Decision Making for Diseases by using Bayesian Classifier

Than Than Myint, Soe Hay Mar
University of Computer, Monywa, Myanmar
thandar.monywa@gmail.com, drsoehaymar@gmail.com

Abstract

Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. In Classification techniques, Naïve Bayesian Classifier is one of the simplest probabilistic classifiers. Bayesian Classifier is based on Bayes Theorem. Bayesian Classifiers are useful in predicting the probability that a sample belongs to a particular class. This paper studies the Naïve Bayesian Classifier and to classify class label of diseases data using Naïve Bayesian Classifier. This paper focuses on tooth care dataset and decides that tooth is care or prevention or care and prevention. This example of dataset contains 2160 instance and 8 attributes from UCI machine learning repository.

1. Introduction

Data Mining refers to extracting or mining knowledge from large amounts of data. Data Mining or Knowledge discovery is the computer- assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides [3].

Classification is an important technique in data mining. Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification can be used for predicting the class label of data objects. Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data ts and it predicts categorical labels [1]. For example, a classification model may be built to categorize bank loan applications as either safe or risky. Many classification methods have been proposed by researchers and are important area of research and of practical application in a variety of fields, including pattern recognition and artificial intelligence, statistics, vision analysis, medicine and so on [6].

A classification model can also be used to predict the class label of unknown records. A classification model can be treated as black box that automatically assigns a class label when presented with the attribute set of an unknown record.

2. Background of the system

The system can be classified every disease. But this paper is used tooth training dataset attributes. There are eight attributes in the dataset. They are ID, Age, Location, Frequency, Material, Method, CPI, Oral Habit, Cure-frequency and Decision. Age contains 5>=, 14>=, 15>=, 44>= and 65>=. There are periurban, Rural and Urban in the location. Frequency contains 1, 2 and 3 times. Materials contain tooth paste and salt. There are horizontal and vertical in the Method tuple. There are Bleeding, Calculus and Healthy in the CPI. There are 1, 2, 3…., 9 times in the Cure-frequency. There are cure, cure and prevention and prevention in the Decision. There are about 2300 records in the tooth training dataset.

3. Classification and prediction

Classification is the process of finding a set of models (or functions) that described and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of object whose class label is unknown. The derived model id based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects, users may wish to predict some mission or unavailable data values [3].

4. Bayesian classification

Bayesian Classifiers are statistical classifier. The naïve Bayes classifier is a Bayesian learning method that has been found to be useful in many practical applications. It is called "Naïve" because it incorporates the simplifying assumption that attribute values are conditionally independent, given the classification of the instance [4]. The researchers shown that the Naïve Bayes classifier is competitive with other learning algorithms, including decision tree and neural network algorithms in many cases and that in some cases it outperforms these other methods [5].

Bayesian classification is based on Bayes Theorem. Bayesian classifiers are useful in predicting the probability that a sample belongs to a particular class or grouping. This technique tends to be highly accurate and fast, making it useful on large databases. Bayesian Classifiers have also exhibited high accuracy and speed when applied to large database. Bayesian Classifiers have also the minimum error rate in comparison to all other classifiers. Bayesian classifiers use:

- Probabilistic Learning: Calculate explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems,
- Incremental: Each training sample can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data,
- Probabilistic prediction: Predict multiple hypotheses, weighted by their probabilities and
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

4.1. Naive Bayesian

The Naïve Bayesian Classifier, or simple Bayesian classifier, works as follows:

1. Each data sample is represented by n-dimensional feature vector, $X=(x_1, x_2 \ldots x_n)$ depicting n measurements made on the sample from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample X. the naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if
   $P(C_i \backslash X) > P(C_j \backslash X)$ for $1 \leq j \leq m, j \neq i$.
   Thus we maximize $P(C_i \backslash X)$. The class C_i for which $P(C_i \backslash X)$ is maximized is called maximum posteriori hypothesis. By Bayes theorem,
   $$P(C_i \backslash X) = P(X \backslash C_i)\, P(C_i)/ P(X)$$
3. As P(X) is constant for all classes, only $P(X \backslash C_i)$ $P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots\ldots = P(C_m)$ and we would therefore maximize $P(X \backslash C_i)\, P(C_i)$.
4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X \backslash C_i)$. In order to reduce computation in evaluating $P(X \backslash C_i)$, the Naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, that is, there are no dependence relationships among the attributes. Thus,
   $$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i)$$
   The probability $P(x_1 \backslash C_i), P(x_2 \backslash C_i), \ldots P(x_n \backslash C_i)$ can be estimated from the training samples.
5. In order to classify an unknown sample X, $P(X \backslash C_i)\, P(C_i)$ is evaluated for each class C_i. Sample X is then assigned to the class C_i if and only if
   $P(X \backslash C_i)\, P(C_i) > P(X \backslash C_j)\, P(C_j)$ for $1 \leq j \leq m, j \neq i$.
   In other words, it is assigned to the class C_i for which $P(X \backslash C_i)\, P(C_i)$ is the maximum.

4.2. Characteristic of naive Bayesian classifiers

Naïve Bayesian Classifiers generally have the following characteristics.

They are robust to isolated noise points because such points are averaged out when estimating conditional probabilities from data. Naïve Bayes Classifiers can also handle missing values by ignoring the example during model building and classification,

They are robust to irrelevant attributes. If X_i is an irrelevant attribute, then $P(X_i \backslash Y)$ becomes almost uniformly distributed. The class conditional probability for X_i has no impact on the overall computation of the posterior probability.

Correlated attributes can degrade the performance of Naïve classifiers because the conditional independence assumption no longer holds for such attributes.

4.3. Naive Bayes assumption

$$P(a_1, a_2 \ldots a_n \backslash V_j) = \prod_i P(a_i \backslash V_j)$$

which gives Naïve Bayes Classifer:

$X_{NB} = \text{argmax}_{vj \text{ in } V} P(V_j) \prod_i P(a_i \backslash V_j)$

4.4. Naive Bayes algorithm

For each target value V_j

$\hat{P}(V_j) \leftarrow$ estimate $P(V_j)$

For each attribute value ai of each attribute a

$\hat{P}(a_i \backslash V_j) \leftarrow$ estimate $P(a_i \backslash V_j)$

Classify_New_Instance(X)

$X_{NB} = \text{argmax}_{vj \text{ in } V} P(V_j) \prod_i P(a_i \backslash V_j)$

Algorithm : Naïve Bayesian Classification. Predict class membership probabilities, such as the probability that a given Sample belongs to a particular class.

Input : Database, C, of the selected training samples dataset and Unknown data X.

Output : Predict class membership

Method:

```
k = total record count of training sample dataset C:

for (i = 0; i < Ck-1; i++)

{  if (C.record (i).cell (Result).value == C)

        Cure Count ++;

    else if (C.record (i).cell (Result).value == P)

        Pre Count ++;

    elseif (C record (i) cell (Result) value== C and
    P)

        Cure and Pre count ++;

}

    total CureProb = Cure Count/k;

    total PreProb =Pre Count/k;

    total Cure and Pre Pro= Cure and Pre count/k;

    n= total cells count in each record of training
    sample dataset T expect ID and Result fields

    for (i = 0; i < Tk-1; i ++)

    {

        C Prob = 1;

        P Prob = 1;

        C and P Prob=1;

        for (j = 0; j <Tn-1; j ++)

        {

C Count = getCount (j,T.record(i).Cell(j).value,
C);

C Prob *= C Count / Cure Count;

PCount = getCount (j,T.record(i).Cell(j).value,
P);

PProb *= PCount / Pre Count;

C and P count = get count;

(j T record (i) cell (j) value, C and P);

C and P Prob*= C and P count/ Cure and Prd
count; }

if (Cprob*total Cure prob>Pprob*total Preoprob
&&c prob* total Cure Prob> C & P prob* total
Cure and pre prob)

display result as decision Cure;

else if (Pprob*total Per oprob> Cprob* total
Cure prob && P Prob* total Pre Prob> C and P
Prob* total Cure and Pre Prob)

display result as decision Pre;

elseif ( C and P  Prob* total Cure and Pre Prob>
C Prob*total Cure Prob &&C and P Prob* total
Cure and Pre Prob> P Prob* total Pre Prob)

Display result as decision Cure and Pre;

}

Procedure getCount (colIndex, colVal, result)

    k = total record count of training sample
    dataset C;

    count = 0;

    for (i = 0; i < Ck-1; i ++)
```

```
{
        if (C.record (i).cell
        (colIndex).value == colVal &&
        C.record (i).cell (i).cell
        (Result).value == result)

        count ++;

}

return count;
```

5. **Overview of the system**

The researcher can import the user's requirement dataset into the system. And then user can add/modify sample data record sets as necessary. The researcher needs to import training sample for classification system. The researcher may view the selected training samples. To classify the unknown record set, the researcher must put the attributes value of the unknown symptoms of user's import dataset into the system. In this system, there are three main processes. They are:

(1)Process 1

When researcher entries training samples, process 1 trains the system by calculating each fact's prior probability and then saves training samples data and abbreviation facts data to training samples file and abbreviation facts file.
(2)Process 2

Process 2 compares facts and get pre-calculated probabilities for each attribute of data. Then it enters probabilities of each attribute of data to process 3.

(3)Process 3

Process 3 calculates total and maximum predictable probabilities or conditional probabilities for each predictable result and then outputs maximum possible predicted results by comparing such probabilities result.

6. Proposed design

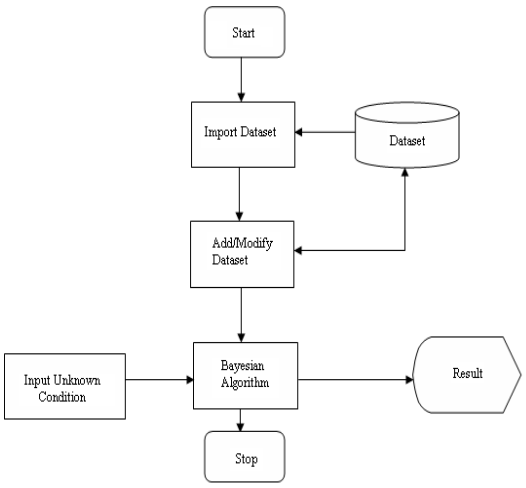

Figure 1. System flow diagram of the system

The researcher can add/modify sample data record sets as necessary. The researcher needs to import training sample and testing sample record sets to test the accuracy of the classification system. The researcher may view the selected training samples and testing samples record sets. Once the training and testing samples have been selected, the system is ready to start testing the classification accuracy. To classify the unknown record set, the researcher must put the attributes value of the unknown mushroom into the system.

7. Effectives of naïve Bayesian classifier

Naïve Bayesian Classifiers have the minimum error rate in comparison to all other classifiers. Bayesian Classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes theorem.

7.1. Holdout method

In the holdout method, the given data were randomly partitioned into two independent sets, training set and a test set as in Figure 2. Typically two third of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set was used to derive the classifier, whose accuracy was estimated with the test set.

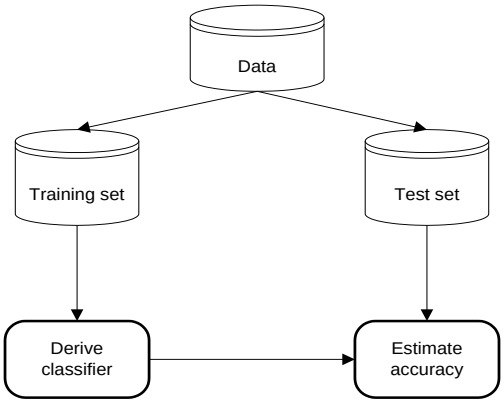

Figure 2. Estimating classifier performances with holdout method

7.2. Accuracy Result

In this paper, 94% classification accuracy after reviewing 200 (id<=200) instances, 92% classification accuracy after reviewing 200 (id>200 and id<= 400) and so on as shown in Table 1.

Table 1. Accuracy table

Training data	Testing data	Accuracy
510 (id>200)	200(id<=200)	94%
510 (id<=200 or id>400)	200 (id>200 and id<=400)	92%
510 (id<=400 or id>600)	200 (id >400 and id<=600)	93%
510 (id<=510)	200 (id>510)	93%
476 (id<=476)	243 (id>476)	94%
476 (id>234)	243 (id<=243)	94%
476 (id<=200 or id>434)	243 (id>200 or id<=434)	95%
476 (id<=400 or id>634)	234 (id>400 or id<=634)	96%

8. Implementation

There are four main categories in the system. They are "Classification System", 'View', 'Windows' and 'Help'. In "Classification System" contains "Add/Modify training dataset", "Record sets Selection", "Naïve Bayesian Classifier" and "Exit" functions. In "View" menu contains the function of toolbar controls. In 'Windows' menu contains the function of windows style and 'Help' menu described about the information of the system.

User can select the desire dataset from the combo box and then clicked the preview button. So, the system will display the training records. The user can select Import samples from dataset. In this form, the user can choose Training Samples and then import training sample data using Filter box. If the user chooses "Tooth" dataset" button, the tooth training dataset will display as shown in Figure 3.

User wants to classify unknown patient suffers toothache can be prevented or it is class label, choose Classify Unknown from Classify Menu or click Classify Unknown button. It will be indisable state after opening unknown classification window. It may take less than a minute or take several minutes depending on the machine, the amount of input unknown diseases facts record and the amount of machine's training sample records.

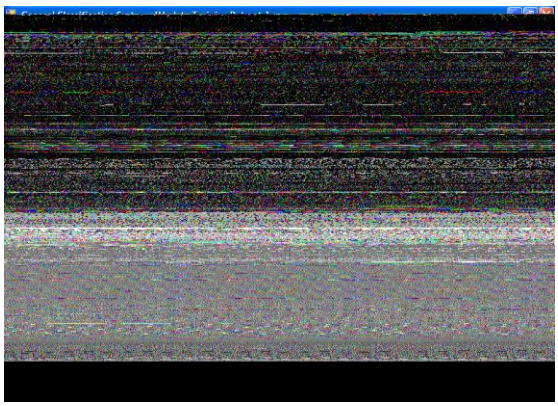

Figure 3. Training dataset

While classification, the progress bar will show the progress of calculation of probabilities by Naïve Bayesian classification method and predicting result progress. The user can view each diseases fact affected on prevention, cure and cure and prevention probabilities on conditional probabilities grid and by double clicking on each probability record will show total prevention, cure/cure and prevention.

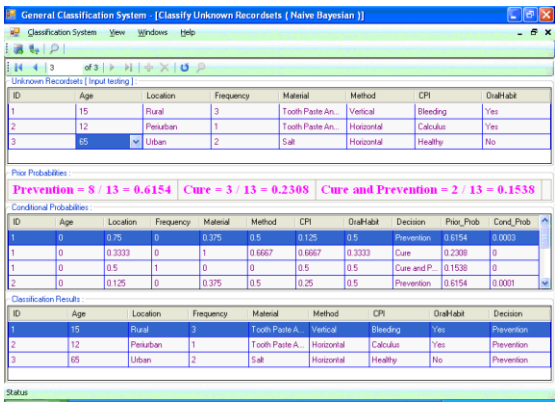

Figure 4. Naïve Bayesian classification

If maximum probability is chosen from probabilities, then application (Naïve Bayesian classification method) put those diseases as resulting prevention.

9. Conclusion

Naïve Bayesian classification is based on Bayes theorem of posterior probability. The Bayesian classifier uses the naïve Bayesian formula to calculate the probability of each class given the values of all attributes. This paper is predicted the class label of unknown sample given the training data. The relative performance of the naïve Bayesian classifier can serve as an estimate of the conditional independence of attributes. This system is presented generating of classification from large datasets. This approach demonstrates efficiency and effectiveness in dealing with the users requirement data for classification. This paper is considered the classification problem of tooth training dataset by using Naïve Bayesian classifier.

References

[1] J. Han and M.Xin, "*Discovering web Access patterns and trends by applying OLAP and data mining technology on Web*", 1997.

[2]J.Hamilton,"*Correlation-based Feature Selection for Machine Learning*" (PhD-Thesis) April, The University of Waikato, NewZealand, 1999.

[3] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, 2001.

[4] D.A. Keim, "*Knowledge Discovery and Data Mining*", Newport Beach, USA, 1997.

[5] H. Lu, R. Setino and H. Liu, "*Neurorule: A connectionist approach to data mining*", VLDB, Switzerland, 1995.

[6] P. Ning, T. M.Steinbach and V.Kumar, " *Introduction to Data Mining*", 1998.