ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 5, Issue 4, April 2016

Prediction of Probability of Chronic Diseases and Providing Relative Real Time Statistical Report using data mining and machine learning techniques

Dhruvi Ragesh Parikh, Yavnika Rajendra Bhagat, Nutan Ramesh Ghanwat.
Department of Computers,
K.J. Somaiya College Of Engineering, Mumbai - 400077, Maharashtra, India.

***Abstract*—Chronic diseases are growing to be one of the prominent causes for deaths worldwide. Fatality rates owing to chronic diseases are accelerating globally, growing across every region, encompassing all socio-economic classes and thus contributing to financial burden. According to the World Health Report, by 2020 their contribution is estimated to rise to 73% of all deaths and contribute to 60% of the global burden of disease. Moreover, 79% of the deaths owing to these diseases occur in the developing countries. Hence, in this paper we present a way which focuses on aiding people get an estimate or probability of having any of the chronic diseases to prevent any future risk at an earlier stage, by considering health test report values and symptoms faced. We employ data mining and machine learning techniques, by using a hybrid intelligent system comprising Naïve Bayesian and Decision tree algorithms in a complimentary fashion to ensure accuracy and efficiency in prediction and provide real time summarized update regarding the particular disease.**

I. INTRODUCTION

The paper comprises of two modules. The first module focuses on proposing a system which can estimate the probability of a user having any of the chronic diseases under consideration. The chronic diseases for which empirical analysis is performed are Chronic Kidney Disease, Coronary Artery Disease, and Hepatitis. The system will aid the user in diagnosing any potential chronic and hence, possible fatal threat at an early stage.

Data mining is discovering hidden knowledge from large amount of data. The problem statement for probability estimation is a classification problem since, the target class is categorical (discrete). Classification is data mining technology which learns relationship(creates a mapping) between set of features (attributes) and final class(label) using supervised learning technology that is its training data set has known label(target class) and predicts label(target class) for set of

features(attributes) whose label (target class) is unknown. Efficiency of a particular classifier varies from application to application. Hence, we cannot claim that one classifier is best for all problems. Using an ensemble (combination) of multiple classifiers proves to generate most accurate results. Here, we employ stacking which is ensemble method for combining predictions of two classifiers namely, Decision Tree and Naïve Bayesian in which prediction of the two classifiers is given to meta-classifier to get final prediction.

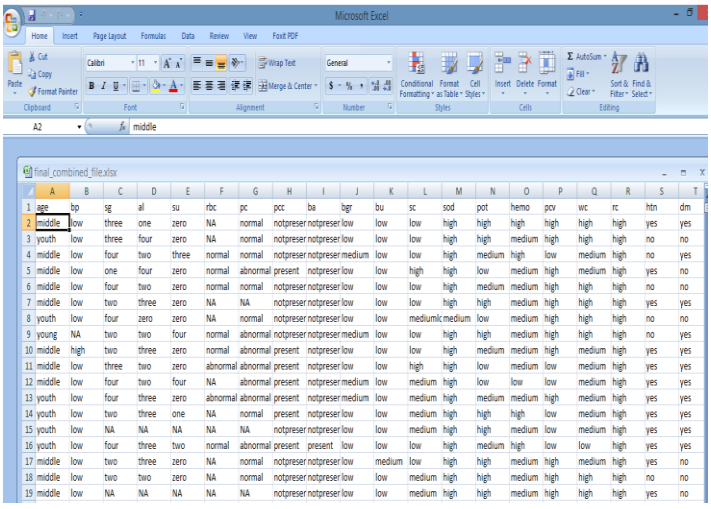
The second module focuses on providing the user dynamic (real-time) overall summarized report and statistics for the disease under consideration. It provides real-time (on the fly) summarized health news, dynamic summarized disease summary (About, causes, treatment, etc.) and dynamic statistics (number of persons inflicted with disease worldwide, mortality rate, success rate, etc.) for which web scraping and text summarization techniques are employed.

II. PROPOSED APPROACH

A. Constructing multi-label classifier for predicting disease probability

1. Gathering training dataset

For construction of classifier and for testing classifier we have used data set from UCI machine learning repository. Three datasets are used: Chronic Kidney Disease, Coronary Artery Disease, and Hepatitis



All Rights Reserved © 2016 IJSETR

2. Processing missing values:

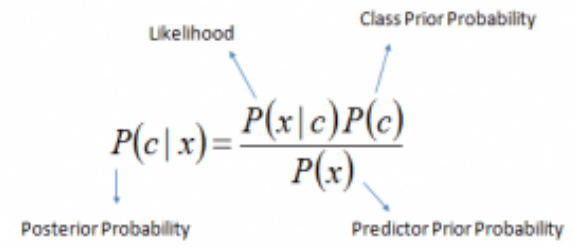
A missing value can represent different things in the data. There can be varied reasons for the presence of missing values such as unavailability, inapplicability, absence of the event, data entry error. Missing values may have effect on prediction of classifier hence it is necessary to process missing values. Processing missing values means replacing missing values with values such that effect of missing values on prediction of classifier minimize. There are broadly two categories of algorithms for processing missing values – single imputation and multiple imputations. In this paper we use multiple imputations in which instead of single imputation we impute multiple values for a single missing value. In summary, the missing value is filled by a value obtained by analyzing other similar values. We have implemented multiple imputations in R language using Amelia library.

3. Brief introduction to individual heterogeneous base classifiers and combination technique for constructing ensemble of heterogeneous classifiers

3.1 Naïve Bayesian Classifier:

Naive Bayesian classification technique uses Bayes' Theorem. Naive Bayesian algorithm is easy to employ and has proven to be a sophisticated and competent method in comparison to varied complex algorithms. It assumes the independent predictor behaviour.

All of the attributes in the dataset independently contribute to the probability of the person having a Chronic Disease. It considers that the presence of a particular predictor in a class is not related to the presence of any other predictor. For instance, Blood Pressure predictor value is not dependant on Red Blood Cells predictor value. All of these attributes independently contribute to the probability of the person having Chronic Kidney Disease.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Working of Naive Bayesian Algorithm

Step I: Use the data set and converts it into a frequency table

- Step II: Find the probabilities and Create Likelihood table.
- Step III: Now, apply the equation to calculate the posterior probability for each class. The posterior probabilities of all classes are compared and the one with the maximum posterior probability is the outcome of prediction.

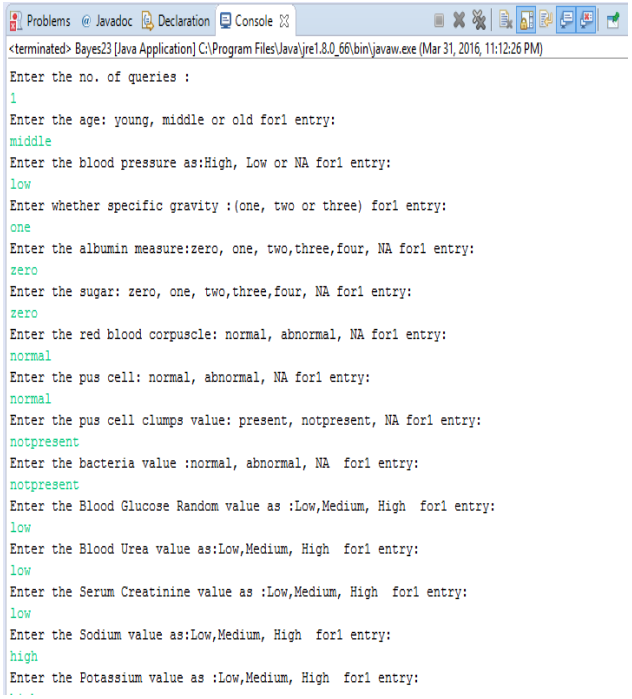
Advantages:

a. Fast
b. Robust to irrelevant attributes.
c. Induced classifiers are easy to interpret.
d. Uses evidence from many attributes.

Disadvantages:

a. Assumes independence of attributes
b. Low performance ceiling on large databases

Below diagram shows sample input and output for Naïve Bayesian Algorithm:



All Rights Reserved © 2016 IJSETR

```
ckd likelihood probability: 5.755142424850702E-7
not ckd likelihood probability: 5.956737830643961E-6
ckd probability: 8.81034957074753
not ckd probability: 91.18965042925248
```

```
ckd likelihood probability: 5.755142424850702E-7
not ckd likelihood probability: 5.956737830643961E-6
ckd probability: 8.81034957074753
not ckd probability: 91.18965042925248

---FINAL RESULT---
not ckd
```

```
ckd likelihood probability: 1.4517822632389132E-6
not ckd likelihood probability: 3.268992552495605E-14
ckd probability: 99.9999977482901
not ckd probability: 2.2517098890533925E-6

---FINAL RESULT---
ckd
```

3.2 J48 classifier (Decision Tree):

J48 classifier uses C4.5 decision tree algorithm as base algorithm. This algorithm generates classification trees whose leaf node represents final class labels and internal node represents features of our problem which contribute to prediction and each branch associated with feature node represents possible number of outcome of features. Path from root to leaf node becomes classification rule.C4.5 algorithm uses top-down greedy approach for generating tree. From set of available feature at node, we select feature for representing that node which best contributes to prediction, we find best feature using entropy and information gain calculation and based on possible outcomes of best selected feature we do splitting and recursively apply this procedure until leaf node is identified. Decision tree algorithm is a fast classifier for larger data sets. Also as compared to other algorithms it provides competent efficiency.
Below diagram shows some set of classification rules generated by C4.5 algorithm on data set provided by UCI for Chronic Kidney Disease.

Advantages:

a. Fast
b. Segmentation of data
c. Decision trees perform feature selection

Disadvantages:

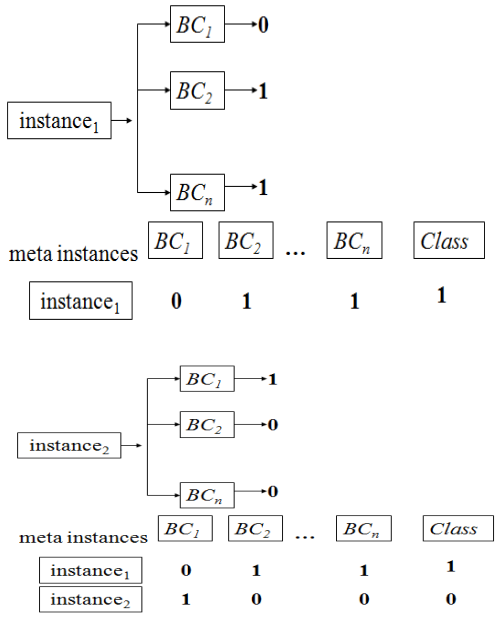
a. Fragmentation as number of splits becomes large
b. Interpretability goes down as number of splits increase

```
if hemo <=12.9 than ckd
if hemo >12.9 if sg two than ckd
if hemo >12.9 if sg three than ckd
if hemo >12.9 if sg NA if sc <=1 than nockd
if hemo >12.9 if sg NA if sc >1 than ckd
if hemo >12.9 if sg five than nockd
if hemo >12.9 if sg one than nockd
if hemo >12.9 if sg four if sc <=1.2 if al two than nockd
if hemo >12.9 if sg four if sc <=1.2 if al three than nockd
if hemo >12.9 if sg four if sc <=1.2 if al NA than nockd
if hemo >12.9 if sg four if sc <=1.2 if al zero than nockd
if hemo >12.9 if sg four if sc <=1.2 if al one than nockd
if hemo >12.9 if sg four if sc <=1.2 if al four than ckd
if hemo >12.9 if sg four if sc <=1.2 if al five than nockd
if hemo >12.9 if sg four if sc >1.2 if age <=66 than ckd
if hemo >12.9 if sg four if sc >1.2 if age >66 than nockd
```

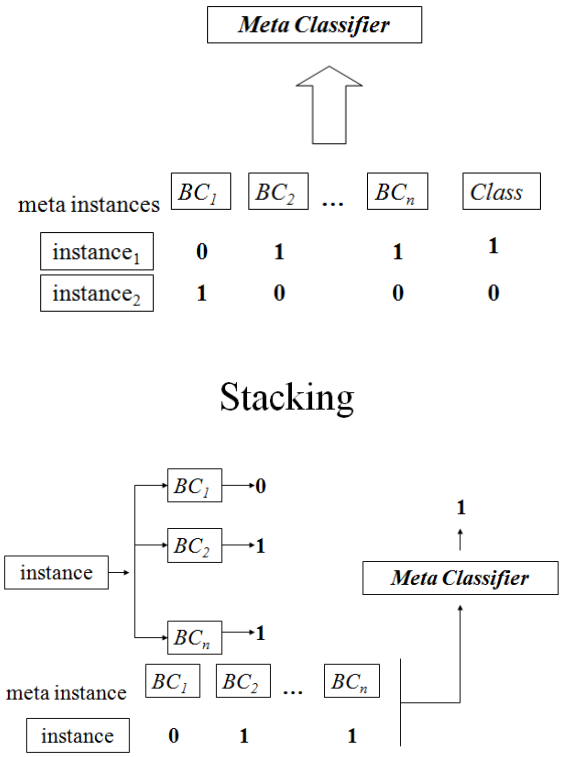
3.3 Combination Technique:

In presented paper we have used ensemble of heterogeneous classifier and for final prediction we combine prediction of each base classifier. For combining classifier there are mainly three categories;
Cascading, voting and stacking
In the proposed paper we use stacking technique for combining classifiers; in this approach it finds relationship between classifier prediction and actual prediction. Stacking technique first trains base classifier with available data and then prediction of base classifier becomes input for combiner classifier that is meta-classifier.

In below diagram BC denotes base classifier and it shows process of stacking. In this instance 1 represents available training data which is input for base classifier and then prediction of this base classifier becomes instance that is input for meta-classifier.



All Rights Reserved © 2016 IJSETR



3.4 Software and Language used for Implementation:

We have used,

1. Java Programming Language for implementation of Naïve Bayesian and C4.5 algorithm and used Eclipse IDE, jdk1.8.0_60, JRE7.

2. R language for multiple imputations algorithm for data pre-processing.

3. Jsoup Library for implementation of Web Scraping. Implementation in Java with Eclipse IDE.

4. Text Summarization Algorithm implementation in Java.

B. Introduction to web scraping technology used for providing real time updated information of disease:.

Automatic summarization is used to create summary from text document using computer program which contains important points from original document. Automatic data summarization uses machine learning and data mining approach for summarization. it is the process of finding subset of the data from original data, which gives the information of the entire set.

Automatic Summarization provides two approaches:

- Extraction - Extractive method identifies important words, sentences, sections of the text from the original document and gives the summary.
- Abstraction: It applies Natural Language processing technique to create a summary which is similar to what a human might produce. It uses its own word to summarize the content means that sometimes words are not present in the original document.

Machine Learning Method: Naive-Bayesian Methods

$$P(s \in S \mid F_1, F_2, ..F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in S) \cdot P(s \in S)}{\prod_{i=1}^{k} P(F_i)}$$

Let s denotes particular sentence
S denotes the set of sentences that make up the summary,
F1,….,Fk denotes the features.

Naïve Bayesian gives a score by using the formula and then selects only the n top sentences and that sentences are extracted and provide as summary to our user.

III. Results

(ckd: Presence of Chronic Kidney Disease,
notckd: Absence of Chronic Kidney Disease,
yes: Presence of Hepatitis,
no: absence of Hepatitis,
live: Presence of Coronary Artery Disease,
die: Absence of Coronary Artery Disease)

J48 accuracy:

Correctly Classified Instances	85.083
Incorrectly Classified Instances	14.917
Kappa statistic	0.5269
Mean absolute error	0.1971
Root mean squared error	0.325
Relative absolute error	25.759
Root relative squared error	77.062
Total Number of Instances	774

All Rights Reserved © 2016 IJSETR

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area	class
	0.94	0.02	0.987	0.94	0.963	0.963	ckd
	0.993	0.06	0.909	0.993	0.949	0.974	notckd
	0.915	0.655	0.806	0.915	0.857	0.644	yes
	0.345	0.085	0.576	0.345	0.432	0.644	no
	0.943	0.625	0.853	0.943	0.896	0.679	live
	0.375	0.057	0.632	0.375	0.471	0.679	die
Weighted average	0.752	0.250	0.794	0.752	0.761	0.764	

Naïve Bayesian Accuracy:

Correctly Classified Instances	84.452
Incorrectly Classified Instances	15.548
Kappa statistic	0.582
Mean absolute error	0.16
Root mean squared error	0.3234
Relative absolute error	45.343
Root relative squared error	77.894
Total Number of Instances	774

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area	class
	0.919	0.02	0.987	0.919	0.952	0.981	ckd
	0.993	0.068	0.898	0.993	0.943	0.997	notckd
	0.866	0.509	0.835	0.866	0.85	0.807	yes
	0.491	0.134	0.551	0.491	0.519	0.807	no
	0.854	0.313	0.913	0.854	0.882	0.882	live
	0.688	0.146	0.55	0.688	0.611	0.882	die
Weighted average	0.802	0.198	0.789	0.802	0.793	0.893	

Ensemble Classifier accuracy:

Correctly Classified Instances	86.254
Incorrectly Classified Instances	13.745
Kappa statistic	0.544
Mean absolute error	0.195
Root mean squared error	0.291
Relative absolute error	55.125
Root relative squared error	70.067
Total Number of Instances	774

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.964	0.02	0.988	0.964	0.976	0.981	ckd
	0.993	0.036	0.943	0.993	0.968	0.994	notckd
	0.963	0.745	0.794	0.963	0.871	0.782	Yes
	0.255	0.037	0.7	0.255	0.373	0.782	No
	0.935	0.563	0.865	0.935	0.898	0.817	Live
	0.438	0.065	0.636	0.438	0.519	0.817	Die
Weighted average	0.758	0.244	0.821	0.758	0.77	0.862	

IV. FUTRURE WORK

Our system gives accurate prediction for Chronic Diseases. Currently, we provide prediction for Chronic Kidney Disease, Coronary Heart Disease and Hepatitis. In our future work, we intend to generalize prediction to encompass more chronic diseases. For summarizing information about particular disease, we will take URL from user and give summarized information of user desired URL using web mining.

We intend to develop an adaptive health management system. We will use health care API's to provide user- centric health management features for healthy living such as tracking the calories intake and suggest diet food/ gym routine/ exercise based on user capability.

We will implement centralized (all-at-one-place) test result management system, which we will use to store health report of numerous users. User can share his reports with doctors,

All Rights Reserved © 2016 IJSETR

other patients, specialists, hospitals. Thereby, relieving the user from keeping track of both, the older and recent hard copies of test results.

We also propose to perform continuous social web mining and notify the users in case of occurrences of any natural or man-made disaster. Thereby, enabling the user tackle any emergency situation tactfully.

V. CONCLUSION

In the paper, we presented an efficient technique for predicting the probability of chronic diseases. We employed stacking technique making use of meta-classifier Bayesian Multinomial which combines the predictions of the individual base classifiers Naïve Bayesian and Decision Tree to provide efficient and accurate results. The statistical details for the same have also been presented. It is evident that the combination/ensemble of heterogeneous classifiers gave better accuracy then the individual base classifiers alone for the diseases under consideration. We also Use web Scraping and Text Summarization techniques to provide real time relative Summarized Statistical Report for the particular disease.

VI. REFERENCES

1. https://archive.ics.uci.edu/ml/datasets.html
2. http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf
3. www.saedsayad.com
4. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=974467&queryText=meta%20classifier%20for%20stacking&newsearch=true
5. http://idb.csie.ncku.edu.tw/tsengsm/COURSE/DM/Paper/ec45.pdf

All Rights Reserved © 2016 IJSETR