Swine flu prediction using data mining techniques: A Review

Dr. Gurmanik kaur[1], Er. Arwinder kaur[2]
[1]A. P., Deptt. of Electrical Engineering
[2]M.tech scholar, Deptt. of Electronics and Communication Engineering
Sant Baba Bhag Singh University, Khiala, Distt: Jalandhar, Punjab, INDIA

Abstract : Data mining is the process of extracting hidden interesting patterns from massive database. In terms of science, this industry is 'information rich' yet 'knowledge poor'. However, data mining with its various analytical tools and techniques plays a major role in reducing the use of cumbersome tests used on patients to detect a disease. This paper highlights the various data mining techniques used for predicting swine flu disease.

IndexTerms - Data mining, swine flu prediction.

I. INTRODUCTION

Swine flu is a contagious viral infection that is spread from pigs to human. It is also called swine influenza, pig influenza, hog flu and pig flu. When a swine flu patient coughs or sneezes, the millions of tiny droplets coming out of the nose and mouth contain the swine flu virus. A healthy person can catch H1N1 swine flu if that person comes in contact with these droplets or touch surface that infected person has recently touched. The swine flu also called the H1N1, is a new strain of common influenza virus. Most commonly, swine flu is of H1N1 influenza subtypes. Swine flu is on the rise and presently thousands of people are dying of it [1]. According to medical practitioners the differentiation with swine flu and normal flu is only possible through pathological tests and these laboratory tests are unnecessary for swine flu [2]. An exhaustive case study was carried out on the detection of swine flu wherein various doctors were interviewed and it was found that out of 10 cases of suspected swine flu, it was very difficult for the doctors to categorize the various flu only and only on the basis of symptoms [3], but there are ways by which this can be done. These are the various data mining techniques which can be applied into biomedical field to extract knowledge from the data. This knowledge can be used to increase revenue, reduce cost, or both [4]. The aim of this paper is to review different data mining techniques used for the prediction of swine flu disease through extraction of interesting patterns from the dataset using vital parameters.

II. RELATED WORK

There have been very few studies conducted on prediction of swine flu by means of data mining techniques .In this context, Thakkar et al. [4] have developed prototype intelligence swine flu prediction software (ISWPS). They have used 17 symptoms of Swine flu and collected 110 symptoms sets from various hospitals and medical practitioners. Naïve bayes classifiers used for classifying the patients of swine-flu had classified the patients into three categories (least possible, probable or most probable). It was reported that the efficiency of results can be further improved by increasing the number of data set, attributes or by selecting weighted features. Borkar and Deshmukh [5] have proposed naïve bayes classifier algorithm for diagnosis of swine-flu disease from its symptoms. The proposed approach showed promising results which may lead to further attempts to utilize information technology for diagnosing patients for swine flu disease. Shinde and Pawar [6] used clustering algorithm K mean to make a group or cluster of Swine Flu suspects in a particular area. The Decision tree algorithm and Naive Bayes classifier were 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 when compared, the naive bayes classifier performed better than decision tree algorithm in finding the accurate count of suspects. Amit tate et al [7]. proposed random forest algorithm for prediction of swine flu from input symptoms taken from patient or user. It was concluded that random forest algorithm maintained best accuracy as compare to other classification systems. The proposed algorithm is extendible to deal with mobile/online solutions to support patients as well for medical diagnostics.

III. DATA MINING TECHNIQUES USED FOR PREDICTION OF SWINE FLU

1. *Naïve Bayes*

Naïve bayes classifier is based on the bayes theorem. The bayes theorem as follows [9]:

$D=\{D_1, D_2, D_3 \ldots . D_n\}$ be a set of n attributes.
D, is considered as an evidence.
H, be hypothesis mean.
The data of D belongs to specific class C.
P(H\D), the probability that the hypothesis H holds given evidence i.e. data sample D.
According to Bayes theorem,

P(H\D) = P(D\H)P(H)/P(D)

Naïve bayes algorithm is built upon the strong assumption that the effect of a variable value on a given class is independent of the values of the other variables. This assumption is also known a class conditional independence.

2. C4.5

C4.5:-It is a well-known algorithm used to generate a decision trees. The basic idea of this tree is to built trees from a group of training data using the concept of information entropy [10]. It has following advantages over ID3 algorithm [11]:

-Handles training data with missing attributes
- Handles differing cost attributes
- Pruning the decision tree after its creation
- Handling attributes with discrete and continuous values

Let the training data be a set $S=s_1, s_2, s_3….$of classified samples. Each sample $S_i=x_1, x_2, x_3….$ is a vector that represents attributes of the sample. The training data is a vector $C=c_1, c_2, c_3…$ represent the class to which each sample belongs to. At each node of the tree, C4.5 selects one attribute of the data that most effectively splits data set of samples S into subsets that can be one class or the other. It is the normalized information gain that results from selecting an attribute for splitting the data. The attribute factor with the maximum normalized information gain is considered to make the decision. The C4.5 algorithm then continues on the smaller sub-lists having next highest normalized information gain [12].

3. Random Forest

Random Forest (RF) is an ensemble machine learning algorithm, which is best defined as a "combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [13]. The random forest algorithm generates k different training data subsets from an original dataset using a bootstrap sampling approach, and then, k decision trees are built by training the subsets. Finally, a random forest is developed from the decision trees. Each sample for the testing dataset is estimated by all decision trees and the final prediction result depends upon the votes of these trees.

IV.CONCLUSION

Data mining provides good results in disease diagnosis when appropriate tools and techniques are applied. Hence data mining is the promising field for healthcare predictions.The data mining has played in an important role in healthcare industry, especially in predicting various types of diseases. In order to obtain the highest prediction accuracy researchers need to design hybrid models to enhance the swine flu prediction. Furthermore, the efficiency of results can be further improved by increasing the size of data set.

V.REFERENCES

[1] Manish Sinha, "Swine flu", Journal of infection and Public Health, Vol. 2, Issue 4, 2009, 157-166.
[2] Vincent C. C. Cheng, Kelvin K. W. To, Herman Tse, Ivan F. N. Hung and Kwok-Yung Yuen, "Two Years after Pandemic Influenza A/2009/H1N1: What Have We Learned?", Clinical Microbiology Reviews, vol. 25 no. 2, 223-263, 2012.
[3] http://www.pref.aichi.jp/global/en/living/medical/influenza.htmlJ.
[4] https://journalofbigdata.springeropen.com/articles/10.1186/2196-1115-1-2.
[5] Thakkar. Hasan and Desai, "Health care decision support system for swine flu prediction using Naïve bayes classifier", International Conference on Advances in Recent Technologies in Communication and Computing.,IEEE,2010..
[6] Ms. Ankita R. Borkar*, Dr. Prashant R. Deshmukh," Naïve Bayes Classifier for Prediction of Swine Flu Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, Issue 4, April 2015, April-2015, pp. 120-123.K. Elissa, "Title of paper if known," unpublished.
[7] Mangesh J. Shinde1, S. S. Pawar2," Comparative Study Of Decision Tree Algorithm And Naive Bayes Classifier For Swine Flu Prediction", International Journal of Research in Engineering and Technology, eISSN: 2319-1163 | pISSN:2321-7308.
[8] Amit Tate1, Ujwala Gavhane2, Jayanand Pawar3, Bajrang Rajpurohit4 , Gopal B. Deshmukh5,"Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm",International Research Journal of Engineering and Technology ,Vol. 04 Issue: 06 | June -2017,Y.
[9] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," Cybernetics and Information Technologies, vol. 13, 2013.
[10] Dunham, M.H., "Data Mining: Introductory and Advanced Topics", Pearson Education Inc(2003).
[11] Xiaoliang, Z., Jian, W., Hongcan Y., and Shangzhuo, W., (2009) "Research and Application of the improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement (ICTM), Vol. 2,pp184-187.
[12] Lakshmi.B.Na, Dr.Indumathi.T.Sb , Dr.Nandini ,"Review A study on C.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy Procedia Technology "24 (2016) 1542 – 1549.
[13] Breiman L, "Random Forests, Machine Learning", Vol. 45, No. 1, pp. 5-32, 2001