TYPE) AS

  vi1w1 xv.i1w1%TYPE;
  vi1w2 xv.i1w2%TYPE;
  vi1w3 xv.i1w2%TYPE;

  vi2w1 xv.i2w1%TYPE;
  vi2w2 xv.i2w2%TYPE;
  vi2w3 xv.i2w3%TYPE;

  vi3w1 xv.i3w1%TYPE;
  vi3w2 xv.i3w2%TYPE;
  vi3w3 xv.i3w3%TYPE;

  vi4w1 xv.i4w1%TYPE;
  vi4w2 xv.i4w2%TYPE;
  vi4w3 xv.i4w3%TYPE;

  vh1w1 xw.h1w1%TYPE;
  vh2w1 xw.h2w1%TYPE;
  vh3w1 xw.h3w1%TYPE;

  IH1 number;
  IH2 number;
  IH3 number;

  OH1 number(6,4);
  OH2 number(6,4);
  OH3 number(6,4);

  io number(6,4);
  oo number(6,4);
  errorout number(7,5);
d number(9,7);
d1 number(9,7);
d2 number(9,7);
d3 number(9,7);
d4 number(9,7);

y1 number(9,7);
y2 number(9,7);
y3 number(9,7);

w1 number(9,7);
w2 number(9,7);
w3 number(9,7);

dw1 number(9,7);
dw2 number(9,7);
dw3 number(9,7);

v11 number(9,7);
v12 number(9,7);
v13 number(9,7);
v21 number(9,7);
v22 number(9,7);
v23 number(9,7);
v31 number(9,7);
v32 number(9,7);
v33 number(9,7);

v41 number(9,7);
v42 number(9,7);
v43 number(9,7);

cursor cv is
select i1w1,i1w2,i1w3,i2w1,i2w2,i2w3,i3w1,i3w2,i3w3,i4w1,i4w2,i4w3 from xv where wt=m;
cursor cw is
select h1w1,h2w1,h3w1 from xw where wt=m;

begin
open cv;
open cw;
--loop
fetch cv into vi1w1, vi1w2, vi1w3, vi2w1, vi2w2, vi2w3, vi3w1, vi3w2, vi3w3, vi4w1, vi4w2, vi4w3;
fetch cw into vh1w1, vh2w1, vh3w1;
--exit when cv%NOTFOUND;

IH1:=pi1*vi1w1+pi2*vi2w1+pi3*vi3w1+pi4*vi4w1;
IH2:=pi1*vi1w2+pi2*vi2w2+pi3*vi3w2+pi4*vi4w2;
IH3:=pi1*vi1w3+pi2*vi2w3+pi3*vi3w3+pi4*vi4w3;

OH1:=1/(1+exp(-IH1));
OH2:=1/(1+exp(-IH2));
OH3:=1/(1+exp(-IH3));

io:=oh1*vh1w1+oh2*vh2w1+oh3*vh3w1;
oo:=1/(1+exp(-io));
--oo:=ROUND(oo);

errorout:=power((po-oo),2);

d:=(po-oo)*oo*(1-oo);

y1:=oh1*d;
y2:=oh2*d;
y3:=oh3*d;
dw1:=y1*0.6;
dw2:=y2*0.6;
dw3:=y3*0.6;
w1:=dw1+vh1w1;
w2:=dw2+vh2w1;
w3:=dw3+vh3w1;
insert into xw values(m+1,w1,w2,w3,errorout);

d1:=oh1*(1-oh1)*(d*vh1w1);
d2:=oh2*(1-oh2)*(d*vh2w1);
d3:=oh3*(1-oh3)*(d*vh3w1);

v11:=pi1*d1*0.6;
v12:=pi1*d2*0.6;
v13:=pi1*d3*0.6;
v21:=pi2*d1*0.6;
v22:=\pi^2 d^2 * 0.6;
v23:=\pi^2 d^3 * 0.6;

v31:=\pi^3 d^1 * 0.6;
v32:=\pi^3 d^2 * 0.6;
v33:=\pi^3 d^3 * 0.6;

v41:=\pi^4 d^1 * 0.6;
v42:=\pi^4 d^2 * 0.6;
v43:=\pi^4 d^3 * 0.6;

v11:=v11+vi1w1;
v12:=v12+vi1w2;
v13:=v13+vi1w3;

v21:=v21+vi2w1;
v22:=v22+vi2w2;
v23:=v23+vi2w3;

v31:=v31+vi3w1;
v32:=v32+vi3w2;
v33:=v33+vi3w3;

v41:=v41+vi4w1;
v42:=v42+vi4w2;
v43:=v43+vi4w3;

insert into xv values(m+1,v11,v12,v13,v21,v22,v23,v31,v32,v33,v41,v42,v43);

DBMS_OUTput.PUT_LINE('First loop');
CLOSE CV;
CLOSE CW;

END;
Declare
n number(2);
m integer(2):=1;
x integer(2);

vi0m ftraining.i0m%TYPE;
vi0a ftraining.i0a%TYPE;
vi0b ftraining.i0b%TYPE;

vi1m ftraining.i1m%TYPE;
vi1a ftraining.i1a%TYPE;
vi1b ftraining.i1b%TYPE;

vi2m ftraining.i2m%TYPE;
vi2a ftraining.i2a%TYPE;
vi2b ftraining.i2b%TYPE;

voutput ftraining.output%TYPE;

Begin

dbms_output.put_line('Training of BPN is in progress.....');

select count(*) into x from ftraining;
n:=1;
for n in 1..10 loop
    select i0m,i0a,i0b,i1m,i1a,i1b,i2m,i2a,i2b,output into
    vi0m,vi0a,vi0b,vi1m,vi1a,vi1b,vi2m,vi2a,vi2b,voutput from ftraining where data=m;
    fcalculation_4(n,vi0m,vi0a,vi0b,vi1m,vi1a,vi1b,vi2m,vi2a,vi2b,voutput);
    m:=m+1;
    if(m>x) then
        m:=1;
    end if;
end loop;
dbms_output.put_line('Training of BPN is compleated.....');
end;
create or replace procedure fcalculation_4(
    m    ftraining.data%type,
    pi0m ftraining.i0m%TYPE,
    pi0a ftraining.i0a%TYPE,
    pi0b ftraining.i0b%TYPE,
    pi1m ftraining.i1m%TYPE,
    pi1a ftraining.i1a%TYPE,
    pi1b ftraining.i1b%TYPE,
    pi2m ftraining.i2m%TYPE,
    pi2a ftraining.i2a%TYPE,
    pi2b ftraining.i2b%TYPE,
    po  ftraining.output%TYPE) AS

    vv00m fv.iv00m%TYPE;
    vv00a fv.iv00a%TYPE;
    vv00b fv.iv00b%TYPE;

    vv01m fv.iv01m%TYPE;
    vv01a fv.iv01a%TYPE;
    vv01b fv.iv01b%TYPE;

    vv02m fv.iv02m%TYPE;
    vv02a fv.iv02a%TYPE;
    vv02b fv.iv02b%TYPE;

    vv10m fv.iv10m%TYPE;
    vv10a fv.iv10a%TYPE;
    vv10b fv.iv10b%TYPE;

    vv11m fv.iv11m%TYPE;
    vv11a fv.iv11a%TYPE;
    vv11b fv.iv11b%TYPE;

    vv12m fv.iv12m%TYPE;
    vv12a fv.iv12a%TYPE;
    vv12b fv.iv12b%TYPE;

    vv20m fv.iv20m%TYPE;
vv20a fv.iv20a%TYPE;
vv20b fv.iv20b%TYPE;

vv21m fv.iv21m%TYPE;
vv21a fv.iv21a%TYPE;
vv21b fv.iv21b%TYPE;

vv22m fv.iv22m%TYPE;
vv22a fv.iv22a%TYPE;
vv22b fv.iv22b%TYPE;
vh00m fw.h00m%TYPE;
vh00a fw.h00a%TYPE;
vh00b fw.h00b%TYPE;

vh01m fw.h01m%TYPE;
vh01a fw.h01a%TYPE;
vh01b fw.h01b%TYPE;

vh02m fw.h02m%TYPE;
vh02a fw.h02a%TYPE;
vh02b fw.h02b%TYPE;

net11 number;
net12 number;
net13 number;

net21 number;
net22 number;
net23 number;

net31 number;
net32 number;
net33 number;

net1x number;
net2x number;
net3x number;
netx number;
ox number;

net1y number;
net2y number;
net3y number;
nety number;
oy number;

net1z number;
net2z number;
net3z number;
netz number;
oz number;

x number;
y number;
z number;
p number;
q number;
r number;

o11 number;
o12 number;
o13 number;
o21 number;
o22 number;
o23 number;
o31 number;
o32 number;
o33 number;

net1 number;
net2 number;
net3 number;

-----
dem00 number;
dea00 number;
deb00 number;
dem01 number;
dea01 number;
deb01 number;
dem02 number;
dea02 number;
deb02 number;
dwm00 number;
dwa00 number;
dwb00 number;
dwm01 number;
dwa01 number;
dwb01 number;
dwm02 number;
dwa02 number;
dwb02 number;
dm0   number;
da0   number;
db0   number;
dewm00 number;
dewa00 number;
dewb00 number;
dewm10 number;
dewa10 number;
dewb10 number;
dewm20 number;
dewa20 number;
dewb20 number;
dewm01 number;
dewa01 number;
dewb01 number;
dewm11 number;
dewa11 number;
dewb11 number;
dewm21 number;
dewa21 number;
dewb21 number;
dewm02 number;
dewa02 number;
dewb02 number;
dewm12 number;
dewa12 number;
dewb12 number;
dewm22 number;
dewa22 number;
dewb22 number;
cvv00m number;
cvv00a number;
cvv00b number;
cvv01m number;
cvv01a number;
cvv01b number;
cvv02m number;
cvv02a number;
cvv02b number;
cvv10m number;
cvv10a number;
cvv10b number;
cvv11m number;
cvv11a number;
cvv11b number;
cvv12m number;
cvv12a number;
cvv12b number;
cvv20m number;
cvv20a number;
cvv20b number;
cvv21m number;
cvv21a number;
cvv21b number;
cvv22m number;
cvv22a number;
cvv22b number;
--oo number;

O10 number;
--O11 number;
--O12 number;
OO1 number;
OO2 number;
OO3 number;
oo number(6,4);

erro rut fw.err_sqr%TYPE;
cursor cv is
select iv00m, iv00a, iv00b,
iv01m,iv01a,iv01b,iv02m,iv02a,iv02b,iv10m,iv10a,iv10b,iv11m,iv11a,iv12a,iv12b,iv20m,iv20a,iv20b,iv21m,iv21a,iv21b,iv22m,iv22a,iv22b from fv where wt=m;
cursor cw is
select h00m,h00a,h00b,h01m,h01a,h01b,h02m,h02a,h02b from fw where wt=m;

begin
open cv;
open cw;
fetch cv into
vv00m,vv00a,vv00b,vv01m,vv01a,vv01b,vv02m,vv02a,vv02b,vv10m,vv10a,vv10b,vv11m,vv11a,vv12m,vv12a,vv12b,vv20m,vv20a,vv20b,vv21m,vv21a,vv21b,vv22m,vv22a,vv22b;
fetch cw into vh00m,vh00a,vh00b,vh01m,vh01a,vh01b,vh02m,vh02a,vh02b;
x:=pi0m;
y:=pi0a;
z:=pi0b;
p:=vv00m;
q:=vv00a;
r:=vv00b;

if(x>=0 and p>=0) then
  net11:=(x*p);
  net12:=(x*q)+(p*y);
  net13:=(x*r)+(p*z);
end if;

if(p>=0 and x<0) then
  net11:=(x*p);
  net12:=(p*y)-(x*r);
  net13:=(p*z)-(x*q);
end if;

if(p<0 and x<0) then
  net11:=(x*p);
  net12:=(-p*z)-(x*r);
  net13:=(-p*y)-(x*q);
end if;

x:=pi1m;
y:=pi1a;
z:=pi1b;
p:=vv01m;
q:=vv01a;
r:=vv01b;
if(x>=0 and p>=0) then
  net21:=(x*p);
  net22:=(x*q)+(p*y);
  net23:=(x*r)+(p*z);
end if;

if(p>=0 and x<0) then
  net21:=(x*p);
  net22:=(p*y)-(x*r);
  net23:=(p*z)-(x*q);
end if;
if(p<0 and x<0) then
net21:=(x*p);
net22:=(-p*z)-(x*r);
net23:=(-p*y)-(x*q);
end if;
x:=pi2m;
y:=pi2a;
z:=pi2b;
p:=vv02m;
q:=vv02a;
r:=vv02b;
if(x>=0 and p>=0) then
net31:=(x*p);
net32:=(x*q)+(p*y);
net33:=(x*r)+(p*z);
end if;
if(p>=0 and x<0) then
net31:=(x*p);
net32:=(p*y)-(x*r);
net33:=(p*z)-(x*q);
end if;
if(p<0 and x<0) then
net31:=(x*p);
net32:=(-p*z)-(x*r);
net33:=(-p*y)-(x*q);
end if;
net1x:=net11+((1/3)*(net12*net13));
net2x:=net21+((1/3)*(net22*net23));
net3x:=net31+((1/3)*(net32*net33));
O10:=1/(1+exp(-net1x));
O11:=1/(1+exp(-net2x));
O12:=1/(1+exp(-net3x));

--DBMS_OUTPUT.PUT_LINE('*********************************');
---------------------------------------------SET 2------------------------------------------
x:=O10;
y:=O11;
z:=O12;
p:=vh00m;
q:=VH00A;
r:=vh00b;

OO1:=O10 *vh00m+O11*vh01m+012*vh02m;
OO2:=O10 *vh00a+O11*vh01b+012*vh02b;
OO3:=O10 *vh00b+O11*vh01b+012*vh02b;

OO:=OO1+((1/3)*(OO2+OO3));

OO:=1/(1+exp(-OO));
--OO:=ROUND(OO);

errorout:=power((po-oo),2);

---We Now Proceed to Compute the change of weights for the input, hidden and output layer.

---First for changing the weight for hidden to output layer
dem00:= -(po-oo)*(oo)*(1-oo)*1*O10;
dea00:= -(po-oo)*(oo)*(1-oo)*(-1/3)*O10;
deb00:= -(po-oo)*(oo)*(1-oo)*(1/3)*O10;

dem01:= -(po-oo)*(oo)*(1-oo)*1*O11;
dea01:= -(po-oo)*(oo)*(1-oo)*(-1/3)*O11;
deb01:= -(po-oo)*(oo)*(1-oo)*(1/3)*O11;

dem02:= -(po-oo)*(oo)*(1-oo)*1*O12;
dea02:= -(po-oo)*(oo)*(1-oo)*(-1/3)*O12;
deb02:= -(po-oo)*(oo)*(1-oo)*(1/3)*O12;

dwm00:= -0.9*dem00;
dwa00:= -0.9*dea00;
dwb00:= -0.9*deb00;

dwm01:= -0.9*dem01;
dwa01:= -0.9*dea01;
dwb01:= -0.9*deb01;

dwm02:= -0.9*dem02;
dwa02:= -0.9*dea02;
dwb02:= -0.9*deb02;

----So the updated wt, for entry into fw table, from hidden to output is

vh00m:=vh00m+dwm00;
vh00a:=vh00a+dwa00;
vh00b:=vh00b+dwb00;

vh01m:=vh01m+dwm01;
vh01a:=vh01a+dwa01;
vh01b:=vh01b+dwb01;

vh02m:=vh02m+dwm02;
vh02a:=vh02a+dwa02;
vh02b:=vh02b+dwb02;

DBMS_OUTPUT.PUT_LINE('++++++++++++++++++++++Values to be inserted into
fw +++++++++++++++++++++++++++++++++');
insert into fw
values(m+1,vh00m,vh00a,vh00b,vh01m,vh01a,vh01b,vh02m,vh02a,vh02b,errorout);

---Now change the weight from input to hidden layer

dewm00:=(dem00 * VH00M)*(O10)*(1-O10)*pi0m;
dewa00:=(dea00 * VH00A)*(O10)*(1-O10)*pi0a;
dewb00:=(deb00 * VH00B)*(O10)*(1-O10)*pi0b;

dewm10:=(dem01 * VH00M)*(O10)*(1-O10)*pi1m;
dewa10:=(dea01 * VH00A)*(O10)*(1-O10)*pi1a;
dewb10:=(deb01 * VH00B)*(O10)*(1-O10)*pi1b;

dewm20:=(dem02 * VH00M)*(O10)*(1-O10)*pi2m;
dewa20:=(dea02 * VH00A)*(O10)*(1-O10)*pi2a;
dewb20:=(deb02 * VH00B)*(O10)*(1-O10)*pi2b;

dewm01:=(dem00 * VH01M)*(O11)*(1-O11)*pi0m;
dewa01:=(dea00 * VH01A)*(O11)*(1-O11)*pi0a;
dewb01:=(deb00 * VH01B)*(O11)*(1-O11)*pi0b;

dewm11:=(dem01 * vh01m)*(O11)*(1-O11)*pi1m;
dewa11:=(dea01 * vh01a)*(O11)*(1-O11)*pi1a;
dewb11:=(deb01 * vh01b)*(O11)*(1-O11)*pi1b;
dewm12:=(dem02 \times vh01m)\times(O11)\times(1-O11)\times\pi2m;
dewa12:=(dea02 \times vh01a)\times(O11)\times(1-O11)\times\pi2a;
dewb12:=(deb02 \times vh01b)\times(O11)\times(1-O11)\times\pi2b;

dewm02:=(dem00 \times VH02M)\times(O12)\times(1-O12)\times\pi0m;
dewa02:=(dea00 \times VH02A)\times(O12)\times(1-O12)\times\pi0a;
dewb02:=(deb00 \times VH02B)\times(O12)\times(1-O12)\times\pi0b;

dewm12:=(dem01 \times vh02m)\times(O12)\times(1-O12)\times\pi1m;
dewa12:=(dea01 \times vh02a)\times(O12)\times(1-O12)\times\pi1a;
dewb12:=(dem01 \times vh02b)\times(O12)\times(1-O12)\times\pi1b;

dewm22:=(dem02 \times vh02m)\times(O12)\times(1-O12)\times\pi2m;
dewa22:=(dea02 \times vh02a)\times(O12)\times(1-O12)\times\pi2a;
dewb22:=(deb02 \times vh02b)\times(O12)\times(1-O12)\times\pi2b;

---The change in wt is given by

cvv00m:=-0.9*dewm00;
cvv00a:=-0.9*dewa00;
cvv00b:=-0.9*dewb00;

cvv01m:=-0.9*dewm10;
cvv01a:=-0.9*dewa10;
cvv01b:=-0.9*dewb10;

cvv02m:=-0.9*dewm20;
cvv02a:=-0.9*dewa20;
cvv02b:=-0.9*dewb20;

cvv10m:=-0.9*dewm01;
cvv10a:=-0.9*dewa01;
cvv10b:=-0.9*dewb01;

cvv11m:=-0.9*dewm11;
cvv11a:=-0.9*dewa11;
cvv11b:=-0.9*dewb11;

cvv12m:=-0.9*dewm21;
cvv12a:=-0.9*dewa21;
cvv12b:=-0.9*dewb21;
cvv20m:= -0.9*(dewm02);
cvv20a:= -0.9*(dewa02);
cvv20b:= -0.9*(dewb02);

cvv21m:= -0.9*(dewm12);
cvv21a:= -0.9*(dewa12);
cvv21b:= -0.9*(dewb12);

cvv22m:= -0.9*(dewm22);
cvv22a:= -0.9*(dewa22);
cvv22b:= -0.9*(dewb22);

---Now the updated weight for insertion into fv table

vv00m:= vv00m+cvv00m;
vv00a:= vv00a+cvv00a;
vv00b:= vv00b+cvv00b;

vv01m:= vv01m+cvv01m;
vv01a:= vv01a+cvv01a;
vv01b:= vv01b+cvv01b;

vv02m:= vv02m+cvv02m;
vv02a:= vv02a+cvv02a;
vv02b:= vv02b+cvv02b;

vv10m:= vv10m+cvv10m;
vv10a:= vv10a+cvv10a;
vv10b:= vv10b+cvv10b;

vv11m:= vv11m+cvv11m;
vv11a:= vv11a+cvv11a;
vv11b:= vv11b+cvv11b;

vv12m:= vv12m+cvv12m;
vv12a:= vv12a+cvv12a;
vv12b:= vv12b+cvv12b;

vv20m:= vv20m+cvv20m;
vv20a:= vv20a+cvv20a;
vv20b:= vv20b+cvv20b;

vv21m:= vv21m+cvv21m;
vv21a:= vv21a+cvv21a;
vv21b:= vv21b+cvv21b;
vv22m:=vv22m+cvv22m;
vv22a:=vv22a+cvv22a;
vv22b:=vv22b+cvv22b;

insert into fv
values(m+1,vv00m,vv00a,vv00b,vv01m,vv01a,vv01b,vv02m,vv02a,vv02b,vv10m,vv10a,
vv10b,vv11m,vv11a,vv11b,vv12m,vv12a,vv12b,vv20m,vv20a,vv20b,vv21m,vv21a,vv21b,vv22m,vv22a,vv22b);

--DBMS_OUTPUT.PUT_LINE('First loop');

CLOSE CV;
CLOSE CW;

END;
APPENDIX 3.1

Methodology Framework

Following are some important attributes in the study of spatial data mining:

Rules: With the help of combined approach of SDM techniques, various rules can be discovered. such as spatial association rule, spatial characteristic rule, deviation and evolution rule, and discriminate rule etc.

Thematic Map: It is a map that shows a theme, which is a single spatial distribution or a pattern, using a specific map type [46]. It presents the spatial distribution of a single or a few attributes. Spatial classification is one of the techniques that analyze spatial and non-spatial attributes of the data objects to partition the data into a set of classes that generates maps representing groups of related data objects. Thematic maps are represented by two ways: raster map and vector map. The raster image thematic maps have pixels associated with the attribute values. In the vector form the spatial objects are represented by its geometry i.e. boundary representation and thematic attributes.

Image Databases: These are special kind of spatial databases which consists of images and pictures stored in the form of grid array. It represents the image intensity in one or more spectral ranges.
There are various architectures proposed for spatial data mining. Some of the important architectures are J Han & Y. Fu’s[74] architecture DBLEARN/DBMINER, M. Holsheimer and M. Kersten’s [8] parallel architecture and C. J Matheus & Chan’s[75] multi component architecture. These are the general data mining prototypes but they can be used or extended to handle spatial data mining. Matheu’s architecture is very general and has been used by other researchers in spatial data mining, including M. Ester etc. [33]. The important spatial operations like spatial joins, map overlays, nearest neighbor queries are some important spatial operators. Thus in order to work efficiently, the operators requires an efficient spatial access method (SAM) and appropriate spatial data structure. The basic building blocks of spatial data structure are points, lines, rectangles etc. To build indices for these data, multidimensional trees such as R-tree, R*-tree, quad tree[62], k-d tree etc. have been proposed.
INTELLIGENCE IN SPATIAL DATA MINING

The effort of developing data mining algorithms for spatial database [76] is mainly motivated by the large amount of data collected through various applications, ranging from geographical information system (GIS), computer cartography, remote sensing, social and environmental assessment and planning etc. Interestingly, the geo-coding of epidemic infected people in combination with his/her address (location), the economic and social statistics of that location creates very large spatially related database. The increasing amount of spatial data and the necessity of intelligent decision support systems have created a new area of research i.e. intelligent spatial data mining. The intelligent spatial data mining is the subfield of data mining which deals with the extraction of implicit knowledge, spatial relationships, or other interesting patterns [44]. It combines research in spatial database, spatial reasoning, statistics, and machine learning.

As stated earlier, in the last few years algorithm for spatial association [44], spatial clustering[77], characteristics of spatial cluster[78], clustering[32], generalized spatial distribution[37] and other were analyzed. Now in order to get
meaningful information and identify meaningful pattern from the accumulated spatial and non-spatial database. In this work the task is divided into the following steps:

1. Using Spatial database to filter out the irrelevant attributes and using it in computing the distance measure parameter based on spatial dataset. The geographic data consists of spatial objects and non-spatial description of these objects. In the earlier approach of creating spatial database, the non-spatial description of spatial objects were stored in a different relational database where one attribute is a pointer to the spatial description of the object[79]. In my work I have modeled the database which is capable of storing spatial and non-spatial attributes together which improves the performance of analytical algorithms. Such model can be utilized for any kind of applicative analysis on large spatial database.

2. Identifying appropriate spatial object classification method which has not enough been explored. For getting about the rule that describes the partition of the database into a given set of classes, the database is analyzed. In the relational database the tuples are treated as an object and each object is assumed to have a pre defined class. The class is determined by one of the attribute which is called the class label attribute. The conventional classification methods, from statistics and machine learning
point of view, consider only relational data. In the process of spatial classification one wants to find rules that partition a set of classified objects into a number of classes using not only non spatial properties of the classified objects, but also the spatial relations of the classified objects to other objects in the database. In the present work the issues regarding classification of spatial data and building classification rule based on spatial and non spatial predicate values has been addressed. For this work I have analyzed the problem of spatial classification. For that I considered various thematic layers of non spatial attributes and their spatial relationships to other objects in the database. Here an algorithm has been proposed that handles spatial relations as well. Experimental results, of spatial classification of spatial database consisting of state location as a spatial factor and HIV/AIDS prevalence data, socio economic attribute of India as non spatial factors, has been presented here.

3. In order to work on actual numerical value of the attributes and establish non linear relation between antecedents (non-spatial and spatial) and consequent an appropriate neural network BPN model has been proposed. The model has been implemented using PL/SQL code with data in Oracle RDBMS. The proposed model will be able to work on large relational database for any applicative area.
Spatial epidemiology and spatial data mining

Elliott and Wartenberg [1] described “Spatial epidemiology is the description and analysis of the geographic, or spatial, variations in disease with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors”. The spread of infectious disease is closely associated with the concept of spatial or spatio-temporal proximity. The individuals who are linked in a spatial and temporal sense are at a high risk of getting infected[2]. Thus the knowledge of spatial and temporal variations of disease and characterizing its spatial structures is essential for the epidemiologist to understand better the population’s interaction with its environment [3]. The history of spatial epidemiology dates back to 1800s, when maps of disease rates in different countries began to emerge to characterize the spread and possible causes of outbreak of infectious disease such as yellow fever and cholera [4]. Proximity to environmental risk factors is therefore important. Spatial epidemiology analysis comprises of wide range of methods. Now it is a big challenge to determine which one to use [5].
APPENDIX 4.2

SPATIAL DATABASE DESIGN ISSUES

Spatial data are the data related to objects and that occupy. The spatial data contains various distance and/or topological information. They are often organized by spatial indexing structures and accessed by spatial access methods. In spatial database the spatial objects are having implicit relationship among them[80]. This poses challenge and bring opportunities for mining information from spatial data [15,81]. *Knowledge discovery from database refers to the extraction of implicit knowledge, spatial relation, or other patterns not explicitly stored in spatial database*[44].

The first hypothesis of the model considers interstate migration distance information. Here each record of the data table is related to one state which contains all the non spatial and spatial information of the state. Therefore each state record represents one object. Now the geographical distance i.e. the distance measure of one object from another definitely contains some implicit relation. Now in terms of the study of epidemiology distribution this distance measure has been taken into account. Moreover it has been found in the study of HIV epidemiology (Appendix A) that he worker class migration and migration distance counts a lot in the spread of HIV infection in India. Migrants mainly migrate to certain metropolitan cities to get employment. Those metropolitan cities are the places where most of the migrant workers, gets in contact with the HIV high risk
group population. The study further says that the migrants from far distance place are away from their families for a longer period of time and are more prone to get in contact with HIV high risk group people. So the distance measure from migration source state to destination state plays an important spatial factor in the study.

The computational framework of getting this distance measure is as follows:

1. A database is created with the non spatial data of the states and the scaled vector data of state boundary. When we use the inbuilt distance function for computing the distance between two polygons it considers the nearest boundary location of the states. But for accuracy of the distance measure the centroid of the states has been taken to measure the distance between source state and destination state.

2. The states like Delhi, Maharashtra, Tamil Nadu, West Bengal and Gujarat area taken as major destination states while the other states are the source states.

3. Now with the help of a procedure the average distance of all the source states from destination states are computed.
PUBLISHED PAPERS IN INTERNATIONAL JOURNALS