Implementing Clinical Decision Support System Using Naïve Bayesian Classifier

Trupti S. Mokati
Dept. of Computer Sci. & Engineering, P.R.Pote (Patil)
College of Engineering and Management, Amravati
tmokati@gmail.com

Prof. Vijay B. Gadicha
HOD Comp. Sci. Dept.,
P.R.Pote College, Amravati
v_gadicha@rediffmail.com

Abstract: To speed up the diagnosis time and improve the diagnosis accuracy in today's healthcare system, it is important to provide a much cheaper and faster way for diagnosis. This system is called as Clinical Decision Support System (CDSS). With various data mining techniques being applied to assist physicians in diagnosing patient diseases with similar symptoms, has received a great attention now a days. The advantages of clinical decision support system include not only improving diagnosis accuracy but also reducing diagnosis time. In this paper, the data mining technique name Naïve Bayesian Classifier, which offers many advantages over the traditional methods of data mining is used that opens a new way for clinicians to predict patient's diseases. As the system is built on the sensitive data for patient privacy it is necessary to add some features that meets the security requirement. Specifically, with large amounts of data related to healthcare is generated every day, the classification can be utilized to excavate valuable information that improve clinical decision support system. Here the fuzzywuzzy string matching algorithm of naïve bayesian classifier is used to perform prediction from large number of symptoms data. The Result analysis perform in the last section on live data of five patient gives that by using proposed technique we try to make the Clinical Decision Support System more helpful for providing diagnosis of deceases more accurately and efficiently.

Keywords: *Clinical Decision Support System (CDSS), Privacy Preserving, Naïve Bayesian Classifier, Fuzzywuzzy algorithm.*

*****

I. INTRODUCTION

Today's Healthcare industry has the global scope to provide health services for patients. One of the part of it is Clinical Decision Support System (CDSS) has a massive amounts of electronic data and experienced such a sharp required and growth rate. However, it is necessary to design and develop as appropriate technique to find great potential economic values from large amount of data and to speed up the diagnosis time and improve the diagnosis accuracy [1]. It is a new system in healthcare industry that is workable to provide a much cheaper and faster way for diagnosis. As the Clinical Decision Support System (CDSS) has huge amount of data it is necessary to apply various data mining techniques to assist physicians in diagnosing patient diseases with different data mining classification functions, and has received a great attention recently [2] [3] [4]. Out of different data mining classification techniques available Naive Bayesian classifier, is one of the popular machine learning tools, has been widely used to perform prediction [5]. Despite its simplicity, it is more appropriate for medical diagnosis in healthcare than some sophisticated techniques [6] [7].

The CDSS with naive Bayesian classifier has offered many advantages over the traditional healthcare systems and opens a new way for clinicians to predict patient's diseases. However, one of the main challenges is how to keep patient's medical data away from unauthorized disclosure. The usage of medical data can be of interest for a large variety of healthcare stakeholders [8]. Without good protection of patient's medical data, patient may feel afraid that his medical data will be leaked and abused, and refuse to provide his medical data to CDSS [9]. Therefore, to develop the clinical decision support system along with address the privacy issues, this paper propose a Privacy Preserving Patient-Centric Clinical Decision Support System, called PPCDSS. For preserving the privacy of patient's medical data, here the sensitive data gets encrypted first by using cryptographic approach and then stored to the data base.

Along with this some of the Objectives which are targeted to achieve are perform efficient prediction of disease on the basis of existing data-set. For that the system introduce a new classification and aggregation approach called fuzzywuzzy, which allows service provider to build naive Bayesian classifier [10]. This helps in reducing diagnosis time for prediction of diseases. And the encryption technique used helps for preserving privacy of patient's data. As the symptoms are vary from patient to patient and may not be present in CDSS database for that on new mechanism is proposed as to review and retrain the CDSS dataset [11]. The remaining paper is organized as, Section II give the implementation procedure of the PPCDSS. Along with, the working and all functionality required for use of fuzzywuzzy algorithm. Section III gives the result analysis that performed by tacking actual symptoms for five different patients. This validate the efficiency of the proposed PPCDSS and gives the advantages of the proposed and develop system. Finally Section IV, concludes the paper.

II. IMPLEMENTATION STRATEGIES

This paper tries to improve the existing system using Clinical Decision Support System based on Naïve bayesian classifier. The system uses are using Data Mining classification technique for Clinical Decision Support System (CDSS). The system will work faster and efficient using this technique [12] [13]. It is widely used in real-life applications because of its simplicity and good performance both in theory and practice. However, in large-scale problems, where huge training data are available, such as road sign detection, the method's training and test phases might be prohibitively demanding in terms of computations. Thus, for large-scale problems the reduction of computational complexity is essential. For the security purpose encryption techniques with AES algorithm for preserving privacy of patient's data is used [14][15]. The complete flow of working system in step-by-step manner is as follows:

A. Stepwise Work Flow of System:

Step 1: Doctor Register with the System.

Step 2: Doctor has to login the system with his authentic email-id and password.

Step 3: Doctor can add / edit / update /delete any number of disease, their symptoms, and their prescription information.

Step 4: Doctor add patient information along with the symptoms he is suffering from to the database and check for diagnosis.

Step 5: Using Database will provide the historical medical data present in our database and processing with the help of Naïve Bayesian classifier fuzzywuzzy search algorithm.

Step 6: After calculation, the predicted result will be send to the next level. On this level the probability of predicted disease risk will be calculated and top three disease having probability of more than 50% are displayed. In this algorithm the maximum probability disease risk will be calculated.

Step 7: Now doctor check the patient symptoms once again and from the result generated in step 6, he suggest most suitable prescription for patient. Finally, proper predicted diseases will be diagnose, this will help to give proper prescription to the patients more effectively.

Step 8: For more proper CDSS designing, doctor review his prescription suggested to the patient. Here he checks that, the patient gets cure form his provided prescription or not. If the patient gets cure then go to step 9 or stop otherwise.

Step 9: Check for any new symptoms that the patient is suffering from and already our CDSS data have. If any new symptoms are identified, Retrain database by adding symptom to that particular disease.

B. Algorithm Used

FUZZY SERACHING with NAIVE BAYES CLASSIFIERS

Here the system uses fuzzywuzzy searching algorithm for diagnosis of patient based on naïve bayesian classifier. The complete description is as follows [16]:

FuzzyWuzzy Algorithm:

It is simple library and command-line general regular expression like utility which could help you when you are in need of approximate string matching or substring searching with the help of primitive regular expressions.

About "approximate" or "fuzzy" string comparison and its need:

Just imagine that you deal with information (like orders) which is sent to you by many people. When these people mention names of places or persons, they could bring to you problems of two kinds:

- they make nasty typos;
- they use different variants of names;

1. For example if you are responsible for checking incoming mail in, you may want to find letters addressed to Indian president. You try to find all which contains words "Narendra Modi" on envelope. But you soon discover that sometimes people address this person as "Priminister Narendra Modi" and sometimes like "Mr. N. Modi" and also "Narendrabhai Modi" (note typos).
2. In another example, if you read google and wikipedia and found that you can compare "Barak" with "Barack" and "Baarck" etc (may be the name of former US president Barak Obama). With the help of "approximate string matching algorithm", also called "fuzzy string matching". But after you use or implement some of algorithms you found that it is not sufficient. You need "approximate" substring search, and ability to specify some complex patterns (for example country could be specified like "Russia" and like "Russian Federation" - but it should not be mixed with "Belarussian Republic" etc.

Naïve Bayesian classification with fuzzy matching:

Fuzzy matching is a general term for finding strings that are *almost* equal, or *mostly* the same. Of course *almost* and *mostly* are ambiguous terms themselves, so it is necessary have to determine what they really mean for your specific needs? The best way to do this is to come up with a list of steps before starting to write any fuzzy matching code. Once you have perform all steps, then it's much easier to tailor your fuzzy matching code to get the best results. These steps are summarized as follows [17]:

A. Normalization

The first step before doing any string matching is *normalization*. The goal with normalization is to transform the input strings into a normal form, which in some cases may be need to do. The most basic normalization you can do is to lowercase and strip whitespace. And in some cases, one can also removes all punctuation in a string.

Example:

'Happy Days' != ' happy days ', with simple normalization you can get 'Happy Days'.lower() == ' happy days '.strip().

Output : 'happy days'

Beyond just stripping whitespace from the ends of strings, it's also a good idea replace all whitespace occurrences with a single space character. The regex function for doing this is re.sub('\s+', s, ' '). This will replace every occurrence of one or more spaces, newlines, tabs, etc, essentially eliminating the significance of whitespace for matching.

B. Regular Expressions

It is helpful to use regular expressions to identify significant parts of a string, or perhaps split a string into component parts for further matching. Create a *simple* regular expression to help with fuzzy matching, because any other code if present to do fuzzy matching will be more complicated, less straightforward, and probably slower.

FuzzyWuzzy algorithm functions

Get match ratios with the help of following fuzzy search Ratio expressions / functions:

1. Simple Ratio

FuzzySearch.ratio("mysmilarstring","myawfullysmilarstirng") - 72

FuzzySearch.ratio("mysmilarstring","mysimilarstring") - 97

Along with this there are different techniques of this as Partial Ratio, Token Sort Ratio, Token Set Ratio, Weighted Ratio etc, Extract Result from calculations using following fuzzy extract expression / functions are:

1. Extract One:

FuzzySearch.extractOne("cowboys", ["Atlanta Falcons", "New York Jets", "New York Giants", "Dallas Cowboys"])

(string: Dallas Cowboys, score: 90)

Extract Top:

FuzzySearch.extractTop("goolge", ["google", "bing", "facebook", "linkedin", "twitter", "googleplus", "bingnews", "plexoogl"], 3)

[(string: google, score:83), (string: googleplus, score:63), (string: plexoogl, score:43)]

In the proposed application, the above ExtractTop method is used that gives the top three predicted diseases from which patient may be suffering from. Result generated from it makes easy for the doctor to calculate the actual disease from only three most probable options.

III. RESULT AND DISCUSSION

The Step-by-step workflow of the proposed system and the algorithm that perform its result extraction is also explain in the above sections. Now the Experimental result shows that Clinical Decision Support System Based on Naïve Bayesian classifier using fuzzy string matching approach gives better prediction result is checked experimental real time data. We have taken actual symptoms from five patients from which they are suffering and provide to our system. The result generated from the system are as follows:

Patient 1: This patient is actually suffering from Thyroid Disease

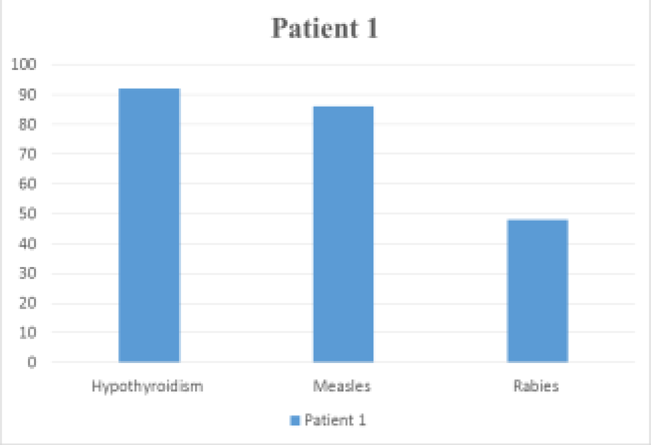
Symptoms given by patient 1 are:

Selected Symptoms: Fatigue, weight gain, Cold intolerance, Dry or brittle hair, memory problem, irritability and depression,

Extra added Symptoms: cholesterol level is high, slow heart rate

Result given by the system:

Hypothyroidism (92%)	Measles (86%)	Rabies (also called hydrophobia) (48%)



Graph 1: Accuracy of prediction for Patient 1

From the above graph we find out that, in terms of diagnosis accuracy of our system with the naïve Bayes fuzzywuzzy algorithm find out most accurate result. That is as the patient is already suffering from thyroid and the diseases predicted from our system is also Hypothyroidism with highest probability of having that disease is 92%.

Patient 2: This patient is actually suffering from Osteoporosis

Symptoms given by patient 2 are:

Selected Symptoms: easy bone fractures, stress fractures of feet at walking or stepping

Extra added Symptoms: lower back pain, pain in legs,

Result given by the system:

Osteoporosis (91%)	Chicken Pox (86%)	Diabetes Mellitus (86%)

Patient 3: This patient is actually suffering from Dengue
Symptoms given by patient 3 are:
Selected Symptoms: easy bone fractures, stress fractures of feet at walking or stepping
Extra added Symptoms: lower back pain, pain in legs,
Result given by the system:

Dengue (88%)	Measles (86%)	Chicken pox (57%)

Patient 4: This patient is actually suffering from Cardio vascular diseases
Symptoms given by patient 4 are:
Selected Symptoms: Persistent high blood pressure (BP), Suger level is high,

Extra added Symptoms: damage of arteries, fatigue

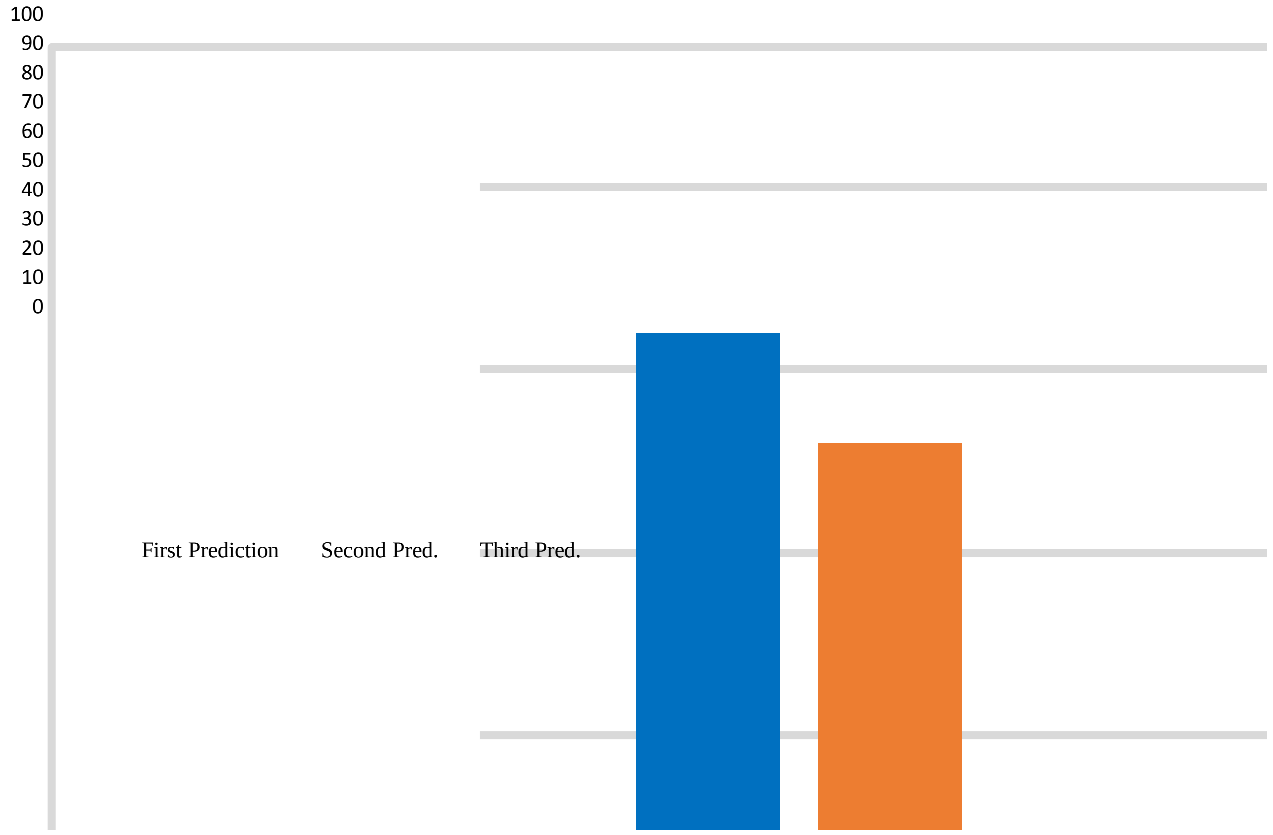
Result given by the system:

Cardio vascular diseases (89%)	Chicken pox (86%)	Dengue (86%)

Patient 5: This patient is actually suffering from Tuberculosis
Symptoms given by patient 5 are:
Chest pain and blood comes out with the sputum.,Cough last from 2-3 weaks,loss of appetite,Coughing up blood,
Extra added Symptoms: general weakness,
Result given by the system:

Tuberculosis (91%)	Hepatitis (49%)	Chicken pox (48%)

As in case of patient 2, the probability of prediction is not vary in most cases. That is first prediction is 91%, second is 86% and third is 86%. This is because, the symptoms given by the patient are generalized and that may cause all three types of disease that obtained in the result. But instead of that the first prediction is given 91% to Osteoporosis that is the actual disease with which patient 2 is suffering from. Therefore the prediction rate of our proposed system is high. The overall analysis of five patient is given in the graph 2 below.



Graph 2: Accuracy of Prediction

Graph 2 above shows the prediction accuracy of our proposed system. As we have collected the actual systems from five patient suffering from different diseases. All the symptoms given to our developed system and the result obtained are given in tabular format above. And in the above graph the overall accuracy of prediction of our system is shown. For all the time the system show the highest percentage to the disease to the actual disease with which the patient is suffering from as its first prediction. This gives our proposed and develop system for CDSS using naïve bayesian classification is most accurate. That is success rate of our proposed system is much high.

Advantages:

By designing the system like this we are able to

- Improving diagnosis accuracy for any critical diseases
- Reducing diagnosis time gives proper prescription in much less time
- Top 3 disease prediction out of which you can choose most probable. The success rate for the first prediction is most of the time high
- Reducing communication overhead
- Preserving privacy of patient’s sensitive data

IV. CONCLUSION

In this paper, the Clinical decision support system using classification technique of data mining with naïve Bayes algorithm is proposed. This naive Bayes with fuzzywuzzy algorithm has excellent performance in generalization so it can produce high accuracy in classification for diagnosis. The patient can securely retrieve TOP three diagnosis result according to their own preferences. With the advantage of encryption technique, the privacy of patient sensitive data is achieved. The processing is done on the encrypted data, so that there is no loss in the privacy of patient’s data while training the CDSS. These results analysis performed evidentially proved that proposed method shows the nice performance of classification accuracy with the proposed algorithm.

REFERENCES

[1] V.B.Gadicha ,“Enhanced Authentication Scheme using Image fusion & Multishared Cryptography”, International Journal of Modern Computer Science (IJMCS), vol.02, issue 04,Aug 2014, ISSN : 2320-7868

[2] V.B.Gadicha, “Data Integrity Proofs in Document Management System under Cloud with Multiple Storage”, International Journal of Engineering & Computer Science (IJECS), vol.03 issue12, Dec2014, ISSN:2319-7242

[3] V.B.Gadicha,“Privacy-Preserving System for shared data in cloud Environment using public auditing scheme : A Review”, International Journal of Advancement in Engineering Technology Management & Applied Sciences (IJAETM&AS), Vol.03, Issue 11,Nov2016, ISSN: 2349-3224.

[4] V.B.Gadicha ,“A survey towards patient centric clinical decision support system using Navie Bayes classification system”, International Journal of Innovative Research in Computer & Communication Engineering, (IJIRCCE), Vol.04, issue 12, Dec 2016 and ISSN: 2320-9798.

[5] V.B.Gadicha ,“A survey Cipher Text policy attribute based encryption & time specified approach”, International Journal of Innovative Research in Computer & Communication Engineering (IJIRCCE), Vol.05, issue 02, Feb 2017 and ISSN: 2320-9798.

[6] V.B.Gadicha, Dr A. S. Alvi, “A Novel Approach towards Authentication by Generating Strong Passwords”, ACM Digital Library”, International Conference on Information & Communication Technology for Competitive Strategies (ICTCS-2016), Udaipur, March 2016.

[7] Ximeng Liu, Rongxing Lu, Jianfeng Ma, Le Chen, and Baodong Qin, “Privacy- Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification”, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. XX, NO. XX, DECEMBER 2014.

Assad Abbas, Samee U. Khan, “e-Health Cloud: Privacy Concerns and Mitigation Strategies”, North Dakota State University, USA.

[8] Jim Basilakis, Bahman Javadi, Anthony Maeder, “The Potential for Machine Learning Analysis over Encrypted Data in Cloud-based Clinical Decision Support – Background and Review”, Proceedings of the 8th Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2015), Sydney, Australia, 27 - 30 January 2015.

[9] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, “Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review”, International Journal of Computer Applications (0975 – 8887), Volume 136 – No.2, February 2016.

[10] Luis Tabares, Jhonatan Hernandez, Ivan Cabezas member IEEE, “Architectural approaches for implementing Clinical Decision Support Systems in Cloud: A Systematic Review”, First Conference on Connected Health: Applications, Systems and Engineering Technologies, 978-1-5090-0943-5/16 $25.00 © 2016 IEEE.

[11] Shreya Anand, Ravindra B Patil, Krishnamoorthy P, “An Analytics based Clinical Decision Support

System for CVD Risk Assessment and Management”, 978-1-4577-0220-4/16/$31.00 ©2016 IEEE.

[12] Kulwinder Singh Mann, Navjot Kaur, “Cloud-deployable health data mining using secured framework for Clinical decision support system”, 978-1-4799-6908-1/15/$31.00 ©2015 IEEE.

[13] Jussi Mattila, Juha Koikkalainen, Arho Virkki, Mark van Gils, Member, IEEE, and Jyrki L¨otj¨onen, “Design and Application of a Generic Clinical Decision Support System for Multiscale Data”, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 59, NO. 1, JANUARY 2012.

[14] C. Schurink, P. Lucas, I. Hoepelman, and M. Bonten, “Computer- assisted decision support for the diagnosis and treatment of infectious disease s in intensive care units,” The Lancet infectious diseases, vol. 5, no. 5, pp. 305–312, 2005.

[15] Tzu-cheng Chuang, Okan K. Ersoy, Saul B. Gelfand, Boosting Classification Accuracy With Samples Chosen From A Validation Set, ANNIE (2007), Intelligent ` Engineering systems through artificial neural networks, St. Louis, MO, pp. 455-461.

[16] Fuzzy-String-Matching, [Online] available: https://stackoverflow.com/questions/21057708/java-fuzzy-string-matching-with-names.