COMPARATIVE STUDY OF DECISION TREE ALGORITHM AND NAIVE BAYES CLASSIFIER FOR SWINE FLU PREDICTION

Mangesh J. Shinde[1], S. S. Pawar[2]

[1]*Student, Computer Department, D. Y. Patil COE, Akurdi, Pune-44, Maharashtra, India*
[2]*Assistant professor, Computer Department, D. Y. Patil COE, Akurdi, Pune-44, Maharashtra, India*

Abstract

The modeling and analysis of the epidemic disease outbreaks in huge realistic populations is a data intensive task that requires immense computational resources. Such effective computational support becomes useful to study disease outbreak and to facilitate decision making. Epidemiology is one of the traditional approaches which are being used for studying and analyzing the outbreaks of epidemic diseases. Although useful for obtaining numbers of sick, infected and recovered individuals in a population, this traditional approach does not capture the complexity of human interactions. The model is only limited to the person to person interaction in order to track the surveillance of disease and it also having performance issues with large realistic data. In this paper we propose, the combination of computational epidemiology and modern data mining techniques with their comparative analysis for the Swine Flu prediction. The clustering algorithm K mean is used to make a group or cluster of Swine Flu suspects in a particular area. The Decision tree algorithm and Naive Bayes classifier are applied on the same inputs to find out the actual count of suspects and predict the possible surveillance of a Swine Flu in a nearby area from suspected area. The performances of these techniques are compared, based on accuracy. In our case the Naive Bayes classifier performs better than decision tree algorithm while finding the accurate count of suspects.

Keywords: *Computational Epidemiology, Aerosol-borne Disease, Clustering, Predictive analysis.*

----------***----------

1. INTRODUCTION

The computational epidemiology is one of the interdisciplinary fields which is preparing and utilizing Computer models for analysis and controls the spatial, temporal diffusion of disease through populations [1]. The models may range from descriptive parameters, for example, the correlations within large databases, to general parameters, for example, analyzing, and computing the spread of disease via person to person interactions in a large population. The epidemic disease, including mankind, creatures and plants, and such diseases having different transmission properties. Similarly, the interactions depend on the disease and the populations, including certain factors like physical proximity to aerosol-borne disease, disease transmitted by sexual contacts, transmitted by mosquito borne diseases, etc [2]. The Mathematical epidemiology has traditionally relied on rate-based differential equation model. In SIR model, researchers partition a population into subgroups based on various criteria, such as demographic characteristics and disease states which are susceptible (S), infective (I), and recovered (R), and use the models to describe disease dynamics across these groups. This model is only limited with the factor of human interaction to study the disease outbreaks. The computational epidemiology enhances the scope of study to analyses disease outbreaks in huge realistic population with the help of modern data mining approaches.

In the computational epidemiology, researchers are harnessing computer power to crack the complicated mystery of how the epidemic diseases spread [4]. The epidemic diseases can be Swine Flu, cholera, jaundice, etc., which can spread under certain environmental conditions. In this paper, the disease Swine Flu and its related parameters are being studied by the help of different data mining approaches [5] [6] [7] [8]. The clustering algorithm k mean and Google map GUI is implemented as an initial phase of the prediction model. The clustering algorithms sort out the area wise suspects of Swine Flu and map them on a Google map. The mapping of suspects on Google map gives the actual idea about the surveillance of a swine flu. The clusters show the list of suspects in a particular area. The statistics of past data records are used to generate classification rules for the prediction algorithms. Such data contains record of victims of swine flu in a particular area and weather parameters at that time. Temperature, humidity and wind speed with their standard count are observed to analyze the current and possible outbreaks by swine flu. The past data records with such observations are used as a trained data set. Decision tree C4.5 algorithm and Naive Bayes classifier is applied on current clustered records. The predictive result shows the actual count of suspects from the primary identified suspects. It also tracks the new locations which will possibly be affected from the region of maximum count of suspects. The result also shows the possible time of emergence of disease in the new area. The web services of the national weather forecast are used to get the standard environmental parameters which are useful to predict the surveillance of Swine Flu. The comparative analysis of Decision tree algorithm and Naïve Bayes classifier has made on the basis of their performance in order to find out correct and actual count of suspects from the primary list. The

weather parameters are always varying so the decision tree algorithm shows limitation with the accuracy. In this prediction model as a probabilistic algorithm the Naïve Bayes classifier outperforms the decision tree algorithm.

2. RELATED WORK

The traditional approach for analyzing the epidemic disease outbreaks rely on past data records known as Epidemiology, which has poor scalability issues. Nevertheless, recent advances in data mining, computing technology, machine learning, and network science make it possible to develop new approaches for producing effective estimates of such research.

P. J. Dionne [1] has identified a simple problem with epidemiology work. The purpose is to describe how the computational work sorts out the complex system regarding epidemic diseases. If the computer simulation does not get the proper result, the moderator must have more than one iterations by revising the process. These iterative processes are used for the predictive analysis of epidemic diseases.

M. Naresh Kumar [10] has developed a robust and effective decision tree based approach for predicting dengue disease. His analysis is based on the clinical characteristics and laboratory measurements of the diseased individuals. He has developed and trained an alternating decision tree with boosting and compared its performance with decision tree C4.5.

S. Kosakovsky Pond, [11] has shown the study of HIV evaluation by using machine learning and network analysis approaches at the level of a single individual, populations of infected individuals, and the a geographic epidemic. Several examples of computational techniques are used for HIV-1 evaluation within-host, between-host and global epidemiological analysis. The author introduced the new computational tools for the analysis of next generation sequence data for studying population structure, molecular evolution, and viral correlates of clinical outcomes of HIV.

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu and Benyuan Liu [12] have described their approach to achieve faster, near real time detection and prediction of the emergence and spread of influenza epidemic, through continuous tracking of flu related tweets originating within United States. They have shown that applying text classification on the flu related tweets significantly enhances the correlation (Pearson correlation coefficient 0.8907) between the Twitter Data and the ILI rates from CDC.

S. Volkova and W. Hsu, [13] have made the system for animal disease outbreak analysis by automatically extracting relational information from on-line data. In this paper author aimed to detect and analysis infectious disease outbreaks by extracting information from unstructured databases. The information extraction component performs document analysis of animal disease. With the use of geospatial information and supports timeline representation of animal disease outbreaks, the visualization component plots extracted events into Google Maps.

H. Qin, A. Shapiro, and L. Yang [14] have designed and implemented a simple spatial computational multi agent model that can be used as a tool to analyze and predict the behavior of emerging infectious diseases. The multi agent spatial-temporal model contributes to Epidemiology, computational simulation and public health in several fields.

Shweta Kharya [15] has discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. This study paper summarizes various review and technical articles on breast cancer. The author focused on current research being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis.

Jyoti Soni, Ujma Ansari and Dipesh Sharma [16] have introduced different data mining techniques for the different databases, particularly in heart disease prediction. He has compared different data mining approaches namely Decision Tree, Naive Bayes classifier, KNN and Neural network and concluded that Decision Tree and Naive Bayes classifier outperforms other approaches after applying genetic algorithm to reduce the actual data size, sufficient for prediction of heart disease.

3. IMPLEMENTATION DETAILS

3.1 Mathematical Model

In this paper we are forming clusters of primary suspects of a Swine Flu and map those on Google map. The prediction algorithm is used to find out the count of actual suspects of Swine Flu from the different clusters. The input and their respective outcome are described below in form of set theory.

S= {I, F, O}

Set S contains the inputs, functions and their respective outputs which are described below in form of set theory.

1) Input:

I = {D, P, Sy}

D= Dataset
P= Weather Parameter
Sy= Symptoms

- *{D= di| di is a set of all data records of patients}*
- *{Sy= Si| Si is a set of Swine Flu symptoms }*
- *{P= Pi| Pi is a set of weather parameters}*

2) ***F= {K, D, B}*** is a set of algorithms used in the system.

***K*= K mean:**

- *{C= Ci | Ci is a randomly selected cluster center }*

- *Distance= (di, Ci) is a calculated distance between each data point and its centroid.*
- *Recalculate distance from new cluster center using:*

$$V_i = \frac{1}{C_i}\sum_{j=1}^{C_i} X_i \qquad (1)$$

***D*= Decision Tree:**

- *N → is a process to create node*
- *As = {m, s}*
- *m → multi way node creation*
- *s → simple node creation*
- *Mc → process to attach leaf node, to the majority class.*

***B*= Bayesian Theorem:**

- *X = {X1,X2,X3...Xn} → set of n attributes*
- *X → evidence*
- *H → hypothesis means*
- *Probability P(H | X) = P(X | H) P(H)/ P(X)*

3) Output:

***O= {Ok, Od, Ob}* is a set of outputs**

- *{Ok = k | k is a set of all primary suspects found by k mean algorithm}*
- *{Od = d | d is a set of actual suspects found by Decision tree algorithm}*
- *{Ob = b | b is a set of actual suspects found by Bayesian theorem}*

3.2 System Overview

The system architecture generally consists of three sections: Data collection, clustering and prediction.

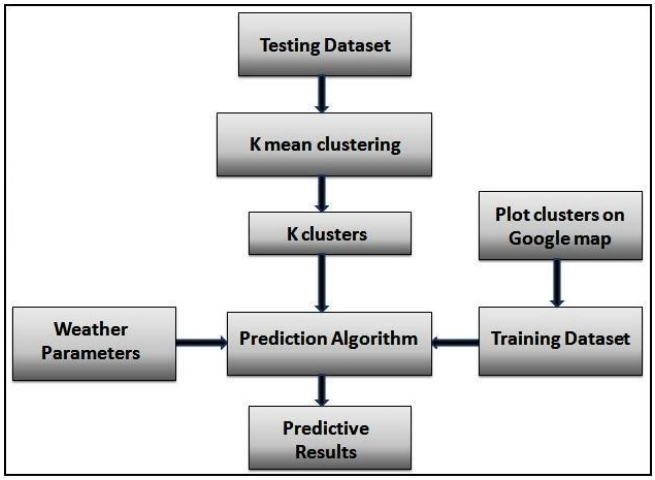

Fig-1: System Architecture

The data record involves list of patients, which shows the primary suspects of Swine Flu. The primary suspects are classified according to the symptoms by which the patient is suffering. The different clusters of primary suspects are formed and map them on Google map with their respective regions. The prediction algorithm then sorts out the actual count of suspects. The architecture of the proposed system is containing different modules explained as follows:

Testing Dataset: The testing dataset is a current data records which involves list of different patients who are suffering from primary symptoms of Swine Flu. The testing dataset is an input for the k mean clustering algorithm.

K mean Clustering: The clustering algorithm K mean is applied on the testing data records to find out region wise suspects of Swine Flu.

'K' Clusters on Google map: The different clusters of suspects are formed by the k mean algorithm and plot them on a Google map to specify the region of Swine Flu surveillance.

Training Dataset: The past data records are used as a training dataset. The prediction algorithm forms the prediction rules by the help of training dataset to sort out actual count of suspects from current clustered record.

Predictive analysis: The decision tree algorithm and Naive Bayes classifier is used to track the actual suspects from the current list of patients with the use of training dataset. The possible surveillance of Swine Flu from current area to nearby area is tracked, with the help of current environment parameters.

3.3 Algorithms

K mean: The clustering algorithms sort out the area wise primary suspects of Swine Flu and map them on a Google map. The mapping of suspects on Google map gives the actual idea about the surveillance of a Swine Flu [5].

- **Input:**
  T= Training Data Set
  Apply K mean algorithm on T.
  X = X1, X2, X3 ... Xn.
  V = V1, V2, V3 ...Vn be the center cluster.

Algorithm 1: K Mean for Making Cluster of Primary Suspects

1. Randomly select 'C' as a cluster center
2. Calculate the distance between each data point and cluster center.
3. Assign data point to the cluster where distance from cluster center is minimum of all cluster centers.
4. Reevaluate distance between number of data points and new obtained cluster center using,

$$V_i = \frac{1}{C_i}\sum_{j=1}^{C_i} X_i$$

5. Recalculate distance between number of data points and new obtain cluster center

Decision Tree algorithm (C4.5): The Decision Tree algorithm is a classification and regression algorithm for use in predictive modeling of both discrete and continuous attributes [17]. The algorithm is used to find out actual suspects from the primary list to analyze different areas with the maximum surveillance of Swine Flu.

- **Input:**
  T: Training Dataset.
  D: Set of Training Tuples.
  S = S1, S2, S3....., Sn Symptoms.
  C = C1, C2, C3....., Cn Classes.

Algorithm 2: Decision Tree for Finding Actual Suspects

1. ***ALGO C4.5*** *(D, S)*
2. *t = CreateNode()*
3. ***IF*** *D Є C then*
4. *return(t)* ***end IF***
5. *IF Si = φ then*
6. *return(t)* ***ELSE***
7. *Attribute Selection Method(Di, Si)*
8. *label(t) = MajorityClass(D,Target)* ***end IF***
9. ***FOREACH*** *outcome j*
10. ***IF*** *Dj Є D then disease found* ***end IF***
11. *IF Dj = φ then*
12. *t'= CreateNode()*
13. *label(t')= MajorityClass(D,Target)* ***ELSE***
14. ***ALGO C4.5*** *(Dj,Si)* ***end IF end FOR***
15. *return(t)*

Bayesian Theorem: Naive Bayes classifier is based on Bayes theorem. The algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes [7] [8]. The algorithm works on same input to find out actual count of suspects.

The Bayes theorem is as follows:

- *X = {X1, X2, X3...Xn}* be a set of n attributes.
- *X*, is considered as evidence.
- *H*, be some hypothesis mean.
- The data of X belongs to specific class C.
- *P(H|X),* the probability that the hypothesis *H* holds given evidence i.e. data sample *X*.
- According to Bayes theorem,

$$P(H|X) = P(X|H)\,P(H)/\,P(X)$$

3.4 Experimental Setup

In this system, Asp .Net framework is used on Windows operating system. It requires the basic hardware constraints. The web services of National Digital Forecast are used to analyze the environmental conditions. High speed Internet connection is required in order to show the clusters of suspects on the Google map.

4. DATA SET

The sample database is taken from government hospital Kolhapur, Maharashtra, in which list of all suspected Swine Flu patients in Kolhapur district in the year 2010 and 2011 is mentioned. This data contains the following information,

Sr. no, Name, Age, Sex, Date of admission, Date of discharge, Address, Date of sample collected. The fields from dataset which are important for predictive analysis are:

- Age: It observes the age group.
- Address: It shows the location of observed patients with respect to environmental conditions out there.
- Symptoms: The Symptoms of a patient.

The weather parameters are observed with their standard counts at which spreading chance of Swine Flu is maximum.

- Temperature: 18 to 23 ◦C
- Humidity: 26 to 34 ◦C
- Wind Speed: 0.005 to 0.2 m/s

5. RESULT SET

The suspects of Swine Flu are classified from the list of 1500 patients. These suspects are classified according to the symptoms by which the patients are suffering. As a clustered result these suspects are mapped on a Google map. The different clusters are formed, which has the cluster number and its mean, the area within that cluster and total number of suspects in that area.

Table -1: Clustering Results

Cluster	Mean	Total Suspects
C1	Area 1	150
C2	Area 2	180
C3	Area 3	300
C4	Area 4	340
C5	Area 5	160

The predictive result shows the actual count of suspects from the primary list of suspects. The result also finds out the possible surveillance of Swine Flu from the suspected area to nearby areas with possible time of surveillance.

Table-2: Predictive Results by Naive Bayes Classifier

Cluster	Mean	Actual Suspects	New Area	Time (hr)
C1	Area 1	130	Area 1.1	48
C2	Area 2	155	Area 1.2	36
C3	Area 3	280	Area 1.3	24
C4	Area 4	314	Area 1.4	40
C5	Area 5	146	Area 1.5	50

Table-3: Predictive Results by Decision tree algorithm

Cluster	Mean	Actual Suspects	New Area	Time (hr)
C1	Area 1	122	Area 1.1	48
C2	Area 2	140	Area 1.2	36
C3	Area 3	272	Area 1.3	24
C4	Area 4	300	Area 1.4	40
C5	Area 5	140	Area 1.5	50

Results are shown in the form of precision recall, which are calculated for comparative analysis. Precision and recall results are different for different set of clusters. Precision is the fraction of retrieved instances which are relevant while

Recall is the fraction of relevant instances which are retrieved [18] [19].

Precision= Correctly Identified Suspects/ Total Identified Suspects (2)

Recall= Correctly Identified Suspects/ Actual Suspects (3)

The comparative results in the form of precision and recall values are shown in following graphs considering number of clusters of Swine Flu suspects on the x axis.

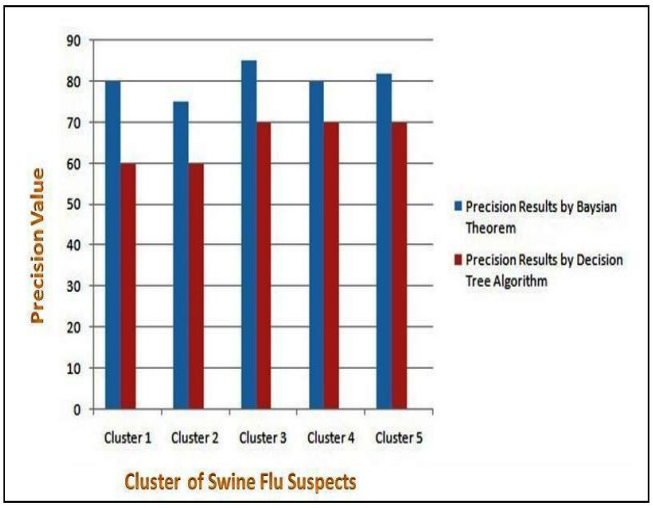

Chart -1: Graph for Precision Values

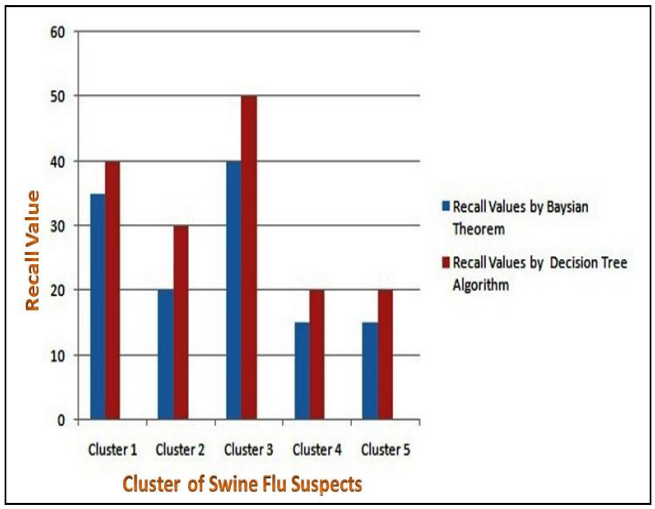

Chart -2: Graph for Recall Values

6. CONCLUSION

The proposed work has outlined how the data mining approaches enhance the computational thinking so that it can be effectively used to support public health epidemiology. The clustering algorithm k mean is designed to plot the patients according to the Swine Flu symptoms on Google map. Whereas the prediction algorithms are used to get an actual count of suspects and possible hazard. As comparative analysis the traditional decision tree algorithm such as C4.5 have been successful in generating classification rules but while processing with the varying weather parameters it creates complex decision tree structure. The Bayesian Theorem is a simple probabilistic algorithm that only deals with the probability of different possible outputs hence in case of changing weather parameters it performs better than decision tree algorithm. In future, this work can be extended by creating prediction model for the different epidemic and air born diseases by analyzing the social health and environmental parameters.

REFERENCES

[1]. P. J. Dionne, "Epidemiology," Biomedical Engineering, *IEEE Transactions on*, vol. BME-19, no. 2, pp. 126–128, 1972.

[2]. K. Bisset, A. Aji, M. Marathe, and W. chun Feng, "High-performance biocomputing for simulating the spread of contagion over large contact networks," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pp. 26–32, Feb 2011.

[3]. S. E. Christopher L. Barrett and M. V. Marathe, "An interaction-based approach to computational epidemiology," in *Proceedings of the Twenty- Third AAAI Conference* on *Artificial Intelligence (2008).*

[4]. K. Bisset, A. Aji, E. Bohm, L. Kale, T. Kamal, M. Marathe, and J.-S. Yeom, "Simulating the spread of infectious disease over large realistic social networks using charm++," in *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2012 IEEE 26th International conference*, pp. 507–518, May 2012.

[5]. J. Xie and S. Jiang, "A simple and fast algorithm for global k-means clustering," in *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*, vol. 2, pp. 36–40, 2010.

[6]. M. Cazzolato and M. Ribeiro, "A statistical decision tree algorithm for medical data stream mining," *in Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pp. 389–392, June 2013.

[7]. X. Yang, Y. Guo, and Y. Liu, "Bayesian-inference-based recommendation in online social networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, pp. 642–651, April 2013.

[8]. K. Sankaranarayanan and M. K. K, "Prediction of different dermatological conditions using nave bayesian classification," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, January 2014.

[9]. C. Griffin and R. Brooks, "A note on the spread of worms in scale-free networks," Systems, *Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 1, pp. 198–202, 2006.

[10]. M. N. Kumar, "Alternating decision trees for early diagnosis of dengue fever," in National Remote Sensing Centre (ISRO), India, June 2013.

[11]. S. Kosakovsky Pond, "Computational analysis of hiv-1 evolution and epidemiology," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pp. 60–63, Nov 2011.

[12]. R. L. S.-H. Y. Harshavardhan Achrekar, Avinash Gandhe and B. Liu, "Twitter improves seasonal influenza prediction," 2011.

[13]. S. Volkova and W. Hsu, "Computational knowledge and information management in veterinary epidemiology," in *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, pp. 120–125, May 2010.
[14]. H. Qin, A. Shapiro, and L. Yang, "Emerging infectious disease: A com- putational multi-agent model," *BioMedical Computing (BioMedCom), 2012 ASE/IEEE International Conference on*, pp. 28–33, Dec 2012.
[15]. D. S. Jyoti Soni, Ujma Ansari, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," in *International Journal of Computer Applications (0975 8887)*, vol. 17, March 2011.
[16]. S. Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease," in *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, Vol.2, No.2, April 2012.
[17]. F. Chen, X. Li, and L. Liu, "Improved c4.5 decision tree algorithm based on sample selection," in *Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on*, pp. 779–782, May 2013.
[18]. D. Kumar and Suman, "Performance analysis of various data mining algorithms: A review," *in International Journal of Computer Applications (0975 8887) on*, pp. 389–392, October 2011.
[19]. M. Doerr and M. Papagelis, "A method for estimating the precision of placename matching," *IEEE Transactions on*, vol. 19, pp. 1089–1101, Aug 2007.

BIOGRAPHIES

Mangesh J. Shinde received the BE degree in Computer Science and Engineering from Shiwaji University, Kolhapur in 2011 and pursuing ME in Computer Engineering from University of Pune at D. Y. Patil College of Engineering, Akurdi, Pune. His research interests include Data Mining and Information Retrieval.

Soudamini S. Pawar received the BE degree in Computer Science and Engineering from University of Pune in 2000 and ME in Computer Engineering from University of Pune in 2010 and has 10 years of teaching experience. She is currently working as Assistant professor at D. Y. Patil College of Engineering, Akurdi, Pune.