Classification of butterfly's sub families system using fuzzy-bayesian decision method

Nway Yu Aung, Thandar Aung
Computer University (Mandalay)
nwayuaung@gmail.com

Abstract

Data Mining refers to extracting or mining knowledge from large amounts of data. Classification is an important technique in data mining. Classification is the process of finding a set of models that describe and distinguish data classes, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. In this paper, Fuzzy-Bayesian Classifier is one of simplest probabilistic classifiers. It is based on Bayes Theorem. A fuzzy logic-based methodology for classification is developed and used in supervised learning. In this paper, Fuzzy-Bayesian is used to build a classifier using a set of butterfly training dataset and to test the unknown dataset. The evaluation of the performance of the classifier is based on the feature set by using the hold-out method as the evaluation criteria. The experiment is performed on sub family of butterfly dataset from UC Irvine Repository of Machine Learning Databases.

1. Introduction

Decision-making problems meet in the daily life or working environment. Sometimes it is very difficult to make good decision. In practice, decision maker usually use the past experiences to make a decision. These past experiences can see as a form of performing experiments to come to a correct decision. Fortunately, the developments of computer technologies and automatic learning techniques can make this easier and more efficient.

Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends and it predicts categorical labels. In the domain of machine learning where it always lets computers decide or come up with suggestions for the right decision there exist many approaches of decision making techniques, such as decision trees, artificial neural networks and Bayesian learning and so on [6].

This system focuses on the Fuzzy Bayesian approach to solve decision making problems. Decision problems called multiple objective decisions (MODs). This class of problems often involves many vague and ambiguous (and thus fuzzy) goals and constraints. This system developed one possible Fuzzy Bayesian decision-making model and proved to be useful to decision makers in many "real-world" problems.

This paper is organized as follows: after this short introduction, Section 2 describes Motivation, in Section 3 describes Background Theory, in Section 4 describes Case Study of Butterfly Dataset, in Section 5 describes System Overview and Section 6 describes Implementation, and then Conclusion describes in Section 7.

2. Motivation

The majority of real-world classification problems require supervised learning where the underlying class probabilities and class-conditional probabilities are unknown, and each instance is associated with a class label. In real-world situations relevant features are known a priori. Therefore, many candidate features are introduced to better represent domain. To construct a supervised classification system, one requires a data set of labeled examples with which to train and test the system. Each example is made up of a class label. A classification system contains a model of how these features describe the different classes, and it is hoped that this model will allow novel examples, for which no label is available, to be correctly classified.

3. Background Theory

3.1. Preprocessing

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size, often several gigabytes or more. Data processing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation and data reduction. This system may be applied preprocessing step to the data before this system builds the classifier. This refers to the preprocessing of data in order to remove or reduce noise.

3.2. Classification

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science [1]. Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications [3].

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes.

In the second step, the model is used for classification. First, the predictive accuracy of the model or classifier is estimated. If the accuracy of the classifier is considered acceptable, the model can be used to classify future data tuples for which the class label is not known. Such data are also referred to as "unknown" or "previously unseen" data [2].

The classifier design can be performed with labeled or unlabeled data. Using a supervised learning method the computer is given a set of objects with known classification and is asked to classify an unknown object based on the information acquired by it during the training phase. 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.

3.3. Fuzzy-Bayes Approach

Product concept selection belongs to multi criteria decision-making (MCDM) problems. In MCDM problems, a decision maker has to pick the best concept among a set of alternatives or product concepts based on a set of criteria or attributes.

However, in case of conflicting alternatives, the task of picking the best concept becomes extremely difficult due to the imprecise or ambiguous data, which is norm in this type of decision problems. In the absence of complete and precise information, the fuzzy set theory becomes an effective tool for modeling complex system.

3.4. Fuzzy Bayesian Decision Method

Classical statistical decision making involves the notion that the uncertainty in the future can be characterized probabilistically. When the user wants to make a decision among various alternatives, the choices is predicated on information about the future, which is normally discredited into various "states of nature [5].

In the context of classification systems, the information typically includes details about learning material, the tasks and the objectives, the attributes information, the contact information and other records. When perceived in the technical terms of the underlying information system, it soon becomes very difficult to manage, integrate, and access different kinds of information. In this article, seek means to model the imprecision of information and simplify the access to information systems, in terms of fuzzy modeling [7].

$$\text{Let } S = \{s_1, s_2 \ldots s_n\},$$

be a set of possible states of nature; and the probabilities that these states

$$P = \{p(s_1), p(s_2), \ldots, p(s_n)\} \text{ Where } \sum_{i=1}^{n} p(s_i) = 1 \qquad (1)$$

Assume that the decision maker can choose among m alternatives,

$$A = \{a_1, a_{2\ldots} a_m\},$$

and for a given alternative a_j assign a utility value, u_{ji}, if the future state of nature turns out to be state s_i. These utility values should be determined by the decision maker for each alternative-state pair as shown in Table 1.

Table1. Utility Matrix

state s_i action a_j	s_1	s_2	…	s_n
a_1	u_{11}	u_{12}	…	u_{1n}
. . .	. . .	. . .	. . .	. . .
a_m	u_{m1}	u_{m2}	…	u_{mn}

The conditional probabilities, denoted $p(x_k | s_i)$, are also called likelihood values. The likelihood values are then used as weight on the previous information, the prior probabilities $p(s_i)$.To find update probabilities, known as posterior probabilities, denoted $p(s_i | x_k)$. The posterior probabilities are equivalent to this statement [4]:

$$p(s_i \mid x_k) = \frac{p(x_k \mid s_i)}{p(x_k)} p(s_i) \qquad (2)$$

$p(x_k)$, is the marginal probability and it calculate for this equation (3).

$$p(x_k) = \sum_{i=1}^{n} p(x_k \mid s_i).p(s_i) \qquad (3)$$

Now the expected utility for the j^{th} alternative is determined from the posterior probabilities,

$$E(u_j \mid x_k) = \sum_{i=1}^{n} u_{ji}\, p(s_i \mid x_k) \qquad (4)$$

And then, the maximum expected utility is now given by

$$E(u^* \mid x_k) = \max_j E(u_j \mid x_k) \qquad (5)$$

3.5. Normalizing Values

Normalization, that the attributes data are scaled within the small range such as -1.0 to 1.0, or 0.0 to 1.0. Normalization is particularly useful for classification algorithms for classification mining. Normalizing the attributes the input values for each attributes measured will help speed up the learning. Normalization helps prevent attributes with initially large range from outweighing attributes with initially smaller range [2].The value for an attribute A are normalized based on mean and standard deviation of A. A value ν of A is normalized to ν' by computing

$$\nu' = \frac{\nu - \bar{A}}{\sigma_A} \qquad (6)$$

Where, $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attributes A. The standard deviation of A is

$$\sigma_A = \sqrt{\frac{\Sigma(A - \bar{A})^2}{n-1}} \qquad (7)$$

3.6. Classifier Accuracy

Estimating classifier accuracy is important since it determined to evaluate how accurately a given classifier will label future data, on which the classifier has not been trained. Accuracy estimates also help in the comparison of different classifiers.

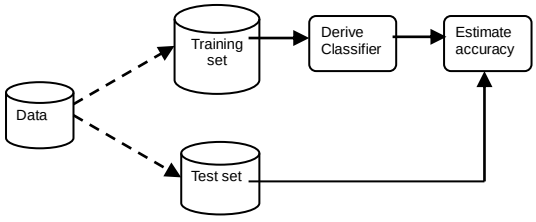

Figure 1. Estimating Classifier Accuracy with Hold-out Method

This system calculates the accuracy with holdout method. In the holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set.

The Sensitivity and specificity measures can be used to determine the accuracy measures. These measures are defined as in equation (8), (9) and (10):

$$\text{Sensitivity} = \frac{t - pos}{pos} \qquad (8)$$

$$\text{Specificity} = \frac{t - neg}{neg} \qquad (9)$$

$$accuracy = sensitivity \frac{pos}{(pos + neg)} + specificity \frac{neg}{(pos + neg)} \qquad (10)$$

4. Case Study Of Butterfly Dataset

Suppose a user, to help make a decision about the whether the family type of the butterfly. Can determined to attempt at the decision process that they are seven states of nature – Anternnae(s_1) ,Head(s_2), Thorax(s_3), Abodmen(s_4), WingShape(s_5), WingColor(s_6), Vein(s_7) and four alternatives- Danaidae(a_1), Rapilionidae(a_2), Nynphalidae(a_3), Pieridae(a_4). This system uses about 1000 number of datasets.

Firstly, calculate the probabilities of state of nature. The sum of these probabilities is equal 1. And then, user wants to consider the utility values. The user collect new information by taking samples from the conditions being considers for classification of family type. The system examines the results of these samples test and gets the researcher's opinions about the conditional probabilities in the form of a matrix. Let X be each data sample.

In this system, the states of the attributes are text data. So, the system converts the related weight values as shown in follow. The weights reflect the significance, importance or occurrence frequency attached to their respective values [2].

$$u_{ij} = \begin{bmatrix} 9 & 6 & 7 & 11 & 10 & 2 & 11 \\ 4 & 9 & 7 & 17 & 13 & 7 & 13 \\ 12 & 5 & 1 & 1 & 9 & 4 & 10 \\ 18 & 13 & 7 & 13 & 16 & 4 & 12 \end{bmatrix}$$

After that, need to normalize the weighted values. So, this system use z-score normalization (or zero-mean normalization) are as follow.

$$u_{ij} = \begin{bmatrix} -0.3522 & -0.7634 & 0.6255 & 0.0861 & -0.7845 & -1.5213 & -1 \\ -1.3585 & 0.2545 & 0.6255 & 1.1189 & 0.3922 & 1.8593 & 3 \\ 0.2516 & -1.1026 & -1.8766 & -1.6353 & -1.1767 & -0.169 & -3 \\ 1.4591 & 1.6116 & 0.6255 & 0.4303 & 1.5689 & -0.169 & 1 \end{bmatrix}$$

5. System Overview

Suppose a user who has made a decision about the family type of butterfly. This paper has four stages. In the first stage, the data are partitioned into training

and testing using the hold-out method over the whole data set. Training set is used to train the classifier. Testing data sets is used to estimate the performance during construction of the system.

In the second stage, fuzzy-Bayesian is used for classification of the instances. This second stage consists of the following stages. Firstly, user needs to consider the utility values. User provides the utility matrix. So, user inputs the main significant conditions of the relative family types. This system calculates the "weighted values" and then "normalizes" it. Further, user chooses the new information by taking samples data from the predefined dataset. The sample tests are capable of providing perfect information. After that, the system calculates the prior probabilities of the samples data and the posterior probabilities and the marginal probabilities. And then, the conditional expected utilities are calculated. Finally, the system will decide what alternative to choose.

The features sets are tested using the testing data set in the third stage. Finally, evaluate the performance of classifier.

5.1. Experimental Methodology

Fuzzy-Bayesian classifier is one of the simplest probabilistic classifier. Fuzzy-Bayesian is based on the Bayes Theorem. The prior and posterior probability of each class is calculated, given feature values present in the instance; the instance is assigned to the class with the highest probability.

Also, calculate the weight value of the related feature by using the utility matrix before prior and posterior is calculated, and then normalized to the weight value. Accuracy of classifier is measured using the hold-out method of a given datasets into train and test sets.

6. Implementation

6.1. Fuzzy Bayesian Classifier

There are four main processes in this system. They are Utility Matrix, Conditional Probabilities, Prior Probabilities and Decision Making.

Firstly, user needs to open "Utility Matrix" process. User need to provide or define the utility values. So, user must choose the significant of the feature of each attributes from the combo box in the related family types and then convert text value to their related weighted value. The weights reflect the significance, importance or occurrence frequency attached to their respective values. After that, need to normalize the weighted values. The attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0. This system used z-score (or zero-mean normalization) normalization method. So, user needs to press "Normalization with Z-score method" button. The system converts the weighted values to normalize values.

In "Conditional Probabilities" process, user needs to collect new information by taking sample from the possible conditions being considered for classify family type. Examine the results of these tests, and get the expert's opinions about the conditional probabilities in the form of the matrix. So, user can choose the sample test from the pre-collected dataset. The system converts the text features of the attributes values to the relative weighted values and the construct the conditional probabilities matrix. In this matrix the summation of the row value must be one.

The expected utility of making the decision is on the basis of just the prior probabilities, before any new information is acquired. The system shows calculations for the new prior probabilities, the probabilities for the new information, the expected conditional utilities and the expected alternatives.

The posterior probabilities are calculated with the prior probabilities, and the marginal probabilities. And then calculate the overall unconditional expected utility information, which is actually the sum of pairwise products of the values.

User needs to decide what alternatives to choose, the totally utility favoring of four family types. Hence, the system chooses alternatives the largest values as shown in Figure 2.

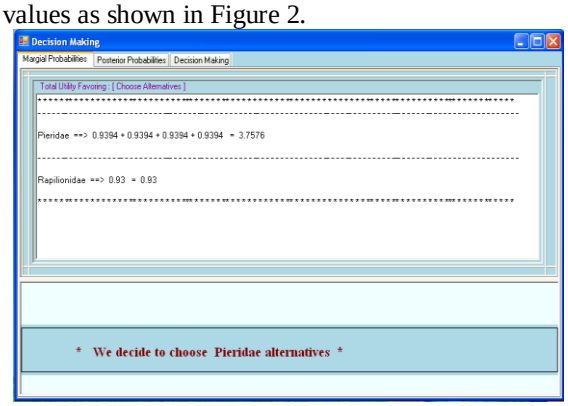

Figure 2. Decision Making Form

Finally, user needs to press the accuracy button. The system gives the accuracy by calculating the hold-out method. This calculated accuracy as shown in Figure 3.

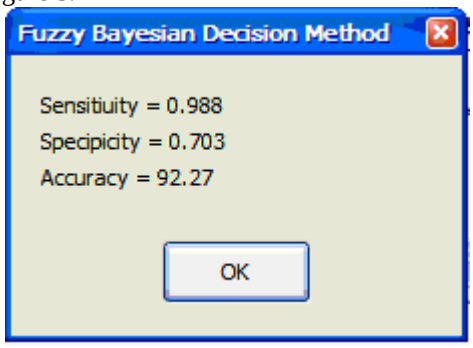

Figure 3. Fuzzy Bayesian Classifier Accuracy Form

6.2. Naive Bayesian Classifiers

The Naïve Bayesian algorithm employed a simplified version of Bayes formula to decide which class a novel instance belongs to. The posterior probability of each class is calculated, given the feature values present in the instances; the instance is assigned the class with the highest probability. Equation (11) shows the Naïve Bayes formula, which makes the assumptions that feature values are statistically independent within each class.

$$p(C_i \setminus x_1, x_2, ..., x_n) = \frac{p(C_i)\prod_{j=1}^{n} p(x_j \setminus C_i)}{p(x_1, x_2, ..., x_n)} \quad (11)$$

The left side of equation (11) is the posterior probability of class C_i given the feature values, $<x_1,x_2,\ldots x_n>$ observed in the instance to be classified. The denominator of the right side of the equation is often omitted because it is a constant which is easily computed if one requires that the posterior probabilities of the classes sum to one. The calculate accuracy of Bayesian Classifier are as in Figure 4:

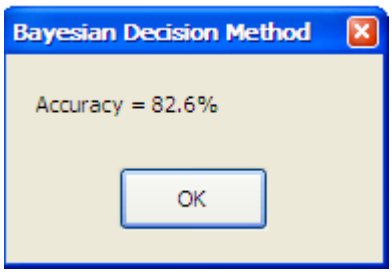

Figure 4. Bayesian Classifier Accuracy Form

7. Conclusions

Classification is a form of data analysis that can be used to extract models describing improtant data classes or to predict future data trends. Fuzzy-Bayesian classifier uses the Bayes formula and Bayes theorem represented knowledge in the form of probabilistic summaries.

This paper predict the class label of butterfly with unknown sample. This paper has presented generating of classification from butterfly datasets. This paper is focused to consider the classification problem of butterfly by using Fuzzy-Bayesian classifier. Throughtout this paper, two machine learning algorithems are used as a basis for comparing the performance of classifiers. The result shows that the Fuzzy Bayesian Classification method gives more efficient accuracy rate than Naïve Bayesian Classification method. And then, it can calculate the other dataset by using Fuzzy-Bayesian classifier.

8. References

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

[2] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001.

[3] Pang-Ning, Tan Michael Steinbach and Vipin Kumar, "Introduction to Data Mining".

[4] Timothy J. Ross, "Fuzzy Logic" , With Engineering Applications, Second Edition .

[5] W.B.Vasantha Kandasamy and S.Devakumar, "Use of Modified Fuzzy Baye's Method in Medical Expert Systems", Indian Institute of Technology (Madras), Chennai, TN, India.

[6] Matias Alvarado, Marisol Vázquez., "Decision Making automation based on fuzzy event-condition-action rules", Departament of Computing, CINVESTAV-IPN, A.P. 14-740, Av.Instituto Politecnico.

[7] Peter V. Golubtsov Stepan S. Moskaliuk, "Bayesian Decisions and Fuzzy Logic", The Erwin Schrodinger International Institute for Mathematical Physics.