Applying Naïve Bayesian Classifier for Getting Probability Based Result for E-Knowledge Services In Healthcare

Vaishali V Kaneria Research Scholar, KSV University, Gandhinagar,Gujarat, India Asst. Prof., Department of MCA, AITS, Rajkot, Gujarat, India

Ms. Richa Mehta Research Scholar-Singhania University, Rajasthan, India Dr. N. N. Jani Dean CS, KSV University, Director, SKPIMCS, Gandhinagar

Gautam Kamani Research Scholar, KSV University, Gandhinagar, Gujarat, India

Abstract

In the existing health system various systems of medicine are available. Here a special EHR has been prepared to integrate multiple SOM. By applying some data mining technique result can be derived intelligently from the existing available data in EHR. Naïve Bayesian classification technique is selected for getting the result because it can give probability based evaluation. Weka tool is used for the selected algorithm. Different test options are used in Weka to dissect the algorithm at better extent.

Index Terms— SOM (System of Medicine), EHR (Electronic Health Record), Naïve Bayesian

I. INTRODUCTION

There are many systems of medicine available to get cured. Here Allopathic and Ayurvedic system of medicine is taken as consideration. This research is focused on building a novel architecture integrating two SOM and a special care is taken to develop an EHR database involving medical experts in both the system in order to transform the architecture into a prototype and its testing to justify the research objectives.

The EHR database can have any number of attributes but here selected 5 attributes are shown in the below sample EHR data set. Further the data needs a classification before the processing. The classification is based on set of condition evaluation. The classifier which can take care of this conditional classification is supplemented with the use of Weka tool [1].

This work is targeted to obtain condition based probability results. To satisfy this purpose the Naïve bayes classification algorithm supported in Weka is selected for the purpose of computation of results.

II. SAMPLE EHR DATASET

Age	Gender	SOM	Disease	Cured
<=20	М	AYU	Allergy	NO
<=20	М	ALO	Allergy	NO
21-40	М	AYU	Allergy	YES
>40	М	AYU	RA	YES
>40	F	AYU	fever	NO
>40	F	ALO	fever	YES
21-40	F	ALO	fever	YES
<=20	М	AYU	RA	NO
<=20	F	AYU	fever	YES
>40	F	AYU	RA	YES
<=20	F	ALO	RA	YES
21-40	М	ALO	RA	YES
21-40	F	AYU	Allergy	YES
>40	М	ALO	RA	NO

III. SELECTION OF SUITABLE DM METHODOLOGY: For getting the result some of the Data mining algorithms were studied. From that Naïve Bayesian is selected because in this algorithm probability based

selected because in this algorithm probability based conditions can be evaluated and result can be get based on that.

Naïve Bayesian Classifier computations [2]:

- P(C_i): P(Disease Cured = "yes") = 9/14 = 0.643 P(Disease Cured = "no") = 5/14= 0.357
- Compute P(V|C_i) for each class

```
P(age = "<=20" | Disease Cured = "yes") = 2/9 = 0.222
```

U.222

P(Diagnosis = "RA" | Disease Cured = "yes") = 4/9 = 0.444

P(Gender = "F" | Disease Cured = "yes) = 6/9 = 0.667 P(SOM = "AYU" | Disease Cured = "yes") = 6/9 = 0.667

P(age = "<= 20" | Disease Cured = "no") = 3/5 = 0.6 P(Diagnosis = "RA" | Disease Cured = "no") = 2/5 =0.4 P(Gender = "F" | Disease Cured = "no") = 1/5 = 0.2P(SOM = "AYU" | Disease Cured = "no") = 2/5 = 0.4

Test Case V = (age <= 30, Diagnosis = RA, Gender = F, SOM = AYU)

Calculating the probability of disease being cured based on the parameter given in V:

P(V|C_i) : P(V|Disease Cured = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

Calculating the probability of disease not being cured based on the parameter given in V:

 $P(V|C_i) : P(V|Disease Cured = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019$

 $P(V|C_i)*P(C_i) : P(V|Disease Cured = "yes") * P(Disease Cured = "yes") = 0.028....(a)$

P(V|Disease Cured = "no") * P(Disease Cured = "no") = 0.007.....(b)

Decision:

Case V belongs to class ("Disease Cured = yes") as the probability=0.028......(a) > probability=0.007.....(b)

IV. NAÏVE BAYESIAN ALGORITHM IN WEKA:

The implementation of naïve bayes algorithm is implemented in Weka environment with initially sample data and later on the prototype is to be implemented with EHR having huge datasets.

Weka supports four Naive Bayesian Algorithm Testing Options are:

- 1) Use training set
- 2) Supplied test set
- 3) Cross validation Folds
- 4) Percentage Split

V. APPLYING NAÏVE BAYESIAN

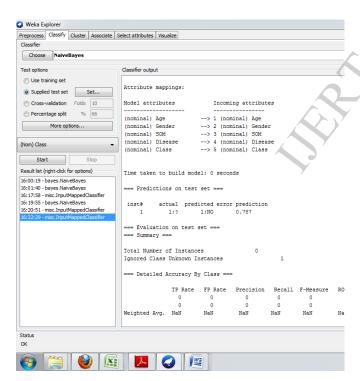
The EHR dataset is subjected to all four options of Naïve Bayesian algorithm.

1) Testing using Training set

reprocess Certify Guster Associat	E Select attributes Visualize				
letter					
Choose NaiveBayes					
est options	Cleanfier output				
Use torroaet	12 2:115 2:115	5.769			
Suppled text set	13 2:1ES 2:1ES 14 1:50 1:50	0.007			
Coss-valdation Frida 10					
O Percentage split % (5)	Evaluation on training set	*			
Have options					
	Correctly Classified Instances	12	85.7143 %		
Nan() Cass	Incorrectly Classified Instances Recon statistic	2	14,2857 \$		
Bart Dro	Hean absolute error	0.3109			
	Root mean equated error	0.3626			
leault list (right-sick for options)	Seletive absolute error	66.9543 \$			
6.00129 - Dayes NaireBayes	Boot relative squared error	15.6254 8			
2011 O SPECIAL CONTRACT	Coverage of cases (0.95 level) Nean rel. region size (0.95 level)				
	Total Rumber of Instances	14			
	Detailed Accuracy by Class				
	77 Rate 77 Rate		7-Beasure BOC -		
	0.8 0.111	0.8 0.8	d.8 0.1		
	0.889 0.2	0.889 0.889	0.889 0.1		
	Weighted Roy. 0.857 0.168	0.857 0.857	0.857 0.1	14	
	Confusion Metrix				
	a b < classified as				
	4 1 1 a = 10				
	1.5 (b = 155				
bts X					10
			_		
	2 🔢 📙 🙆				- P B - 4 41 54/2012

Correctly Classified Instances: 85.7% Incorrectly classified instances: 14.3%

2) Testing using supplied test set



Correctly Classified Instances: 60% Incorrectly classified instances: 40%

3) Testing using Cross validation Folds

Classifier Choose ZenaR										
Choose ZenaR										
Test options	Classifier output									
 Use training set 	1 1:8		UPASU	0.533						
E Suppled test set	1 1:4		UYA:	0.533						
Cross-validation Folds 10	1 114	10	TATU	0.933						
Percentage spit 15 66	Stratified	cross-va	idation ==	-C.						
	Summary									
Nare options										
	Correctly Class			8		57.1429				
Nor) Cized •	Incorrectly Cla		instances	6		42.8571				
	Xappa statistic Mean absolute e			0.50						
Start Stop	Root mean squar			0.50						
esuit list (right-click for options)	Relative absolu			100						
2.55:00 - rules Zerolt	Root relative a		ror	100	1					
2:58:29 - rules ZeroR	Coverage of cas	88 (0.95	level}	100	4					
22:58:31 - rules ZeroR	Mean rel. regio			100	1					
12:58:32 - rules ZeroR	Total Sumber of	Instano	18	14						
12:58:33 - rules, ZeroR 12:58:40 - rules, ZeroR										
(2.58:40 - rules Zerolt (2.58:45 - sules Zerolt	Detailed Ac	curacy B	Class							
12:58:59 - ndes 2008										
2.59:04 - rules ZeroR		IF Sace				F-Measure				
12:59:13 - rules ZeroR		-	1	0.571	1	0.727	0.147	ANU ALO		
12:59:38 - rules Zerolt	Weighted Avg.	0.571	0.571	0.327	0.571		0.167	MAN		
2.59-48 - rules Zerok	serviced and.	web'll	v.3/1	11-341	4+871	V. 618	9,287			
	Confusion H	atrix								
	a b < clas	sified a	02							
	8 0 1 e = AYU									
	6 0 b = ALO									
ana										
x									la	2 40
										-

Correctly Classified Instances: 57.14% Incorrectly classified instances: 42.9%

4) Testing data using Percentage split

Weka Explorer	COLUMN CONTROL	
process Clessify Cluster Associate	Select attributes Houalce	
ssfer		
Choose ZeroR		
st options	Classifier output	
Use training set	2 1:80 2:YES + 0.667	
Suppled test set Set	3 1:50 2:125 + 0.667	
Cross-validation Folds 10	4 2:YES 2:YES 0.667	
	Evaluation on test solit	
	Sumary	
Nore options		
	Correctly Classified Instances 2 50 %	
on) Cured 🔹	Incorrectly Classified Instances 2 50 W Rome statistic 0	
	Kapa statistic 0 Mean absolute error 0.5	
Start Stop	Root mean squared error 0.527	
suit list (right-click for options)	Relative absolute error 100 %	
AND TRACTOR	Root relative squared error 100 %	
58:31 - rules ZeroR	Coverage of cases (0.95 level) 100	
58:32 - rules ZeroR 58:33 - rules ZeroR	Mean rel, region size (0.95 level) 100 %	
58:40 - rules ZeroR	Total Number of Instances 4	
SR-45 - rules ZeroR		
58:59 - rules ZeroR	Detailed Accuracy By Class	
59:04 - rules ZeroR		
:59:13 - rules.ZeroR	TP Rate FP Bate Precision Recall F-Heasure ROC Area Class	
59:38 - rules ZeroR	0 0 0 0 0.5 10	
59:48 - rules ZeroR	1 1 0.5 1 0.667 0.5 YES	
05:27 - rules ZeroR II 05:38 - rules ZeroR	Weighted Avg. 0.5 0.5 0.25 0.5 0.333 0.5	
105:36 - rules Zerok 105:46 - rules Zerok		
05:55 - rules ZeroR	Confusion Matrix	
05:46 - rules ZeroR	a b < classified as	
07:29 - rules. ZeroR.	a D < CLESSITIED as 0.2 a = ND	
19:18 - rules ZeroR	0.2 6 = 90	
32:04 - rules. ZeroR	02 0 = 10	
: 10:07 - rules ZeroR : 10:17 - rules ZeroR		
12.17 - FDES ZIFOK		
tus -		
		~ ~
		- P 🗄 👍 🖕 1100
9 🗀 🚺 I		 I* II: 4 19 3010

Correctly Classified Instances: 50% Incorrectly classified instances: 50%

VI. CONCLUSION:

Here the results are shown which are taken using Weka tool. Using all the types of Naïve Bayesian options results has been taken. The Results from Naïve Bayesian training set, Naïve Bayesian supplied test set, Cross validation folds and Percentage split are not consistent. Improved algorithms are needed to be build for more consistent and improved result. This can be further utilized in terms of services for the users.

REFERENCES

[1] <u>http://www.cs.waikato.ac.nz/ml/weka/</u>
[2] <u>http://www.inf.u-szeged.hu/~ormandi/ai2/06-naiveBayes-example.pdf</u>

[3] http://www.ijetae.com/Volume2Issue2.html

