An Empirical Model of Supervised Learning for Electronic Health Records

S. Sai Kavya
M.Tech Scholar,
Department of CSE, AITAM, Tekkali

Gorti Satyanarayana Murty
Professor
Department of CSE, AITAM, Tekkali

Abstract:-Examining the health records is an interesting research issue in the field of medical knowledge and data engineering. Electronic health records are basic sources which maintains the patient health information that contains vitals, demographics and encounter or episode information. We propose an empirical model of classification approach for analyzing the test samples with training samples of electronic health records. We use improved supervised learning model to classify the health records. Our proposed model gives efficient results than traditional approaches.

******

I. INTRODUCTION

Nowadays, So Many Clinics maintain the medical information of the patient in digital format i.e. personal health records. It is an efficient and better way to understand the patient diagnosis, procedures or any kind of reports. Usually real time entities of personal health records may not be perfect and may not have complete information and we are unable to collect automatically. Generally this information can be provided by the domain experts who deal with personal health records. When some vital information like laboratory tests and medications are collected and diagnoses performed some conditions. . However, the process of class label identification are very time and space consuming in medical domain when data is complex.Usually most of the attributes are dependent based on the class labels and respective practical deployment

To address this issue , various researchers proposed various supervised and unsupervised learning models. Cluster based model may not give the optimal results because they can'tidentify the new testing sample, they can only identify the existing behavior of the objects. The major issue with classification models is consistency of data. Data should be valid and should not containnull values. Dataset usually consists of diagnosis information or vital information. Usually we have two types of diagnosis codes ICD9 and ICD10.ICD 10 is the latest diagnosis information which used to identify the diagnosis of the patient.

To address this current issue, previous working models proposes good machine learning frameworks in which it uses class labels for separation with soft label information calculated and it is based on the attributes of the dataset. Most of the domain experts suggests soft label information that computes actual result instead of random classification. Binary classification models assign label to the attributes based on the probability. Generally, classification models preprocess the dataset, and computes initial probability based on the soft label information and computes probability with respect to all attributes of the testing dataset.

Although Electronic Health Records (EHRs) have attractive and increasing research attention in the data mining and machinelearning communities in recent years,

Knowledge extraction from massive volumes of data especially from the medical domain is not well explored area. It is not simple to analyze the testing sample behavior in medical domain due to risk factor. Initially total dataset attributes can be unlabeled and major issue with medical domain is, accuracy because automation of results are very important incase of patient evaluation during the diagnosis of the patient. Various models proposed to handle labeled and unlabeled but every model has its own advantages and disadvantages.

II. RELATED WORK

Traditional models like semi-supervised heterogeneous graph-based methods are complex to classify the unknown labels. Tree traversing is very complex to search an element or position to insert or get element.Evidence based risk prediction model is a complex procedure in mining longitudinalhealth examination records. To handle the non-homogeneous data, traditional model proposes a graph based Heterogeneous model. It is complex to analyze personal health records or electronic examination records due to different types of data.Various cluster-based models proposed by researchers. Group of categorical information is complex because it is very difficult to analyze the new sample of data through existing samples, even though semi supervised learning models utilizes both labeled and unlabeled data, they are not efficient and high-risk factor [4][5].

The major contribution of the paper is to identify the decision labels of the testing sample by analyzing with training samples with computation of posterior probability. Missing value injection is an interesting feature in our current model. Usually classification model should have all attributes to validate the testing samples otherwise it fails to analyze the decision label.Our model improved the classification with improved naïve Bayesian classification model and artificial missing value injection is an important feature which improves the reliability of the classification model[6][7].

Even though various traditional models proposed by various authors from years of research, every model have its own advantages and disadvantages. Traditional graph based approach is complex to traverse and manage the structure, takes time while adding node element for large set of data. Time and space complexity are the major issues in the traditional model. Artificial measures are required to insert global missing values instead of constant insertion of the values. Domain experts provide those missing values to maintain the dataset consistency prior to the classification of the testing dataset. Each set of attributes moves to different categories

A two-stage evaluation strategy to evaluate theproposed SHG-Health.In the first stage, we evaluated an algorithm's ability toidentify high-risk cases, regardless of their disease category.It can be understood as testing the algorithm's ability topredict mortality risk. All predicted disease cases wereregarded as predicted positive cases and the true diseasecases were regarded as (true) positive cases. As there is noground truth for the negative or healthy cases, we used measuresthat focus on positive predictions, namely precision,recall/sensitivity, and F-score. While precision measureshow correct the positive predictions of an algorithm are,recall shows its ability to catch the positive cases. F-scorecalculates the harmonic mean between precision and recall.In the second stage, we looked into a method's abilityto predict the correct disease class given that it predicted acase to be in one of the high-risk classes. This is a conditionalevaluation that only considers cases that were predicted asone of the disease classes. Macro-precision and macro-recallmeasures were used. Macro-averaging takes the average ofprecision or recall scores computed from individual classesIt assumes that all classes are equally important, so thatthe performance of minority classes can be reflected in themacro-averaged scores[9][10].

III. PROPOSED WORK

We propose an empirical model of supervised learning model for classification of electronic health records. We analyze the attribute values of the physical health records and assign the class label. Class label can be assigned to health record based on the posterior probability of the record and improves the traditional classification issue like missing values, missing data can be imputed based on the average value of the row or global cost. Patient vital information can be considered as testing sample and forwarded to training dataset to classify efficiently. Our algorithm handles missing value problem in the classification and multi-class classification problem with both labeled and unlabeled data and risk prediction is simple and efficient.

Health records in electronic format can be considered as personal health records. It contains the patient demographic information like name, gender and contact details. While doing encounters to the patient ,they gather the vital information like sugar levels, blood pressure and etc.. This measure can be usually analyzed by the physician. These measures can be analyzed with supervised learning model without user intervention.

Missing value computation with Artificial Measures:

Generally in traditional model of classification,classification fails if there is an empty or null value in the vital information because complete posterior probability can be calculated based on the prior probability of the vital information or health records of the patient. In our model we provide an average constant value based on the artificial measures, it maintains some decision rules and injects the respective value in to the missing field and continues the process of classification.Here, considered vital information are temperature, respiratory rate, pulse and blood pressure. The sample training dataset as

Temperature (C)	**Respiratory rate(breathes per min)**	**Pulse(beats per min)**	**(BP)Systolic pressure**	**(BP)Diastolic pressure**	**Status**
32	**18**	**80**	**120**	**80**	**T**
40	**18**	**90**	**110**	**90**	**F**
35	**70**	**70**	**100**	**100**	**F**
40	**30**	**60**	**130**	**90**	**F**
29	**19**	**90**	**105**	**80**	**T**

The above measures shows list of training dataset standard measures, we use some sample set of training measures to analyze testing sample behavior by computing the posterior probability. If testing sample has complete attribute values, it is easy to compute the measures of the testing sample but if we miss the some attribute value, we need to replace with artificial measures with decision rules. Status variable indicates that status of the patient is normal or not.

Ex: if temp >=40 && pulse <=120 && pulse>=80 then

Systolic pressure :=avg(pulse+temp)/2

To analyze the vital information of the patient we use improved Bayesian classifier. It analyzes the behavior of the sample with existing training datasets of medical information (vital information which is calculated previously).

Novel Bayesian classifier is defined as series 'C' of classes and a set 'A' of attributes. A common class that belongs to'C' and it is denoted by c and a common property or attribute belongs to 'A' as A_i. Let us assume a database 'D' with a series of properties or vita information values and decision value of the set. The training dataset of the Naïve

Bayesian Classifier consists of initial probability followed by conditional probability of all attributes and posterior probability

Let us consider a sample dataset which consists of various health records vital information blood pressure, sugar levels etc. class labels, and it is considered as attribute set C ($c_{1,}c_c$,……$c_{n)}$ for training dataset and computestotal probability for positive decision label and negative decision label and followed by the computation of posterior probability with respect to all attributes ,finally calculate the probability of the diagnosis based on vital information.

Algorithm to classify the vital information of the patient:

Sample space: set of patient health record

H= Hypothesis that X is an vital information

P(H/X) is our confidence that X is an vital information or health record

P(H) is Prior Probability of H, i.e., the probability that any given data sample is an agent regardless of its behavior

P(H/X) is based on more information, P(H) is independent of X

Estimating probabilities:

P(X), P(H), and P(X/H) may be estimated from given data

Bayes Theorem

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

Steps Involved:

1. Each data sample is of the type

X=(x_i) i =1(1)n, where xi is the values of X for attribute A_i

2. Suppose there are m classes C_i, i=1(1)m.

X belongs to $C_i$iff

P(C_i|X) > P(C_j|X) for 1<= j <= m , j!=i

I.e. BC assigns X to class C_i having highest posterior probability conditioned on X

The class for which P(C_i|X) is maximized is called the maximum posterior hypothesis.

From Bayes Theorem

3. P(X) is constant. Only need be maximized.

- If class prior probabilities not known, then assume all classes to be equally likely
- Otherwise maximize P(X|C_i)P(C_i)

P(C_i) = Si/S

Problem: computing P(X|C_i) is unfeasible!

4. Naïve assumption: attribute independence

$P(X|C_i) = P(x_1,\ldots,x_n|C) = \pi P(x_k|C)$

5. In order to classify an unknown sample X, evaluate P(X|C_i)P(C_i) for each class C_i. Sample X is assigned to the class $C_i$iffP(X|C_i)P(C_i) > P(X|C_j) P(C_j) for 1<= j < m, j!=i

In the above model it computes the initial probability with set of attributes in the training samples dataset, followed by conditional probability and then decides the class label with posterior probabilities of the sample.

IV. CONCLUSION

We have been concluding our current research work with efficient supervised learning model like classification approach to analyze the samples or input behaviors of the personal health record vital information by forwarding to the large training dataset to compute the posterior probabilities of the samples. We improve the major drawback of missing attribute value of the classification with artificial measures instead of general constant. Our proposed model gives efficient results than traditional models.

REFERENCES

[1] “Cognitive impairment assessed at annual geriatric health examinations predicts mortality among the elderly,” Preventive Medicine, byC. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li,.

[2] “An Efficient TCM Supervised Learning Approach With Naïve Bayesian Classifier“ byNaveen Gosu , VakacharlaDurgaprasadarao , N. Tulasi Radha

[3] “Health checks for the over-65s,” http://www.nhs.uk/Livewell/

[4] “General health checks in adults for reducing morbidity and mortality from disease (Review),” by L. Krogsbøll, K. Jørgensen, C. Grønhøj Larsen, and P. Gøtzsche

[5] “A relative similarity based method for interactive patient risk prediction,” by B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang

[6] “Ranking-based classification of heterogeneous information networks,” by M. Ji, J. Han, and M. Danilevsky,

[7] “A general graph-based semi-supervised learning with novel class discovery,” Neural Comput. Appl., vol. 19, pp. 549–555, 2010. By F. Nie, S. Xiang, Y. Liu, and C. Zhang,

[8] “Compact graph based semi-supervised learning for medical diagnosis in alzheimer’s disease,” by M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang,

[9] “Mining Health Examination Records—A Graph-Based Approach” by Ling Chen, Xue Li, Member, IEEE, Quan Z. Sheng, Member, IEEE, Wen-Chih Peng, Member, IEEE, John Bennett, Hsiao-Yun Hu, and Nicole Huang,

[10] “Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network,” by Y. Li and J. C. Patra