CHAPTER 1
INTRODUCTION
CHAPTER 1
INTRODUCTION

1.1 Motivation

Many people were admitted to hospitals every year due to some contagious diseases because of climatic changes. In order to find out the remedies of these situations, the inhabitants in those areas approach doctors and other experts to seek their advices regarding the treatment and suggestions for the diseases incident on them and suggestions for their control. Every disease has its own symptoms. Many organisms like virus, bacteria and fungi etc. cause different diseases. Sudden climatic changes may also be another reason for the disease hit.

Spatial analysis and Geo-Spatial Information System (GIS) research areas are developed absolutely independent to each other. The volumes of doing research on spatial analysis preceding to the evolution and popularization of GIS technology clearly express that spatial analysis, and more generally, geographical analysis, can be undertaken without the aid of a GIS. It is equally clear that GIS have proliferated as display devices without claiming any spatial analytical capabilities other than spatial overlay and perhaps some basic descriptive statistics related to area and distance computations.

Despite their independent histories, there are several advantages to integrating spatial analytical capabilities within a GIS framework. From a GIS point of view, there is a rising demand for systems that "do something" other than show maps and organize data. It would be useful to some users to have the capability of analyzing data once they have been displayed. For instance, it might be useful to know something about the statistical relationship between the spatial distribution of welfare recipients and the distribution of welfare offices after displaying the two data sets on one map. Similarly, it would be useful to have some knowledge of the statistical impacts of changing the units for which data are display. From the spatial analytical point of view, there are some advantages to linking statistical methods to the database and exhibit capacities of a GIS. In this way the GIS can act as an "enabling technology" that is, whilst the GIS may not be necessary for the analysis of spatial data, it can facilitate such analysis. It is also possible that the display and data organizational capabilities of the GIS could even produce insights into the analysis
that would otherwise be missed. This might be especially the case if increasing access to highly disaggregate data sets is possible.

Spatial analysis deals with two quite distinct types of information. One concerns the attributes of spatial objects, which include measures such as area, population, rainfall, or soil type, etc. The other concerns location information about the spatial objects which are generally described by means of their positions on map or by geographic coordinate systems. The spatial objects concerned in most analyses are polygons, which correspond to measurement zones, statistical reporting areas such as census tracts and school districts, and points, which correspond to sampling points. For some types of spatial analysis it is common to represent polygons by points.

Interoperability within GIS is disturbed with not only data but also operations. It is very essential that, focuses of research should be made on more exclusively on the access to multiple, scattered, heterogeneous and autonomous information resources and on their combination. The special needs of applications concerned with geospatial data, like:

- The vast amounts of data considered.
- The complex and greatly structured data is existence.
- The co-existence of numerous dissimilar geographic formats.
- The increasing tendency towards reprocess of geographic data all around the world.
- The spread nature of geographic data.

Moreover, temporal aspects play important role in many geospatial applications. For instance, in geo-marketing applications, it is useful to keep many versions of data for analysis over time. Applications such as traffic control or intelligent transportation require fast access to highly dynamic data.

GIS is concerned with many tasks, such as input and store, query and analyze, display and select geographic information. The participating information sources may vary widely, from legacy systems to geospatial system repositories. The three main actions focused on are the following: (1) collecting data from the information sources, (2) organizing data in a geospatial store or keeping track of data organization in different systems and (3) querying data. In the sequel, the process is referred to as 3-step data handling.

Storing the database for such critical applications require building of an efficient data warehouse, which is an eager approach. Data is integrated, non-volatile and time-variant collection of data. It allows the multidimensional storage of data. A
common integrator component extracts data sets from different sources and integrates them before sending them to the warehouse.

The process of collecting data is known as data warehousing. New data is added to the data warehouse from various sources, which generally handle data in different formats. Metadata describes how data to be transformed and how it is integrated. A common integrator component is used, that can extract data sets from different sources and integrates them before sending them to data warehouse. Before data is added to the data warehouse, consistency checks are performed and a mechanism that guarantees all data is loaded exactly once. Query processing typically includes processing various types of queries on the database of data warehouse, ad-hoc queries and regular queries. To answer these typical queries which are very quickly processed, the queries are logged in a query log so that other or later time if they occur or posed the query process need not take much time of processing. What-if analyses are other fold of queries in the data warehouse. These queries really fetch more important scenarios to understand by the geo-spatial user in the GIS.

1.2 Spatial Data

Spatial data is the data associated to objects that inhabit space. Spatial objects stored in spatial database signified by spatial data types as well as spatial relationships between such objects. Spatial data holds topological or/and distance data and are frequently organized by spatial indexing configurations and accessed by methods of spatial access. This situation demands new technologies for knowledge discovery in large spatial databases, or spatial data mining, that is, extraction of implicit facts, spatial relations, or other moulds not explicitly stocked in spatial databases. The effectiveness of spatial data mining algorithms solves critical defy to spatial data mining owing to the enormous amount of spatial data and the density of spatial data type and the of methods spatial access. Geographic data contain spatial objects with non-spatial metaphors of these objects. Non-spatial description can be hoards in a conventional relational database here one attribute is refer to spatial explanation of the object. Spatial data can be expressed by using two different properties, topological and geometric. Hence, the methods for ascertain knowledge can be determined on the non-spatial or/and spatial objects’ spatial properties.

There are a few key words for the learning of spatial data mining. A spatial characteristic rule describes the general characteristics of spatial data. A spatial discriminate rule describes the features discriminating a class of spatial data from
other classes. A spatial association rule specifies the implication of one set of characteristics by another set of characteristics in spatial databases. Thematic maps show the spatial distribution of certain attributes in the database.

1.3 Knowledge Discovery

Machine learning techniques are learning from examples, generalization and specialization. Algorithms of learning from examples cannot be directly espoused for huge spatial databases since the algorithms are exponential in the number of examples and they do not manage noise and incoherent data very well. Han. et al [1][2][3][4][5][6] customized these practices and gave an attribute-oriented (as contrasting to the machine learning algorithms of tuple oriented) induction algorithm to extract knowledge from huge relational databases and Lu et al [7][8] extended this method to spatial databases. So, the hypotheses for relational databases must be conceded to spatial data mining. The generalization-based information discovery necessitates background knowledge of concept hierarchies, explicitly given by the experts or generated automatically. For example, regions representing counties can be merged to provinces and provinces can be merged to larger regions. Attribute-climbing the simplification hierarchies and briefing the general associations between spatial and non-spatial data at superior concept levels perform oriented induction. Lu et al [7][8] expressed two generalization-based algorithms, non-spatial-data-dominant and spatial-data-dominant generalizations.

1.4 Our Contribution

Data collected from 4 different areas, 4 different villages in each area totally 16 places and also Heart attack, Cholera, Typhoid, Dengue disease are considered for prediction. The data also collected from hospitals, village offices, district health centers and personal interviews with the people. Data synchronization was also done.

The multi dimensional data set is applied to existing data mining prediction algorithms and the comparative study was made among the above algorithms. The prediction algorithms considered for this analysis are the chance of disease hit identification using the classification rule, the disease prediction with clustering algorithm, disease prediction using ID3 algorithm of decision tree, identify the correlation attributes and predict the disease using Apriori concept in association rules.

Validation of these algorithms was also done by using Precision, Recall, F-measure, Error Rate, and Accuracy Rate. The accuracy rate got between 75- 85%.
These validations compared with SPSS software and Weka tools. For getting more accuracy a new hybrid algorithm proposed.

A new specialized hybrid data mining algorithm was implemented using PHP Language to predict the disease incidence. The hybrid algorithm is the combination of classification, clustering, decision tree and association rules. The new proposed algorithm got 93% accuracy for prediction of diseases.

“Geographic data contain spatial objects and non-spatial descriptions of these objects. Non-spatial description can be hoard in a conventional relational database here one attribute is a indicator to spatial explanation of the object. Spatial data can be illustrated using two different properties topological and geometric”.

The spatial data is the definition of the structure of the spatial knowledge. The spatial knowledge consists of the relevant attributes of the spatial data with their degree of intensity. The degree of intensity of the attributes is identified as fuzzy set of values. A fuzzy set indemnification is made to the features of the spatial objects that participate in the spatial knowledge. The more relevant spatial knowledge is the collocation of the spatial objects. The spatial knowledge is interchangeably called as collocation in this work.

The collocation is a set of co-located spatial objects with regard to their common features. The similarity in the set of collocations is identified to be as the pattern form, and thus a collocation is a pattern. The collocation identified as pattern can describe the basic nature of the spread over spatial stretch in the spatial domain. Thus collocation pattern happens to be the spatial knowledge.

The problem of the work basically two fold, one is the creation of proper storage structure to store the collocation and, two, is the temporal idea of collocation. The collocation or collocation pattern is not a constant for a given specific spatial domain, it varies based on various aspects of the spatial characteristics, with regard to these changes, the collocation is said to be changed. The changes occur in the collocation are observed temporally. The changes occurred to the collocation are size; due to change of number of features available in the collocation pattern, change of shape due to the change of the intensity of the features in the collocation. The intensities of the features are identified to be as values of a fuzzy set.

The major aim in the contribution is to pervade the feasibility of integrating several GIS, with simple way of representing more crucial information that is viable for performing data manipulation in an orderly manner.
1.5 Overview of Thesis

The chapter 2 discussed about Epidemics, chronic diseases that are the most important social disasters track strategic-virulent disasters that influence the ecosystem of a spatial zone. So many parameters influence the spread of epidemic; difficult task is to find preventive measures. A probabilistic learning is made on the fitness demographic data, to find out the symptoms of the sick people. A collocation rule is defined as just a syntactic demonstration of the factors in the form of antecedent and consequent. The participation index can boost or reduce the status of the collocation rule and there by its parameters that are used to form further.

The chapter 3 discussed about 10911 samples are collected from 4 different areas (Tribal, Hill, Rural, Urban), and 4 different villages in each area totally 16 places are selected, 361 people are effected by dengue disaster. This data collected from health department, Hospitals, Urban Local Body, inhabitants from the above areas, Doctors from various hospitals, health officers, different records from urban local body, village offices, district health centers, on site observation and personal interviews with the people etc. the data was preprocessed and made the records as 10605 and 344. Data preparation is an essential issue for both data mining and data warehousing, as real-world data have a tendency to be imperfect, noisy, and incoherent. Data integration data cleaning, data transformation, and data reduction are the data preparation methods. Data cleaning programs are used to fill in omitted values, soft noisy data, recognize outliers, and correct data discrepancies. Data integration unites data from multiples resources to form a consistent data store. Metadata, association analysis, data conflict exposure, and the declaration of semantic heterogeneity work towards smooth data combination. Data transformation programs made the data into suitable forms for mining. For example, characteristic data may be normalizing so as to fall between small ranges, like 0 to 1.0. Data diminution techniques such as data dimension reduction, numerosity reduction, cube aggregation, discretization and data compression can be used to attain a reduced version of the data.

The chapter 4 discussed about the proposed new hybrid algorithm is unique from the commonly used prediction algorithms in data mining. The proposed methods overcome the disadvantages of existing methods as the number of frequent items is less. The new algorithm proved to be efficient in terms of time and space complexity and proves to be accurate when compared with a standard statistical analysis tool such
as SPSS the prediction can also reveal distinct group of patients having a common illness with various treatment plans. It can also give common causative factors and other decenties.

The outliers as well as clusters can also be detected graphically using this algorithm and all extreme outliers can be further analyzed using root cause analysis if an assignable cause can be found. The algorithm has been implemented successfully by developing software using the programming language PHP. The results are tested and verified. The proposed new method is unique and different from the ancient and traditional combined methods.

The chapter 5 discussed about the fundamental function of software structural design in self-adaptive schemes and outlines methods are considered for supporting the methodology. A collocation rule with weaker parameters may also exist in the spatial zone, but cannot play important role for the detection of the epidemic. Framework was explained for the application of collocation rules i.e., spatial acquaintance by the health campaign. Spatial knowledge is extracted by applying one of the data-mining concepts called Collocation rule. This Collocation rule gives us the Spatial Knowledge. Such Spatial Knowledge is mined for epidemic as an example. EDMS Architecture was proposed to help the disaster management team. A novel self adaptive disaster management system architecture was also proposed when system has to adapt different disaster components. This architecture can be extended to detect all disaster types in future. The data structure described in this chapter is that fits all the requisites to store the spatial knowledge or collocation rule. Towards the application and integration of GIS, the feasible implementation of collocation rule that forms as a basic element of knowledge base of GIS is a more important concern. The structure has exceptional importance for the application relevant tasks. Particularly the idea of intensification of the meaning “the feature with fuzzy sets” and its representation has drawn more relevance. The representation of the structure is complex, and its various subscribed parts which have achieved not yet been applied by theorists is given with a proof of complexity and its performance evaluation.

The chapter 6 discussed about the probability notations are assumptions for the phenomena of change of collocation. This chapter provokes study of the nature of collocation with only notational significance in probability. The representation of collocation and its allied parts are more probable in a spatial domain. The probability of occurrence of a collocation and its different varying natures are identified and
denoted with several levels of probabilistic equations. Baye’s rule helps in sorting out the indispensable features. The result achieved in this chapter throws light on probabilistic idea that collocations change in their size and shape.

Epidemics, chronic syndromes which are the most important social disasters go after strategic-virulent disasters that are made on the health demographic data. Using the Collocation rule, the affected area of dengue is found and results are obtained. Using the Participation index the symptoms that influence spreading the dengue were found and results are obtained. i.e. the health campaign of spatial knowledge. Identifying dengue effected area using spatial maps as spatial knowledge i.e. to identify spreading of dengue was proposed and proved. Dengue affected Areas were shown on Map. Probabilistic study on Dengue affected Areas of a spatial zone.

The chapter 7 discussed about the inhabitants with the significant characteristics form a classification model for disease prediction. The clustering model allows us to understand different group behavior for the disease hit. The knowledge extracted from the clustering model helped to identify the significant characteristics of inhabitants who formed a particular cluster. This model also allows predicting the diseases well in advance so that the appropriate remedial measures could be taken against the disease hit.

The chapter 8 discussed about the different strong association rules are generated with the data set by applying Apriori algorithm. It reveals that there are certain associations among various parameters in the database for the prediction of a contagious disease. An unsupervised model for identifying the significant characteristics of insolvent customers and are supervised classification model for insolvency prediction was created. The knowledge extracted from that model helped to identify the significant characteristics of insolvent inhabitants which formed a particular group. The supervised classification model was built on a data set.

The chapter 9 discussed about each algorithm has its individual strength and limits. The application of each algorithm may vary based on the data set. By considering the strength and disadvantages of the existing algorithms a new hybrid algorithm (EDPI) was proposed and developed in PHP language. The experimental results of Classification, Clustering, Apriori Algorithm, Decision Tree Models and Hybrid Algorithm are validated with the matrices (Precision, Recall, F-measure, Accuracy and Error rates). The newly developed Hybrid Algorithm proved to be accurate when compared with a standard statistical tool.