Detecting and Describing Disease Outbreaks Using Bayesian Networks

Project Report submitted in partial fulfillment of the requirement

for the degree of

Master of Technology

in

Computer Science & Engineering

under the Supervision of

Dr.Sakshi Babbar

By

Minakshi Shastri(132216)

Jaypee University of Information and Technology
Waknaghat, Solan – 173234, Himachal Pradesh

May 2015

Certificate

This is to certify that project report entitled **“Detecting And Describing Disease Outbreaks Using Bayesian Networks”**, submitted by **Minakshi Shastri** in partial fulfillment for the award of degree of **Master of Technology** in Computer Science & Engineering to J**aypee University of Information Technology, Waknaghat, Solan** has been carried out under my supervision. This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Dr. Sakshi Babbar

Date:

Assistant Professor(CSE)

Acknowledgements

My truthful thanks goes to my supervisor, **Assistant Professor Dr.Sakshi Babbar**, who supported me through out my study with her patience and lot of knowledge. I appreciate all her contributions of time,ideas that make my M.Tech. experience creative and interesting.It is an extreme honor to work under you.

I take this opportunity to express gratitude to all of the Department faculty members for their help and support.I want to thank my panel members **Dr.Deepak Dahiya, Dr.Rajni Mohana, Dr.Pardeep Kumar, Amit Kumar Singh** and **Suman Saha** for their encouragement, insightful comments and hard questions.

I also thank my parents for the continous encouragement, support and attention.I also place on record, my sense of gratitude to one and all, who directly or indirectly helped me during my project work.

Last but not the least, I would like to thank God and my family: my parents for supporting me spiritually throughout my life.

Date: **Minakshi Shastri**

Abstract

Disease outbreak detection systems for early detection monitor emergency data for irregularities by comparing the distribution of recent data against baseline distribution. In this work, a Bayesian model based on disease outbreak system is used to detect the epidemic. The detection of anthrax epidemic refers to detecting possibility of anthrax attack in minimum number of days. Features related to anthrax disease are taken to develop a probabilistic model using bayesian network. Then by exploring bayesian network we became able to extract the features or patterns that were different to common knowledge captured by bayesian model. We implemented a rule-based technique that compares recent health-care data against data from a baseline data and finds subgroups of the recent data which shows trend. We experimentally proved that this approach give the detection time of less than 72 hours.

CONTENTS

Acknowledgements ii

Contents iv

List of Figures vi

List of Tables vii

Abbreviations viii

1 Introduction 1
- 1.1 Related work . . . 2
- 1.2 Challenges . . . 4
- 1.3 Overview of work . . . 5
  - 1.3.1 P-value computation . . . 7

2 Epidemic in India 9
- 2.1 What is epidemic . . . 9
  - 2.1.1 Types of outbreaks . . . 9
  - 2.1.2 Causes of epidemics . . . 10
- 2.2 Epidemic case study . . . 12
  - 2.2.1 Dengue case study in India year 2012[22],[11],[17]: . . . 12
  - 2.2.2 Dengue outbreak of 2013 worst in six years: . . . 14
  - 2.2.3 Anthrax outbreak causes quarantine of Indian village: . . . 16
  - 2.2.4 Anthrax case detected in Andhra Pradesh: . . . 16
  - 2.2.5 Epidemiology of anthrax in India: . . . 17
- 2.3 Automated system for disease outbreak . . . 19
- 2.4 Why Bayesian Network . . . 20
- 2.5 Contributions: . . . 21
- 2.6 Summary . . . 21

3 Bayesian Network 23
- 3.1 Understanding Bayesian Networks . . . 23
- 3.2 Bayes' Rule . . . 24
- 3.3 Joint probability distribution and conditional independence in BN . . . 27

3.3.1 Joint probability distribution 27
3.3.2 Conditional independence: 28
3.3.3 Marginalization: 29
3.4 Inferences in BN: 29
. . . . 31
3.5 Conditional independence relations 31
. . . . 32
3.6 Properties of Bayesian network 32
3.7 Summary 34

4 Research Methodology 35
4.1 Detection of disease outbreak 35
4.2 Description of patterns 37
4.3 System Design: 37
4.3.1 P-value computation 38
4.4 Summary 40

5 Experimental results 41
5.1 Experimental setup 41
5.1.1 Datasets and Bayesian Networks 42
5.1.1.1 Datasets 43
5.1.1.2 Bayesian Network Model for Anthrax disease outbreak 44
5.1.1.3 Bayesian Network Model for Dengue disease outbreak 50
5.2 Results 52
5.2.1 Detection of epidemic: 52

6 Conclusion and future work 65
6.1 Conclusion 65
6.2 Future work 66

Bibliography 67

LIST OF FIGURES

1.1 Daily counts of health care data 5
1.2 Sales of drugs with day of week 6

2.1 Number of dengue cases in 2013 14
2.2 Dengue cases in past six years 15
2.3 Deaths cases from dengue in past six years 15
2.4 Monthwise trend of anthrax disease 17

3.1 A Simple Bayesian network 24
3.2 A simple Bayesian network showing conditional independence . . . 28
3.3 In figure 3.3.a Earthquake is query variable and JohnCalls is evidence,in 3.3.b JohnCalls is query variable and Burglary is evidence,in fig 3.3.c Burglary is query variable and Alarm,Earthquake are evidence.In fig 3.3.d Alarm is query variables and some effects like JohnCalls are known 31
3.4 Example for d-separation where lack of iodine node d-separates the two sets of node 32

4.1 Regions distribution in area 37
4.2 Before randomization 38
4.3 After randomization 39

5.1 Patient Status Bayesian Network 45
5.2 Probability table for weather 46
5.3 Probability table for food condition 46
5.4 Probability table for flu level 47
5.5 Probability table for action 47
5.6 Probability table for gender 48
5.7 Probability table for reported symptom 48
5.8 Probability table for drug 48
5.9 Probability table for dayofweek 49
5.10 Examples of records in patient data set 49
5.11 Bayesian Network for dengue 51
5.12 Recent data set with anthrax attack 57
5.13 Evaluation of algorithm 64

LIST OF TABLES

1.1 Contigency Table for Fisher exact test 8

2.1 Comparison of results for dengue in2012 14

5.1 Attribute set for patient set dataset 43
5.2 Attribute set for patient set dataset 50
5.3 Attack on Mar-01 with detection time 2 days 58
5.4 Attack on Jun-06 with detection time 4 days 58
5.5 Attack on Jul-12 with detection time 3 days 59
5.6 Attack on Apr-04 with detection time 2 days 59
5.7 Attack on Aug-14 with detection time 2 days 59
5.8 Attack on Jan-14 with detection time 1 day 60
5.9 Attack on Feb-21 with detection time 1 day 61
5.10 Attack on May-3 with detection time 2 days 61
5.11 Attack on Sep-02 with detection time 2 days 62
5.12 Attack on Oct-24 with detection time 1 day 62
5.13 Attack on Dec-15 with detection time 2 days 63
5.14 Attack on Nov-14 with detection time 2 days 63
5.15 Detection time and false positive table 64

ABBREVIATIONS

BN	Bayesian network
DAG	Directed acyclic graph
CPT	Conditional probability table
EM	Expected maximization
WSARE	Whats strange about recent events
DV	Dengue virus
ELISA	Enzyme-linked immunosorbent assay
API	Application program interface
PS	Patient status
ED	Emergency department

CHAPTER 1

INTRODUCTION

Detection of disease outbreaks includes detection of anomalous patterns during monitoring of public health data. For disease outbreak detection anomalies are defined as any observations that are different from the normal behavior of the data. Many traditional anomaly detection techniques look at the data records individually, and try to determine whether each record is anomalous with respect to the historical distribution of data. Here we consider the problem of detecting patterns of anomalies in large, multidimensional datasets. We try to consider the anomalies which may be due to noise and interested in finding patterns which are difficult to capture by traditional techniques. In the disease surveillance task, we wish to detect causes such as epidemics or bioterrorist attacks which give rise to patterns of unusual emergency department records. Our objective is to detect the anomalous records, keeping low false positive rate and avoiding detection of anomalies due to noise which are irrelevant. Using rule based approach technique on existing algorithm we experimentally proved lower false positive and lower detection time.

In case of early detection of disease outbreak, it is difficult to find readily available data which is indicative of useful signal in the outbreak. For a definitive diagnosis of disease like lab reports takes several days to weeks after samples for a particular disease are submitted. During this point of time, the outbreak may have already enlarged into a large scale epidemic. So instead of waiting for definite diagnostic data, there is need of pre diagnosis of data like symptoms of waiting for definite diagnostic data, we can monitor pre-diagnosis data, such as the symptoms exhibited by patients at the public heath care center. Pre-diagnosis of data is known as syndrome surveillance. This may lead to rise in false positive rate due to increase in number of patients in a particular field. For e.g. mistakenly attributing an increase in patients exhibiting respiratory problems to an anthrax attack rather than to influenza. For syndrome surveillance we need patient's multivariate database containing all information such as age, gender, symptoms exhibited, home zip code, work zip code, and time of arrival at the ED.When epidemic occur in particular region then naturally there is increase in alarms in number of patients in emergency department. So this dramatic swing in patients is easily noticed during late stages of an epidemic but task is to detect outbreak during early stages and lessen its effects .We try to detect outbreaks as early as possible that are not present in large scale epidemic like which effect a limited number of population.In the field of disease outbreak detection more work has been done.We are going to be explain work related to this field in next section.

1.1 Related work

Detecting disease outbreaks refer to activation of alarm or signal after epidemic in the region. So it becomes necessary to detect the outbreak at earliest time before it leads to harm at higher level. In paper[1] an algorithm based upon rule based approach was proposed for detection of disease outbreak. Here rule refers to combination of attribute and its value as an example season=winter. Previous approaches were able to identify distinct data points that were rare because of particular combinations of features. But this method was not considered to be good solution in case we try to find patterns as groups with specific characteristics whose recent pattern of sickness is anomalous relative to historical patterns. Using proposed approach anomalous patterns were characterized with a rule. The evaluation of algorithm is done based upon number of false positives and detection time. After using rule based approach Bayesian network is used for anomaly detection in disease outbreaks[2].Bayesian network is combination of probabilistic theory and graph theory. It gives the better understandability of the world and it allows us to make valuable predictions about how the system will behave in future. Using specific features of Bayesian network it become better option for detection of disease outbreaks. An algorithm WSARE [3] was proposed to find the significant anomalous groups of patterns. These patterns were taken as anomalous after comparing baseline data with recent data. Deviation from baseline data in recent dataset leads to suspicious event. Bayesian network become capable of encoding historical patterns as in [18] a Bayesian network to study historical epidemiological databases was proposed. In this Bayesian network 38 nodes were connected to each other refers to different districts of the region.Probablities of occurrences of events for all the nodes based upon its inference with other nodes were calculated. Now it becomes necessary to include spatio temporal components during detection of disease outbreaks. Spatio temporal components refer to spatial and temporal factors responsible for spreading disease as an example how a disease depend upon the region or how it spreads over time. Similarly in [14] a spatio temporal method was proposed to convert non-spatio,non-temporal detection system to sptio-temporal one. The evaluation of performance of proposed system was done and acknowledged as better system. Further a bayesian network for Clinical Diagnosis in Veterinary Medicine was developed. In this issue of modeling deviating knowledge for different subpopulations is captured in a single Bayesian network. In [9] a bayesian approach for detecting both specific and non-specific disease outbreaks is introduced. Results shows that this approach helps in detection of unexpected diseases more than it interferes with detecting known diseases. There should be a function for evaluating performance of detection algorithms in the field of surveillance. Similar work is done in [12] a bayesian network framework for reasoning under uncertainty about the performance of outbreak detection algorithms is used. In the field of surveillance system we can use different data mining models for detection system as in paper [21]. Selection of different dengue data attributes are used for modeling and performance comparison is done based on previous existing models. An improved performance is achieved by proposed model. Results related to better performance depends on the selection of attributes in dengue dataset. The selection of attributes and the formulation of outbreak definition enable the classifier to outperform other models.

Though there is huge advancement in the technologies related to disease outbreak detection but still there are some challenges remaining that are being discussed in next section of challenges.

1.2 Challenges

1. **Highest peak can't be date of epidemic release:** If a graph plotted between day-number and count of health care cases according to available data,then according to standard detection algorithm the highest peak in the graph shown in Figure 1.1 would be the date of the anthrax release. But the anthrax release occurs on day index 74,400, which is clearly not the highest peak either graph. Normally the anthrax releases affect such a limited number of people that it is undetected by different detection algorithms. So the problem is to extract those patterns which are suspicious to us and anomalous as compared to the whole data set to detect disease outbreaks.

2. **Highest peak can be false positive:**If there is a graph between sales of drugs and dates then according to this graph the number of sales of drugs increases as the flu cases increases shown in Figure 1.2, but it doesn't mean that highest peak shown in graph below between sales of drugs and number of flu cases will be the date of attack.There is a possibility of false positive in case if we are going to detect influenza ,we know influenza have same symptoms at start as flu so if we are considering this graph then at highest peak it will give the attack release date which is not always true but can be false positive in actual.It can be possible that disease outbreak occur somewhere in between graph which is not necessary to be highest peak.So

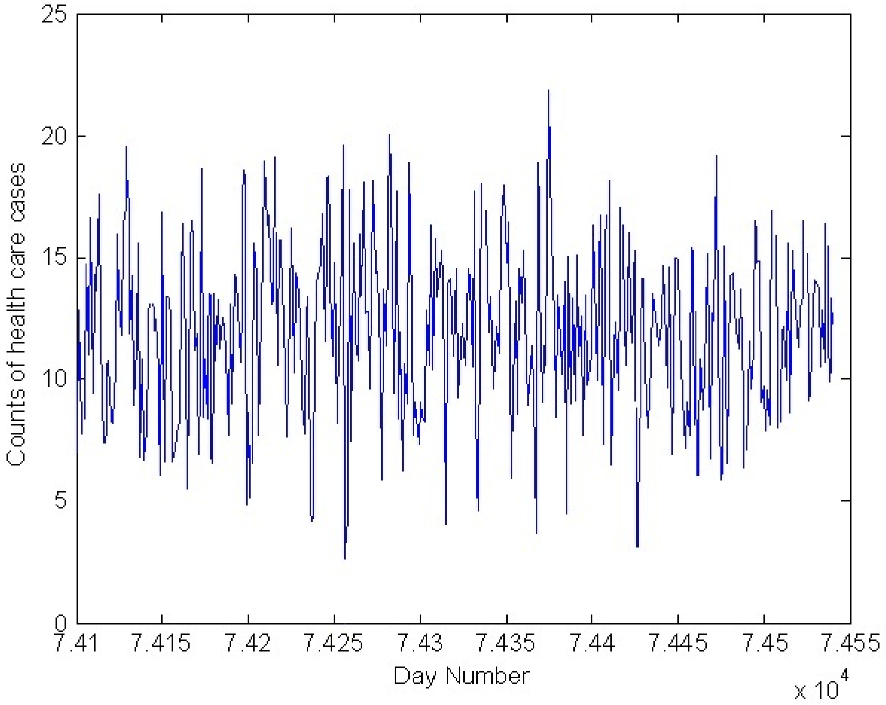


FIGURE 1.1: Daily counts of health care data

model should be able to detect disease outbreak with earliest time and low false positive rate.

A probabilistic model for disease outbreak detection must be needed which enable us to overcome these challenges during the detection process.Next section discussed about the overview of complete work regarding detection of disease outbreak.

1.3 Overview of work

Traditional approach would report an ED case as anomaly if it contains unusual value for some attribute[1]. As an example, a patient over plus hundred years old is considered as unusual record. Traditional approach for detection of disease outbreak working good for single attribute but get fails to identify in case of combinations of attribute which can be unusual together. These techniques fail to

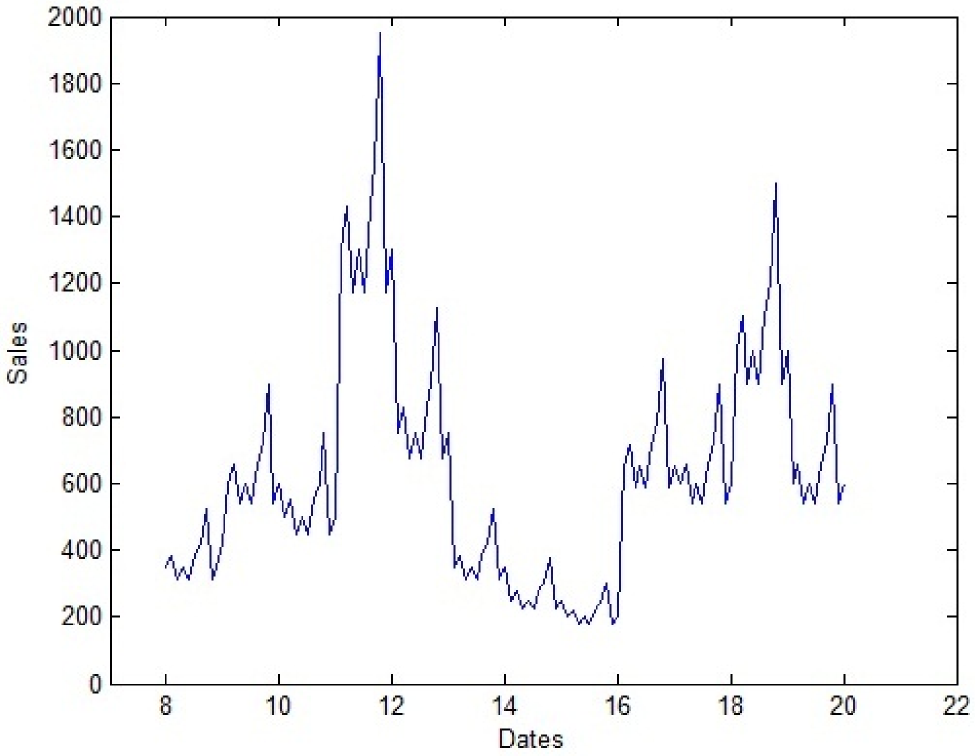


FIGURE 1.2: Sales of drugs with day of week

identify the specific groups in the data which are unusual from normal data. Already there is different detection algorithm in multidimensional space like neural nets, probabilistic models etc. But these algorithms have some limitations also. Anomaly detection for early epidemic detection succeeds at finding anomalies that are unusual based on the underlying population, but these anomalies are treated in isolation from each other. Mainly, we want to find if the recent proportion of a group with particular characteristics is anomalous based on normal proportion. Traditional outlier detection system returns sometimes insignificant groups which are useless for early detection. So to detect anomalous group monitoring of daily counts of single or combinations of attributes are necessary. A training set is used to refer the normal behavior of data.. Then, a threshold would be set based on available data. Whenever the daily counts outgo this threshold, an alarm is generated. This technique gives good results in case of when we have knowledge of monitored features .To capture prior features of the spatial, temporal, and demographic signatures of disease is difficult. So there is need of an algorithm that is able to identify the anomaly instead of pre-defined anomaly. An approach to this problem uses a rule-based anomaly pattern detector in which a anomalous pattern is summarized by a rule, which in our current implementation consists of one or two components. Each component takes the form feature=value. Multiple components uses logical AND operation to combine two components.. For example, a two component rule would be Gender = Male and Season=winter. Benefit of this approach is its understandability. But it has some limitations in case of multiple hypothesis testing due to increase in false positives. Moreover, temporal health care data includes the seasonal variations, so detection system should be able to capture these variations also. We are going to check the deviations of recent data from normal data to identify the specific anomalous groups. Normality of data depends upon the temporal and spatial variations in data. Anomalies are identified as individual data points with a unusual attribute or unusual combination of attributes. Suppose two attributes (Flu level and Action) in Bayesian network model of patient data set for detection of disease outbreak.Flu level attribute having two states namely high and low which can influence the probability of Action attribute in the model having states absent and evisit.There is need to find the anomalous groups From the anomaly detection point-of-view,we may ask which observations are suspicious for this domain? We believe that the observations could be: (1) presence of evisit state of Action attribute when flu level is low, (2) decrease in visiting number of patients in ED(emergency department) when flu level is high.So we try to find such unusual cases which are anomalous relative to typical profile. Thus, in our example of using patient data sets, if there is a dramatic upswing in the number of patients from a particular neighborhood, then an early detection system should raise an alarm.

1.3.1 P-value computation

P-value refers to strength of our hypothesis-value for each possible component rule is calculated using fisher exact test because we are searching for anomalies.P-value is calculated on java platform for simulated dataset.For e.g we construct the

	Match in recent data	**Match in baseline data**
Summer AND Hot	6	496
!(Summer AND Hot)	40	9504

Table 1.1: Contigency Table for Fisher exact test

contingency table including matching and non-matching records,first entry consist of total number of matches for summer AND hot occurrence in the recent dataset.Similarly non-matching for summer AND hot is calculated in both.Null hypothesis is taken as independence of row and column attributes in two contingency table of fisher exact test given in Table 1.1.The hypothesis test measures how different the distribution of recent data to baseline data. This test gives the p-value that decides the significance of rule. So p-value is calculated for each rule to check the strength of null hypothesis. If p-value for particular rule is below threshold level then null hypothesis is rejected and alternate hypothesis is taken under consideration. As an example running hypothesis test for table below we get the p-value for specific two component rule as 0.023.If threshold level is taken as 0.05 then this value indicates that count for recent data for given rule is not independent from count for baseline data for given rule. As we are moving towards epidemic detection firstly we must have knowledge about epidemics,impact of epidemic in India. In the next chapter we will be discussing about the epidemic term,epidemic cases in India and types of outbreaks.We are using bayesian network for the detection of disease outbreak,introducing basic bayesian network in chapter3.Further chapter includes reasearch methodology we have used and experimental results corresponding to bayesian network.In the last chapter6 we have mentioned the conclusion and future work.

CHAPTER 2

EPIDEMIC IN INDIA

2.1 What is epidemic

Affecting large ratio of population at the same time in a region is refers to epidemic, and spreading from person to person in a locality where the disease is not always prevalent[22],[24]. Also epidemic describes a disease that is widespread, disturbing a large number of individuals within a population, in a region at the same time, an epidemic disease can't spread to limited place.

2.1.1 Types of outbreaks

Outbreak is a term used in epidemiology to describe an occurrence of disease greater than would otherwise be expected at a particular time and place. It may affect a localized group or impact upon thousands of people across an entire continent. Outbreaks may also refer to epidemics, which infect a region in a country or a group of countries. Outbreaks can be categorized by the way that the causative agent (usually a microorganism) is spread within the population to reach new susceptible people. For e.g. The place (or object) where people get infected by the microorganism is called 'the source' of the outbreak.

- **Common source outbreaks:** : Infection from the same source for most of the cases lead to kind of outbreak called a common source outbreak.

- **Point Source Outbreaks:** Common source outbreaks where the source has infected cases at one particular region, during a small period of time, are taken as outbreak refers to point source outbreaks. In such conditions the source is said to be set at a single point in time and place. These outbreak have a typical bell shaped epidemic curve that increases suddenly heights and then drop suddenly.Because of this reason, the epidemic curve of such kind of outbreak can help us to identify the moment of attack.
- **Continuing common source outbreaks:** When most of the cases in an outbreak have been infected by the single source, however over a long period of time, then this type of outbreak known as continuing common source outbreak. The shape of the epidemic curve for such kind of outbreaks does not increase sharply or does not reach to peak, yet reaches a flat terrain that is persistent over time, until the source is removed.
- **Propagated outbreaks:** : When an contagious disease is transferred from one person to another then considering a single or common source is responsible for outbreak is not feasible. The causative agent is transmitted within the population through human interaction patterns. The distribution of the epidemic curve in propagated outbreaks can be at variance and depends on the interaction pattern and the proportion of prone individuals. Naturally, when a disease producing agent which could be a virus, bacterium or some other microorganism, pathogen, is injected into a fully naive population with relatively easy communicable transmission, person to person, for example airborne(transmitted by air), the epidemic curve increases sharply with incremental jumps. This raise reflect the generations of cases in the population.

2.1.2 Causes of epidemics

An epidemic develops when new cases of a particular health issue or disease develop in people during a specific period. This condition may not be communicable. If an epidemic spreads throughout the world, it is referred to as pandemic. An epidemic can be caused due to various factors. Following are a list of factors that cause epidemics. The environment and host will be in constant interaction and a disease is caused when there is a disturbance in equilibrium between the environment, agent and host. A disease becomes epidemic when the environmental conditions are not favorable for man and favorable for the disease agent[11]. One might have observed some disasters such as earthquakes, floods, famine and wars are mostly followed by the epidemics of infections. It happens because after the disaster the conditions are favorable for the occurrence of epidemics. There are also other factors supplementing and complementing the factors that are responsible for the outbreak of epidemics.

- **Pre-existent diseases in people:** If the people during a locus area unit already laid low with a disease, the population in this region is additional probably to suffer from epidemics. If there's no pre-existent unwellness in this region, the population within the region is not like to suffer from epidemics when the disasters.

- **Resistance potential of host:** The protection and organic process standing of host population determines the condition of disease. The communicable diseases like infectious disease, diphtheria, respiratory disease and morbilli area unit a lot of inclined in kids with poor nutrition.

- **Temporary settlement of population:** Rehabilitation operations that area unit taken when a disaster area unit typically started within the packed camps quickly. it's tough to supply sanitation, safe potable and lots of different basic amenities at these places. This will increase the incidence of infectious diseases like itch, infectious disease, infectious disease, measles, infectious disease and different skin disorders.

- **Damage to public utility:** If water and sewerage get broken, it will cause contamination in massive scale and afterwards leads to epidemic breakouts. If there's a disruption in health programs, it should result in sickness revival.

2.2 Epidemic case study

2.2.1 Dengue case study in India year 2012[22],[11],[17]:

1. Four serotypes DV-1, DV-2, DV-3, and DV-4 from the family of Flaviviridae contains one more virus referred as Dengue virus.

2. A large scale epidemic of dengue fever happened in year 2012 affecting many districts of West Bengal.

3. The number of dengue cases in 2012 were more than dengue cases in year 2010 and 2011.

4. The age group 11-30 years with male majority were affected mainly during 2012.

5. Outbreak occurred in Aug-Nov time period because of increased vector transmission in the mansoon and post monsoon periods.

6. As the dengue outbreak mainly happened in the months of August to November of 2012, the highest number of cases was reported from the 1st week of September to almost mid-October for both seromarkers of dengue (IgM and the NS1 antigen).

7. Dengue fever is transmitted mainly by Aedes aegypti mosquito and by Ae. Albopictus.

8. Dengue is most common widespreading disease in the world today.

9. With significant morbidity and mortality it is an increasingly widespread tropical arbovirus infection

10. Dengue disease is known to be endemic in India for over two centuries.

11. In case of suspected case of dengue, the presence of ant dengue IgM antibody in patient submits recent infection. IgM antibody is the first immunoglobulin isotype to appear.

12. Anti-dengue IgM detection using enzyme-linked immunosorbent assay (ELISA) machine represents one of the most important advances in medical field and has become one of the tool for routine dengue diagnosis.

13. Mainly, MAC ELISA (IgM antibody capture ELISA)machine is based on detecting dengue-specific IgM.

14. Although outbreaks did not occur, a higher number of cases of suspected dengue infection were reported to referral laboratory in the similar months in 2010 and 2011.

15. The process for the detection of dengue-specific IgM antibodies includes collection of blood from each patient suspected to be suffering from dengue. At least 5 days after onset of fever blood samples were collected from patients. The information related to age and sex of each patient was recorded.

16. Because of the availability of dengue NS1 antigen detection kits supplied by Integrated Disease Surveillance Programmed (IDSP),blood samples were also collected from clinically suspected important dengue fever cases (where fever duration was less than 4 days).

17. For further processing a portion of the dengue NS1-positive sera were well-kept in 80 freezer for further serotyping.

18. Dengue cases were detected more from 13 districts of West Bengal but Kolkata was also affected containing 74.32 % portion for dengue cases.

19. India witnessed widespread dengue fever outbreak in the year 2012.

20. Highest number of cases were recorded in Tamil nadu i.e.9249 followed by West Bangal which reported 6,067 dengue cases.

21. Other states included in dengue epidemic during year 2012 were Maharashtra, Kerala, Karnataka, Odisha, Delhi, Gujarat, Pondicherry, Haryana, and Punjab.

Year	**Total fever cases**	**Total dengue cases**
2010	763	88(11.5%)
2011	1343	328(22.4%)
2012	1518	514(33.8%)

Table 2.1: Comparison of results for dengue in2012

2.2.2 Dengue outbreak of 2013 worst in six years:

1. From year 2008 to oct 2013 total number of dengue cases in India:55063
2. Number of deaths from dengue in year 2008 =80
3. Number of deaths from dengue in year 2012=242
4. Number of deaths from dengue in year 2013=138
5. Reasons: Heavy rains, intense construction activity in cities, lack of surveillance system.
6. Fatality rate in 1996=3.3% for 16517 number of dengue cases.
7. Fatality rate in 2012=0.5%
8. Fatality rate in 2013=0.2%
9. Lower fatality rate means there is improvement in surveillance system more.

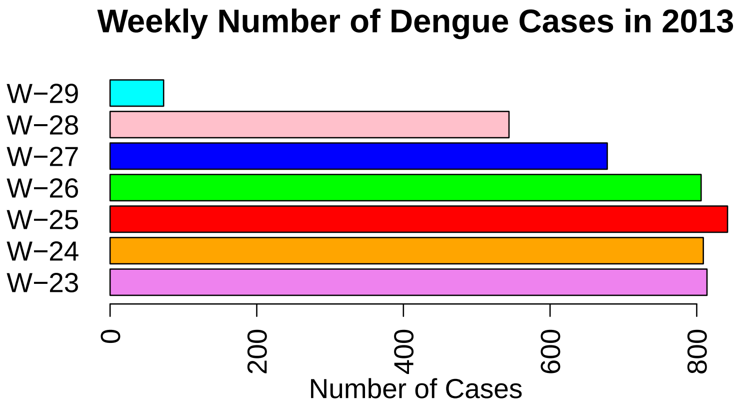


Figure 2.1: Number of dengue cases in 2013

Graph shown in Figure 2.2 is drawn between number of cases weekly cases.X-axis corresponds to number of cases and y-axis represents the week number.A bar chart is drwan based upon the data available of dengue.

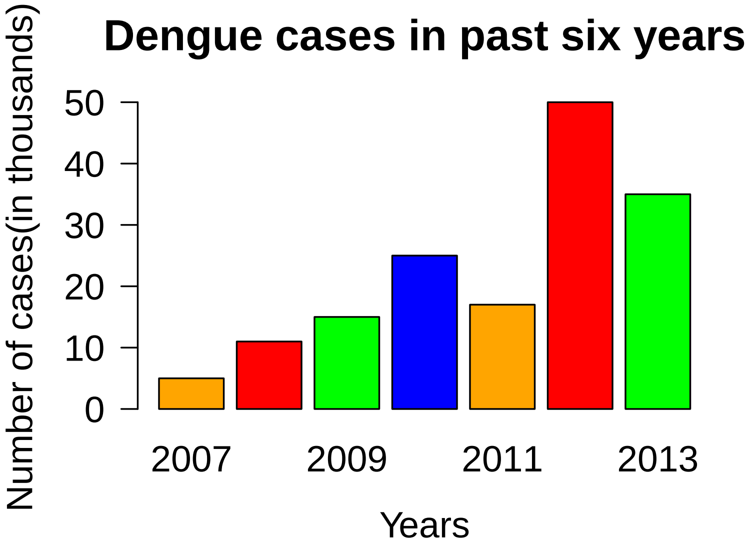


FIGURE 2.2: Dengue cases in past six years

Graph shown in Figure 2.3 is drawn between years and number of cases for dengue.Bar chart is plotted representing different colors for different years.

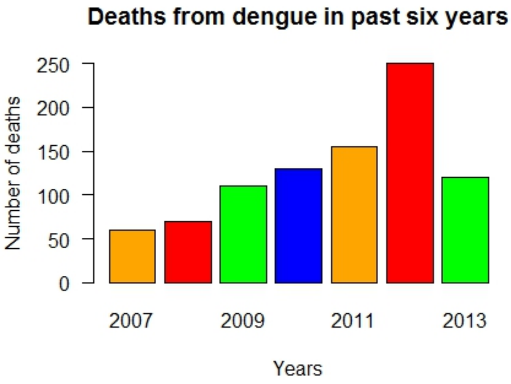


FIGURE 2.3: Deaths cases from dengue in past six years

Above graph shows death rate and number of cases according to available information for dengue disease in past six years.

2.2.3 Anthrax outbreak causes quarantine of Indian village:

1. Health officials in India had recently isolated a village of roughly 30 houses in the Jharkhand's Simdega district because of an alleged Anthrax outbreak.Due to anthrax outbreak in this area seven people were killed.

2. The infection has been traced and found that cow as a source that had spores [32].

3. The one who had eaten or be incontact with cow were infected.

4. Primary infection occurs while handling infected meat [33].

5. After research it has been found that spores are able to survive upto 150 years.These spores are responsible for anthrax disease.

6. Symptoms of Anthrax disease include fever and chills, shortness of breath, nausea and vomiting.

2.2.4 Anthrax case detected in Andhra Pradesh:

1. An anthrax was unearthed in a district of Andhra Pradesh, Anantapur.The anthrax case was found from Chintapalli, a village in outskirts of Karnataka.

2. A class B student caught by a disease and he was admitted in a hospital in Bangalore. The boy was brought to a local hospital namely Hindupur government hospital, there after doctors allegedly denied to provide medical assistance. The boy was kept in isolation in Hindupur government hospital. As anthrax disease is a highly contagious disease so for preventive measures the people surrounding to patient i.e. classmates and family members are also kept under surveillance

3. To find out the source of the infection, district medical and health authorities forced to conduct an epidemiological survey.

4. Anthrax disease is a bacterial disease transmitted to human bodies through animals. It is highly contagious with a high mortality rate

5. In the recent past this was the just another incident of anthrax from the outskirts of AP and Karnataka. And since several citizens and daily goin workers circulate in between two states, it is quite difficult to guard the citizen from suspicious outbreak. Officials in Anantapur are now busy to prevent the spread of anthrax in the area.

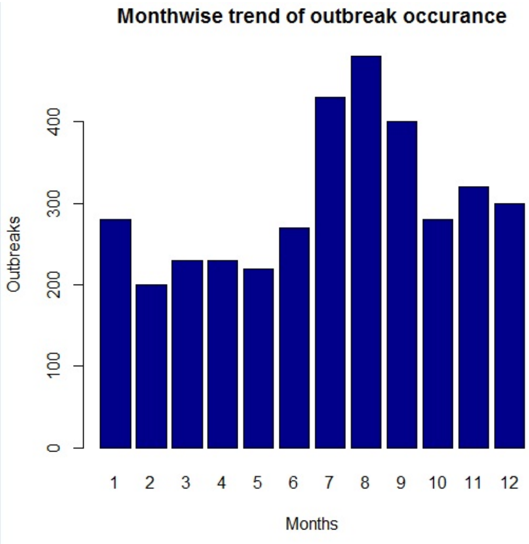


FIGURE 2.4: Monthwise trend of anthrax disease

Graph shown in Figure 2.4 refers to relationship between month and outbreak occurrence.This graph denotes how changing trends for months shows variation with outbreak occurrence.

2.2.5 Epidemiology of anthrax in India:

1. Sensitivity to anthrax varies considerably between different animal species.

2. Higher occurrence was reported in large ruminants(cattle and buffaloes)followed by small ruminants(sheep and goat)

3. Fewer outbreaks have been reported in pigs and elephants.

4. Epidemics have generally been reported between July to September and also November and January, coinciding with the post monsoon months across the country.

5. Outbreaks develop mainly during dry months that follow a prolonged period of rain when the spores are exposed and ruminants have greater access to them.

6. Hot and humid season facilitates the germination of the spores in the environment.

7. Seasonal variation in anthrax occurrence is also seen between different zones of India.

8. Control measures includes vaccination of animals during the latter half of May or early June

9. An important step in the implementation of anthrax control is the acquisition of data or information about the disease.

10. Surveillance can be used to predict areas where natural livestock cases of anthrax are likely to occur.

11. Field veterinarians should have the facility to make on-site diagnosis or good liaison with laboratory services to ensure diagnosis without delay.

12. Vaccines can be deployed strategically in endemic areas and should be administered at least a month prior to the established anthrax outbreak period.

13. Anthrax is more dangerous when it occurs in areas considered to be free and in a typical seasons and climate conditions.

14. Surveillance, vaccination and proper disposal of carcasses are the most efficient ways of preventing and controlling anthrax infection in domestic herds and also limit its transmission to humans.

Next section discusses about the need of automated system in the field of disease outbreak detection.

2.3 Automated system for disease outbreak

Everything in this world is uncertain i.e.we cannot predict what will happen in next few seconds. Therefore, we require a probabilistic model which predicts the future values depending upon the past and present values.We have implemented a probabilistic model to detect the disease outbreak.Already there are some algorithms and techniques which had worked for the same purpose like bayesian network anomaly pattern detection for disease outbreaks[2],it uses the Bayesian Network for construction of baseline distribution by taking the joint distribution of the data and conditioning on attributes that are responsible for the changing trends.Algorithm WSARE 3.0 is used in this work.This algorithm is able to detect outbreak with earliest possible detection time while maintaining a low false positive count.It uses a multivariate approach to improve its timeliness of detection.WSARE based technique was given by Rule based anomaly pattern detection for detecting disease outbreaks[3].WSARE approaches uses bayesian Network to produce baseline distribution that accounts for temporal trends of data.

Epidemic detection using probabilistic model becomes very interesting because of seasonal variations and periodic trends in data.Epidemic within particular area depends upon many factors like population growth,urbanization,seasonal variations,etc.Therefore,role of probabilistic model comes into play which captures seasonal variations, spatial and temporal data sets.Our goal is to propose probabilistic model which will be able to detect epidemic and also having capability of describing epidemic.

Disease outbreak detection is done using bayesian network,but why we move to bayesian network is going to be discuss in next section.

2.4 Why Bayesian Network

BN(Bayesian Network) is a kind of probabilistic graphical model which combines probability theory with graph Theory to compactly represent real world problems.

- Anything can be modeled by a Bayes net.The model can be of any domain.It gives all the possibilities for the existing states in the model. We can model any kind of problem using Bayesian network.As an example if bayes net belongs to health system then a body can be sick or healthy and so on. It returns the states of some part of a specific domain that is being modeled and bayesian network describes how these states are related to each other by probabilities.
- Bayesian network gives the better understandability of the world and it allows us to make valuable predictions about how the world will behave in future.
- Bayesian network is capable for capturing casual relationship among different features in Bayesian model.Using Bayesian network it becomes easy for estimating probabibilies in forward direction as compared to reverse direction like if the patient has lung cancer, what are the chances that their X-ray will be unusual?", rather than in reverse direction, "if the X-ray is unusual, what are the chances of lung cancer being the cause?
- Bayesian network is capable for encoding unusual behavior of existing features in model.
- Bayesian network has the capability of encoding historical patterns to show strong deviation to identify the anomalies.
- Bayesian inference is helpful in case of description of derived anomalies.As an example if we have visible data as symptoms and hidden nodes as disease then Bayesian network is used to infer the posterior probability of each disease given all the symptoms.

- Bayesian networks are adaptable to domain knowledge i.e.we can start modeling from limited knowledge about the domain and grow the model as new knowledge acquired by us.

2.5 Contributions:

In this work we implement anomaly detection based on domain knowledge captured by a bayesian network models specifically for categorical data sets. We summarize our contributions as follows. In order to determine important and unusual outliers the integration of domain knowledge is necessary. The domain specific knowledge here can be collected from historical data. In this work, bayesian networks is used to capture domain knowledge. Bayesian networks provide ability of causal interactions among attributes whose existence is in the domain. Using these causal interactions, we can determine unusual anomalies and also becomes able to provide related information for the learned outlier. Outlier are treated as data points which shows unusual behavior to the normal data.

2.6 Summary

In this chapter we have studied about epidemic,environmental factors responsible for the epidemic occurrence.The environment conditions changes rapidly so if these conditions are not favourable according to human being then an emerging disease can lead to epidemic.A disease affecting limited number of people can also we considered as epidemic if grows rapidly in the population.A complete overview of epidemic is given in this chapter.Section 2.1 presented the types of outbreaks,causes of epidemic(causative agents).Status of epidemic in India since last six years is discussed in section 2.2.Epidemic occurrence due to dengue and anthrax attack at Indian population is discussed in section 2.2.Though with the passage of time there is control on the number of death cases because of epidemic but still there is need of lot of improvement.In this section we have also discussed about the number of cases related to these diseases and reason of occurrence. In the next chapter we will study about bayesian network.Properties of bayesian network that can be applied to detect the epidemic.Modelling of bayesian network for disease outbreak detection system.

CHAPTER 3

BAYESIAN NETWORK

3.1 Understanding Bayesian Networks

BN(Bayesian Network) is a kind of **probabilistic graphical model** which combines **Probability with Graph Theory** to compactly represent real world problems[5].Graphical models are used to describe knowledge of complex problems.They are also called **Belief Network**, **Bayesian Belief Network**, and **Casual Probabilistic Network**.Probability gives advantage of dealing with '**uncertainty**'.Here uncertainty means anything which is not certain.In many problem domains, it is not always possible to create complete consistent model of world.Thus to achieve relevancy,it is necessary not only to deal with what is possible,but also what is probable.Bayesian network is a kind of directed acyclic graphs(DAGs)in which the edges in the graph have direction and there is no cycle within the graph.They enable an effective representation and computation of the joint probability distribution over a set of random variables.

The structure of a DAG is defined by two sets: the set of nodes and the set of directed edges.The nodes represent random variables and are drawn as circles labelled by the variable names[5]. The edges represent direct dependencies among the variables and are represented by arrows between nodes.Bayesian network provides graphical representation of probabilistic relationships among the set of random variables.Bayesian networks are graphical models of casual interactions among

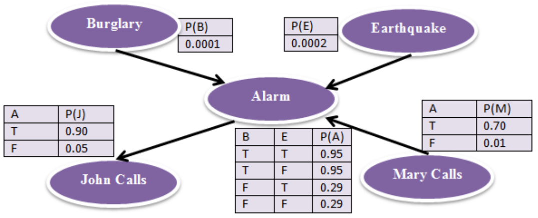


FIGURE 3.1: A Simple Bayesian network

the set of nodes present in the network, these nodes are considered as variables or we can say vertices of a graph and the interactions as directed links denoted by links and arcs between the nodes.Nodes represent events. Any pair of unconnected or nonadjacent nodes of such a graph indicates conditional independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Hence, probabilistic net- works capture a set of dependence and independence properties associated with the variables represented in the bayesian network.

Example: Consider a simple Bayesian Network as shown in Figure 3.1 containing five events burglary,earthquake,alarm,johncalls and marycalls.There exist casual relationship among two events.Conditional probability table(CPT) attached with each node.CPT is used to determine the joint distribution of collection of variables.Directed path from one event to other shows cause-to-effect relationship.In Figure 3.1 there is a directed path from earthquake to alarm,which indicates that earthquake is the cause for occurrence of alarm event.Similarly, for other events also there is existence of casual relationship.Each event or node contains the probability table,number of entries in the table depend upon the parent by which that node infer.

3.2 Bayes' Rule

Bayes'rule is a formula for determining conditional probability named after 18th-century British mathematician Thomas Bayes[5] The formula for bayes'rule given in equation below provides a way to study existing predictions or theories given new or ad- ditional evidence.It is a statistical principle for combining prior knowledge of the classes with new evidence gathered from data.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{3.1}$$

It is given in Equation[1] that we have to find X event in the presence of evidence Y . Specifically, posterior belief P(X | Y) is calculated by multiplying our prior belief P(X) by the likelihood P(Y | X).Where P(Y | X) represents that Y will occur if X is true. Direct computation of P(X | Y) turns out to be difficult.Thus advantages of Bayes'rule comes into existence as it enables us to compute P(X | Y) in terms of P(Y | X).

For example, suppose that we are interested in diagnosing patients with flu cases who visited ED(emergency department). Let X and Y represent the events that person has flu and event of season respectively. On the basis of past data, let we have prior probability of events P(X) = 0.1 and P(Y) = 0.5. Suppose, We are interested in knowing probability of person suffering from flu given season is winter, i.e., P(X | Y). It becomes difficult to solve this statement directly. However, we are likely to know P(Y | X) from records specifying the proportion of flue cases among those diagnosed. Suppose P(Y | X) = 0.8. Based on this information, we can compute our statement using Bayes' rule,Equation[1]. Thus,in the presence of evidence that the season is winter we revise our prior probability from 0.1 to a posterior probability of 0.16 as calculated below.

$$P(X|Y) = \frac{0.8 \times 0.1}{0.5} \tag{3.2}$$

$$P(X|Y) = 0.16 \tag{3.3}$$

Let us consider another example. Suppose that Bob can decide to go to work by one of three modes of transportation,car, bus, or commuter train. Because of high traffic, if he decides to go by car, there is a 50 % chance he will be late. If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20 %. The commuter train is almost never late, with a probability of only 1 %, but is more expensive than the bus.

- Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car. Since he does not know which mode of transportation Bob usually uses, he gives a prior probability of 1/3 to each of the three possibilities. What is the boss' estimate of the probability that Bob drove to work?

We have the following information given in the problem:

$$P(Bus) = P(Car) = P(Train) = \frac{1}{3}$$

$$P(Late|Car) = 0.50$$

$$P(Late|Train) = 0.01$$

$$P(Late|Bus) = 0.20$$

$$P(Car|late) = \frac{P(Late|Car)P(Car)}{P(Late|Car)P(Car)+P(Late|Bus)P(Bus)+P(Late|Train)P(Train)}$$

$$P(Car|late) = \frac{0.5\times\frac{1}{3}}{0.5\times\frac{1}{3}+0.2\times\frac{1}{3}+0.01\times\frac{1}{3}}$$

$$P(Car|late) = 0.7042$$

3.3 Joint probability distribution and conditional independence in BN

3.3.1 Joint probability distribution

Conditional probabilities like P(X|Y) which are used to determine how much the occurrence of Y event influences the occurrence of X event.Joint probabilities determines the likelihood of two separate events simultaneously. The joint probability for two events X and Y is expressed mathematically as P(X,Y).

$$P(X_1, X_2, X_3, ..., X_n) = P(X_n|X_1......X_{n-1})P(X_{n-1}|X_1....X_{n-2})P(X_2|X_1)P(X_1) \tag{3.4}$$

This equation can be understand with the help of example shown in Figure 3.1.Let we have to calculate the joint probability of all events i.e. P(JC,MC,A,B,E).Where, JC,MC,A,B and E refers to JohnCalls,MaryCalls,Alarm,Burglary and Earthquake respectively.

$$P(JC, MC, A, B, E) = P(JC|MC, A, B, E)$$

$$P(JC, MC, A, B, E) = P(JC|A)P(MC|A, B, E)P(A, B, E)$$

$$P(JC, MC, A, B, E) = P(JC|A)P(MC|A)P(A|B, E)P(B, E)$$

$$P(JC, MC, A, B, E) = P(JC|A)P(MC|A)P(A|B, E)P(B)P(E)$$

$$P(JC, MC, A, B, E) = 0.90 \times 0.70 \times 0.95 \times 0.0001 \times 0.002$$

$$P(JC, MC, A, B, E) = 0.000000119$$

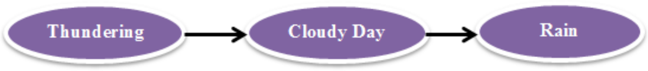


FIGURE 3.2: A simple Bayesian network showing conditional independence

3.3.2 Conditional independence:

Bayesian theory moves around key concepts of dependency, independency and conditional independence among variables. In very general terms, two events are said to be dependent if knowledge of one provides predictive value for other. For example,consider a simple BN in which Rain is dependent on cloudy day . Knowing day is cloudy or not, we can predict possibility of rain. However there are situations when knowledge of one variable provides no predictive value for the knowledge of other.For example, knowledge number of people in a specified area provides no prediction on probability of rain. More formally, independent events can be described using the definition below.

Definition: A variable X is independent of another variable Y with respect to a probability distribution P, if equation given below holds.

$$P(x|y) = P(x); \forall x \epsilon dom(X), \forall y \epsilon dom(Y) \tag{3.5}$$

Let us consider an example of simple Bayesian network Thundering to cloudy day and cloudy day to rain casual relationship as shown in Figure 3.2 .In this case, while it is true that knowledge of thundering provides predictive value for rain because no thundering means it is not necessary to be cloudy day and also not raining.So this predictive value is entirely comes through Cloudy day. If we already know about cloudy day, knowing thundering or not does not help further predict its raining or not. Here, the two variables Rain and Thundering are conditionally independent given knowledge of Cloudy day.

3.3.3 Marginalization:

We might want to compute P(X) from a joint distribution P(X,Y,Z). This is computed by summing over all possible combinations of values Y and Z to solve P(X)as given in Equation[6]. P(X) can be computed from the joint distribution as below.

$$\begin{aligned} P(X,Y,Z) &= \sum_{Y}\sum_{Z} P(X,Y,Z) \\ &= P(X,Y=y_1,Z=z_1)+P(X,Y=y_2,Z=z_1) \\ &+ P(X,Y=y_1,Z=z_2)+P(X,Y=y_2,Z=z_2) \end{aligned}$$

3.4 Inferences in BN:

Inference refers to the conclusion reached on the basis of evidence and reasoning.Inference in Bayesian Networks has four catagories.

1. Diagnostic inference
2. Casual inference
3. Inter-casual inference
4. Mixed inference

Considering the simple Bayesian Network example as shown in Figure 3.1 types of inferences are explained below.

1. **Diagnostic inferences:**Abductive reasoning or Diagnostic inference states the inference from effects to causes as shown in Figure 3.1.Let we have given JohnCalls and query is to find the probability for burglary event infer

by JohnCalls as shown in equation below.As shown in Figure 3.4 JohnCalls is E(evidence)and Burglary is query variables(Q).

$$P(B|JC)$$

2. **Casual inferences:**Deductive reasoning or Casual inference states the inference from causes to effects as shown in Figure 3.1.Let we have given Burglary and query is to find the probability for JohnCalls event infer by burglary as shown in equation below.Let we have given that Burglary,infer that probability of JohnCalls and MaryCalls as represented below.As shown in Figure 3.4 Burglary is evidence(E)and JohnCalls or MaryCalls are query variables(Q).

$$P(JC|B), P(MC|B)$$

3. **Intercasual inferences:** Intercasual inference states the inference between causes of common effect as shown in Figure 3.1.Let we have given both Alarm and Earthquake as evidence and query variable as burglary.Given evidence alarm and earthquake then infer the probability of Burglary as represented below.Here alarm and evidence together are E(evidence)and burglary is query variables(Q) as shown in Figure 3.4.

$$P(B|A \cap E)$$

4. **Mixed inferences:**Mixed inference states the inference in which Some causes and some effects are known as shown in Figure 3.1.Some causes and some effects are known.Given JohnCalls and absence or Earthquake event at same time as shown in Figure 3.3 then infer the probability of alarm as represented

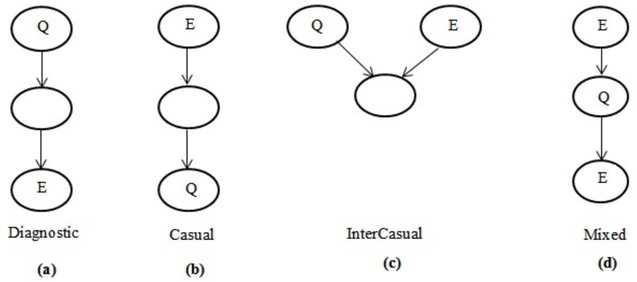


FIGURE 3.3: In figure 3.3.a Earthquake is query variable and JohnCalls is evidence,in 3.3.b JohnCalls is query variable and Burglary is evidence,in fig 3.3.c Burglary is query variable and Alarm,Earthquake are evidence.In fig 3.3.d Alarm is query variables and some effects like JohnCalls are known

below.

$$P(A|JC \cap \neg E)$$

Here Q=query variable. E=Evidence

3.5 Conditional independence relations

To determine whether two variables in a Bayesian network are conditionally independent a simple algorithm is used called d-separation X and Z are d-separated by a set of evidence variables E iff every undirected path from X to Z is "blocked".

1. If every undirected path from the nodes in X and nodes in Y is D-separated by a given set of evidence nodes E then X and Y are conditionally independent given E.

2. A set of node E, D-separates two sets of nodes X and Y if every undirected path from a node in X to a node in Y is blocked given E.

3. A path is blocked given set of nodes E if there is a node Z on the path for which one of the conditions holds.

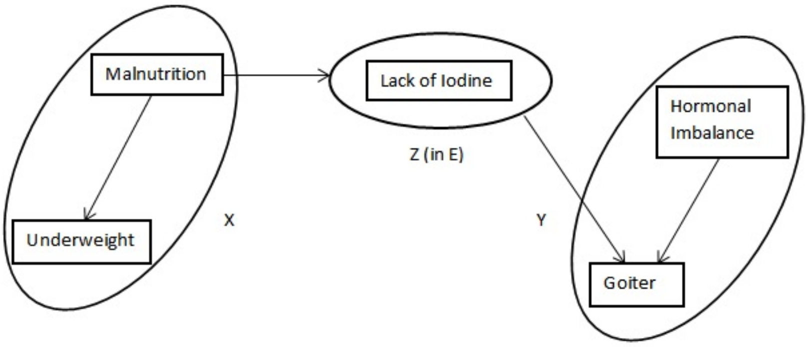


FIGURE 3.4: Example for d-separation where lack of iodine node d-separates the two sets of node

Let us consider a simple Bayesian network Figure 3.3 in which malnutrition leads to lack of iodine,lack of iodine leads to goiter and so on.Now lack of iodine node is very crucial because it d-separates malnutrition and goiter.Because In order to reason whether a person has goiter or not it is sufficient the presence of hormonal imbalance there is no need to look into malnutrition is present and not.So lack of iodine node d-separates these two sets of nodes as shown in Figure 3.4.

3.6 Properties of Bayesian network

1. **Parameter learning:**To fully specify the Bayesian Network (BN) and represent the joint probability distribution, probability distribution for node A conditional upon A's parents is computed. The distribution of A conditional upon its parents can be of any form. These distributions consist of some unknown parameters that can be estimated from data. Sometimes maximum likelihood technique is used for estimation of these parameters. Direct maximization of likelihood is difficult normally when there exit undetected variables.A standard approach to this problem is expected maximization technique which helps to compute expected values of undetected variables conditional on detected data,with maximizing the complete likelihood and taking assumption that expected values are correct computed earlier. Under insignificant regularity conditions this process converges on maximum

likelihood values for parameters.It is well understood that the key to the appearance of posterior probability is to make use of conditional independence and work with factorizations of joint probabilities rather than joint probabilities themselves.

2. **Structure learning:**Bayesian Network (BN) can be specified by expert knowledge or network structure and the parameters taken from local distributions must be learned from data, or both. Learning the graph structure of a BN is based on the distinction between the three possible types of adjacent triplets allowed in a directed acyclic graph (DAG).Optimization based search is an alternative technique used for structure learning of Bayesian network.Posterior probability is taken as common score during the structure learning of training data.To check the quality of bayesian network a score is calculated for that Bayesian network.To improve the score of structure an incremental local search technique is used.Now the R bnlearn application raises several possible algorithms for learning the network structure.

3. **Casuality and Bayesian networks:**Research questions motivating many scientific studies are causal in nature. Causal questions arise in medicine (e.g., treatment of diseases), management (e.g., the effects of management style on problem solving efficiencies), risk management (e.g., causes of risk events), customer satisfaction (e.g., drivers of customer satisfaction), and many other fields. Causal inference is used to measure effects from experimental and observational data.. The framework has two key parts. First, causal effects are viewed as comparisons of potential outcomes, each corresponding to a level of the treatment and each observable, had the treatment taken on the corresponding level with at most one outcome actually observed, the one corresponding to the treatment level realized. Second, the assignment mechanism is explicitly defined as a probability model for how units receive the different treatment levels. In this perspective, a causal inference problem is thus viewed as a problem of missing data, where the assignment mechanism is explicitly modeled as a process for revealing the

observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects.

3.7 Summary

In this chapter we have presented the overview of bayesian netwok.Bayesian network is combination of probabilistic theory and graph theory.It is the network of variables or nodes which are connected to each other by directed links.These nodes can be any random variable whose probability of occurrence changes accordingly.Basic structure of bayesian network is given in section 3.1.Basic bayes'rule for bayesian network is explained in section 3.2.Conditional independence relations between nodes are discussed in section 3.3,section 3.4.Inference is the main property of bayesian network by which we can describe the behaviour of nodes present in bayesian network have been discussed in section 3.4.With the help of example d-separation between nodes is given in section 3.5. Bayesian network shows the cause-effect relationship between nodes,this refers to casuality property of bayesian network given in section 3.6. In the next chapter we will discuss about the method used for detection of epidemic.Steps used to detect the epidemic,identifying the anomalous patterns.After identifying anomalous patterns detection time is calculated based upon alerts generated at each suspicious event.

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Detection of disease outbreak

Disease surveillance is the collection, analysis, understanding and distribution of health care data for avoiding health related issues.The main aim of surveillance system for disease outbreak is early detection of outbreak for avoidance of further illness and mortality. For early detection of disease outbreak detection a rule based detection system is used.It is rule based detector for disease outbreak system which compares recent data with baseline data to find the rules which are treated as anomalies.Anomaly shows unusual behavior in the dataset as compared to normal behavior of data.In this work disease outbreak detection is done using bayesian network.

The modelling of bayesian model started by identifying features relevant to Anthrax.Number of features are used to model the bayesian network for anthrax detection,all features are divided into two subcategories one is taken as environmental features and other one is response features.Environmental features are responsible for changing trends in data due to seasonal variations and all non-environmental features are taken as response features. There is casual relationship between both i.e. response and environmental features.With the property of bayesian network best relationship between two features is encoded.Based on the domain knowledge we defined probability of existence of events in bayesian model trained on the year.

For example, we know by common knowledge that bad food condition influences number of patients in emergency department to rise.

Datasets generated for two year time period from learned bayesian network.One year data is taken as training set and second year data taken as testing set.So one year is for training and other one for evaluation. To identify anything strange in testing set i.e.recent events,current day under observation is taken for evaluation.Whereas baseline data is taken under training set. The events that comes under certain rule in recent data are the number of cases matching the rule, Similarly number of cases matching for certain rule in baseline data is also calculated.Randomization of both the datasets is done to check the valid rules during evaluation.The rules which are consistent before and after randomization are valid patterns.Null hypothesis is taken as independence of recent data with baseline data i.e.there is no relation between recent data and baseline data.To find the strength of null hypothesis p-value is calculated.Here we used fisher exact test for p-value computation.A 2X2 contingency table is formed based upon current rule under observation.P-value computation for all the two component rules before randomization is taken as score to differentiate from the p-value after randomization.Now a count for each rule is calculated after randomization based upon score calculated before randomization for corresponding rule.The count for each rule is incremented after randomization, if its p-value is less than as calculated before randomization for that rule.The maximum count over number of randomization tests leads to rule as anomalous.This process continues till complete year time span and anomalous patterns are identified.

The days with significant P-value after randomization are treated as anomalies. As an example Season = spring AND Food condition = bad is treated as anomalous pattern.Similarly we identify different anomalous patterns for complete year by comparing training set with testing set..An epidemic attack is declared when number of patients having anomalous patterns are increased under certain threshold.

4.2 Description of patterns

After the detection of anomalous patterns it becomes necessary to describe these patterns,means to find the reason of spreading of epidemic in an area.Based on the symmetry of patterns in the datasets and common behaviour of patterns we can easily identify the reason of spreading.For e.g if we find patterns like season=winter,food condition=bad and food condition=bad,day-of-week=Sunday in consecutive days then bad food condition and weekend becomes the reason of spreading disease because of unusual proportion of persons against baseline.High rise in number of people with these patterns leads to epidemic in the area.

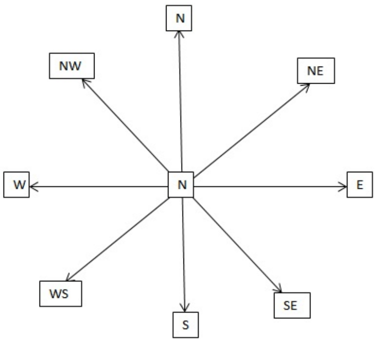


Figure 4.1: Regions distribution in area

Probabilistic model for anthrax detection should be able to detect anomalous pattern using spatio- temporal characteristics of a region.Like in bayesian model region is considered individually to analyse the effects within adjacent regions. For e.g let us consider there is anthrax attack on north(N)region the what is the probability of anthrax attack to the adjacent region of north i.e.north-east north-west as shown in Figure 4.1

4.3 System Design:

Algorithm used for the detection of disease outbreak is completed in two steps.One contains calculation of p-value for all the possible rules in current day which is treated as before randomization process.Second one contains the calculation of

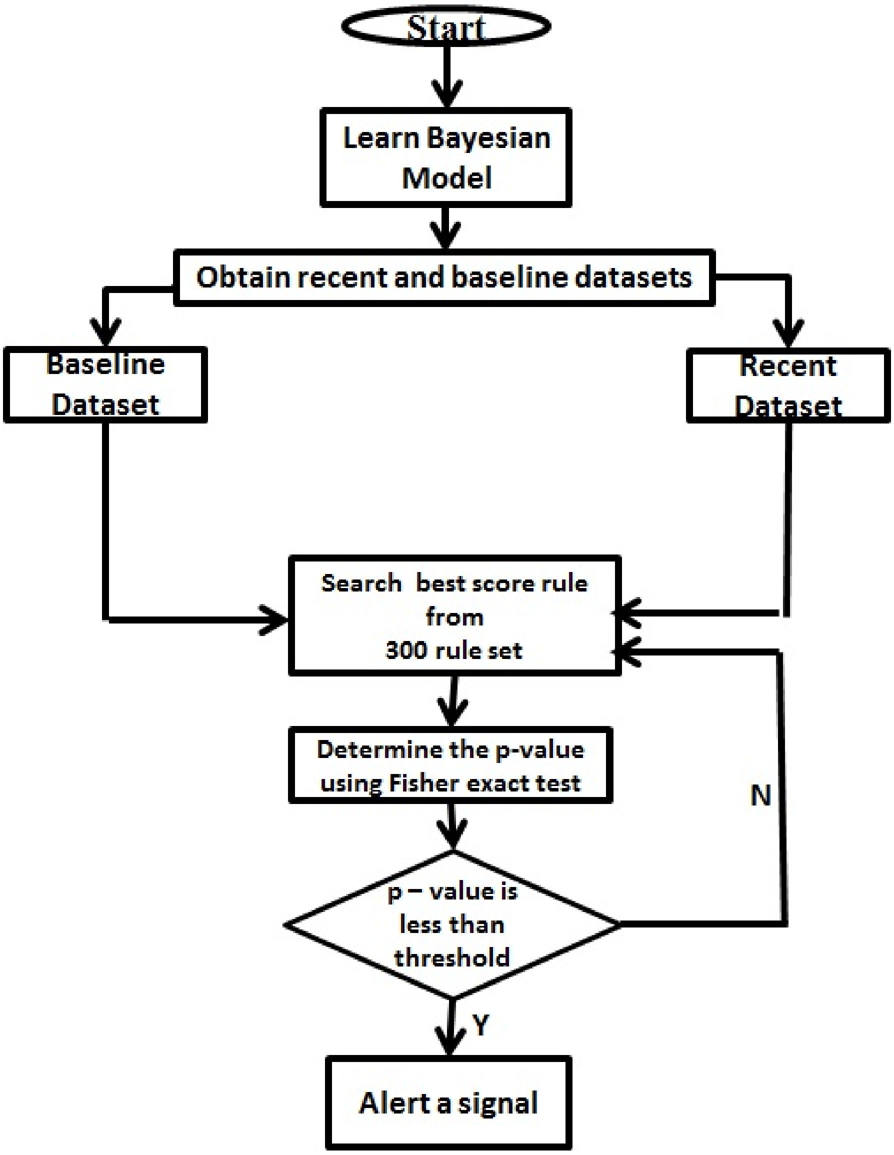

Start
Learn Bayesian Model
Obtain recent and baseline datasets
Baseline Dataset
Recent Dataset
Search best score rule from 300 rule set
Determine the p-value using Fisher exact test
p – value is less than threshold
N
Y
Alert a signal

Figure 4.2: Before randomization

p-value after randomization of datasets. Flow of the algorithm for before randomization and after randomization as shown in Figure 4.2 and Figure 4.3 respectively.

4.3.1 P-value computation

P-value refers to strength of our hypothesis-value for each possible component rule is calculated using fisher exact test because we are searching for anomalies. Null hypothesis is taken as independence of row and column attributes in two contingency table of fisher exact test given in Table 1.1.The hypothesis test measures

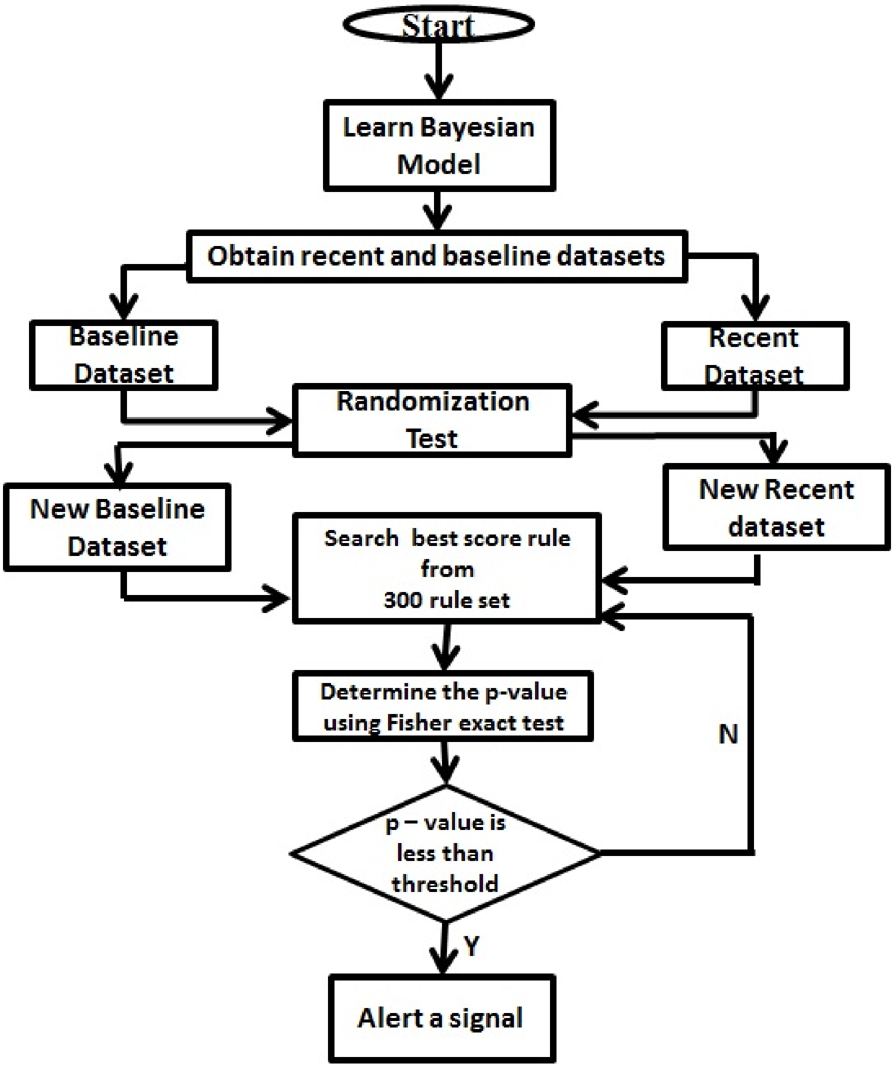

Start
Learn Bayesian Model
Obtain recent and baseline datasets
Baseline Dataset
Recent Dataset
Randomization Test
New Baseline Dataset
New Recent dataset
Search best score rule from 300 rule set
Determine the p-value using Fisher exact test
N
p – value is less than threshold
Y
Alert a signal

Figure 4.3: After randomization

how different the distribution of recent data to baseline data. This test gives the p-value that decides the significance of rule. So p-value is calculated for each rule to check the strength of null hypothesis. If p-value for particular rule is below threshold level then null hypothesis is rejected and alternate hypothesis is taken under consideration. As an example running hypothesis test for table below we get the p-value for specific two component rule as 0.023.If threshold level is taken as 0.05 then this value indicates that count for recent data for given rule is not independent from count for baseline data for given rule. In this way p-value is calculated for each possible two component rule, before and after randomization both.The rule having lowest p-value for the current day before randomization is considered as having higher value.P-value for the rule having higher value is calculated using randomized test.If p-value is lower than specified threshold level then alert is raised.Lower p-value refers to lower false positive and higher detection time.P-value below threshold value is taken under consideration.The rules having lower p-value are considered and deviation of recent data to baseline data becomes reason for alarm generation.P-value is calculated to check how different the baseline data to recent data.

4.4 Summary

In this chapter basic method used for identification of anomalous patterns is presented.To check the deviation in current day,current day is compared with baseline data which is encoded by learning bayesian network.Null hypothesis is considered as independence of recent data and baseline data i.e.there is no relationship between recent and baseline data.P-value is calculated to check the strength of null hypothesis is given in section 4.3.P-value is calculated using fisher's exact test by constructing contingency table.If calculated p-value is lower than threshold level then alarm is generated.Lower p-value refers to rejection of null hypothesis and acceptance of alternate hypothesis.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Experimental setup

This chapter discusses about the experimental setup needed to implement the detection process. Bayesian network is based on domain knowledge for anthrax detection.Trained Bayesian network is learned in Netica.Simulation of datasets is done on Netica. Netica Java API is used to implement all the steps of detection process

- **Netica:**Netica is useful in learning the desired Bayesian network with the help of the given data, the conditional probability tables were developed by Netica. Which are very useful in the later course. Netica is a powerful, easy-to-use, complete program for working with Bayesian networks and influence diagrams. It has an intuitive and smooth user interface for drawing the networks. Once a network is created, the knowledge it contains can be transferred to other networks by cutting and pasting, or saved in modular form by creating a library of nodes with disconnected links. Netica can use the networks to perform various kinds of inference using the fastest and most modern algorithms. The Netica-J API is a complete library of Java classes for working with Bayesian networks (also known as Bayes nets, belief networks, graphical models or probabilistic causal models) and influence diagrams (also known as decision networks). It contains functions to build, learn from data, modify, transform, performance-test, save and read nets, as well as a powerful inference engine. It can manage "cases" and sets of cases, and can connect directly with most database software. Bayes nets can be used for diagnosis, prediction, classification, sensor fusion, risk analysis, decision analysis, combining uncertain information and numerous probabilistic inference tasks. The Netica API has been designed to be easily extended in the future without changing what already exists. Many new features are currently under development, and it will continue to be extended for years to come.

- **Netbeans:**NetBeans IDE is the official IDE for Java 8. With its editors, code analysers, and converters, we can quickly and smoothly upgrade our applications to use new Java 8 language constructs, such as functional operations, and method references.Batch analysers and converters are provided to search through multiple applications at the same time, matching patterns for conversion to new Java 8 language constructs. With its constantly improving Java Editor, many rich features and an extensive range of tools, templates and samples, NetBeans IDE sets the standard for developing with cutting edge technologies out of the box. In this project work Netbeans was used for all the coding part of p-value computation for one component rule as well for the two component rule.Further anomalous patterns are determined.

5.1.1 Datasets and Bayesian Networks

Simulation of bayesian network is done to generate datasets for two year period.The environmental conditions are changing rapidly i.e.they are not static,in terms of weather,flu levels and food conditions in the area changing as the time passes.Flu level starts to rise during the fall and goes low in the spring and summer[2].Flu level strike in winter,leads to highest flu levels during the year.Weather has only two values namely hot and cold,containing property of remaining the same as it was on previous day.Food condition has values good and bad,bad food condition causes outbreak in the area due to food poisoning.

5.1.1.1 Datasets

Datasets are generated for two year time span.First year is taken as training set and second year is taken as testing set.Recent dataset falls under training set while baseline dataset falls under training set.We identify the unknown classes of testing set using training set.These unknown classes refers to attack class present in the test.To identify the records having unknown class(attack or non-attack)can be classified using detection process.

The detection algorithms trained on data from the first year until the day being monitored, while the second year is used for evaluation. The anthrax release is randomly chosen in recent dataset which is choosen for evaluation period.
Bayesian network is used to obtain the simulated data for anthrax disease outbreaks.Bayesian network generates a simulated data set called PS(patient status) data set. The number of cases the PS network generates daily is typically in the range of 30 to 50 records as shown in Figure 5.10 .Each line in a data set represents a health care record where a person in the simulation city did one of three things:visit an ED department, purchase medication, or was absent from school/-work. The fields in the PS data set can take on the following values as shown in Table 5.1.

Attribute	States
Region	NW, N, NE, W, C, E, SW, S, SE
Gender	male or female
Flu	none, low, high or decline
Day-of-week	sat, sun or weekday
Weather	hot or cold
Season	winter, spring, summer or fall
Action	purchase, evisit or absent
Reported-symptom	none, respiratory, nausea or rash
Drug	none, nyquil, aspirin or vomit-b-gone
Date	Jan-01-2012 to Jan-01-2013

Table 5.1: Attribute set for patient set dataset

5.1.1.2 Bayesian Network Model for Anthrax disease outbreak

Bayesian network consists of two group of attributes i.e.response attributes and environmental attributes which affect the upswings and downswings of anthrax disease.This network is used to generate records for individual patients i.e.patient status data set is generated using bayesian network shown in Figure 5.1 .Environmental attributes mainly include attributes namely weather,food condition,flu level and day of week.We differentiate between environmental attributes,which are attributes that cause trends in the data.The environmental attributes are specified by the user based on the user's knowledge of the problem domain.The environment of the city is not static, with weather, flu levels and food conditions in the city changing from day to day. Flu levels are typically low in the spring and summer but start to climb during the fall. We make flu season strike in winter, resulting in the highest flu levels during the year. Weather, which only takes on the values of hot or cold, is as expected for the four seasons, with the additional feature that it has a good chance of remaining the same as it was yesterday. Each region has a food condition of good or bad. A bad food condition facilitates the outbreak of food poisoning in the area.

Conditional Probability table is attached with each attribute namely with Weather(W) ,FluLevel(FL),FoodCondition(FC),Gender,Action,ReportedSymtom,Drug and DayOfWeek.The values are generated probabilistically for these attributes using some domain knowledge.The calculation of probababilites given in CPT table is computed using NETICA tool.Automatically generation of values is done in NETICA for each variable or node of bayesian network.The environment of the city is not static, with weather, flu levels and food conditions in the area changing from day to day. Flu levels are typically low in the spring and summer but start to climb during the fall. We make flu season strike in winter, resulting in the highest flu levels during the year. Weather, which only takes on the values of hot or cold, is as expected for the four seasons, with the additional feature that it has a good chance of remaining the same as it was yesterday. Each region has a food condition of good or bad. A bad food condition facilitates the outbreak of food poisoning in the

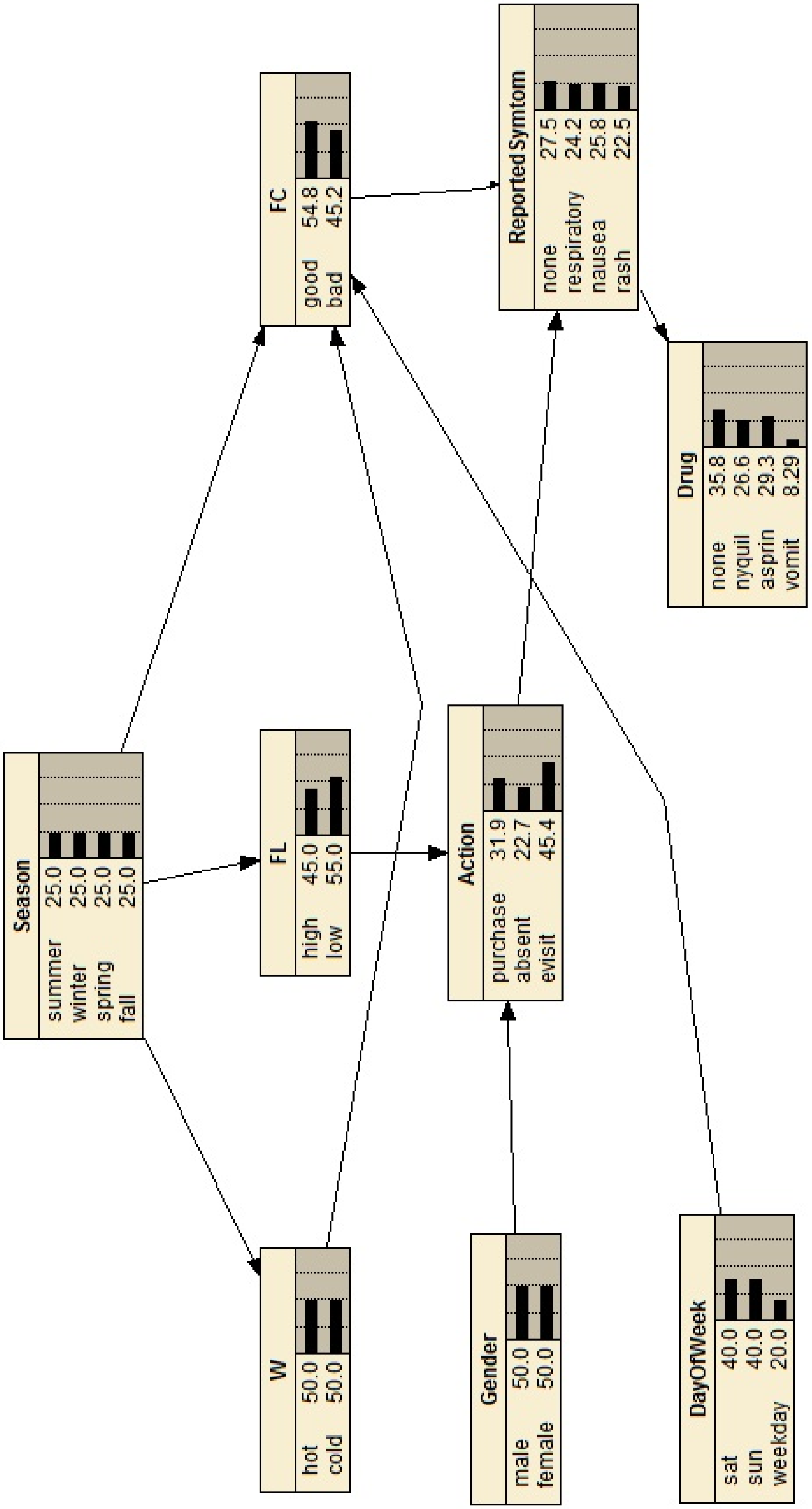
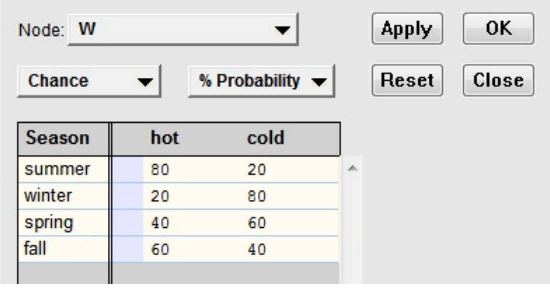


Figure 5.1: Patient Status Bayesian Network

area. Different probability tables for each node on the basis of above statements are as given below:

1. Weather(W): hot or cold are two states of weather,weather variable has one parent namely Season as shown in Figure 5.2 so its CPT contain four entries to four values of Season.



Node: W

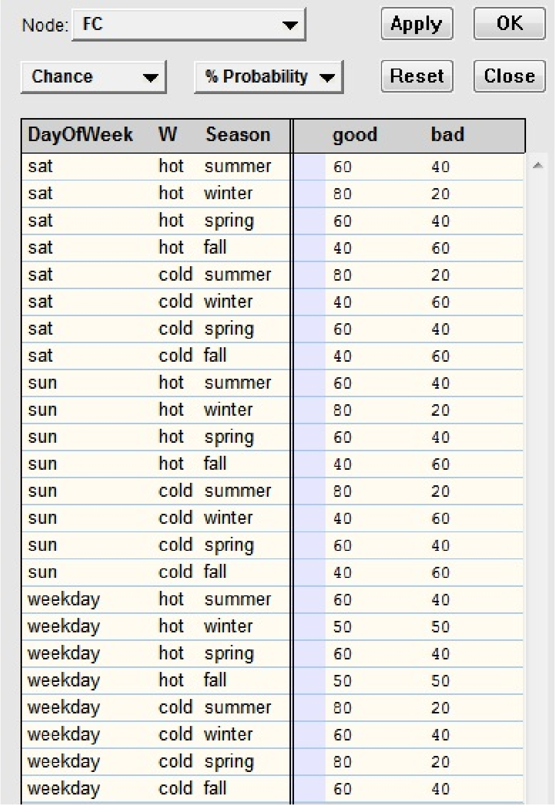
Chance

% Probability

Season	hot	cold
summer	80	20
winter	20	80
spring	40	60
fall	60	40

FIGURE 5.2: Probability table for weather

2.Food condition(FC): good or bad are two states of food condition,this variable has three parents and containing all the possible entries in its CPT as shown in Figure 5.3



Node: FC

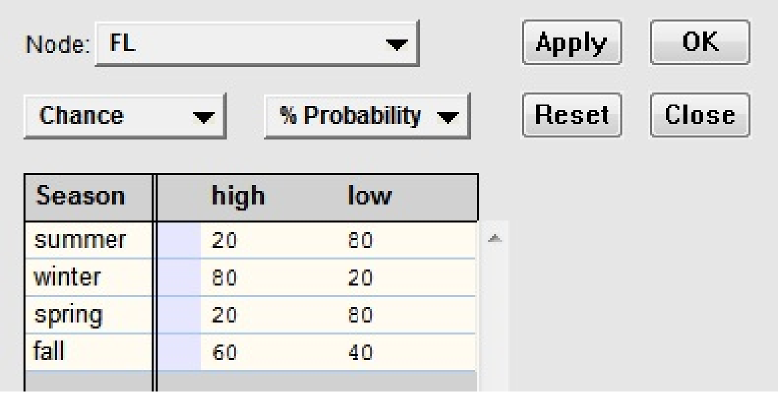
Chance

% Probability

DayOfWeek	W	Season	good	bad
sat	hot	summer	60	40
sat	hot	winter	80	20
sat	hot	spring	60	40
sat	hot	fall	40	60
sat	cold	summer	80	20
sat	cold	winter	40	60
sat	cold	spring	60	40
sat	cold	fall	40	60
sun	hot	summer	60	40
sun	hot	winter	80	20
sun	hot	spring	60	40
sun	hot	fall	40	60
sun	cold	summer	80	20
sun	cold	winter	40	60
sun	cold	spring	60	40
sun	cold	fall	40	60
weekday	hot	summer	60	40
weekday	hot	winter	50	50
weekday	hot	spring	60	40
weekday	hot	fall	50	50
weekday	cold	summer	80	20
weekday	cold	winter	60	40
weekday	cold	spring	80	20
weekday	cold	fall	60	40

FIGURE 5.3: Probability table for food condition

3. Season: winter, spring, summer or fall are the four states of season variable having no parent. 4.Flu level(FL): low or high are two states of flu level.It contain one parent only so corresponding entries in its CPT given in shown in Figure 5.4



Node: FL

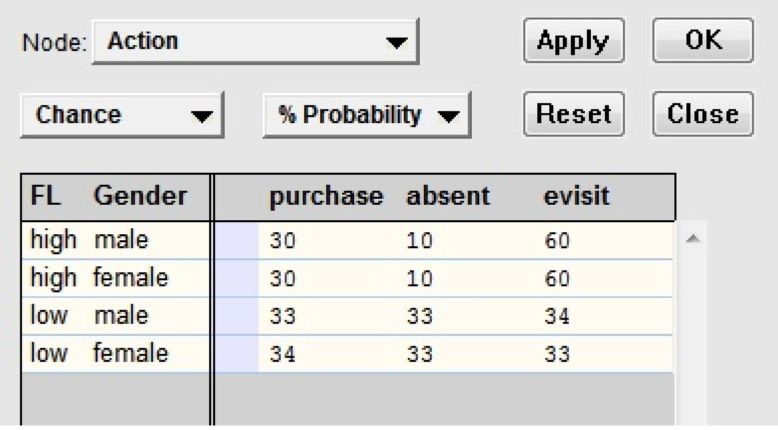
Chance

% Probability

Season	high	low
summer	20	80
winter	80	20
spring	20	80
fall	60	40

FIGURE 5.4: Probability table for flu level

4. Action: purchase, evisit or absent are three states of action variable.Action variable has two parents as shown in bayesian network given in shown in Figure 5.5.Conditional probability table contain entries corresponding to its parents.



Node: Action

Chance

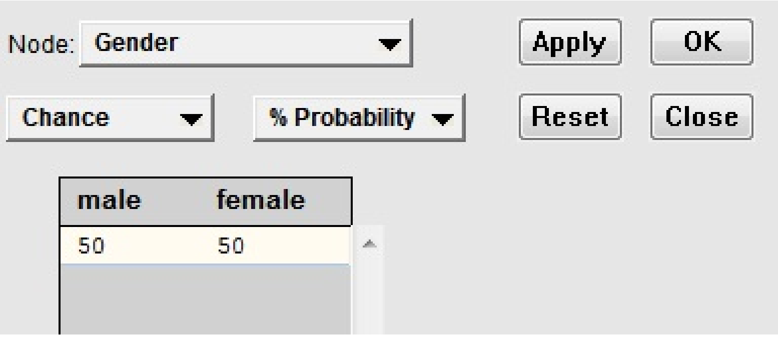
% Probability

FL	Gender	purchase	absent	evisit
high	male	30	10	60
high	female	30	10	60
low	male	33	33	34
low	female	34	33	33

FIGURE 5.5: Probability table for action

5. Gender: male or female are two states of gender variable as shown in Figure 5.6.This variable has no parents.

6. Reported symptom: none, respiratory, nausea or rash are four states of reported symptom,containing possible entries in its CPT as shown in Figure 5.7 .



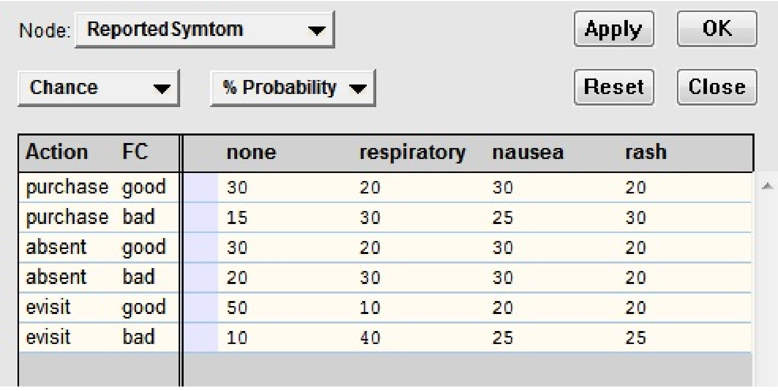
Node: Gender

Chance

% Probability

male	female
50	50

FIGURE 5.6: Probability table for gender



Node: Reported Symtom

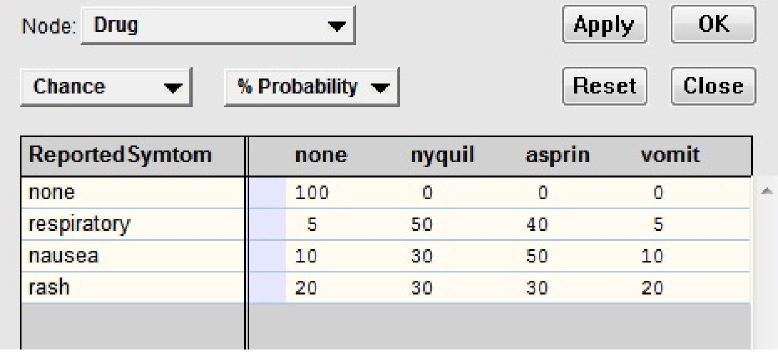
Chance

% Probability

Action	FC	none	respiratory	nausea	rash
purchase	good	30	20	30	20
purchase	bad	15	30	25	30
absent	good	30	20	30	20
absent	bad	20	30	30	20
evisit	good	50	10	20	20
evisit	bad	10	40	25	25

FIGURE 5.7: Probability table for reported symptom

7. Drug: none, nyquil, aspirin or vomit-b-gone are four states of drug having no parent.Its CPT is given in Figure 5.8



Node: Drug

Chance

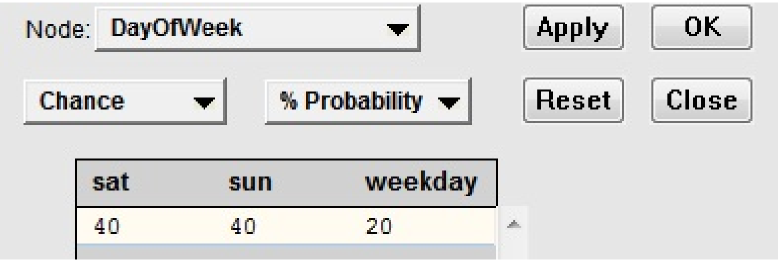
% Probability

Reported Symtom	none	nyquil	asprin	vomit
none	100	0	0	0
respiratory	5	50	40	5
nausea	10	30	50	10
rash	20	30	30	20

FIGURE 5.8: Probability table for drug

8. Day of week: sat, sun or weekday are three states of day of week having no parent.The condition probability table for this variable is given in Figure 5.9

Bayesian network used in our simulation produces individual health care cases.Figure 5.10 shows the Patient Status network.On each day,for each person in each region we simulate the individual's values from this bayesian network.The number of cases



Node: DayOfWeek

Chance

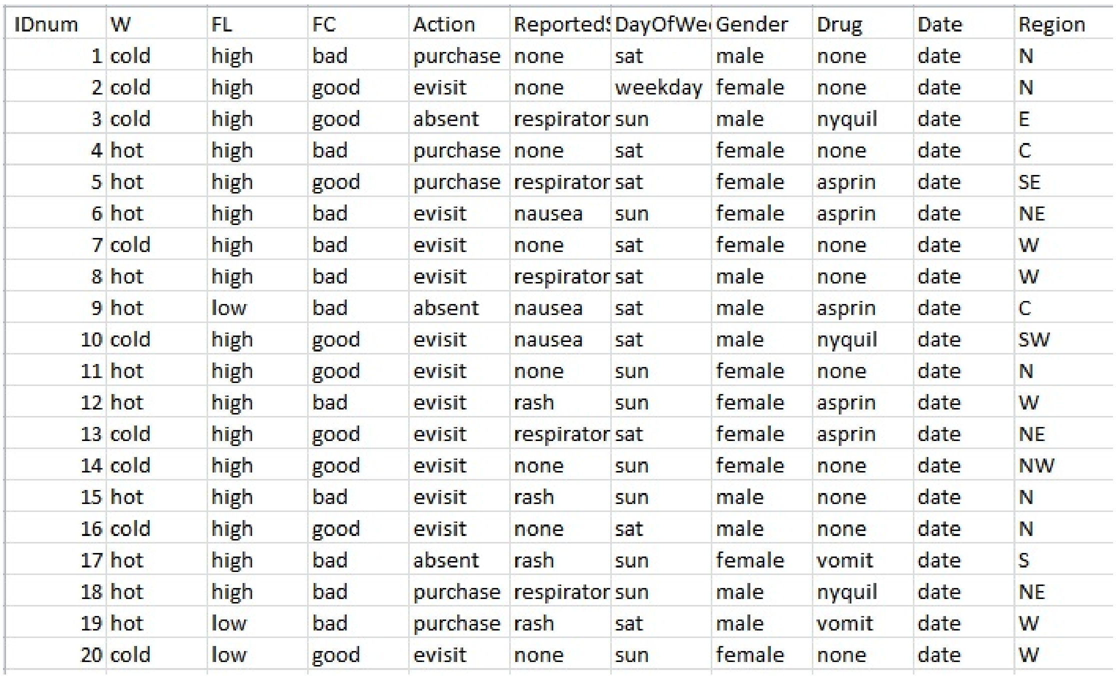
% Probability

sat	sun	weekday
40	40	20

FIGURE 5.9: Probability table for dayofweek

the Patient Status network generates daily is in the range of 30 to 50 records.Figure 5.10 contains 20 records in patient status data set after simulation.

There are some records where we put anthrax attacks based upon low p-value.Again for two data sets from original dataset i.e.baseline dataset and recent dataset we calculate p-value using randomization test and finds those subsets whose proportions have changed the most between recent data and baseline.The days with these subsets are returned as anomalies. Further there is another bayesian network for dengue disease as given below.



IDnum	W	FL	FC	Action	ReportedS	DayOfWe	Gender	Drug	Date	Region
1	cold	high	bad	purchase	none	sat	male	none	date	N
2	cold	high	good	evisit	none	weekday	female	none	date	N
3	cold	high	good	absent	respirator	sun	male	nyquil	date	E
4	hot	high	bad	purchase	none	sat	female	none	date	C
5	hot	high	good	purchase	respirator	sat	female	asprin	date	SE
6	hot	high	bad	evisit	nausea	sun	female	asprin	date	NE
7	cold	high	bad	evisit	none	sat	female	none	date	W
8	hot	high	bad	evisit	respirator	sat	male	none	date	W
9	hot	low	bad	absent	nausea	sat	male	asprin	date	C
10	cold	high	good	evisit	nausea	sat	male	nyquil	date	SW
11	hot	high	good	evisit	none	sun	female	none	date	N
12	hot	high	bad	evisit	rash	sun	female	asprin	date	W
13	cold	high	good	evisit	respirator	sat	female	asprin	date	NE
14	cold	high	good	evisit	none	sun	female	none	date	NW
15	hot	high	bad	evisit	rash	sun	male	none	date	N
16	cold	high	good	evisit	none	sat	male	none	date	N
17	hot	high	bad	absent	rash	sun	female	vomit	date	S
18	hot	high	bad	purchase	respirator	sun	male	nyquil	date	NE
19	hot	low	bad	purchase	rash	sat	male	vomit	date	W
20	cold	low	good	evisit	none	sun	female	none	date	W

FIGURE 5.10: Examples of records in patient data set

5.1.1.3 Bayesian Network Model for Dengue disease outbreak

Bayesian network for Dengue disease outbreak consist of two group of attributes i.e.response attributes and environmental attributes as for above bayesian model which affect the upswings and downswings of dengue disease as shown in Figure 5.11.This bayesian network is used to generate records for individual patients i.e.patient status data set is generated using bayesian network.Environmental attributes mainly include attributes namely month,rain fall,wind speed,humidity and temperature.We differentiate between environmental attributes,which are attributes that cause trends in the data.The environmental attributes are specified by the user based on the user's knowledge of the problem domain.Similarly as we have implemented anomalous patterns for anthrax disease we can implement WSARE algorithm on bayesian network given below and can identify the anomalous patterns.

The fields in the PS data set can take on the following values as shown in Table 5.2: After simulation of datasets, detection process for anthrax attack is completed

Attribute	States
Region	NW, N, NE, W, C, E, SW, S, SE
month	Jan-Dec
rainfall	high,low,none
wind-speed	fast,slow,none
humidity	high,low,none
temperature	minimum,meximum
growth of virus	large-scale,small-scale
headache	low,high,extreme,none
heart-problems	minor,big,none
cold	yes,no
fever	high,low,none
action	evisit,absent,purchase,none
drug	aspirin,ibuprofen,none
date	Jan-01-2014 to Dec-31-2014
age	child,youth,senior

TABLE 5.2: Attribute set for patient set dataset

which we is going to be discuss in next section of results.

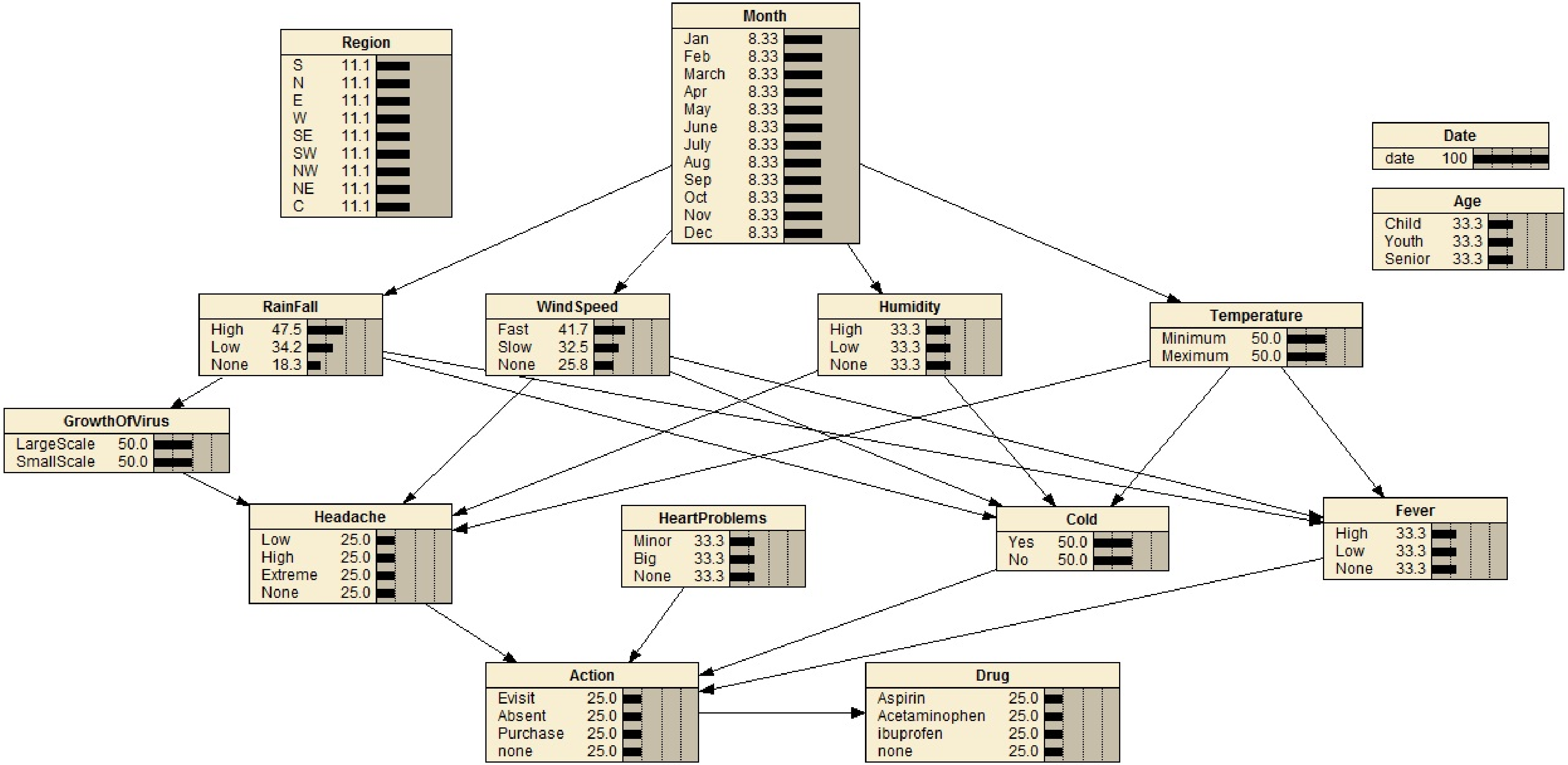


Figure 5.11: Bayesian Network for dengue

5.2 Results

5.2.1 Detection of epidemic:

We ran detection algorithm on simulated PS data set comprising both baseline data as well as recent data. Data set is created for a two year period, beginning on January 1, 2013 and ending December 31, 2014. The algorithm used for detection is trained on data from the first year until the current day while the second year is used for evaluation. The anthrax release was randomly chosen in the period between January 1, 2014 to December 31, 2014.We tried to simulate anthrax attacks that are not trivially detectable.Our evaluation measures tests the algorithm$'$ s performance in terms of detection time against false positives over p-values. The lower p-values refers to lower false positives and higher detection times while the vice-versa is true with higher p-values. Any alert arising before the start of the simulated anthrax attack is taken as a false positive. Detection time is considered as the first alert raised after the release date of attack. We ran the algorithm in two steps one containing before randomization and other one containing after randomization process.After comparing current day with baseline data according to given two component rule p-value is computed.Th threshold level is 0.05.The rules having p-value less than threshold level are taken as suspicious before randomization process.List of rules having p-value less than 0.05 is given below.

1. summer AND hot=7.618278092420458E-4
2. winter AND cold=0.0025595597704729323
3. spring AND cold=0.00530704550878507
4. fall AND hot=7.001794762730397E-27
5. fall AND cold=0.02365738383266181
6. summer AND low=7.33091525249078E-4

7. winter AND high=0.0014588599565338938

8. spring AND low=0.00110838641666191

9. fall AND high=6.99414261818419E-10

10. fall AND low=4.383640594026901E-5

11. summer AND good=0.0034681290818585012

12. winter AND good=0.011095118968813784

13. winter AND bad=0.038313230446704385

14. spring AND good=0.0025339396842104767

15. fall AND good=6.135978854863944E-6

16. fall AND bad=5.524660535355697E-8

17. fall AND female=5.440188890512035E-41

18. fall AND male=5.440188890512035E-41

19. summer AND purchase=0.019247001435938296

20. winter AND purchase=0.04687363102332742

21. spring AND purchase=0.031080955835171016

22. fall AND purchase=2.1688815240942265E-6

23. fall AND absent=0.006151365052825792

24. fall AND evisit=3.2570887502029585E-5

25. fall AND respiratory=3.090456408385384E-5

26. fall AND nausea=8.053604788467149E-4

27. fall AND rash=0.020291742684940214

28. summer AND sat=0.03326265442994574

29. winter AND sun=0.032653278985975934

30. spring AND sat=0.0306994665754416

31. spring AND sun=0.0407332730885923

32. fall AND sat=1.249758936599114E-4

33. fall AND sun=8.819478144712508E-7

34. fall AND weekday=0.0034607143143102416

35. fall AND nyquil=7.393969865705662E-5

36. fall AND asprin=1.9475388982412563E-4

37. hot AND high=4.707858011484178E-10

38. cold AND high=3.6239894064302197E-5

39. cold AND low=8.727599598981109E-4

40. hot AND good=0.04000523573820274

41. hot AND bad=8.513095440383821E-6

42. cold AND good=2.6142976203460893E-5

43. cold AND bad=0.0011672088250315065

44. hot AND female=5.440188890512035E-41

45. hot AND male=5.440188890512035E-41

46. hot AND purchase=0.005811939458317348

47. hot AND evisit=0.0057205193057705975

48. cold AND purchase=9.854817008403101E-4

49. cold AND absent=0.012484592437651197

50. cold AND evisit=0.006147437197063805

51. hot AND respiratory=0.013462965225366233

52. hot AND nausea=0.013927143392330445

53. cold AND respiratory=0.023219101467950597

54. cold AND nausea=0.0159354355277956

55. hot AND sat=0.027862719843519842

56. hot AND sun=9.103504251521231E-4

57. hot AND weekday=0.04999345100753667

58. cold AND sat=0.0025339396842104767

59. cold AND sun=8.34147760925804E-4

60. cold AND weekday=0.033571425681740

61. hot AND nyquil=0.02640761984397836

62. hot AND asprin=0.004003441551151521

63. cold AND nyquil=0.012604139339554963

64. cold AND asprin=0.020620490770695304

65. high AND female=1.4433316915886504E-32

66. high AND male=4.5238680940358676E-27

67. low AND female=1.8691961348250648E-8

68. low AND male=1.1002829114824745E-13

69. high AND purchase=0.03671085907820535

70. low AND purchase=0.04350794347301583

71. high AND respiratory=0.021179217176438

72. high AND asprin=0.03524743279126759

73. good AND female=5.537745934447231E-19

74. good AND male=5.537745934447231E-19

75. bad AND female=1.166691485537654E-21

76. bad AND male=1.166691485537654E-21

77. female AND purchase=2.5191925611220813E-16

78. female AND absent=4.6211824457061524E-11

79. female AND evisit=1.1002829114824745E-13

80. male AND purchase=5.537745934447231E-19

81. male AND absent=7.2911681120882E-6

82. male AND evisit=2.5191925611220813E-16

83. female AND respiratory=2.5191925611220813E-16

84. female AND nausea=7.2911681120882E-6

85. female AND rash=7.2911681120882E-6

86. male AND respiratory=4.6211824457061524E-11

87. male AND nausea=1.1002829114824745E-13

88. male AND rash=7.2911681120882E-6

89. female AND nyquil=4.6211824457061524E-11

90. female AND asprin=1.1002829114824745E-13

91. female AND vomit=0.0027462673407803472

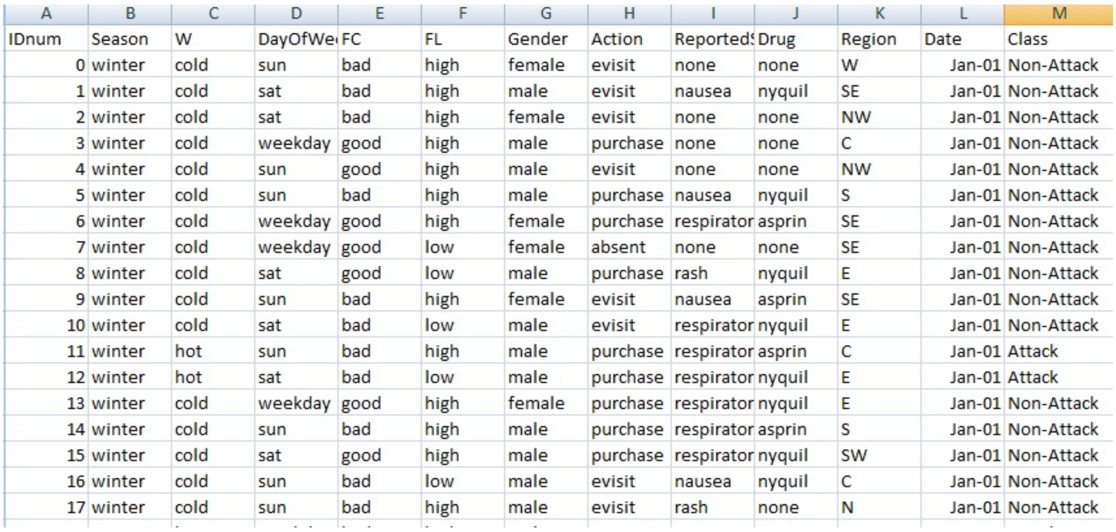
92. male AND nyquil=2.5191925611220813E-16

93. male AND asprin=1.8691961348250648E-8

94. male AND vomit=0.0027462673407803472

Now these rules are important for us in after randomization process because of lower p-value.We ran the randomization process 1000 times for each rule given in above list calculated before randomization.After randomizaton the rules which will remain consistent in most of the times are taken as valid and anomalous patterns.

We are observing minimum 3 days for deciding for anthrax attack.If three consecutive days having occupance of same patterns then last day alarm is generated. To check false positive count some attacks on the recent dataset are artificially inserted i.e. records having class label Attack are the days on which anthrax attack considered to be occur whereas records having class label Non-Attack are the days on which anthrax attack not to be occur. Now compare each day of recent dataset with baseline to check the deviation i.e. having lower p-value compared to some threshold value. We ran the algorithm for approx. 300 two component rule sets in our dataset and check anomalous pattern for each two component rule. Artificially inserted attacks in the recent dataset can be specified by two class labels Attack and Non-attack as shown in the Figure 5.12.If the algorithm detects attack on days labeled as non-attack then it is considered as false positive,whereas attack on days labeled as attack as shown in Figure 5.12 are considered as detection time in days.



A	B	C	D	E	F	G	H	I	J	K	L	M
IDnum	Season	W	DayOfWe	FC	FL	Gender	Action	Reported	Drug	Region	Date	Class
0	winter	cold	sun	bad	high	female	evisit	none	none	W	Jan-01	Non-Attack
1	winter	cold	sat	bad	high	male	evisit	nausea	nyquil	SE	Jan-01	Non-Attack
2	winter	cold	sat	bad	high	female	evisit	none	none	NW	Jan-01	Non-Attack
3	winter	cold	weekday	good	high	male	purchase	none	none	C	Jan-01	Non-Attack
4	winter	cold	sun	good	high	male	evisit	none	none	NW	Jan-01	Non-Attack
5	winter	cold	sun	bad	high	male	purchase	nausea	nyquil	S	Jan-01	Non-Attack
6	winter	cold	weekday	good	high	female	purchase	respirator	asprin	SE	Jan-01	Non-Attack
7	winter	cold	weekday	good	low	female	absent	none	none	SE	Jan-01	Non-Attack
8	winter	cold	sat	good	low	male	purchase	rash	nyquil	E	Jan-01	Non-Attack
9	winter	cold	sun	bad	high	female	evisit	nausea	asprin	SE	Jan-01	Non-Attack
10	winter	cold	sat	bad	low	male	evisit	respirator	nyquil	E	Jan-01	Non-Attack
11	winter	hot	sun	bad	high	male	purchase	respirator	asprin	C	Jan-01	Attack
12	winter	hot	sat	bad	low	male	purchase	respirator	nyquil	E	Jan-01	Attack
13	winter	cold	weekday	good	high	female	purchase	respirator	nyquil	E	Jan-01	Non-Attack
14	winter	cold	sun	bad	high	male	purchase	respirator	asprin	S	Jan-01	Non-Attack
15	winter	cold	sat	good	high	male	purchase	respirator	nyquil	SW	Jan-01	Non-Attack
16	winter	cold	sun	bad	low	male	evisit	nausea	nyquil	C	Jan-01	Non-Attack
17	winter	cold	sun	bad	high	male	evisit	rash	none	N	Jan-01	Non-Attack

FIGURE 5.12: Recent data set with anthrax attack

Anomalous patterns and false positive for 12 datasets with their date of attack are given below.Each table contain false positive and anomalous patterns for each

Index	**Pattern**	**False Positive**
1	winter and bad	5
2	summer and hot	3
3	spring and low	8
4	summer and bad	4
5	fall and bad	3
6	spring and M	7
7	fall and sat	9
8	winter and asprin	4
9	hot and bad	7
10	hot and F	5
11	hot and respiratory	3
12	col and asprin	6

Table 5.3: Attack on Mar-01 with detection time 2 days

Index	**Pattern**	**False Positive**
1	summer and bad	9
2	fall and hot	4
3	spring and low	7
4	fall and bad	5
5	fall and sat	3
6	hot and bad	8

Table 5.4: Attack on Jun-06 with detection time 4 days

tested dataset.False positive per month can be calculated as total occurrences of false positives to the number of months in which alarm generated.As an example in Table **??** total number of false positive for all the patterns is 64 and alarm generated over 12 months so false positive per month is calculated as 64/12=5.3 as false positive with 2 days as detection time.Similarly for all the datasets this process continues.

Index	**Pattern**	**False Positive**
1	spring and good	5
2	spring and low	4
3	winter and bad	8
4	summer and M	9
5	spring and M	4
6	winter and asprin	3
7	hot and bad	7
8	hot and F	5

Table 5.5: Attack on Jul-12 with detection time 3 days

Index	**Pattern**	**False Positive**
1	winter and absent	6
2	spring and hot	10
3	winter and evisit	5
4	winter and bad	8
5	fall and bad	4
6	winter and respiratory	9
7	winter and nausea	3
8	spring and asprin	6
9	hot and bad	11

Table 5.6: Attack on Apr-04 with detection time 2 days

Index	**Pattern**	**False Positive**
1	winter and bad	4
2	summer and hot	6
3	fall and low	7
4	winter and rash	2
5	summer and M	5
6	fall and sat	12
7	winter and asprin	6
8	hot and bad	14
9	hot and F	8

Table 5.7: Attack on Aug-14 with detection time 2 days

Index	**Pattern**	**False Positive**
1	low and asprin	9
2	summer and hot	5
3	cold and asprin	11
4	winter and bad	13
5	hot and nyquil	3
6	winter and asprin	10
7	hot and bad	6
8	hot and respiratory	9

Table 5.8: Attack on Jan-14 with detection time 1 day

Index	Pattern	False Positive
1	winter and bad	12
2	winter and cold	9
3	spring and low	4
4	fall and hot	7
5	fall and bad	10
6	spring and M	3
7	winter and low	5
8	hot and bad	9
9	hot and high	12
10	hot and respiratory	7

Table 5.9: Attack on Feb-21 with detection time 1 day

Index	Pattern	False Positive
1	winter and bad	4
2	summer and hot	6
3	spring and low	11
4	spring and good	9
5	spring and M	7
6	winter and absent	9
7	spring and purchase	11

Table 5.10: Attack on May-3 with detection time 2 days

False positive and detection time is calculated to check the performance of algorithm.As we are dealing with an alarm threshold of 0.05.We considered all the alarms generated by algorithm.If alarm generated before actual day of attack then it is treated as false positive,whereas alarm generated after actual day of attack is taken as detection time in days. Suppose this algorithm generates 6 alarms before actual attack and one alarm after 3 days after attack,then 6 becomes false positive and 3 is taken as detection time.We plot the graph between false positive per month and detection time in days.X-axis measures false positive per month and Y-axis measures detection time in days.False positive per month is calculated as in [34] Suppose we have total 16 false positive corresponding to 5 anomalous patterns over 5 months then false positive per month is calculated as 16/5 i.e.3.2,similarly

Index	**Pattern**	**False Positive**
1	winter and bad	5
2	winter and respiratory	7
3	summer and bad	6
4	winter and rash	4
5	fall and bad	14
6	winter and sat	7
7	winter and asprin	5
8	spring and sat	4
9	fall and sun	9
10	fall and sat	10

TABLE 5.11: Attack on Sep-02 with detection time 2 days

Index	**Pattern**	**False Positive**
1	winter and bad	9
2	fall and nyquil	7
3	fall and asprin	10
4	cold and high	6
5	hot and bad	11
6	spring and M	3
7	hot and evisit	13
8	winter and purchase	5
8	cold and evisit	9
8	cold and respiratory	3

TABLE 5.12: Attack on Oct-24 with detection time 1 day

we have calculated false positive per month for 12 datasets and corresponding time for detection.Table for detection time and false positive is given in Table 5.15.

Figure 5.13 plot the curve over the 12 data sets, with an alarm threshold 0.05.In the graph x-axis denotes the false positive per month and y-axis denotes detection time in days.Graph is plotted between detection time and false positive. Any alert occurring before the start of the simulated anthrax attack is treated as a false positive. Detection time is calculated as the first alert raised after the release date.With the graph plotted we can say that as false positive increases the

Index	**Pattern**	**False Positive**
1	winter and bad	11
2	hot and sun	4
3	spring and low	3
4	cold and sat	9
5	spring and M	5
6	hot and weekday	11
7	winter and asprin	8
8	cold and weekday	4
8	cold and nyquil	6
8	cold and asprin	3
8	hot and F	7

Table 5.13: Attack on Dec-15 with detection time 2 days

Index	**Pattern**	**False Positive**
1	low and F	3
2	summer and hot	5
3	hot and purchase	9
4	winter and bad	7
5	high and evisit	10
6	spring and M	6
7	low and asprin	4
7	low and asprin	11
8	high and asprin	2

Table 5.14: Attack on Nov-14 with detection time 2 days

detection time decreases and vice versa.

With the help of graph shown in Figure 5.13 we can conclude that with the increase in false positives there is decrease in detection time and vice-versa.

Dataset	Detection time	False-positive
A	4	3.6
B	3	3.7
C	2	5.1
D	2	5.3
E	2	5.7
F	2	5.9
G	2	6.2
H	2	7.1
I	2	7.3
J	1	7.6
K	1	7.8
L	1	8.2

Table 5.15: Detection time and false positive table

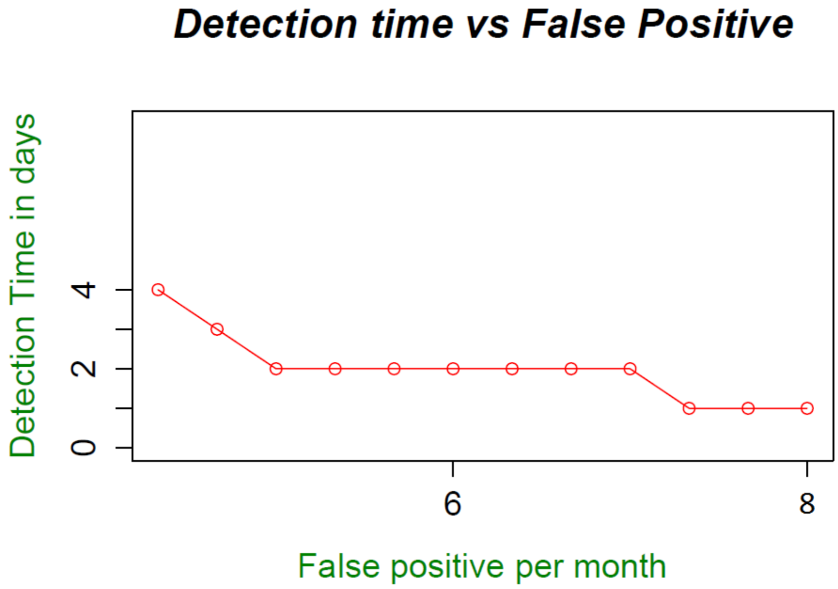


Figure 5.13: Evaluation of algorithm

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

Algorithm approaches the problem of early outbreak detection on multivariate surveillance data using two key components. The first component is association rule search, which is used to find anomalous patterns between a recent data set and a baseline data set. The second major component of algorithm is the use of a Bayesian network to model a baseline that changes due to temporal fluctuations such as seasonal trends and weekend versus weekday effects. When the bayesian network structure is learned, the environmental attributes are not permitted to have parents because we are not interested in predicting their distributions. Instead, we want to determine how the environmental attributes affect the distributions of the response attributes. This algorithm operates on an assumption that the environmental attributes account for the majority of the variation in the data. This work implemented a bayesian model for detection of anthrax epidemic. Bayesian model signifies domain knowledge of anthrax disease and captures seasonal variations. Anomalous patterns can be mined by bayesian network that are different to common knowledge encoded in the model. This model raises an alarm if such anomalous patterns are present in three consecutive days, and provide explanation on possible causes of spread.In future,we will work on real data sets for disease outbreak detection in the field of biosurveillance.

6.2 Future work

Further more in existing system WSARE [3] algorithm is implemented for detection of disease outbreaks using bayesian network. Comparing recent data with baseline data leads to alarm generation based upon rule based approach.Alarm generated before actual attack of epidemic is taken as false positive and alarm generated after actual attack of day is counted as detection time.So our main goal is to detect epidemic keeping low false positive and earliest detection in terms of days. An optimization technique named as Ant Colony optimization can be used for optimization of existing algorithm. By optimization we mean to lower the false positive rate. Ant colony optimization technique basically used to find optimal path, it could be effective in well known algorithms for example in trevelling salesman problem,job scheduling algorithm etc. We can use this technique in our work to lower false positive as compared to existing technique. Time complexity of existing technique, epidemic detection, can be minimized using ant colony optimization. The function used for optimization in ant colony algorithm is termed as criterion function.In our work the criterion function would be to make dataset unique by appending another attribute in existing dataset. Since if we inject attack on a particular day with a certain pattern say x AND y , all those pattern similar to x AND y will be marked as suspicious patterns thereby increasing count of false positive. Now using Ant Colony Optimisation we can add another attribute in the data set , the basis of newly added atribute would be based on some criteria, further at the time of anomolaus pattern detection we will consider the newly added attribute as well. While ants could represent a function used to traverse the records of files both in baseline and recent dataset.Records are taken as particles for ants.

BIBLIOGRAPHY

[1] W.K. Wong, A. Moore, G. Cooper, and M. Wagner, "Rule-based anomaly pattern detection for detecting disease outbreaks," in AAAI/IAAI, 2002, pp. 217-223.

[2] W.K. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in ICML, 2003, pp. 808-815.

[3] W.K. Wong, A. Moore, G. Cooper, and M. Wagner, "What's strange about recent events (WSARE): An algorithm for the early detection of disease outbreaks," The Journal of Machine Learning Research, vol. 6, pp. 1961-1998, 2005.

[4] G. F. Cooper, D. H. Dash, J. D. Levander, W.-K. Wong, W. R. Hogan, and M. M. Wagner, "Bayesian biosurveillance of disease outbreaks," in Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004, pp. 94-103.

[5] P. L. Geenen and L. C. Van Der Gaag, "Developing a Bayesian network for clinical diagnosis in veterinary medicine: from the individual to the herd," in Proceedings of the Third Bayesian Modelling Applications Workshop; Edinburgh, 2005.

[6] U. B. Kjærulff and A. L. Madsen, "Probabilistic networks-an introduction to bayesian networks and influence diagrams," Aalborg University, 2005.

[7] D. L. Buckeridge, H. Burkom, M. Campbell, W. R. Hogan, and A. W. Moore, "Algorithms for rapid outbreak detection: a research synthesis," Journal of biomedical informatics, vol. 38, pp. 99-113, 2005.

[8] P. Sebastiani, K. D. Mandl, P. Szolovits, I. S. Kohane, and M. F. Ramoni, ”A Bayesian dynamic model for influenza surveillance,” Statistics in medicine, vol. 25, pp. 1803-1816, 2006.

[9] Y. Shen and G. F. Cooper, ”A Bayesian biosurveillance method that models unknown outbreak diseases,” in Intelligence and Security Informatics: Biosurveillance, ed: Springer, 2007, pp. 209-215.

[10] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, ”Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project,” PLoS medicine, vol. 5, p. e151, 2008.

[11] F. C. Coelho, C. T. Codeço, and C. J. Struchiner, ”Complete treatment of uncertainties in a model for dengue R0 estimation,” Cadernos de saúde pública, vol. 24, pp. 853-861, 2008.

[12] M. Izadi, D. Buckeridge, A. Okhmatovskaia, S. W. Tu, M. J. O‘connor, C. Nyulas, and M. A. Musen, ”A Bayesian network model for analysis of detection performance in surveillance systems,” in AMIA Annual Symposium Proceedings, 2009, p. 276.

[13] M. MUBANGIZI, E. MWEBAZE, and J. A. QUINN, ”Computational Prediction of Cholera Outbreaks,” Kampala. ICCIR, 2009.

[14] X. Jiang and G. F. Cooper, ”A Bayesian spatio-temporal method for disease outbreak detection,” Journal of the American Medical Informatics Association, vol. 17, pp. 462-471, 2010.

[15] B. M. Althouse, Y. Y. Ng, and D. A. Cummings, ”Prediction of dengue incidence using search query surveillance,” PLoS neglected tropical diseases, vol. 5, p. e1258, 2011.

[16] M. R. Nikoo, R. Kerachian, S. Malakpour-Estalaki, S. N. Bashi-Azghadi, and M. M. Azimi-Ghadikolaee, ”A probabilistic water quality index for river water quality assessment: a case study,” Environmental monitoring and assessment, vol. 181, pp. 465-478, 2011.

[17] Dr. H. Rahman."Epidemiology of anthrax in india",Indian Council of Agricultural Research (ICAR),oct-2012.

[18] A. Beresniak, E. Bertherat, W. Perea, G. Soga, R. Souley, D. Dupont, and S. Hugonnet, "A Bayesian network approach to the study of historical epidemiological databases: modelling meningitis outbreaks in the Niger," Bulletin of the World Health Organization, vol. 90, pp. 412-417a, 2012.

[19] V. Racloz, R. Ramsey, S. Tong, and W. Hu, "Surveillance of dengue fever virus: a review of epidemiological models and early warning systems," PLoS neglected tropical diseases, vol. 6, p. e1648, 2012.

[20] M. Karim, S. U. Munshi, N. Anwar, and M. Alam, "Climatic factors influencing dengue cases in Dhaka city: a model for dengue prediction," The Indian journal of medical research, vol. 136, p. 32, 2012.

[21] N. D. Ahmad Tarmizi, F. Jamaluddin, A. A. Bakar, Z. A. Othman, S. Zainudin, and A. R. Hamdan, "Malaysia Dengue Outbreak Detection Using Data Mining Models," Journal of Next Generation Information Technology, vol. 4, 2013.

[22] Y. Maeno, "Detecting a trend change in a multiregional epidemic," arXiv preprint arXiv:1307.5300, 2013.

[23] T. Jombart, D. M. Aanensen, M. Baguelin, P. Birrell, S. Cauchemez, A. Camacho, C. Colijn, C. Collins, A. Cori, and X. Didelot, "OutbreakTools: A new platform for disease outbreak analysis using the R software," Epidemics, vol. 7, pp. 28-34, 2014.

[24] X. Jiang, "A Bayesian Network for Estimating and Predicting Epidemic Curves."

[25] T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson, "Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data," PLoS computational biology, vol. 10, p. e1003457, 2014.

[26] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data," Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 5, pp. 1-129, 2014.

[27] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: A survey," ACM Computing Surveys, 2007.

[28] D. Cecilia, "Current status of dengue and chikungunya in India," 2014.

[29] B. Bandyopadhyay, I. Bhattacharyya, S. Adhikary, J. Konar, N. Dawar, J. Sarkar, S. Mondal, M. Singh Chauhan, N. Bhattacharya, A. Chakravarty, A. Biswas, and B. Saha, "A Comprehensive Study on the 2012 Dengue Fever Outbreak in Kolkata, India," ISRN Virology, vol. 2013, p. 5, 2013.

[30] Himanshi Dhawan,Dengue outbreak of 2013 worst in 6 years :http://timesofindia.indiatimes.com/india/Dengue-outbreak-of-2013-worst-in-6-years/articleshow/25146647.cms,Nov 3, 2013

[31] Daniel Daw,Anthrax outbreak causes quarantine of Indian village:http://bioprepwatch.com/biological-threats/anthrax-outbreak-causes-quarantine-of-indian-village/339832/,October 30, 2014

[32] David Easley and Jon Kleinberg.(2010).Networks, Crowds, and Markets: Reasoning About a Highly Connected World(1st edition).Available:http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch21.pdf

[33] L. Gunaseelan, R. Rishikesavan, T. Adarsh, R. B. E. Hamilton, and J. Kaneene, "Temporal and geographical distribution of animal anthrax in Tamil Nadu state, india."

[34] Y. Shen, C. Adamou, J. N. Dowling, and G. F. Cooper, "Estimating the joint disease outbreak-detection time when an automated biosurveillance system is augmenting traditional clinical case finding," Journal of biomedical informatics, vol. 41, pp. 224-231, 2008.