Disease Prognosis and Forestalling System

Utkarsh Bajpai[1*] and Jai Prakash Bhati[2]

[1]M. Tech (Scholar), Department of Computer Science & Engineering,
School of Engineering and Technology, Noida International University,
Plot no-1, Sector- 17 A, Yamuna Expressway, Gautam Buddh Nagar, U.P.-201310

[2]Asst. Professor, Department of Computer Science & Engineering,
School of Engineering and Technology, Noida International University,
Plot no-1, Sector- 17 A, Yamuna Expressway, Gautam Buddh Nagar, U.P.-201310.

Abstract

In today's era the diseases are the biggest threat and disaster to the mankind. In order to detect the diseases in early stage, we need a much accurate analysis of medical data in biomedical and health care organization, which is only possible by implementing the concept of big data. Moreover, the accuracy of the analysis is reduced when the proportion of data is inaccurate or incomplete. There is a wide spectrum of diseases spread over the different regions of the country, which shows a specific medical drawback of prediction which leads to disease outbreak, and there are no such successful models in order to fight back against the spreading disease, which leads to the colloidal damage in the society. My efforts in this research is to stop spreading of any disease in its early stages so that disease outbreak can be stopped and infected people can be cured within a short period. Moreover, it will freeze the spreading of disease further.

Keywords: Disease outbreak, Bio-medical and health care, Disease Prognosis, Disease Forestalling.

1. Introduction

In this research I focused on streaming the concept of 'Machine learning algorithm' for the prognosis of disease outbreak and alarming the medical communities about the disease outbreak and providing forestalling method to the infected patients to stop mass scale infection and disaster[1].

In case of disease forestalling our prime task is to identify the risk factor of disease, so that mob can know how to deal with the threat and fight back to neutralize its affect. In this project for establishing disease prevention or forestalling system will be using extended form of "Beatties Model" (1991) which I named as Utkarsh star forestalling model.

2. Disease Prognosis Implementation

2.1 Data Source

Hospital Database have accumulated large data about patients and their medical status. The disease plays the major role of threat to the mankind. In India most of the deaths are caused by diseases. As per the report of "Centre for global health research". Every year in India there are deaths due to

- Respiratory Diseases (9%)
- Perinatal conditions (6.3%)

- Pneumonia (6.2%)
- Neoplasm (5.7%)
- Malaria (3%)

Record sets with medical attributes are obtained and analyzed, from various hospitals in India. With the help of framed data set the pattern significant to the specific disease is extracted. The record data was split equally into practice data set and trial data set, the records of each set we took randomly.

The attributes "Diagnosis" is identified as the predictable attribute with value "1" for the patients with any kind of disease and value "0" for the patient with no disease (Healthy patient with good immune system).
"Patient Id" is used as the key; the rest are input attributes. Let us assume that there no such missing details and information submitted by patients.

<u>Predictable Attributes</u>

- Diagnosis (Value 1: (Disease Found));
(Value 0: (No Disease Found));

<u>Prime or key Attribute</u>

- Patient Id: Patient identification number

<u>Input Attributes</u>

- Age (in years)
- Sex (Value 1: Male; Value 0 : Female)
- Disease Type (Value 1: Fever; Value 2: High Fever; Value 3: Chest pain; Value 4: Cough and cold)
- Increase in Blood sugar (Value 1: >120mg/dl; Value 0: <120mg/dl)
- ECG- Electrographic result (Value 0: Normal; Value 1: having ST-T wave abnormality; Value 2: left ventricular hyper therapy)
- Smoking (Value 1: Past; Value 2: Current; Value 3: Never)
- Obesity (Value 1: Yes; Value 2: No)
- Drinking (Value 1: Past; Value 2: Current; Value 3: Never)

2.2 Implementation of Naïve Bayesian classification algorithm for prognosis of diseases.

The Baye's Postulates are the fundamentals for many machine learning and data mining methods. The algorithm is used to create models with predictive capabilities. This technique is particularly used when the dimensionality of the input is high. It represents the probability of each input attributes from the record to the predictable conditions [2].

Bayes Rule

A conditional probability is the chances of some conclusion(x) on the basis of some observation (θ), where the dependence relationship between (x) and (θ) is denoted by $P\left(\frac{x}{\theta}\right)$,

Where

$$P(x/\theta) = \frac{P(\theta/x)\,P(x)}{P(\theta)}$$

Bayesian classification algorithm works as follows:

1) Let (T) be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional vector

$$Y = (y_1, y_2, y_3, \dots\dots, y_n)$$

Depicting n measurement made on the tuple from (n) attributes respectively.

$$C = (C_1, C_2, C_3, \dots\dots, C_n)$$

2) Suppose there are (z) classes, $B_1, B, B_3, \dots\dots, B_z$. Given a tuple, Y that classifier will predict that Y belongs to the class having higher posterior probability condition on Y. That is, the Bayesian classifier predicts that tuple(y) belong to the class (B_i)

If and only if

$P(B_i/Y) > P\,(B_j/Y)$

For $j \neq i \; and \; 1 \leq j \leq z$

Thus, we maximize $P(B_i/Y)$ then the class B_i for which $P(B_i/Y)$ is maximised is called the maximum posterion hypothesis. By Baye's Theorem

$$P(B_i/\mathrm{Y}) = \frac{P(\mathrm{Y}/B_i)\,P(B_i)}{P(\mathrm{Y})}$$

3) As, P(Y) is constant for all classes, only $P(\mathrm{Y}/B_i)\,P(B_i)$ need to be maximised. If the class prior probabilities are not known then it is commonly assumed that the classes are equally likely that is,

P(B_1) = P(B_2) = …….= P(B_z) and we would therefore maximize $P(\mathrm{Y}/B_i)$; otherwise we will maximize $P(\mathrm{Y}/B_i)\,P(B_i)$

Note: The class prior probability may be estimated by

$P(B_i) = |B_i\,, T/T|$

where $|B_i\,, T|$ is the number of training tuples of classB_i *in* T.

4) Given data that may have many attributes would be expensive to calculate$P(\mathrm{Y}/B_i)$; therefore in order to reduce calculation in computing$P(\mathrm{Y}/B_i)$, in this Bayesian assumption of class conditional independence is made. This presumes that the value of the attributes is conditionally independent of one another, given the class label of the tuple (i.e. that there is no dependence relationship among the attributes),

Thus,

$$P(\mathrm{Y}/B_i) \;= \prod_{k=1}^{n} P(Y_k/B_i)$$

$$= P(Y_1/B_i)\;\; P(Y_2/B_i), \dots\dots, P(Y_n/B_i)$$

We can easily estimate the probabilities $P(Y_1/B_i),\; P(Y_2/B_i), \dots\dots, P(Y_n/B_i)$ from the training tuple.

Recall that here y_k refers to the value of attribute (A_k) for the tuple (Y). For each attribute we look at whether the attribute is categorical or continuous valued. For instance, to calculate$P(\mathrm{Y}/B_i)$, we consider the following

a. If (A_k) is categorical, then $P(Y_k/B_i)$ is the number of tuples of classes (B_i) $in\ T$.

b. If (A_k) is continuous values, then we need to do a bit a more work, but the calculation is quiet simple. A continuous valued attribute is typically assumed to have a Gaussian distribution with a mean (μ) and standard deviation (σ)

$$g(y, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^\wedge 2}}$$

So that $P(Y_k/B_i) \;=\; g(Y_k,\;\mu B_i, \sigma B_i)$

We need to compute(μB_i) and (σC_i), which are the mean and standard deviation, of the value of attribute (A_k) for training tuples of class (B_i). We then plug these two quantities into the above equation.

5) In order to predict the class label of (Y), $P(\mathrm{Y}/B_i)\ P(B_i)$ is evaluated for each class(B_i). The classifier predicts that the class label of the tuple Y is the class B_i)

If and only if

$$P(\mathrm{Y}/B_i)\ P(B_i)\ >\ P(\mathrm{Y}/B_j)\ P(B_j)$$

$$for\ 1 \leq j \leq z\ , j \neq i$$

Or simply we can say that the predicted class label is the class B_i) for which $P(\mathrm{Y}/B_i)\ P(B_i)$ is maximum.

Implementation of Naïve Bayesian classifies within an example

The below example is a demonstration of applying the classifier. It shows how to calculate the probability of any disease using Naïve Algorithm.

S. No.	*Age Group*	*Blood Group*	*Any Disease*	*Health Status*	*Need AID*
1	Young	A+	Y	Critical	Yes
2	Young	AB+	Y	Critical	Yes
3	Mid-age	B+	N	Normal	No
4	Mid-age	A+	N	Normal	No
5	Mid-age	O+	N	Normal	No
6	Senior Citizen	AB+	Y	Healthy	No
7	Senior Citizen	B+	Y	Healthy	No
8	Senior Citizen	AB+	N	Healthy	No
9	Senior Citizen	AB+	N	Healthy	No
10	Senior Citizen	O+	Y	Healthy	No

Table1: **Class labelled training tuples from Divya Jyoti Hospital**

Predicting a class label using Baysian classification. I predict the class label of a tuple using a Baysian classification from the above training data in the table mentioned above.

The data types are expressed by the attributes age_group, blood_group, any_disease, health_status, need_aid.

The class label attribute: need_aid has two distinct values (namely, {Yes, No})

Let B_1 correspond to the class need_aid= Yes

B_2 Correspond to the class need_aid= No

The tuple I wish to classify is

Y= (age_group = young, blood_group= AB, any_disease= N, health_status= Healthy)

We need to maximize

$P(Y/B_i)\ P(B_i)\ ;\ for\ i = 1,2\ \ P(B_i)$

The prior probability of each class can be computed based on training tuples:

P (need_aid= Yes) = 2/10 = 0.2

$= \frac{\text{Total nos. Yes}}{\text{Total number}}$

P (need_aid= No) = 8/10 = 0.8

$= \frac{Total\ nos.\ No}{Total\ number}$

To compute $P(Y/B_i)\ ; for\ i = 1{,}2.\ We\ need\ to\ compute\ the\ value\ of\ a\ conditional\ probability.$

P(age_group= young \ need_aid= Yes) = $\frac{2}{2}$ =1

$= \frac{Total\ young\ who\ need\ aid}{Total\ need\ aid}$

P(age_group= young \ need_aid= No) = $\frac{0}{8}$

= 0

$= \frac{Total\ young\ who\ don't\ need\ aid}{Total\ who\ don't\ need\ aid}$

P(blood_group= A+\ need_aid= Yes) = $\frac{1}{2}$

= 0.5

$= \frac{Total\ A{+}who\ need\ aid}{Total\ no.of\ need\ aid}$

P(blood_group= A+ \ need_aid= No) = $\frac{1}{8}$

= 0.125

$= \frac{Total\ A{+}who\ do\ not\ need\ aid}{Total\ no.do\ not\ \ need\ aid}$

P(Any_Disease= Y \ need_aid= Yes) = $\frac{2}{2}$ =1

$= \frac{Total\ no.\ of\ any\ disease\ (Y)\ in\ need\_aid(Yes)}{Total\ nos.who\ need\_aid(Yes)}$

P(Any_Disease = Y \ need_aid= No) = $\frac{3}{8}$ =0.375

$= \frac{Total\ no.\ of\ any\_disease\ (Y)\ in\ need\_aid(No)}{Total\ nos.who\ need\_aid(No)}$

P(health_status= Healthy\ need_aid= Yes) = $\frac{0}{2}$ =0

$$= \frac{Total\ nos.of\ healthy\ who\ need\ aid}{Total\ nos.of\ need\ aid}$$

P(health_status= Healthy\ need_aid= No) = $\frac{5}{8}$ =0.625

Now by using the above probability, we obtain

P(Y\ need_aid= Yes) =

P(age_group=Young\need_aid=Yes) * P(blood_group= A\ need_aid= Yes) * P(any_disease= Yes\need_aid= Yes) * P(health_status= healthy\ need_aid= Yes)

= 1*0.5*1*0

=0

Similarly P(Y\ need_aid= No)

=0.1*0.125*0.375*0.625

=0.0029

Therefore, Inorder to find the class (B_i) that maximize P(Y|B_i) P(B_i); We calculate

= P(Y\need_aid= Yes) P(need_aid= Yes)

= 0*0.2

=0

=P(Y\need_aid= No) P(need_aid= No)

= 0.0029*0.8

= 0.00232

Now considering the maximum value from both

Therefore, as per the Bayesian classifiers prediction (need_aid= No) for tuple Y

JETIR

2.3 Implementation of Analyzed Data

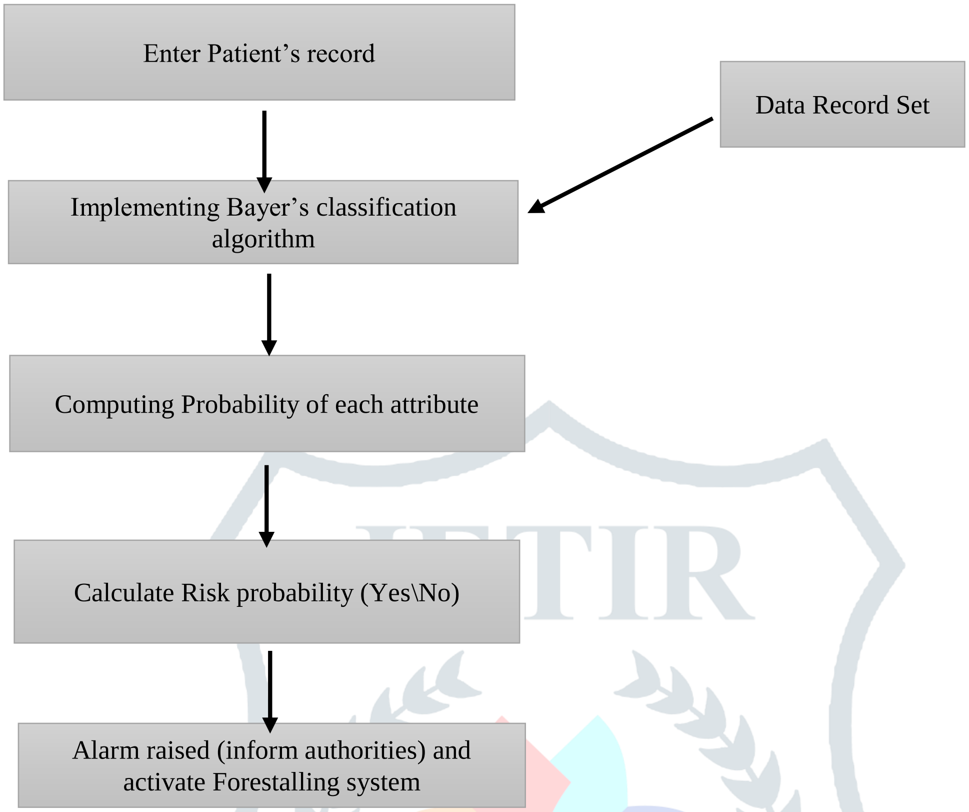
Enter Patient's record
Data Record Set
Implementing Bayer's classification algorithm
Computing Probability of each attribute
Calculate Risk probability (Yes\No)
Alarm raised (inform authorities) and activate Forestalling system

Figure 1: Flow diagram of Implementation of analyzed data

** This technique Naïve Bayer's classification is particularly suited when the dimensionality of the inputs is maximum.

3. Disease Forestalling System

3.1 Beatties Model (1991)

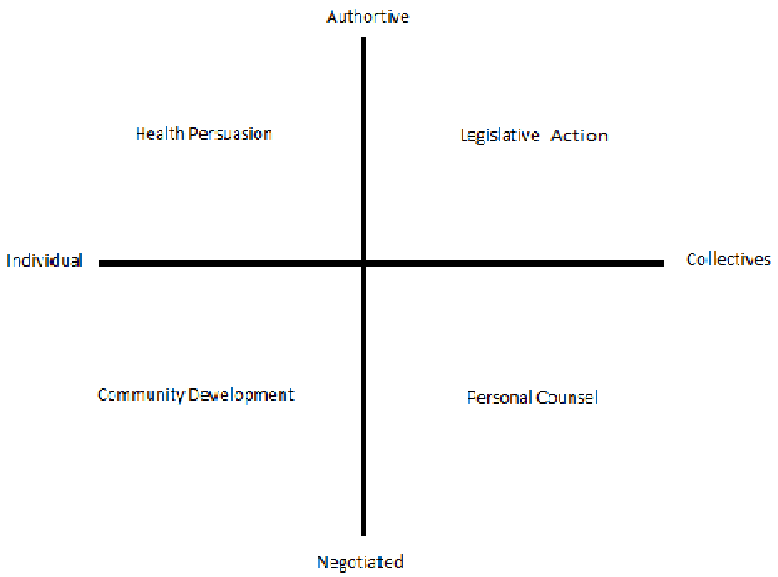

Figure 2: Diagrammatic Representation of Beatties Model

- Beatties model of health promotion and prevention is the complex analytical model that acknowledges whether the health promotions are embedded in a social practice.
- It has an efficiency to analyze current and previous health promotion strategies.
- The Beatties model functioning is divided in to four sections as explained below:
  - ➢ Health persuasion: It deals in focusing on why such behavior is happening. Example: Launching a campaign of eating 5 fruits and vegetables a day on TV for spreading general awareness.
  - ➢ Legislative Actions: It deals with the focusing on acts, resources and policy. Example: framing of law that subsidies the price of basic food stuffs.
  - ➢ Personal Counseling: It emphasize to ensure the greater control over the disease. Example: working with dietician on food and physical individual personal plans and goals.
  - ➢ Community Development: It deals with inculcating of community level skills. Example: Communities producing and distributing food themselves.

The main drawback of this model is that it does not explain about the behavioral changes and does not focus on the research implementations. Therefore, in order to overcome this drawback of Beatties model, I present Utkarsh Star forestalling model.

3.2 Utkarsh Star Forestalling Model

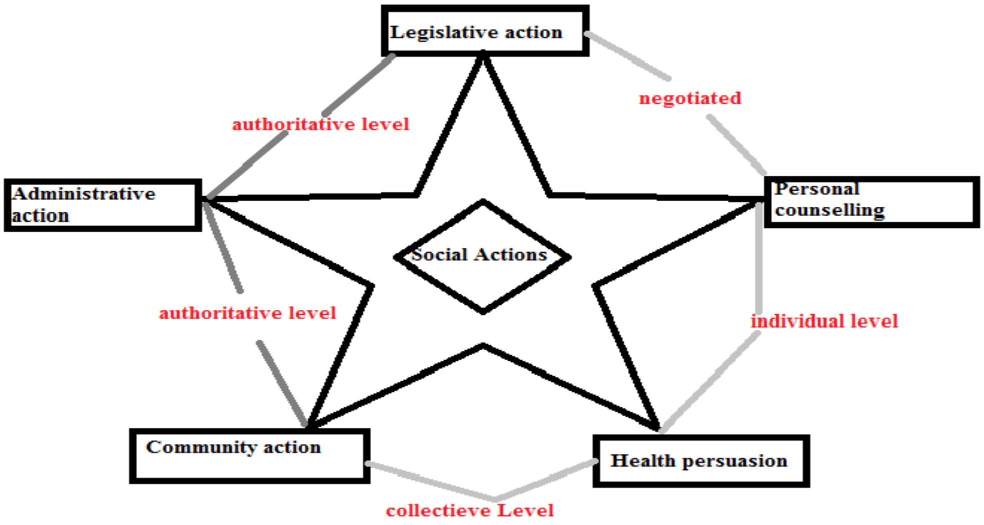

Figure 3: Diagrammatic representation of Utkarsh Star Forestalling Model

The Concept of Utkarsh Star Forestalling model is based on five pillars of health preventions that are:

- Medical
- Behavioral
- Educational
- Client Centered
- Societal changes

The prime components of this model are:

- **Legislative Action:** This is sub divided in to two responsibilities that are Framing of Acts and implementing the acts in the situation of epidemic.
- **Administrative Action:** This is sub-divided in to two responsibilities that are Implementing control measures and forcing the acts framed by the legislative body.
- **Health Persuasion:** This is further sub-divided into two responsibilities that are commitment towards consistency and health stability.
- **Community Development:** This is further sub divided into two responsibilities that are awareness campaigns and Research and Development.
- **Personal counseling:** This is further sub divided into two responsibilities that are Educating individual about how to maintain personal hygiene and informing person about do’s and don’ts to fight back with disease.

Moreover, Utkarsh Star Forestalling model also focuses on Area of health Promotion Activity such as:

- Health Education Programs
- Economic and regular activities
- Environmental health measures

- Healthy Public Policy
- Organization Development
- Community based work
- Preventive Health services

4. CONCLUSION

Diseases are always an unwanted threat to mankind. Early prognosis of disease leads to an immediate treatment, reduces the treatment cost and the chances of life threat. Using the concept of Naïve Bayesian classification algorithm for prognosis of disease. The opportunity in prognosis of future disease on top of current disease of a certain patient is provided. The results obtained after implementing Naïve Bayesian classification algorithm had high accuracy in diagnosis for patients before or at early stage of the disease. When talking about disease forestalling system I emphasize on prevention model to stop, further spread of disease. In this research work I have studied about the various Prevention models but the most I preferred for this research is Beatties model (1991), It was the advance prevention model of that time with minor drawbacks, which I already discussed above, in order to overcome that drawback I presented Utkarsh star forestalling model which can be treated as an extended form of Beatties Model. It has an ability to sustain the current requirement and ability to cope up with the future technology.

5. FUTURE RECOMMENDATION

As a future recommendation, the proposed work may be applied with the help NLP (Natural language Processing), which will be a boon to medical research NLP is an integrated part of Artificial intelligence and the combination of my research in disease prognosis and forestalling with Artificial intelligence will be next-gen technology.

REFERENCES

[1] Shantakumar B.Patil, Y.S. Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009

[2] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology.

[3] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering

[4] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[5] J. Sandhya et al., "Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques."