Evaluation of Diagnosis according to Myanmar Traditional Medicine by using Expectation Maximization

Hnin Wai Wai Hlaing, Myint Thuzar Tun
Computer University (Maubin)
hninwaiwaihlaing1111@gmail.com, myintthuzartun@gmail.com

Abstract

Nowadays, computer based medical system is playing a role in assisting both diagnosis and treatment. Thus, this system intends to provide information for junior traditional medicine practitioners and user who interested traditional medicine. Before evaluating, this system stores the knowledge of traditional medical experts and medical records from previous cases as training database. And, it produces the generate rules from training data set by using Naïve Bayesian Classifier. When user inputs symptoms, this system analyzes corrected diagnosis and suitable dosage. If user inputted symptoms are not evolved by NB classification, we use Expectation Maximization (EM) step that computes maximum likelihood estimation of unlabeled data. This EM step probabilistically evaluates unlabeled data by using available labeled data which is training by NB. As a result, in this paper, we evaluate corrected diagnosis and proper dosage by using semisupervised learning method (EM with NB classification) in order to improve correctness of classifier.

Keywords: traditional medicine, naïve bayesian classifier, expectation maximization step, symptoms

1. Introduction

In all countries of the world, there exists traditional knowledge related to the health of humans. Traditional medicine is known as "the health practices, approaches, knowledge and beliefs incorporating plant, animal and mineral-based medicines, spiritual therapies, manual techniques and exercises, applied singularly or in combination to treat, diagnose and prevent illnesses or maintain well-being"[10]. The Traditional Medicine has existed in Myanmar since time immemorial. In Myanmar Traditional Medicine, the diagnosis is classified as appropriate result which is based on the patients' suffering symptoms in order to give the suggested medicine. Knowing the corrected diagnosis from patient's symptoms can save the life of many patients. As well, information is required to decide appropriate conditions, how the required information can be extracted and criteria are needed for the users to get the relevant decision. Therefore, this system is intended to give information about the diagnosis and the related dosages for user inputted symptoms by using a semisupervised learning method- that's the combination of Naïve Bayesian classifier and Expectation Maximization step.

Combining the Expectation Maximization with Naive Bayesian classifier is one of the promising approaches for making use of unlabeled data [7]. The Naive Bayesian classifier is a simple but it is an effective classifier which has been used in numerous applications of information processing. But one key difficulty with Naive Bayesian classifier is that it requires a large number of labeled training examples to learn accurately. Labeling must often be done by a person; this is a painfully time-consuming process. The accuracy of Naïve Bayesian classifier also depends on training data. If user inputted data pattern isn't evaluated by Naïve Bayesian classifier, the classification may degrade the accuracy for the inputted data. The accuracy of classifier can be improved by using EM step.

The expectation maximization (EM) step computes maximum likelihood estimates of unlabeled data in probabilistic models involving missing values. The unlabeled data are considered incomplete because they come without class labels. Accordingly, this system first trains a NB classifier with only the available labeled data, and uses the classifier to assign probabilistically class labels to each unlabeled data by calculating the expectation of the missing class labels. In its maximum likelihood formulation, EM finds the maximum likelihood of all the data - both the labeled and unlabeled. [5] Consequently, this paper describes to evaluate a diagnosis based on the symptoms of the patients by using the combination of NB and EM.

We describe six sections in this paper. This section introduces the evaluating of diagnosis according to Myanmar Traditional Medicine using Expectation Maximization. The remaining sections of this paper are organized as follows. In Section 2, the knowledge of Traditional Medicine and the hypothesis of NB classification and EM step will be discussed for this system. In this Section 3, we mention the process of assessment diagnosis. The next Section 4 presents the symptoms' data set in training database and computing the result for this system. In Section 5, we present the implementation

of the system by using the combination of Naive Bayesian classifier and Expectation Maximization step. The last section describes the conclusion of this system.

2. Evolution Methodology

Data mining refers to extracting or mining knowledge from large amount of data. It is a process that uses a variety of data analysis tools to discover patterns with relationships in data and to make valid classification. Classification is the form of data analysis that can be used to extract models describing import data class or to predict future data trends. Classification predicts categorical (discrete, unordered) labels.[4] Classification is used different methods based on the nature of the data such as Naïve Bayesian classification is used for the dataset that has class label, the combination of Naïve Bayesian classifier and Expectation Maximization step is used for the dataset that contains unlabeled data, etc.

The combination of NB classifier and EM step in data mining techniques plays an important role in information retrieve process for extraction the most useful data from several records [5]. In this system, the combination of NB and EM is used to develop a diagnosis system for Myanmar Traditional Medicine based on the symptoms of the patients. Furthermore, the database in this system is composed with a variety of symptoms information that can be used for evaluating a diagnosis. Therefore, we initially explain about the knowledge of Traditional Medicine. We also give the hypothesis of Naïve Bayesian classifier and Expectation Maximization method.

2.1 Knowledge of Traditional Medicine

Traditional Medicine (also known as indigenous or folk medicine) comprises medical knowledge systems that developed over generations within various societies before the era of modern medicine [10]. The totals of 232 different kinds of raw materials are used to formulate the traditional drugs. Among them, 183 (79%) are from plant origin and other from animal origin, minerals and aquatic origin. Main parts used of medicinal plants are stem and root. Flower, leaf, bark, rhizome and fruits are also used in formulation.

Traditional medicine currently practiced in Myanmar has four main components:

The Desana Method: It is based on natural phenomenon such as hot and cool. Its concepts largely depend on Buddhist Philosophy, with the therapeutic use of herbal, mineral compounds and diet.

The Bethitzza Method: It is based on Ayurvedic concepts with extensive use of herbal and mineral compounds to establish balance among three dosas namely Vata, Pitta and Kapha.

The Astrological Method: It is based on the calculation of zodiac of stars, planets and the time of birth and age. These calculations are linked to prescribe dietary practices.

The Vezzadara Method: It is largely dependent on meditation and practices of alchemy. The skill, know-how and techniques of the drug preparations are such that they are derived from heavy metals such as arsenic and its compounds after they are converted into insert ones by means of series of chemical processes, in order to obtain supernatural power. [2]

2.2 Naive Bayesian Classifier (NB)

Bayesian classifier is a statistical classifier and can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. The Naive Bayesian classifier is a simple but is an effective classifier which has been used in numerous applications of information processing such as image recognition, natural language processing and information retrieval, etc.

The Naïve Bayesian classifier works as follows: Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. Thus this system finds the class that maximizes $P(C_i|X)$. The class C_i for which $P(C_i|X)$ maximizes that is called the maximum posteriori hypothesis. By Bayes' theorem for each class according to equation (1):

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}, \qquad (1)$$

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$ in order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made in equation (2). This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple.

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \qquad (2)$$

In order to predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i in equation (3). The classifier predicts that the class label of tuple X is the class Ci if and only if.

$$P(C_i \mid X)P(C_i) > P(C_j \mid X)P(C_j); 1 \le j \le m., j \ne i \qquad (3)$$

In other words, the predicted class label is the class C_i for which $P(X|C_i)P(C_i)$ is that maximum.

In theory, Bayesian classification has the minimum error rate in comparison to all other classifiers. However, owing to inaccuracies in the assumptions made for it is such as the lack of available probability data. For that reason, we additionally used Expectation Maximization method in this system.

2.3 Expectation Maximization (EM)

The EM is a method that computes a conditional expectation and solves a maximization problem, hence the name Expectation Maximization (EM). The EM consists of the E-step in which the expected values of the missing sufficient statistics given the observed data is computed, and the M-step in which the expected values of the sufficient statistics computed in the E-step are used to calculate complete the data maximum likelihood estimation.

E step: It assigns x_i to classify C_k with the probability of equation (4).

$$P(x_i \in C_k) = P(C_k \mid x_i) = \frac{P(C_k)P(x_i \mid C_k)}{P(x_i)} \quad (4)$$

This step calculates the probability of class label membership of x_i, for each class label. These probabilities are the expected class membership for x_i.

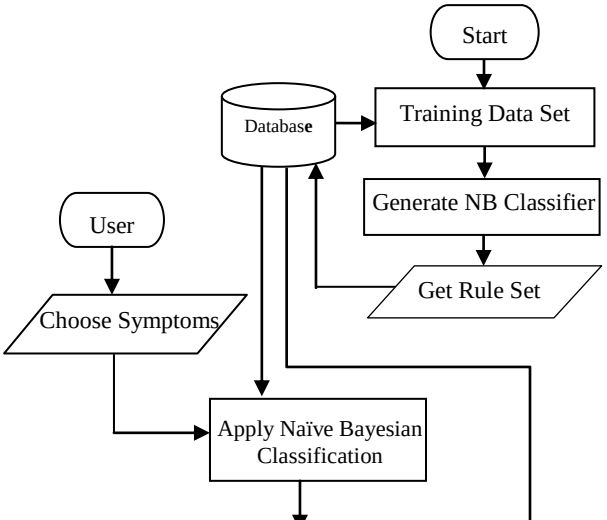
M step: It uses the probability estimates from E-step to re-estimate (or refine) in equation (5).

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i P(x_i \in C_k)}{\sum_j P(x_i \in C_j)} \quad (5)$$

This step is the maximization of the likelihood from the given data. [4]

3. Proposed System Design

In this system, we distinguish into three stages within the process of evaluation diagnosis: (i) prepares the patients symptoms as the variable data for the use of mining techniques; (ii) generates rules by using Naïve Bayesian classifier and (iii) evaluates what kind of diagnosis for the new entire symptoms inputting from user by using the combination of NB classifier and EM method as shown in Figure 1.



(i) Preparing Stage: There are many popular diagnoses such as hypertension, diabetes mellitus and ama vata, etc. In general, hypertension is divided into Tikshna type and Mana type; Diabetes mellitus is separated into Sangahita and Prissava; and Ama Vata is divided into Pitta dominant type and Kapha dominant type according to Myanmar Traditional Medicine. Each kind of diagnosis has evidence symptoms. These symptoms are very useful data by developing computer based medical system. The proposed system collects the most occurrence symptoms and several patients' records that are stored in database. Moreover, this system also collects the standard dosages with related diagnosis in order to give suitable dosage about user inputted symptoms.

(ii) Generating Stage: Each patient's records represent data sets with 16 attributes (sleepiness, giddiness, anger, urine color, constipation, clammy, pain, indigestion, body weight, blood pressure, thirst, appetite, fever, depression, complexion and classvar). In this system, patient records are separated into six different classes by using classification rule sets. Each class represents a distinct returning decision of diagnosis type: hypertension (h1 or h2) or diabetes mellitus (dm1 or dm2) or ama vata (av1 or av2). Naïve Bayesian classifier uses a probabilistic approach to classify and generate classification rules. It attempts to compute conditional class probabilities and predicts the most probable class. For this purpose, the evaluating of diagnosis will be computed by the Naïve Bayesian classifier's rule sets that are based on 16 attributes with each record per patient in training data set.

(iii) Evaluating Stage: The user enters the new symptoms concerning with his/her feeling. The discrete values substitute in the combination of Naïve

Bayesian Classifier and Expectation Maximization Step. This system firstly applies the Naïve Bayesian classification to determine a related diagnosis. If the NB classification is satisfied, then displays the related diagnosis and suitable dosage of the user inputted symptoms. Otherwise, this system operates EM step that finds the maximum probability of diagnoses and displays the possibility of diagnosis and appropriate dosage. Thus, this step evaluates a diagnosis as a result when user inputs the awareness symptoms.

4. Data Set and Result for the System

We use data mining techniques to investigate various diagnoses. This system uses the EM step with NB classification and makes available to support junior traditional medicine practitioners and the user who interested in traditional medicine. Before classify the data, this system takes together and converts of patients' data. The total data is grouped as 200 training data set. In classification data, this system generates the rules from the training data set by using the Naïve Bayesian classifier. The "TrainData" gets symptoms from the "PrepareData" for taking out the generate rules. These generate rules are extracted by using NB based on training data set from "TrainData" table. In this system, the combination of NB and EM method is used for better performance in the evaluating of diagnosis according to traditional medicine.

4.1 Data Set

The data sets have been taken from "Department of Traditional Medicine (Maubin), Ministry of Health". These data sets are included in sixteen fields such as sleepiness, giddiness, anger, urine color, constipation, clammy, pain, indigestion, body weight, blood pressure, thirst, appetite, fever, depression, complexion and classvar. Each field has related features. We describe Table 1 to identify patient's symptoms and classvar with features.

"Table 1: Identification of Patient's Symptoms and Features"

Symptoms of Patients	Features
Sleepiness	high, normal, low
Giddiness	yes, no
Anger	yes, no
Urine color	yellow, straw
Constipation	yes, no
Clammy	yes, no
Pain	neck, chest, joint, no
Indigestion	yes, no
Body weight	normal, less
Appetite	normal, loss
Fever	yes, no
Depression	yes, no
Blood pressure	high, normal, low
Thirst	yes, no
Complexion	yes, no
Classvar	h1,h2,dm1,dm2,av1,av2

4.2Evaluating Result for New Patent data

When the user inputs symptoms, this system calculates the class variable (h1, h2, dm1, dm2, av1and av2). In the class variable, "h1" is Tikshna type of hypertension, "h2" is Mana type of hypertension, "dm1" is Sangahita type of diabetes mellitus, "dm2" is Prissava type of diabetes mellitus, "av1" is Pitta dominant type of ama vata and "av2" is Kapha dominant type of ama vata.

(i) Calculation the Result with NB Classification

In this system, we suppose that there are six classes, "h1", "h2", "dm1", "dm2", "av1" and "av2. Assume that user input symptoms are:

X=(sleepiness=low, giddiness=no, anger =no, urine color=straw, constipation=no, clammy=no, pain=joint, indigestion=yes, bodyweight=normal, appetite=loss, fever=yes, depression=yes, blood pressure= normal, thirst=no, complexion=yes).

We calculate the user inputted symptoms with Naïve Bayesian Classifier. Given a set of training data, $P(C_i)$ can be estimated by counting how often each class occurs in the training data.

P(classvar=h1)=34/200=0.17

P(classvar=h2)=32/200=0.16

P(classvar=dm1)=33/200=0.165

P(classvar=dm2)=33/200=0.165

P(classvar=av1)=42/200=0.21

P(classvar=av2)=26/200=0.13

The probabilities $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)$,.., $P(x_{15} \mid C_i)$ can be estimated from the training data. To reduce the computational expense in estimating $P(X \mid C_i)$ for all possible Xs, the classifier makes a naïve

assumption that the attributes (features) used in describing X are conditionally independent of each other given the class of X by using equation(2), respectively.

P(X|classvar=h1)=0

P(X|classvar=h2)=0

P(X|classvar=dm1)=0

P(X|classvar=dm2)=0

P(X|classvar=av1)= 0.00137548113

P(X|classvar=av2)=0

The classifier predicts these user input symptoms as belonging to the available class having the highest posterior probability conditional on X. As P(X) is constant for all six classes, only $P(X \mid C_i)$ P (C_i) needs

P(X|classvar=h1) P(classvar=h1)=0

P(X|classvar=h2) P(classvar=h2)=0

P(X|classvar=dm1) P(classvar=dm1)=0

P(X|classvar=dm2) P(classvar=dm2)=0

P(X|classvar=av1)P(classvar=av1)= 0.0002888516636

P(X|classvar=av2) P(classvar=av2)= 0

to be maximized by mean of equation (3), respectively.

As a result, the Naïve Bayesian classifier evaluates the result of classvar = "av1" for input sample X. The result of NB classification is tested on 30 cases of patients' data. Afterward, Naïve Bayesian classification is evaluated 20 cases on these cases. Thus, in this system, the accuracy result of NB classification has 66.67 % of these data.

(ii) Calculating the Result with EM step

EM starts with an initial estimate that each object is assigned a probability that it would posses a certain set of attribute values given that it was a member of a given class. NB classification focuses on only the computation of class-conditional probability density. In some calculation, NB classifier could not analyze the best class on the given symptoms during the process of evaluating diagnosis.

In order that, to evaluate the best class in this system, the EM step with NB classifier maximizes the ability to predict the attributes of an object and given the corrected class of the object.

For that reason, we give explanation as an example in this paper that the NB classifier does not evaluate and EM step with NB classifier solves the following user inputted symptoms:

X=(sleepiness=low, giddiness=yes, anger =no, urine-color=yellow, constipation=yes, clammy=yes, pain=no, indigestion=yes, bodyweight=normal, appetite=loss, fever=yes, depression=no, blood pressure=high, thirst=yes, complexion=no).

Initially, we calculate the user inputted symptoms with Naïve Bayesian Classifier:

P(X|classvar=h1)=0

P(X|classvar=h2)=0

P(X|classvar=dm1)=0

P(X|classvar=dm2)=0

P(X|classvar=av1)=0

P(X|classvar=av2)=0

P(X|classvar=h1) P(classvar=h1)=0

P(X|classvar=h2) P(classvar=h2)=0

P(X|classvar=dm1) P(classvar=dm1)=0

P(X|classvar=dm2) P(classvar=dm2)=0

P(X|classvar=av1) P(classvar=av1)=0

P(X|classvar=av2) P(classvar=av2)= 0

We calculate the probability of each symptom in diagnosis. Amount of some probabilities of symptoms in each diagnosis has zero and then the probability of diagnosis concerning these symptoms has zero. After that, this system checks the probabilities of all diagnosis (h1, h2, dm1, dm2, av1 and av2). If all these probabilities have zero, at this time, the Naïve Bayesian classifier can not predict classvar. Thus, we use Expectation Maximization step to predict classvar by applying Equation (4) and (5).

$$m_{h1} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{h1})}{\sum_j P(x_i \in C_j)}\right) = 4.6269$$

$$m_{h2} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{h2})}{\sum_j P(x_i \in C_j)}\right) = 3.684416$$

$$m_{dm1} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{dm1})}{\sum_j P(x_i \in C_j)}\right) = 3.7449$$

$$m_{dm2} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{dm2})}{\sum_j P(x_i \in C_j)}\right) = 3.117577$$

$$m_{av1} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{av1})}{\sum_j P(x_i \in C_j)}\right) = 4.116659$$

$$m_{av2} = (1/15)\left(\sum_{i=1}^{15} \frac{x_i P(x_i \in C_{av2})}{\sum_j P(x_i \in C_j)}\right) = 4.104072$$

Therefore, the Expectation Maximization step computes the maximizations of the six classes based on the probability estimates. As a result, m_{h1} is the most maximization in six classes for the input user symptoms X. With the intention that, EM step evaluates classvar = "h1" to choice the proper diagnosis.

In this system, 10 cases of patients' data in 30 cases are not evaluated by NB classification. Thus, 5 cases of patients' data amount the remaining 10 cases are evaluated by using EM step. At that time, in this system, the accuracy result of EM step has 83.33% of patients' data.

5. Implementation of the System

This system is intended to evaluate diagnosis of patient by using Naïve Bayesian classifier and Expectation Maximization method. It is developed by using C#.NET and also used Microsoft SQL Sever 2005 to build the database.

Foremost, this system evaluates the appropriate diagnosis by using NB classifier. This classifier primarily trains with the information from precedent patients' records and works out the suitable diagnosis

for current circumstances of new patient based on training patients' records. When user chooses the fifteen attributes options for patient's symptoms, he or she clicks a "Calculate Naïve Bayesian" button to find out a diagnosis as shown in Figure 2.

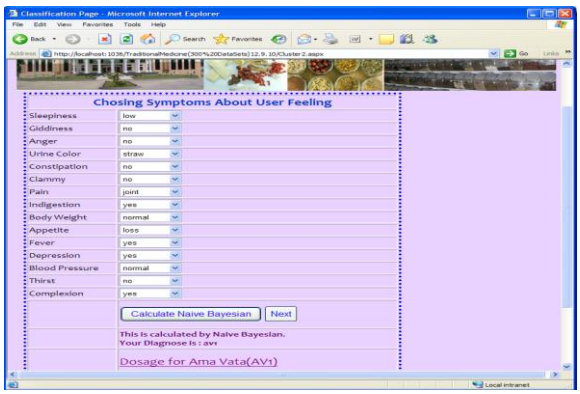


Figure 2: Evaluating of Diagnosis by NB Classifier

If NB classifier can't determine the kind of diagnosis for user inputted symptoms, user can also use "Calculate EM" button to guess a diagnosis for user suffering symptoms as shown in Figure 3.

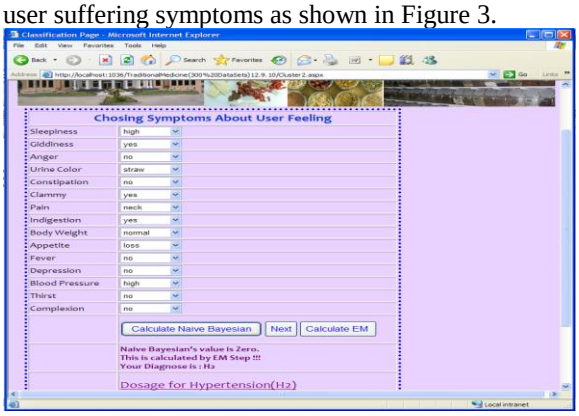


Figure 3: Evaluating Diagnosis by EM Step

In this system, user can also observe suitable dosage of traditional medicine that is related to the diagnosis of user by going to the related link as illustrate as Figure 4.

Figure 4: Dosage for Suitable Diagnosis

6. Conclusion

This system focuses on the evaluation of diagnosis according to traditional medicine by using the combination of NB and EM methods. By the effectiveness of this system, the user can quickly know diagnosis and correctly utilizes suitable dosage. Additionally, this system also gives the knowledge of traditional medicine and usage for user who interested in traditional medicine. Therefore, this system can give effectiveness valuable services for users.

References

[1] A. P. Dempster, N. M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm". Journal of Royal Statistical Society Series.

[2] Compiled and Distributed by Department of Traditional Medicine, Ministry of Health, "Traditional Medicine Handbook". In collaboration with Japan International Cooperation Agency (JICA).

[3] H. Zhang and J. Su: "Naive Bayesian Classifiers for Ranking", Faculty of Computer Science, University of New Brunswick.

[4] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2001

[5] K. Nigam, A. K. Macallum, S. Turun and T. Mitchell: "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning (2000).

[6] V. P. Kamboj, "Herbal Medicine". Central Drug Research Institute. Current Science, VOL 78, NO.1, 10 January 2000.

[7] Y. Tsuruoka and J. Tsujii, "Training a Naïve Bayes Classifier via the EM algorithm with a Class Distribution Constraint", CREST, JST (Japan Science and Technology Corporation)

[8] Z. Ghahramani and M. Jardan, "Supervised Learning from Incomplete Data via an EM Approach". In NIPS6, 1994.

[9]http://en.wikipedia.org/wiki/Naive_Bayesclassifier

[10]http://www.who.int/topics/traditional_medicine/en