Second Student Research Symposium (SRS), International Conference on Advances in Computing, Communications
and Informatics (ICACCI'13), 22 – 25 August 2013, Mysore, India

DOCAID: PREDICTIVE HEALTHCARE ANALYTICS USING NAÏVE BAYES CLASSIFICATION

ZEON TREVOR FERNANDO[*]**, PRIYANK TRIVEDI**[*]**, ABHINANDAN PATNI** [*]**, PRIYAL TRIVEDI**[†]

[*] VIT University, India
zeon.trevor@gmail.com,priyank.trivedi2010@vit.ac.in,abhinandan.patni.2011@gmail.com
[†] AMITY University, India
priyaltrivedi1404@gmail.com

Key words: machine learning, visual analytics, predictive analytics, healthcare.

INTRODUCTION

With the advancement in the field of medical research, there has been a copious increase in the data that is stored in the various public and private hospitals, clinics and other places of medical practice. This large store of data needs to be administered in a proper manner so that we can derive useful insights and conclusions using a proper analysis system. Such large amount of data is aptly handled through analyzing the unstructured or structured data by utilizing machine learning algorithms. Predictive analytics, a key component of machine learning algorithms, helps users make enhanced and supervised decisions. Visual Analytics is a tool to cost-effectively sort the exuberant and rapidly incrementing data in the field of medical research. It helps us cope with the assorted data in an organized manner, which the human brain would be able to visualize easily. This would in turn provide new innovative and potential results. This analytics not only provides structured data but also initiates structured thoughts in the mind of humans. As practitioners analyze certain anomalous situations, the process of visual analytics process would provide sorted relevant data related to it. This would in turn decrease the cost of maintaining huge amount of data.

As known, there are Electronic Health Records (EHR) which are created for medical assistance. Thus, with the onset of such records, there also arises a pressing need for an envisioned analysis of such data. Combining EHR data with visualization and a clinical decision support system would thus help come up with inferences which otherwise might be missed through manual work. Thus, there is an urgent need for the combination of thoughts of an IT professional and a medical clinician.

In this paper, we employ the machine learning techniques to predict diseases for a patient using the symptoms described by them. We utilize the Naïve Bayes Classification [1] algorithm to develop the predictive analytics system and predict aptly the diseases for the patients. In present system aids doctors with five specific diseases with the elaborated feature based classification. In order to aid the doctors and clinicians graphically, the system incorporate visual analytics technique using Cytoscape Web [2]

The rest of the extended abstract is organized as follows. The next section describes the related work in the field of machine learning and visual analytics and motivation behind this work. It is followed by Naïve Bayes Classification implementation to predict diseases for a patient and visual representation for the purpose of graphical analysis.

MOTIVATION AND RELATED WORK

Due to the clinical data overflow, there is a need to include software tools into the field of medical research. The rise in the medical records provides an opportunity to analyze, reconstruct and tread on a path leading to efficient handling of data. Clinicians and other practitioners may not be able to construe the information the medical records contain. The predictive analysis could be used to come up with defined solutions for the situations that are transparent from the data. While the unstructured data would create bemused situations for anybody to understand, predictive analytics is thus, a comprehensible and transparent approach which is apparent to both the patients and the clinicians. The patients can be satiated with the predictive analysis which could not be solely provided by a clinician's prescription.

Ross et al [3] used the visual analytics approach for syndromic-surveillance of data. Many data surveillance systems have been researched in the recent times such as Early Aberration Reporting System(EARS)[4] , Electronic Surveillance System for the Early Notification of Community based Epidemics (ESSENCE II) [5] and Biosense [6], but none of these tools have been extremely successful in catering to all the needs of clinicians, doctors and other practitioners.

While deploying machine algorithms based on the symptoms of the disease incorporated in the system our work incorporates the sophisticated tools for visual analytics, particularly of high dimensional data, which is a perfect add on to the inferences drawn from the machine learning multi-class classification algorithm. Sophisticated machine learning techniques have often been applied by various organizations and are increasingly interested in healthcare policies. A plethora [6,7] of machine learning techniques have been applied to determine transparent, unbiased healthcare policies by government and organizations across the world.

Several works [9,10] have employed machine learning techniques pursuance with visual analytics. Active Learning Intent Discerning Agent (ALIDA) [10] is a framework which has paved the way for future research involving data visualization. Our work employs the machine learning algorithm along with visual analytics to address the issues related to analytics.

PREDICTIVE ANALYTICS

We have developed a disease prediction system, DOCAID, for Typhoid, Malaria, Jaundice, Tuberculosis and Gastroenteritis, which bases its diagnosis on the patient symptoms and complaints using Naive Bayes Classification.

The different diseases we have taken into consideration along with their respective symptoms are shown in Table 1. The input for the system consists of 11 features/input variables which are the patients symptoms or complaints observed either by the doctor or the patient itself (i.e. $X_1, X_2,..., X_{11}$). Here we can view the features as an 11-dimensional vector X, in which each of X_i belongs {0, 1} where 0 indicates no and 1 indicates yes depending upon the input. The outputs of the system can be considered as *y* which contains various diseases as values.

According to the Naives Bayes classification, the probability for a particular disease(*y*) given its symptoms can be estimated using conditional probability model represented as,

$$p(y = \text{disease} \mid X_1, X_2, \ldots.., X_n) \qquad (1)$$

Using Bayes theorem the probability of a disease can be written as,

$$p(y = \text{disease} \mid X_1,....,X_n) = \frac{p(y = \text{disease})\, p(X_1,....,X_n \mid y = \text{disease})}{p(X_1,....,X_n)} \qquad (2)$$

For the purpose of classification, we are interested only in the values produced by the numerator as the denominator does not depend on *y* and hence it is effectively constant. Thus the probability for a particular disease can be rewritten using the chain rule of repeated application of conditional probability,

$$p(\ y = \text{disease} \mid X_1, X_2, \ldots.., X_n) = p(\ y = \text{disease})\ p(\ X_1,\ldots.,X_n \mid y = \text{disease})$$

$$= p(\ y)\ p(\ X_1 \mid y) p(\ X_2 \mid y, X_1) \ldots. p(X_n \mid y, X_1, X_2 \ldots\ldots, X_{n-1})$$

Since it is Naives Bayes classification, it is assumed that every feature X_i is conditionally independent for a given output *y*. This implies that,

$$p(y = \text{disease} \mid \mathrm{X}) = p(y = \text{disease}) \prod_{i=1}^{n} p(X_i \mid y = \text{disease})$$

Here we estimate the values of *p*(*y* = disease) and *p*(Xi | *y* = disease) using the training data set which consists of various patient records, where each individual record consists of the respective symptoms (X) and the disease predicted (*y*). These values are called as the parameters of the Naïve Bayes classifier. Since each of the values of the feature vector X belongs to {0, 1} we can use the maximum likelihood estimates of the probabilities which is represented below,

$$p(\ Xi \mid y = disease) = \frac{\text{Number of entries corresponding to } y = \text{disease}}{\text{Total number of entries in the dataset}} \qquad (4)$$

$$p(\ Xi \mid y = disease) = \frac{\text{Number of entries corresponding to } Xi \text{ where } y \text{ is disease}}{\text{Number of entries corresponding to } y = \text{disease}} \qquad (5)$$

S No	Disease	Symptoms
1	Typhoid	Fever, Headache, Nausea, Anorexia, Diarrhea
2	Malaria	Fever, Headache, Chills, Sweating, Vomiting
3	Jaundice	Fatigue, Headache, Yellow Eyes and Skin, Anorexia, Nausea
4	Tuberculosis	Blood in Cough, Fatigue, Fever
5	Gastroenteritis	Nausea, Vomiting, Soreness of Throat, Fatigue

Table 1 Diseases along with their symptoms

S No	Features/ Input Variable	Symptoms Indicated by the variables
1	X_1	Fever
2	X_2	Headache
3	X_3	Nausea
4	X_4	Anorexia
5	X_5	Diarrhea

6	X_6	Chills
7	X_7	Sweating
8	X_8	Vomiting
9	X_9	Yellow Eyes and Skin
10	X_{10}	Blood in Cough
11	X_{11}	Soreness of Throat

Table 2: Features representing their symptoms corresponding to equation (1)

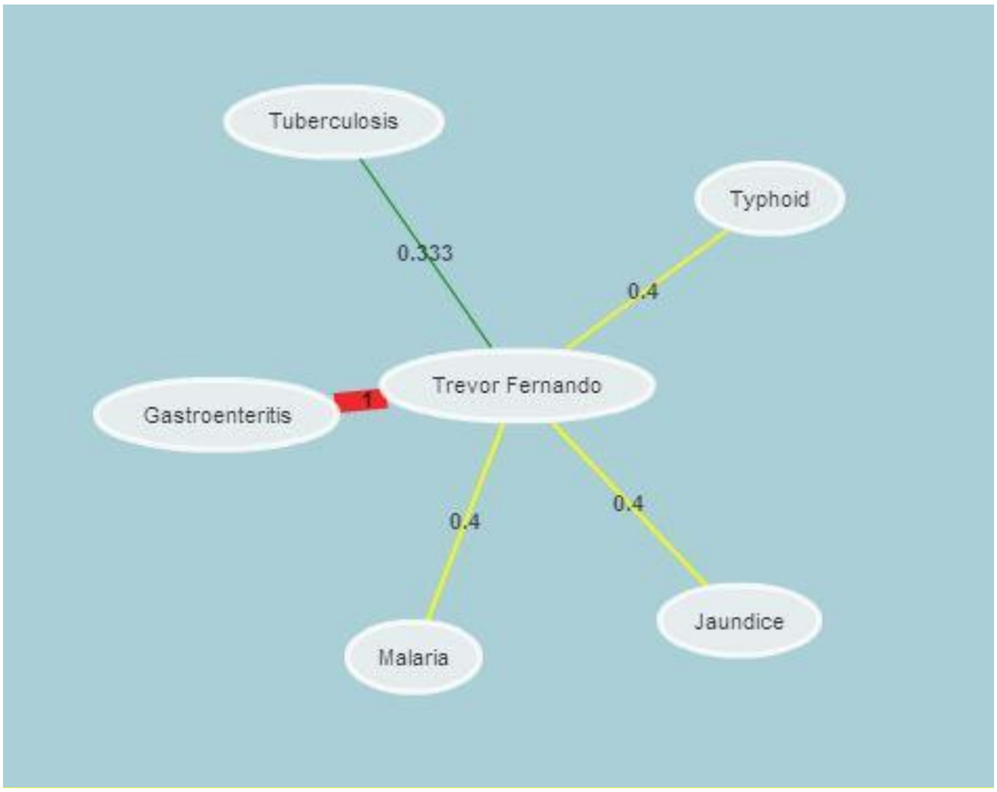

Figure 1: Visual representation of diseases through Naive Bayes Theorem

The Figure 1 depicts the graphical representation of p(y = disease | X) of a patient. We use a data set of 300 records out of which 80% is used as the training set and 20% as the test data set. After simulating the system on the test data set we get an accuracy of 91 percent. In order to perform visual analytics, our system uses Cytoscape Web [2, 8], an open source, web based network visualization tool which helps the clinicians and practitioners fathom the predictive analytics in a more supervised manner. Such a visual representation not only enables the clinicians to observe minute details which might be missed in general physical operation but also make certain insights which may be helpful in biomedical sciences as well.

CONCLUSION AND FUTURE WORKS

In this paper, we employ machine learning techniques, specifically Naïve Bayes classification algorithm to develop a predictive analytics system in the healthcare industry. We test the system with sample data set of records to produce results with 91 percent accuracy. The results obtained show that the Naïve Bayes classification algorithm works appropriately over the clinical data to produce results matching with the data set. Thus, we can conclude that machine learning algorithms are beneficial for the healthcare industry and would largely benefit them to analyze the Electronic Health Records. In order to produce graphical results showing the probable diseases enabling the practitioners analyze their patients better.

We are currently working on the neural network based predictive model and principal component analysis of the feature extraction of the diseases incorporated in the system. We are planning to implement a neural network model because of its various capabilities like non-parametric, non-linearity, input-output mapping and adaptability which makes it a natural choice for modeling complex medical predictive analytics.

REFERENCES

[1] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103-130, 1997

[2] Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T, "A travel guide to Cytoscape plugins", Nature Methods. 2012 Nov;9(11):1069-76. doi: 10.1038/nmeth.2212. Epub 2012 Nov 6.

[3] Ross Maciejewski,Ryan Hafen, Stephen Rudolph, George Tebbetts, William S. Cleveland David S. Ebert, Shaun J. Grannis "Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques", IEEE Computer Graphics and Applications - Special issue on sketching tangible interfaces augmented reality on mobile phones archive Volume 29 Issue 3, May/June 2009 Pages 18-28

[4] Early Aberration Reporting System (EARS), Centers for Disease Control and Prevention.

[5] Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J, "A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II)", J Urban Health. 2003 June; 80(2 Suppl 1):i32-42.

[6] Calvin White, Melicia Brown, Wayne Johnson, Scott Norville, "BIOSENSE: Use of Biosurveillance Software to Detect Clusters of Visits for Diarrhea Fort Worth", 2010.

[7] D Biggs, B Ville, E Suen. A Method of Choosing Multi-way Partitions for Classification and Decision Trees. *Journal of Applied Statistics*, 18(1):49-62

[8] Christian T. Lopes, Max Franz, Farzana Kazi, Sylva L. Donaldson, Quaid Morris, and Gary D. Bader. Cytoscape Web: an interactive web-based network browser, Oxford Journals. Bioinformatics 2010; 26(18): 2347–2348; doi: 10.1093/bioinformatics/btq430

[9] G Cass. An Exploratory Technique for Investigating large quantities of categorical Data. *Applied Statistics*, 29(2):119-127

[10] Green T.M., Maciejewski R., DiPaola S. "ALIDA: Using machine learning for intent discernment in visual analytics interfaces. Visual Analytics Science and Technology", 2010, page(s): 223-224.

[11] George H. John and Pat Langley (1995),"Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345.

[12] Morgan Kaufmann, San Mateo. Kononenko, I. 1990. Comparison of inductive and Naïve Bayesian learning approaches to automatic knowledge acquisition. In Wielinga, B., ed., Current Trends in Knowledge Acquisition. IOS Press.

[13] Caruana, R., Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms", Proceedings of the 23rd international conference on Machine Learning.