Volume 4, Issue 1, January 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification

Manjusha K. K* *Dept. of Computer Science Karpagam University Coimbator, India.*

K. Sankaranarayanan *Sri Rmakrishna Institute of Technology, Coimbatore. India*

Seena P *Dept. of Dermatology Govt. Medical College Kottayam, India*

Abstract - We live in the world of data-rich times and each day, more data are collected and stored in databases. This medical data about large patient population is analyzed to perform medical research. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently. A number of studies have shown that the diagnosis of one patient can differ significantly if the patient is examined by different physicians or even by the same physician at various times. Automated medical diagnosis helps the doctors to predict the correct disease with less time. Dermatological diseases are always neglected but may even lead to death. Prediction of dermatological disease is very difficult because of the number of diseases presentation. So we propose a system which allows obtaining data patterns with the help of Naïve Bayesian theorem. In this paper we have experimented on data gathered from tertiary health care centres which surveys the people from various areas of Kottayam and Alappuzha, Kerala, India. Naïve Bayesian algorithm reveals the chances of different dermatological disease and also finds out the percentage of occurrence of each disease.

Keywords - Data Mining, Naïve Bayesian Classification, Medical Data, Dermatology, prediction.

I. Introduction

The huge amounts of data generated by healthcare transactions are too complex and analyzed by traditional methods. When medical sectors apply data mining on their existing data they can discover new, useful and potentially life saving knowledge. Data mining is the process of extracting or mining knowledge from large amounts of data. In data mining, intelligent methods are applied in order to extract data patterns. The increasing volume of medical science calls for analysis of computer based approaches for extracting useful information and it cannot be done by traditional methods. Data mining is a tremendous opportunity to assist physician deal with this large amount of data. Its methods can help physicians in various ways such as interpreting complex diagnostic tests, combining information from multiple sources, providing support for differential diagnosis. Data mining identifies trends within the data that go beyond simple analysis, through the use of sophisticated algorithms. The discovered trends can be used to find out the disease outbreaks.

II. Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. It is process of discovering interesting patterns trends in large data sets in order to find useful decision-making information. Before a data set can be mined, it first has to be cleaned. This cleaning process removes errors, ensures consistency and takes missing values into account. Then, computer algorithm is used to mine the clean data looking for unusual patterns. Finally, the patterns are interpreted to produce new knowledge.

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list.

a) Characterisation: It is a summarisation of general features of objects in a target class, and produces characteristic rules.
b) Discrimination: It is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.
c) Regression: It is a statistical method often used for numerical prediction.
d) Association: It studies the frequency of items occurring together in transactional databases and based on support and confidence threshold.
e) Classification: It is the processing of finding a set of model that distinguish data classes for the purposes of being able to use the model to predict the class of objects whose class label is unknown.
f) Prediction: It is either used to predict unavailable data values or a class label for some data.
g) Clustering: It is used to place data elements into related groups without advance knowledge of the group definitions.

III. Medical Data Mining

Clinical repositories containing large amount of biological, clinical & administrative data are increasingly becoming available as health care systems integrate patients information for research and utilization objective. Data

mining techniques applied on these databases discover relationships and pattern which are helpful in studying the progression & the management of disease. Data mining refers to extracting or "mining" knowledge from large amounts of data. Knowledge discovery as a process consists of an iterative sequence of Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation. With the rapid advancement in information technology, many different data mining techniques and approaches have been applied to complementary medicine. Statistics provide an impressive background to define and evaluate the result.

Here, we intend to express some of data mining applications dealing with few dermatological conditions.

IV. Dermatological Disease

Skin is the largest organ in the body. Skin separates the inside of our body to the outside world. So protecting skin from diseases is important. Dermatology is a branch of medicine dealing with skin, hair, nail and its disease. Recently, skin diseases have become common to everyone. Many factors including microbes, various drugs, exposure to ultraviolet radiations in sunlight etc. possibly causes skin problems. Although skin diseases are easily detectable and diagnosing symptoms and deciding therapy are easier than other systemic diseases, many people ignore the importance of them.

V. Why Naïve Bayesian Classification

In medical data mining, Naïve Bayes classification has an indispensable role. Naive Bayesian classifier has shown great performance in terms of accuracy so if attributes are independent with each other we can use it in medical fields [1]. For clinical data, missing values always occur. Naive Bayes handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing values in columns with simple data types are interpreted as missing at random.

If we choose to manage our own data preparation, Naive Bayes usually requires binning. Naive Bayes relies on counting techniques to calculate probabilities. Columns should be binned to reduce the cardinality as appropriate. Numerical data can be binned into ranges of values (for example, low, medium, and high), and categorical data can be binned into meta-classes (for example, regions instead of cities). Equi-width binning is not recommended, since outliers will cause most of the data to concentrate in a few bins, sometimes a single bin. As a result, the discriminating power of the algorithms will be significantly reduced. In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. They provide theoretical justification for other classifiers which do not explicitly use Bayes theorem. For example, under certain assumption, it can be shown that many neural network and curve fitting algorithms output the maximum posterior hypothesis, as does the naive Bayesian classifier.

VI. What is Naïve Bayesian Classification

Naive Bayes classifier is a probabilistic classifier based on the Bayes theorem, considering Naive (Strong) independence assumption. Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. Naive Bayes can often perform more sophisticated classification methods. It is particularly suited when the dimensionality of the inputs is high. When we want more competent output, as compared to other methods output we can use Naïve Bayes implementation. Naïve Bayesian is used to create models with predictive capabilities.

Bayes' Theorem:

Probility(B given A) = Probility(A and B)/Probility(A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone

Let X be a data tuple. In, Bayesian terms, X is considered "evidence". Let H be some hypothesis, such as that the data tuple X belongs class C. P(H|X) is the posterior probability, of H conditioned on X. In contrast, P(H) is the prior probability, of H.

Bayes' theorem is

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

$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$$

Similarly, P(X|H) is the posterior probability of X conditioned on H. P(X) is the prior probability of X.

VII. Methodology

The main goal of the research is to analyse the data from the surveys and to decide whether it is suitable to be analyzed with the use of the data mining methods. The analysis performed within this research are based on data surveyed from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala and filled out by registered medical practitioners.

These are the steps we have planned to perform with our data mining environment and data sets as well.
1. Collecting and reviewing the data set.
2. Select appropriate algorithm suitable for the data set.
3. Training the selected algorithm on reduced data set, by removing the attributes that appeared to be uninformative in building and visualizing the data
4. Using the optimal data set formed for each algorithm of the most useful data identified in step 6.
5. Evaluating the results.
6. Randomizing the data set.
7. Evaluating and comparing results as well as algorithms performance.

A. Data Source

Data was collected from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala and filled out by doctors. The research developed on the basis of the survey. Fig 1. shows the selected attributes for predicting eight diseases.

B. Data Set Description

Here, we are predicting the probability of occurring eight dermatological conditions. The medical data set contains the profiles of n=230 patients and has 21 medical attributes corresponding to the numeric and categorical attributes listed in Table I. The data set has medical information like symptoms, epidemiology and anamnesis.

TABLE I
INPUT ATTRIBUTES USED FOR ANALYSIS

1	Extent of ill	(value 0: No; value 1: Moderate; value 2: Severe)
2	Fever	(value 0: No; value 1: Yes)
3	Level of fever	(value 0: Low; value 1: moderate; value 2: high)
4	Duration of Fever	(value 0: short; value 1: long)
5	Morphologie the exanthema	(value 0: Maculopapules; value 1: Vesicular; value 2: Maculopapular rash)
6	Localization	(value 0: face; value 1: neck; value 2: body; value 3: cheeks)
7	Progressive exanthema	(value 0: slowly; value 1: quickly)
8	Painful exanthema	(value 0: No; value 1: Yes)
9	Type of enantheem	(value 0: koplik spot; value 1: Pharyngitis; value 2: blisters; value 3: Aardbeitong)
10	Aardbeitong	(value 0: No; value 1: Yes)
11	Conjunctivitis	(value 0: No; value 1: Yes)
12	Seasons	(value 0: Summer; value 1: Autumn; value 2: Spring; value 3: winter)
13	Age Group	(value 3: Elder; value 2: Younger; value 0: Baby)
14	Prior contact	(value 0<5 days; value 1<15 days; value 2 days <25; value: 3>1 month)
15	Patient Medication	(value 0: No; value 1: Yes)
16-18.	Medicine	(value 0: Medicine A; value 1: Medicine B; value 2: Medicine C)
19	Vaccination	(value 0: No; value 1: Yes)
20	Recent journeys	(value 0: No; value 1: Yes)
21	Contact with Animals	(value 0: No; value 1:Yes)

```
node Rubella {
        kind = NATURE;
        discrete = TRUE;
        chance = CHANCE;
        states = (True, False);
        parents = (Leeftijd, voorafgaande_contacten);
        probs =
                // True        False          // Leeftijd    voorafgaande_contacten
                (((0.6137485,   0.3862515),    // ouder_kind  enkele_dagen
                  (0.8319473,   0.1680527)),   // ouder_kind  enkele_weken
```



```
            ((0.1568952,   0.8431048),    // jonger_kind enkele_dagen
             (0.5957544,   0.4042456)),   // jonger_kind enkele_weken
            ((0.2925547,   0.7074453),    // zuigeling   enkele_dagen
             (0.2701741,   0.7298259)));  // zuigeling   enkele_weken        ;
numcases = 1;
whenchanged = 1144264474;
belief = (0.460179, 0.539821);
visual V2 {
        center = (282, 342);
        height = 9;
        };
};
```

VIII. Result and Discussion

A prototype Naive Bayesian algorithm was proposed to find the chances of occurrence of eight skin diseases on the basis of input variables. The application was built in Java platform using Net beans IDE. An output window predicting the occurrence of the condition on the basis of the input variable is represented in Fig II. As seen in the figure the chances of occurrence of various diseases is presented. Accordingly, with the set of inputs given the predictable chances are more for Scarlet fever and least for the occurrence for Kawasaki disease or chicken pox.

A. Output Screens

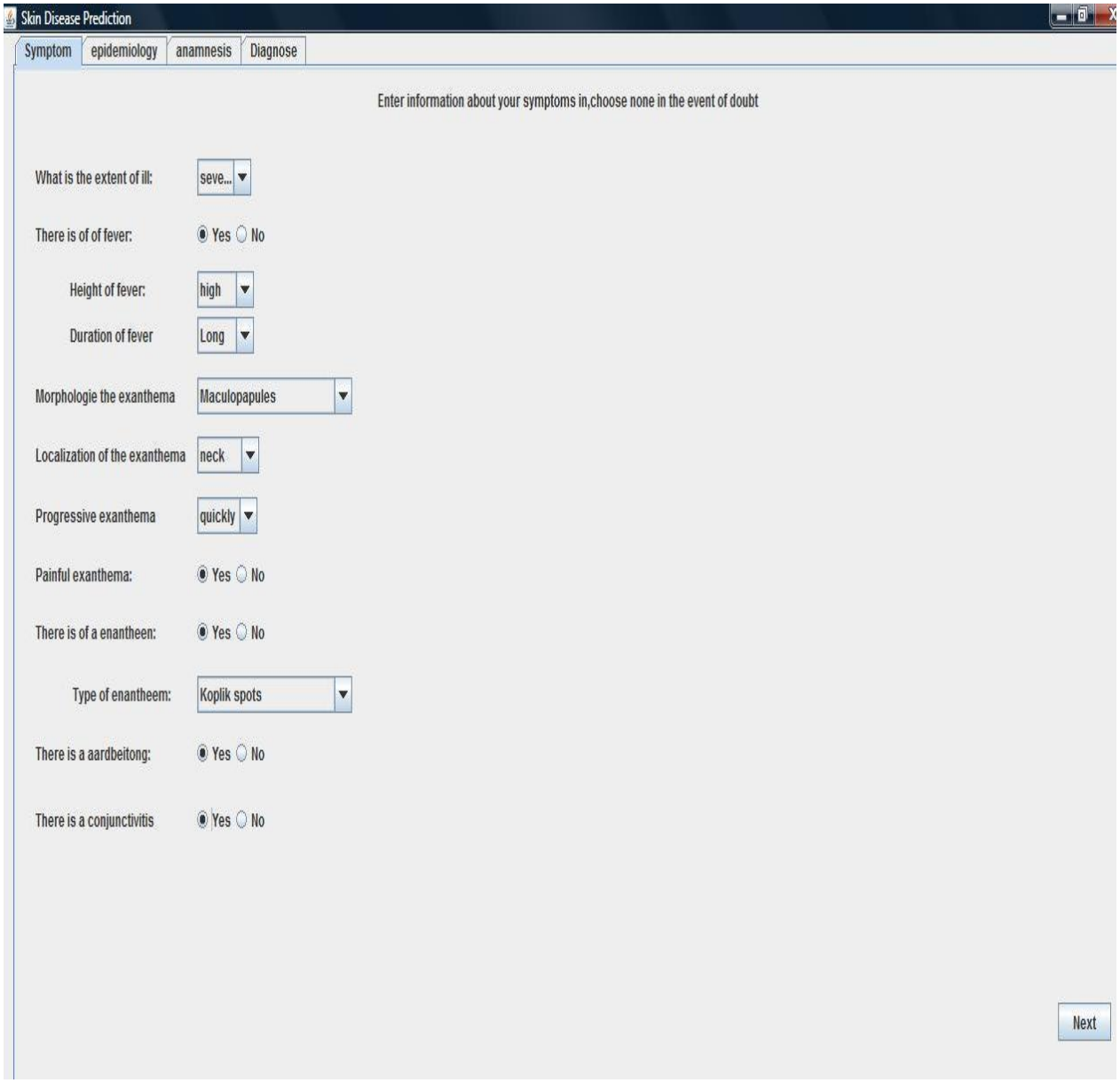

Fig. 1 Data input page

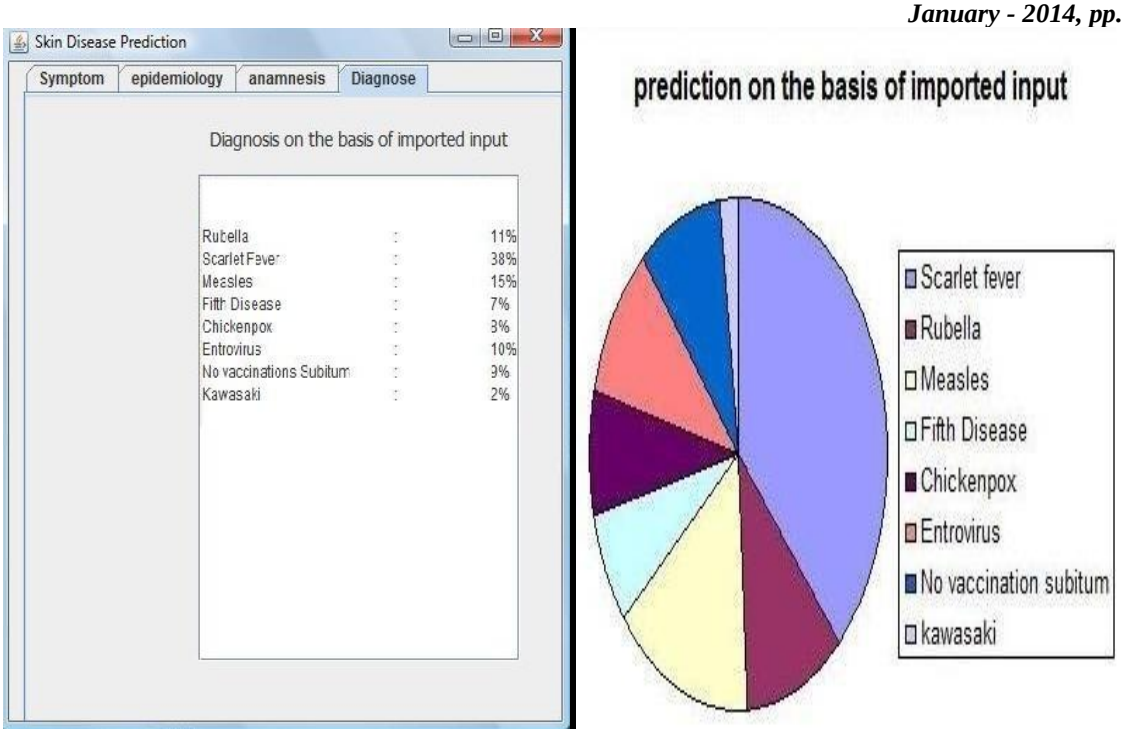

Fig. 2 Prediction window on the basis of imported input.

IX. Conclusion

Prediction of different Dermatological diseases using Naïve Bayesian Classification in data mining technique gives possibilities of eight diseases using patient attributes. The system can extract hidden knowledge from the database. This is effective model to predict dermatological diseases. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. We can extend this work with other data mining techniques and other medical measurements besides the above list. We can also predict other diseases other than dermatological diseases.

REFERENCES

[1] Divya Tomar and Sonali Agarwal, " A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-technology*, Vol.5, No.5, pp. 241 – 266, 2013.

[2] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*,2nd ed.,Morgan Kaufmann Publishers., An Imprint of Elsevier, 2006.

[3] E.Barati, M. Saraee, A.Mohammadi, N. Adibi and M.R. Ahamadzadeh,"A survey on utilization of data mining approaches for dermatological (skin) diseases prediction," *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI):* March Edition, pp.1-11, 2011

[4] G.Subhalakshmi, K.Ramesh and M. Chinna Rao, "Decision support in heart disease prediction system using naive bayes," *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol.2, No.2, pp.170-176, April-May 2011.

[5] Kenneth Revett, Florin Gorunescu, Abdel – Badeesh Salem and El-Sayed El-Dahshan, "Evaluation of the feature space of erythematosquamous dataset using rough sets," *Annals of University of Craiova, Math. Comp. Sci. Ser.* Vol. 36(2), pp. 123 – 130, 2009.

[6] Dariusz Matyja, "Application of data mining algorithms to analysis of medical data," Master Software Engineering thesis, Blekinge Institute of Technology, Ronneby, Sweden, Aug. 2007.

