CHAPTER-2

SPATIAL DATA MINING COLLOCATION TECHNIQUE FOR DENGUE DISASTER
CHAPTER 2
LITERATURE SURVEY

2.1 Introduction

Geography is a combinatorial discipline and geographic information under investigation often extent across several domains. The complication of spatial data and geographic exertions, together with inherent spatial relationships, comprise a massive challenge to conservative data mining methods and call both hypothetical research and development of new methods to help in deriving information from great and heterogeneous spatial datasets. For a long time spatial analysis of health data was restricted to the mapping of individual cases or rates of particular diseases and other health related parameters. More of these ‘health’ maps had become accessible as the use of geographical information systems (GIS) in health associated situations improved [9][10] [11][12]. Many literary research works has been taken place [13][14].

Although disease maps may provide clues of the etiology of diseases and may facilitate decisions concerning planning of health systems, sound analytical methods are needed to assess associations between health events and etiological factors that vary gradually over geographical regions. The last two decades saw therefore a vast and still ongoing development of statistical methods for the analysis of spatial data. A formula has been applied as Hazard science towards Risk Science, to understanding the perils and their consequences (threats), following a probabilistic method using spatial data mining [9][10].

Due to bigger heterogeneity of spatial data, the suppliers of geographic data specify various models for similar spatial objects. Context explicit semantics is the best approaches suggested which contracts with provision of characteristic space derivations. An Ontological analysis need to be done on the fundamentals of the domain space.

A feature space consists of all input data objects, each of which is typically described by many variables (some of which, in a spatial dataset, may represent geographic characteristics and relationships). Unknown and surprising patterns, drifts or relationships can conceal deep in such a huge feature space and make it extremely hard for logical methods or ocular approaches to find.
A theoretical space is produced by all probable configurations of the outfits used to notice patterns in a characteristic space. Characteristically, however, the theoretical space for a big and huge dimensional geographic dataset had an excessive degree of complexity. Various factors caused this. First, every pattern may engage a dissimilar subset of variables from the actual data, and the amount of such subsets (henceforth subspaces), i.e., probable combinations of elements, is vast. Second, within a subspace, latent patterns can be of different forms (e.g., clusters can be several shapes). Third, for a precise pattern form (e.g., cluster of a precise shape), its stricture space is still enormous, i.e., there are various ways to organize its parameters. Fourth, patterns can differ over geographic space, i.e., patterns can be dissimilar from one region to another.

The richness of attributes (variables, or dimensions) in a data set can provide both opportunities and challenges for data analysis. On one hand, the availability of many attributes within the data enables the identification of complex (and preferably unexpected) patterns (e.g., multivariate relationships across domains). On the other hand, it is inevitable that irrelevant attributes exist in the data and the result can be misleading or useless if the analysis method is unable to discriminate relevant and irrelevant attributes.

2.2 Review

Spatial Data Mining

Spatial database systems are database systems for the management of spatial data. Spatial data mining is extraction of knowledge from large databases. Spatial data mining turn into more attractive and significant as more spatial data have been built up in spatial databases. Some of the important aspects in spatial data mining are finding topological-relations, metric-relations, and direction-relations. Spatial Association Rules, Spatial clustering, spatial trend detection and spatial classification are the techniques used on spatial data [3][4]. Mining spatial collocation patterns is an important spatial data-mining task with broad applications. Knowledge discovery in databases is the non-trivial extraction of contained, formerly unknown, and potentially helpful data from databases.

Approaches for discovering collocation rules in the literature can be categorized in to two classes, namely, spatial statistics-based approaches use measures of spatial correlation to characterize the relationship between different types of spatial features. Measures of spatial correlation include the cross-K function with
Mote Carlo simulation; mean nearest- neighbour distance, and spatial regression models. Computing spatial correlation measures for all possible collocation patterns can be computationally expressive due to the exponential number of candidate subsets given a large collection of spatial Boolean features. Data mining approaches can be further divided into a clustering-based map overlay approach and association rule-based approaches.

A clustering based map overlay approach treats every spatial attribute as a map layer and considers spatial clusters (regions) of point–data in each layer as candidates for mining associations. The sets of layers are taken as X and Y, a clustered spatial association rule is defined as X→Y (CS, CC%), for X∩Y = φ, where CS is the clustered support, defined as the ratio of the area of the cluster (section) that assures both X and Y to the whole area of the study section S and CC% is the confidence of cluster, which can be construed as CC% of areas of clusters of X interconnect with areas of clusters (sections) of Y.

Association rule-based approaches can be classified into transaction based methods and distance-based methods. Transaction-based approaches points on significant transactions on space so that Apriori-like algorithms can be used. A distance–based approach defines K-neighbouring class sets. The number of instances for each pattern is referred as the prevalence measure.

Spatial Statistics

Using spatial information measures, enthusiastic methods such as cross k-functions with Monte Carlo replications have been developed to check the two spatial features’ collocation. Some other economic method for finding and evaluation of the collocation is to arbitrarily partition the space into a lattice, counting the number of instances for each spatial feature that are related to each cell of lattice.

Extracting Patterns by Generalization

At the outset the revises include, the spatial mining predicament of how to haul out a special type of closeness relationship namely by distinctive two clusters of faces based on the types of their adjoining features is another study [15][16][17][18][19][20]. Classes of features are prepared into concept hierarchies. Further, the issues of whose discriminators are “better” than other by initiating the notion of maximal discriminators, and also using a ranking system to quantitatively evaluate maximal discriminators from different perception hierarchies [21].
Clustering Techniques Support

A sensible and rather popular method to spatial mining is the use of clustering procedures to analyze the spatial allocation of data. While such procedures are effectual and resourceful in recognizing spatial clusters, they do not support further analysis and detection of the properties of the groups.

Mining Collocation Patterns

Collocation patterns represent subsets of Boolean spatial characters whose instances are frequently located in close geographic propinquity. Boolean spatial features describe the presence or absence of geographic object types at different locations in a two dimensional or three dimensional metric spaces, such as the surface of the earth. Examples of Boolean spatial features include plant species, animal species, road types, cancers, crime, and business types.

Mining collocation patterns give the benchmark of observing the general characteristics of a given spatial region with more pertinent Boolean features with the \( s\% \) (support) and \( c \) (confidence) \[17][18\]. The effort of mining collocation patterns into spatial information approaches and combinatorial methods \[1][2\]. The spatial collocation pattern-mining structure presented in the previous works has bias on admired events. It may overlook some greatly confident but “infrequent” co-location rules by utilizing only “support” based clipping.

In a spatial database \( S \), let \( F = \{f_1,\ldots,f_k\} \) be a group of Boolean spatial characteristics. Let \( I = \{i_1,\ldots,i_n\} \) be a group of \( n \) occurrences in the spatial database \( S \), where each occurrence is a vector containing of \{instance-id, location, spatial features\}. \( \succ \) Neighbourhood relation \( R \) over pair wise places in \( S \) exists \( \succ \) is assumed \[22\]. Collocation rules are models to infer the presence of spatial features in the neighbourhood of instances of other spatial features. For example, “Nile crocodile’s \( \rightarrow \) Egyptian plover” predicates the presence of Egyptian plover birds in areas with Nile crocodiles. The aim of this collocation rule mining is to discover rules in the outline of \( a \rightarrow b \), where \( a, b \) are subsets of spatial features, ‘\( a \)’ describes the group of spatial features that outline the precursor part of the rule and \( b \) describes the action and its significant parts the support and the confidence.

The rule specifies that the chance of the spatial collocation rule soaks up the deed of the rule in the “nearby” places of the spatial items that obey with the collocation rule. To capture the concept of the predicate “nearby”, the concept of
neighbour-set \( L \) is a set of instances such that all pair-wise locations in \( L \) are neighbours. Neighbourhood relation \( R \) may be defined based on Euclidean distance and neighbouring instances are linked by edges. A collocation pattern \( C \) is a group of spatial features, explicitly, \( C \subseteq F \). A neighbour-set \( L \) is called as a row occurrence of collocation pattern \( C \) if every aspect in \( C \) emerges in an occurrence of \( L \), and there exists no suitable subset of \( L \) does so. To indicate all row occurrences of a collocation pattern \( C \) as \( rowset(C) \). In other words, \( rowset(C) \) is the group of neighbour-sets where spatial characteristics in \( C \) are collocated.

The conditional probability is the probability to a neighbour-set in \( rowset(A) \) is a piece of a neighbour-set in \( rowset(A \cup B) \). Instinctively, the conditional probability \( p \) specifies that, the incidences of the spatial characters in \( A \), the probability to discover the incidence of \( B \) in a close by region is \( p \).

**Estimating Symptoms by applying Collocations Rules**

In the invented figure-2.1 the landscape explains two important spatial grades, sea and lake. The contagion spread is renowned in the figure-2.2 the water in the lake is troubled by the lichens and mosses at the west side of the lake as a lot of the water is immobile and covered by marsh.

The people using the water resources at this zone might be affected by various kinds of fecal blemish in water and food. The water in the sea is infected with the crude oil and base products and the high salts, so the people cannot obtain the water for domestic purposes, the climatic changes are affected by the water contents in the shores of sea. People breading their lives at the shores will have the indirect

![Figure-2.1 Important spatial marks, sea and lake](image)
contamination of fecal material in the water and as well as in the form of moisture in the air.

As the lake water is supplied into the agricultural lands surrounding in the adjacent north-east zones, there may people who are affected indirectly by the virulent characteristics. To find out the most probable symptoms or the causative agents in particular that affect people causing disease-deaths, a probabilistic study can be made on the collected demographic-health data.

As it is not easy to identify the stretch of virulent characteristics from the spatial data, the characteristics are tested in the effected people who are resident of the zone.

2.3 Related Work

A few epidemics that are extended due to general sources like stained water and stained food are shown below. (The table 2.1 clearly describes about the contributing agent, sources, and tanks of the disease.)

![Figure –2.2 the water in the lake is effected by the lichens and mosses](image)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Causative Agent</th>
<th>Infection Sources</th>
<th>Reservoirs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dengue</td>
<td>Arthropod-borne virus</td>
<td>Breed in pools of water</td>
<td>Humans</td>
</tr>
<tr>
<td>Bacillary</td>
<td>Shigella disenteriae (B)</td>
<td>Fecal contamination of food and water</td>
<td>Humans</td>
</tr>
<tr>
<td>Cholera</td>
<td>Vibrio cholerae (B)</td>
<td>Fecal contamination of food and water</td>
<td>Humans</td>
</tr>
<tr>
<td>Giardiasis</td>
<td>Giardia spp. (P)</td>
<td>Fecal contamination of water</td>
<td>Wild mammals</td>
</tr>
<tr>
<td>Paratyphoid</td>
<td>Salmonella paratyphi(B)</td>
<td>Fecal contamination of food and water</td>
<td>Humans</td>
</tr>
</tbody>
</table>
The history of the epidemics has got their own influence in the world history which has taken lots of lives together. The epidemic is the result of an infection caused by any of the microorganisms. Every disease is a race between the microorganisms and the host. The microorganism, following the ineradicable rules of progression, strives to survive and replicate, while the host's resistant system raises a hostile protection designed to find, demolish, and get rid of it. An agent that kills its host rapidly cannot be expected to continue to exist long enough to replicate. Thus too much virulence is not selected for in development. Gels, which can replicate and be conceded from one host to any more, are favoured.

Sometimes these epidemics would spread to a larger extent like a continent may turn into a pandemic. The main concentration is on the world’s most deadly disease dengue which was not only an epidemic but also a pandemic which shook the world with the fear of death. The occurrence of dengue fever has improved radically since the 1960s, with around 50–100 million people infected every year. Early descriptions of the situation date from 1779, and its viral source and the conduction were enlightening in the early on 20th century. Dengue disease has turn into a global epidemic since the World War II and is prevalent in more than 110 countries.

The details are shown below: The impact of the disease dengue on the world. The major dengue pandemics are generally listed as: First: 265–420 AD, Second: 15th to 19th centuries, Third: 1779-1780, Fourth: 1906-1907, Fifth: 1953-1954, Sixth: 1970-1971-1972, Seventh: 1981-1982, and some would argue that now it is in the Eighth: 1991 to the till date. Each deadly disease, save the last, was escorted by many thousands of bereavements. Dengue is prevalent in more than 110 countries. It infects 50 to 100 million people worldwide a year, foremost to half a million hospitalizations, and roughly 12,500–25,000 deaths. Unfortunately, despite modern medical technology, dengue remains an efficient killer.

Course of the Disease

Dengue fever (also called as breakbone fever) is a transferable tropical disease, which is caused by the dengue virus. This virus is primarily passed on by Aedes mosquitoes, particularly A. aegypti. Aedes aegypti prefers to put down its eggs in reproduction water containers, to live in slam propinquity to humans, and to supply on people rather than any other vertebrates. Dengue can be spread via fouled blood products and also through organ donation. The Symptoms for this disease include muscle and joint pains, headache, fever, and a distinctive skin rash that is as similar as
to measles. In a small quantity of cases the disease expands into the dengue hemorrhagic fever cause life-threatening, resulting in bleeding, blood platelets count reduced and leakage of blood plasma, or into dengue shock condition, where occurrence of dangerously low blood pressure. The first trace of a case of possible dengue fever is in the Chinese medical encyclopaedia from the Jin Dynasty (265–420 AD) which referred to a "water poison" connected with flying insects.

**Treatment of acute dengue**

Treatment of sensitive dengue is sympathetic, using either oral or intravenous rehydration for meek or sensible sickness, intravenous fluids and also blood transfusion for more stern cases. There are no precise antiviral drugs for dengue, however sustaining proper fluid balance is essential. Treatment depends on the symptoms, altering from oral rehydration therapy with close follow-up at home, to hospital admittance with supervision of intravenous fluids and/or blood transfusion. Persistent medical processes such as nasogastric intubation, and intramuscular injections and major punctures are shunned, in view of the blood loss risk. Paracetamol (acetaminophen) tablet is used for fever and nervousness while NSAIDs such as aspirin and ibuprofen are shunned as they might make worse the risk of bleeding. Blood transfusion is instigated early in patients presenting with unbalanced very important signs in the face of a declining hematocrit, rather than waiting for the haemoglobin attentiveness to decrease to some prearranged "transfusion trigger" level. Stored packed red blood cells or unmodified collected blood are suggested, while platelets and fresh frozen plasma are regularly not. During the recovery phase intravenous fluids are congested to prevent a state of fluid overload. If fluid excess arises and vital signs are steady, stopping further fluid can be all that is necessary. If a person is exterior of the crucial phase, a loop diuretic such as furosemide can be used to get rid of surplus fluid from the transmission.

**The three phases for course of infection:**

(a) **Febrile:** The febrile phase engages high fever, potentially more than 40°C (104°F), and is associated with extensive pain and a headache, this is usually lasts 2 to 7 vomiting may also happen. A rash occurs in 50–80% of those with signals in the 1st or 2nd day of indications as flushed skin, and later in the path of illness (days 4–7), as a measles-like hives. Some petechiae (small red acne that do not vanish when the skin is pushed, which are caused by conked out down capillaries) can appear at this position, as may some mild blood loss from the mucous membranes of the nose and
mouth. The fever itself is characteristically biphasic in nature, violation and then recurring for 1 or 2 days, even though there is wide variation in how frequently this pattern actually happens.

(b) **Critical:** The disease proceeds to a decisive phase around the time fever conclude and typically last within one to two days. At some point in this phase there can be significant fluid growth in the abdominal cavity and chest due to greater than before capillary permeability and outflow. This directs to reduced blood supply and depletion of fluid from the circulation to vital organs. At some stage in this phase, organ dysfunction and stern bleeding, typically from the gastrointestinal tract, may happen. Shock (dengue shock condition) and also hemorrhage (dengue hemorrhagic fever) happen in less than 5 percent of all dengue cases, however persons who have previously been stained with another serotypes of dengue ("secondary infection") virus are at an amplified risk. This critical phase, while unusual, occurs relatively more usually in children and young adults.

(c) **Recovery:** The recovery phase happens next, with resorption of the seeped out fluid into the bloodstream. This frequently lasts two to three days. The development is often striking, and must be accompanied with harsh itching and a slow heart rate. Another hives may occur with moreover a maculopapular or a vasculitic manifestation, which is followed by peeling of the skin. At some point in this stage, the state occurrence of a fluid overload; if it affects the brain, it can made a reduced level of consciousness or seizures. Emotion of fatigue might last for weeks in adults.

### 2.4 The Law of Total Probability

Even though there are a lot of solutions to avert diseases, finding the exact area to apply the deterrence measure with exact inputs turns into the criterion. By

![Figure - 2.3 Dengue virus virions](image)
applying the Law of Total Probability and Bayes’ Theorem, the probabilities of the symptoms that are difficult to find alone are evaluated, by assuming the related symptoms occurrences, the inference is made whether the disease occurs or not. The reverse of conditionality of events is evaluated by the Bayes’ theorem, where the indications and the causative-agents were analyzed and find with a reciprocal equivalence. Table-2.1 shows the most possible indications that cause the contagions.

Let us assume, with a priori probability, a person chosen at random has certain illness that will affect the society (leads to epidemic), given by \( P(I) = 0.001 \). Through the information that the person tested positive for the illness, and the reliability of the test, known to be \( P(Z/I) = 0.92 \) and \( P(Z/I') = 0.04 \), according to the Bayes’ theorem of posterior probability the person was sick with illness. The information that someone had a positive response to the test might be considered as our data to make the collocation pattern. The mining of collocation rule lay in two aspects. Firstly, how to identify and measure confident spatial collocation rules? Secondly, how to mine the patterns efficiently?

The conditional probability of the collocation is the possibility that a neighbour-set describing the characteristics of survival of causative agent, contagion sources, is a piece of the overall neighbour-set in the spatial domain for this endemic application Given a spatial domain in a database outlook \( S \), to measure the inference strength of a spatial characteristic in a collocation pattern, a participation proportion \( Pr(C, f) \) has to be described. A characteristic \( f \) has a participation proportion \( Pr(C, f) \) in pattern \( C \) means every time the characteristic \( f \) is observed, with possibility \( Pr(C, f) \), all other characteristics in \( C \) are also observed in a neighbour-set.

In spatial application domain, as there are no ordinary transactions, for a unbroken space, a participation index is projected to measure the inference strength of a pattern from spatial characteristics in the pattern.

For a collocation pattern \( C \), the participation index \( Pl(C) = \min_{f \in C} \{Pr(C, f)\} \). In other words, wherever a feature in \( C \) is observed, with a possibility of as a minimum \( Pl(C) \), all other characteristics in \( C \) can be observed in a neighbour-set. A high participation index value indicates that the spatial characteristics in a collocation pattern likely prove up together.

2.5 Survey of data management and analysis in disaster situations

The Authors Shu-Ching Chen, Vagelis Hristidis, Tao Li, Yi Deng and Steven Luis published a paper “Survey of data management and analysis in disaster
situations” describes The area of disaster management obtains increasing attention from various disciplines. The computer scientists need to invent better mechanisms to manage and analyze the data created in disaster management conditions. The features identified for the Disaster Management are Prevention, Advance warning, early detention, Analysis of problem and assessment of scope, Notification of the public, Mobilization of response, Containment of damage, relief and medical care for the affected people. In this paper they made an endeavor to survey and manage the current information in the management and investigation of data in climatic conditions. They proposed Disaster management workflow with the disaster related situations like Data integration and ingestion, information retrieval, information extraction, information filtering, data mining and decision support. To efficiently discover information from the huge amount of data collected in disaster management, the pattern recognition, data mining as well as knowledge extraction algorithms need to be efficient and scalable.

The Data types must be unified or scaled in order to allow comparison and combination. The data must be weighted or verified in a quantitative and consistent manner. To utilize the domain knowledge to guide data mining process, improve data mining performance, and make decisions.

2.6 Data and Disaster: The Role of Data in the Financial Crisis

The Authors Louise Francis and Virginia R. Prevosto published a paper “Data and Disaster: The Role of Data in the Financial Crisis” describes the crisis was originally caused by non-payments on subprime loans, helped and assisted by groups of credit derivatives and asset-backed securities, but corporate non-payments, such as like Lehman Brothers, and absolute fraud have also put in to the crisis.

In this paper they explained how data that was obtainable to credit agencies, underwriters, Exchange Commission (SEC) and the Securities, and fund managers might have been used to identify the problems that had lead to financial crisis. They demonstrated that data quality participated a significant role in this mispricing and business intelligence mistakes that caused the disaster, utilized a number of comparatively simple statistics to demonstrate the due diligence that must have, but was not achieved, used the Mad off swindle and the mortgage render down as data excellence case studies, applied simple exploratory procedures to demonstrate simple techniques that might have been used to identify problems. Defined some modeling methods that might have been used to assist underwrite mortgages and discover
indications of fraud. Data quality issues prepared a significant contribution for the international financial crisis. Data Mining Algorithms need to apply for extract the data for the given data base, association rule need to apply for grouping the data and data intelligence to be apply for making of judgment.

2.7 Design Challenges for an Integrated Disaster Management Communication and Information System

The Authors Thomas Luckenbach, Andreas Meissner, Holger Kirchner, Thomas Kirste and Thomas Risse published a paper “Design Challenges for an Integrated Disaster Management Communication and Information System” explained about Disaster response and recovery efforts necessary timely coordination and interaction of public emergency services to save lives and property. There is a marvelous potential for increasing effectiveness and efficiency in surviving with a disaster. In this paper they also draft requirements and inventive technology for an incorporated disaster management communication with an information system, addressing in meticulous network, configuration, data management and scheduling issues throughout the response and also recovery phases. Identified challenges and sketched an integrated communication and information system for disaster response and recovery, addressing in particular networking, service and device configuration, data management, and resource scheduling. Network needs to provide robust communications, integrating heterogeneous networks to allow the rescue effort to proceed smoothly, even in the most difficult communication environments. Security is a foremost concern necessitating solutions for encryption, authentication, data integrity, and non-repudiation. Devices and user interfaces must be tailored to hostile environments and to users who are often not computer literate. The designers find fast-changing environments and resource managements.

2.8 Association rule for classification of Heart-attack patients.

N.Deepika and N.Chandrasekhar through the article”Association rule for classification of Heart-attack patients” presented about the various effective heart attack prediction systems using Pruning-Classification Association Rule (PCAR). A proficient methodology for the generation of association rules from the heart disease warehouses for prediction of heart attack has been represented. Initially, the Pima Indian heart attack data warehouse is useful to get pre-processed in order to make it appropriate for mining process. Once preprocessing gets over, the heart disease data warehouse is binning with the aid of the modified equal width binning interval
approach to discrediting continuous valued attributes. The approximate width of the desired interval is chosen based on the opinion of medical expert and is provided as an input parameter to the model. First, they have converted numeric attributes into categorical form based on above techniques. Accordingly the frequent patterns appropriate to heart disease are excavation with the aid of PCAR algorithm from the extracted data. In addition, the patterns essential to heart attack prediction are chosen on the basis of the computed significant class labels. The classification technique is trained with the selected class labels for the effective prediction of heart attack. Lastly, they have generated the association rules which are useful to identify general associations in the data. The results thus attained have illustrated that the deliberate prediction system is capable of forecasting the heart attack effectively. The paper presents about the various effective heart attack prediction system using PCAR: an Efficient Approach for mining Association Rules. A Proficient methodology for the generation of association rules from the heart disease warehouses for prediction of heart attack has been identified [23].

2.9 Customer compliant management using association rule mining

Behrouz Minaei-Bidgoli, Elham [24] explained a new approach of using data mining tools for customer compliant management using association rule mining. The data of citizens’ complaints on Tehran municipality were analyzed. Using the technique it was possible to find the primary factors those caused complaints in different geographical regions in different seasons of the year. The idea of contrast association rules were also applied to discover the variables that influence complaints occurrence. In order to accomplish this objective, citizens were grouped according to the demographical and cultural characteristics and the contrast association rules were extracted. The results show that there is a strong relationship between citizen gender and education and patterns of complaints occurrence [25].

2.10 The problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction

Sunita Soni, Jyoti Soni, Ujma Ansari analyzed the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction. Their focus was on using different algorithms and combinations of several target attributes for intelligent and effective heart attack prediction using data mining. For predicting heart attack, significantly fifteen attributes were listed. With basic data mining technique other approaches like Time Series, Clustering, Association Rules
and soft computing can also be incorporated. The outcome of predictive data mining technique on the smart dataset exposes that Decision Tree outperforms and another time Bayesian classification is having analogous accuracy as of choice but other prognostic methods like Neural Networks, Classification based on clustering, KNN are not performed fine. The 2nd conclusion is that the correctness of the Bayesian Classification [26] and Decision Tree further improves. This algorithm reduces the actual information size to get the best subset of attribute adequate for heart disease forecasting.

2.11 Short-Term Prediction of Wind Farm Power: A Data Mining Approach

The authors Andrew Kusiak, Haiyang Zheng, and Zhe Song, observed time series models for forecast the power of a wind farm in different time scales, i.e., 10-minutes and hour-long intervals in an article named “Short-Term Prediction of Wind Farm Power: A Data Mining Approach”[27], the time series representations are made with data mining algorithms. Five distinct data mining algorithms have been tested with various wind farm datasets. Out of five algorithms two achieved particularly well. The sustain vector machine regression algorithm executes accurate predictions of wind power and also wind speed at 10-minutes intervals up to 1 hour into the future, while the multilayer perception algorithm is also accurate in forecasting power over one hour-long gaps up to 4 hours ahead. Wind velocity can be forecasted fairly exactly based on its historical values, still, the power cannot be determined accurately given a power curve model and forecasted wind speed. Test computational outcomes of all time series methods and data mining algorithms are conferred. The tests were executed on data produced at a wind farm of 100 turbines. A comprehensive comparative analysis of the four models was reported in the paper.

The multi period ahead of predictions produced by the first three times series models were accurate. The integrated model for power prediction turned out to be less accurate and stable than the time series models.

The time series prediction models accurately predict the wind speed and more importantly the wind farm power at different time scales. The models are applicable to the wind farm and electricity market management and predictive control of individual turbines, both leading to wind power generation optimization.

The ultimate goal of the research presented in this paper was to further improve the accuracy of power prediction of wind farms. This is a fundamental issue in the wind industry. One avenue to be pursued in future research is the transformation.
It is likely that other existing or expanded data mining algorithms will further enhance the power prediction performance. The short-term time series prediction model may become a basis for predictive control aimed at optimizing the wind turbine control setting to maximize the power captured from the wind. One disadvantage of the proposed approach is that the time series model uses its own previously predicted values. As the number of prediction steps increases, the errors get accumulated.

A possible approach for improving prediction accuracy is to build a set of prediction models for each time step. These models would not need their own predicated values as inputs.

2.12 A New Evolutionary Algorithm (EA) for induction of mixed decision tree.

Marek Kretowski. Marek Gizes, Bialystok Technical University, Poland have presented a new evolutionary algorithm for induction of mixed decision tree. In non-terminal nodes of a mixed tree, various types of tests can be positioned, ranging from typical inequality test up to an oblique test based on a splitting hyperactive plane. In contrast to classical top down methods, the proposed system searches for an best possible tree in a global approach, i.e., it learns a structure of tree and finds test in one run of EA. Specialized generic operators are developed, which allow the system to exchange parts of trees, generating new sub trees, pruning existing one and changing the node type and the test. An informed mutation application scheme was introduced and the number of unprofitable modification got reduced.

2.13 Significant relationships between medical factors related to heart disease.

K. Srinivas et.al [28], in their study, briefly examined the possible use of categorization based on data mining techniques. They studied rule based decision tree, Naive Bayes with Artificial Neural Network to huge volumes of health care data. Their study exhibited significant relationships between medical aspects related to heart disease. In this Paper, the authors have presented an intelligent and effective method in predicting heart attack using data mining. Firstly, they have offered an efficient method for the extraction of noteworthy patterns from the heart syndrome data warehouses for the proficient prediction of heart attack. Five mining Goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models. All this models explained complex queries in predicting heart attack.
2.14 Overview of the current research carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis.

Shelly Gupta et al. summarized a variety of assessment and technical articles on the disease breast cancer diagnosis and prediction. In this paper the authors, presented an Overview of the current research conceded using the techniques of data mining to improve the breast cancer diagnosis and prediction. From this study it is pragmatic that the correctness for the diagnosis investigation of various useful data mining classification techniques are extremely acceptable and can assist the medical professionals problem is mainly analyzed under ANNs and its accuracy come advanced in comparison to additional classification techniques applied to the same. But more efficient models could also be provided for prediction problem like by acceding to the best features of distinct models. In both cases we can say that the best model can be obtained for building several different types of models, or by trying different technologies and algorithms.

Summary

Epidemics, chronic diseases which are the foremost social disasters track strategic-virulent disasters that effect the ecosystem of a spatial region. Since so many parameters influence the spread of epidemic, it is a difficult task to find preventive measures. A probabilistic study is prepared on the health demographic data, to find out the symptoms of the sick people. A collocation rule is defined as just a syntactic demonstration of the parameters in the outline 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.