Classification of Fish Based on Naïve Bayesian Classifier

Su Su Latt, Myo Myint
Computer University, Myitkyina
susulatt08@gmail.com , mmyoemyint@gmail.com

Abstract

Classification is the form of data analysis that can be used to extract models describing import data class or to predict future data trends. This system uses the Naive Bayesian Classifier. This system is useful in predicting the probability that a sample belongs to a particular class. It makes the assumption of class conditional independence that is, given the class of a sample the values of attributes are conditionally independent of one another. When the assumption holds true, then the naïve Bayesian Classifier is the Classifier is more Bayesian Classifier and more accurate in comparison with all other classifiers. This system is used to training dataset of fish. And then receiving the user's input testing data of fish. By using these testing data, the system will decide what kinds of edible or poisonous of fish, based on a probabilistic model of the observed data and prior knowledge. This system calculates the accuracy of testing data using holdout method.

1. Introduction

Data Mining refers to extraction or mining knowledge from large amounts of data. Data Mining is task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses or other information repositories [3]. In classification learning problem, the learner is given a set of training examples and the corresponding class labels, and outputs a classifier. Naive Bayesian Classifier has been well-known as an effective and efficient classification algorithm. Naive Bayesian Classification is the optimal method of supervised learning if the values of the attributes of fish are independent given the class of data. Although this assumption is almost always violated in practice, recent work has shown that naïve Bayesian learning is remarkably effective in practice and difficult to improve upon systematically. The Bayesian Learning Algorithms combine training data with a prior knowledge to get a posterior probability of an hypothesis. Bayesian Classifiers are statistics that a given sample belongs to a particular of class. This system focus on fish training dataset. Naïve Bayesian if useful the probability that the hypothesis hold given the observed data sample.[3].

This paper is organized as follows: Section 2 discusses related works; Section 3 explains the theoretical background of classification and Bayesian Classification algorithm. The corresponding design method and the implementation of this thesis are described in Section 4. Finally, Section 5 presents conclusion of the current work and future directions.

2. Related works

Bayesian Classifiers have the minimum error rate comparison to all other classifies. Naïve Bayesian Classifier can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Naive Bayesian Classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. In some applications, the accuracy of a classification or prediction is the only thing that matters.

Yoshimasa Tsuroka and Jun,ichi Jsujii using a Naïve Bayes Classifier and the EM algorithm at the University of Tokyo. Nigan et al.(2000) reported that the accuracy of text classification can be improved by a large pool of unlabeled documents using a naïve Bayesian Classifier and the EM algorithm. They presented two extensions to the basic EM algorithm.Prderen et al.(1998) emplayed the EM algorithm and Gibbs Sampling for word sense disambiguation by using a naïve Bayesian classifier[7].Chotirat" Ann" Ratanamahatana and Dimitrios Gunpulos Scaling up the Naïve Bayesian Classifier. Using Dession Trees for feature selection at University of California. Much work has been done on feature subset selection.Juhn, Kohavi, and Pflegar define the problem of feature subset selection to be the of finding a subset of the original set of features of a dataset [1].

3. Background Theory

3.1. Classification

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of keen interest to researchers in psychology and computer science. 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. Using a supervised learning method the computer is given a set of object based on the information acquired by it during the training phase.

Two types of classification are supervised learning and unsupervised learning [3].

Since the class label of each training sample is provided, this step is also known as supervised learning. In contrast to supervise learning was unsupervised leaning. In this learning, the class label of ach training sample is not known, and the number or set of classes to be learned may not be known in advance [2].

3.2. Naïve Bayesian Classification

The Naïve Bayesian Classifier is a straightforward and frequently used method for supervised learning [1].Bayesian Classifiers are useful in predicting the probability that a sample belongs to particular class. Naive Bayesian Classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes[3].

This assumption is called class conditional indepentance.Bayesian Classifying is based on Bayes Theorem. Each class has a probability P (C_k) the represents the prior probability of classifying an attribute into C_k; the values of P (C_k) can be estimated from the training dataset. For n attribute values, Xi, the goal of classification is clearly to find the conditional probability P(C_k \ $V_{1^\wedge}$ $V_{2^\wedge}$…^ v_n). By Bayes' rule, this probability is equivalent to

$$\frac{P(V_1 \wedge V_2 \ldots\ldots \wedge V_n \setminus C_k) P(C_k)}{P(V_1 \wedge V_2 \ldots\ldots \wedge V_n)} \qquad (1)$$

For classification, the denominator is irrelevant, since, for given values of the V_c, it is the same regardless of the value of C_k. The central assumption of Naïve Bayesian Classification is that within each class, the values V_j are all independent of each other[1].

3.3. Estimating Classifier Accuracy

Estimating classifier was important in that it allows one to evaluate how accuracy a given classifier will label future data, that is, data on which the classifier has not been trained and also help in the comparison different classifiers. Classifier accuracy was defined as the percentage of the test examples correctly classified by the algorithm[4]. Holdout and cross-validation are two common techniques for assessing classifier accuracy, based on randomly sampled partitions of the given data.

3.3.1. Holdout Method. The holdout method is a simple technique that uses a test set of class-label samples.

In the holdout method, the given data were randomly partitioned into two independent sets; training set and test. Typically two third of the data are allocated to the training set. The training set was used to derive the classifier, whose accuracy was estimated with the test set. Formally, let D_h , the holdout set, be a subset of D of size h, and let D_t be D\Dh.The holdout estimated accuracy is defined as

$$\text{acc } h = \frac{1}{h} \sum_{(\acute{y}\acute{y}) \in D_h} \delta(I(D_1 \upsilon_1) y_1) \qquad (2)$$

where $\delta(i, j)$ =1 if i=j and 0 otherwise[4].

3.4. Different kinds of Fish

Aquatic Biomes (or) Fish can be divided into two types: (a) inland or fresh water fish and (b) ocean or salt water. Many fish have edible and some fish has poison [9]. The poisonous of fish can be divided into three groups. There are

1. Fish that contain venomous spines in the spines on the tail or on the operculum.
2. Fish that carry poisonous bite and
3. Fish having poisonous flesh or liver.

Any fishes have 11 attributes: Teeth, Operculum, Barbel, Dorsal Fin, Pelvic Fin, Pectoral Fin, Anus Fin, Caudal Fin, Scale, Lateral Line, and Tail [5].

3.5. Algorithm of Naive Bayesian Classification

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, Edible or Poison.

Method:

```
1. K=total record count of training sample dataset C:
2. for (i=0; i<C_{k-1} , i++)
3. {
4. if (C.record(i).cell(Result).value = = e)
      edible Count ++;
      else
            if(C.record(i).Cell (Result).value = = p)
            poison Count ++;
            }
            totalEdibleProb= edible Count/k;
            totalPoisonProb= poison Count/k;
             m= total record count of testing sample data
            except ID and Result fields;
            n= total cells count in each record of testing
            sample dataset T expect ID and Result fields;
5.  for (i=0; i< T_{m-1}: i++)
6  {
7.  eProb = i;
8.  pProb = i ;
9.  for(j=0; j< T_{m-1}; j++)
10. {
11. eCount =getCount(j,T.record(i).Cell(j) .value,e);
```

```
12. eProb* = eCount/ edible Count;
13. PCount= getCount(j,T.record(i).Cell(j).value,p);
14. pProb*= pCount/poisonCount;
15. }
16. if(ePro> totaledibleProb)
17. display result as edible e;
18. else if (pProb> totalPoisonProb)
19. display result as poison p;
20. }
    Procedure getCount (colIndex, colVal,result)
        1. k- total record count of training sample dataset C;
        2. count =0;
        3. for(i=0; i< Ck-1;i++)
        4. {
        5.  if(C.record(i).cell(colIndex).value==colVal && C.record(i).cell(i).cell (Res ult).value)== result)
        6. count++;
        7. }
        8. return count;
```

Figure 1.Naive Bayesian Algorithm

4. System Design and Implementation

4.1. System Design

In Figure 2, the selected training samples and Unknown data (or) testing dataset are randomly partitioned into independent set. Training data are analyzed by Bayesian Algorithm. Training data also predict result of new data using Bayesian Algorithm. Test data are used to estimate the classifier accuracy.

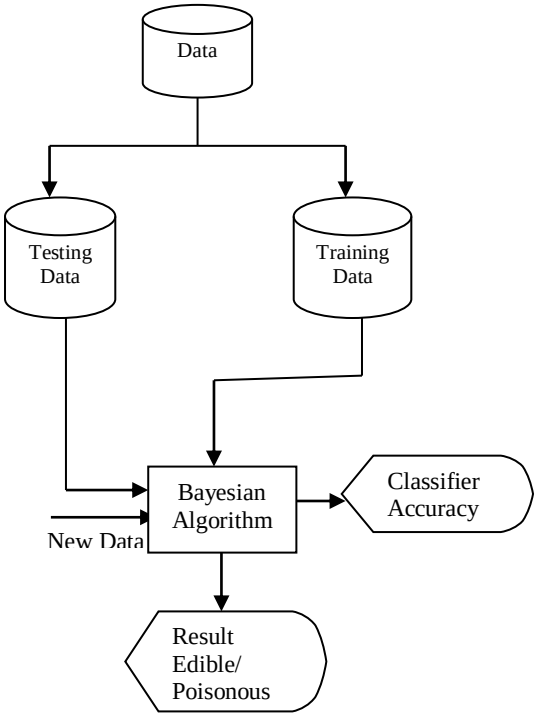

Figure. 2 The System Design for our system

This system have 4 Steps:

In step 1, when fish researcher entries training samples, process 1 trains the system of each attribute's new, insert and delete samples data. In Data Selection, the user can select training data and testing data for calculating Bayesian Classification. If the user want to delete a row by clicking check box in it.

In this step 2 trains the system of training data by calculating each attribute's prior probability and conditional probability of each class label.

At step 3, estimate accuracy of testing data. The predictive 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.

In step 4, calculate the result of the user choice unknown sample data. This process compares facts and gets pre-calculated probilities for each attribute of testing data. Calculates total and maximum predictable probabilities or conditional probabilities for each predictable result and then outputs maximum possible predicted results by comparing such probabilities.

4.2. Attributes Information and Sample Dataset

In this training data set, there are 11 attributes. There are 2 classes, edible and poisonous. The following table describes name of attributes and description of these attributes.

Table 1. Attributes of Information

No.	Attributes	Description
1.	Teeth	normal, absent, sting
2.	Operculum	soft, spinous
3.	Barbel	present, absent
4.	Dorsal Fin	spinous, soft, null
5.	Pelvic Fin	spinous, soft, null
6.	Pectoral Fin	spinous, soft, null
7.	Anus Fin	spinous, soft, null
8.	Caudal Fin	spinous, soft, null
9.	Scale	small, large, absent
10.	Lateral Line	absent, complete, interrupt
11.	Tail	sting, null, incomplete

Table 2. Sample data for Training Dataset

ID	Teeth	Oper - culum	Barbel	Dorsal Fin	Pelvic Fin	Pectoral Fin	Anus Fin	Caudal Fin	Scale	Lateral Line	Tail	Class label
1	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
2	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
3	sting	spinous	absent	null	null	null	null	null	absent	complete	null	poison-ous
4	normal	soft	present	soft	null	spinous	null	soft	absent	complete	null	Edible
5	Absent	soft	present	spinous	soft	soft	soft	soft	large	complete	incom plete	poison-ous
6	absentl	soft	present	soft	null	soft	soft	soft	small	complete	null	edible
7	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
8	absent	soft	present	spinous	soft	soft	soft	soft	large	complete	incom plete	poison-ous
9	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
10	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
11	normal	soft	present	soft	null	spinous	null	soft	absent	complete	null	edible
12	Absent	soft	abesntt	null	null	soft	soft	spinous	Absent	complete	sting	poison-ous
13	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
14	absentl	soft	present	soft	null	soft	soft	soft	small	complete	null	edible
15	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
16	normal	soft	present	soft	null	spinous	null	soft	absent	complete	null	edible
17	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
18	normal	soft	present	soft	null	soft	soft	soft	absent	complete	null	edible
19	Absent	soft	present	spinous	soft	soft	soft	soft	large	complete	incom plete	poison-ous
20	absentl	soft	present	soft	null	soft	soft	soft	small	complete	null	edible

4.2.1. Predicting a class label using naïve Bayesian Classification. The user wish to classify is unknown sample.Condisier a similar measurement after testing data has been portioned in accordance with the n outcomes of a test on the feature X.

$X = (X_1, X_2, \ldots\ldots X_n)$

In this sample X= (Teeth = "absent ", Operculum = "soft", Barbel = "present", Dorsal Fin ="soft", Pelvic ="null", Pectoral Fin = "null", Anus ="null", Cal dual Fin = "soft", Scale = "large", Lateral Line ="large", Tail = "null").

Consider the class labels of training dataset are two classes (edible or poisonous).

$P(C_i \setminus X) > P(C_j \setminus X)$ for $1 \le j \le m, j \ne i$.

C_1 = (Class-label =" Edible ")

C_2 = (Class-label =" Poisonous ")

The user will calculate the Prior probability of training data set.

$$P(C_i \setminus X) = \frac{(X \setminus C_i)P(C_i)}{P(X)} \qquad (3)$$

Sample data values

(for C_1) testing data of edible =77, training data of edible or poisinous=90

(for C_2) testing data of poisonous =30, training data of edible or poisinous=90

$$P(C_1) = \frac{77}{90} = 0.86, \quad P(C_2) = \frac{30}{90} = 0.14$$

The user compares facts and gets pre-calculated Conditional Probability for each attribute of testing data based on training data. Calculates total product of conditional probability each data and maximum predictable probabilities or conditional probabilities for each predictable result.

$P(X \setminus C_i) = \prod P(X_k \setminus C_i)$ (4)

$P(X \setminus C_1) = 0.0179$

$P(X \setminus C_2) = 8.18532$

And then the user outputs maximum possible predicted results by comparing such probilities.

$P(X \setminus C_i)P(C_i) > P(C_j \setminus X)C_j$ for $1 \le j \le m, j \ne i$. (5)

$P(C_1 \setminus X) = P(X \setminus C_1)\, P(C_1) = 0.86 \times 0.00179$

$= 0.0015451203588092$

$P(C_2 \setminus X) = P(X \setminus C_2)\, P(C_2) = 0.14 \times 8.18532$

$= 1.1459461813983156$

Therefore the naive Bayesian Classifier predicts class label = "edible" for sample X.

4.3. Implementation of Naïve Bayesian Classifier

When the user clicks the Training data tab, it will display training data information from. If the user wants to insert new record, the user can choose each of attributes from combo box and the user want to delete the original training data select the row click delete sample menu .When the user clicks the Data Selection tab, it will display training data and testing data for calculation Bayesian Classification as shown in figure 3.

When the user clicks the Accuracy Estimation tab, it will display accuracy result as shown in figure 4. Fish researcher trains the system with the training samples dataset and then testing data accuracy by using the testing samples dataset. The user can choose Estimation Accuracy Sample from File Menu.

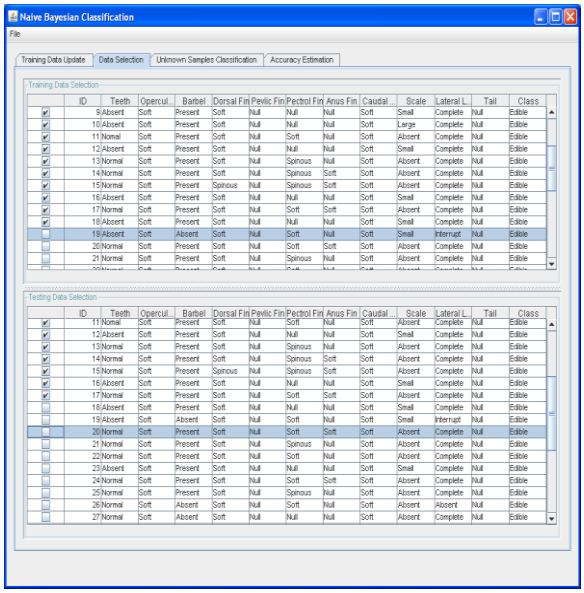

Figure 3. Data Selection Form

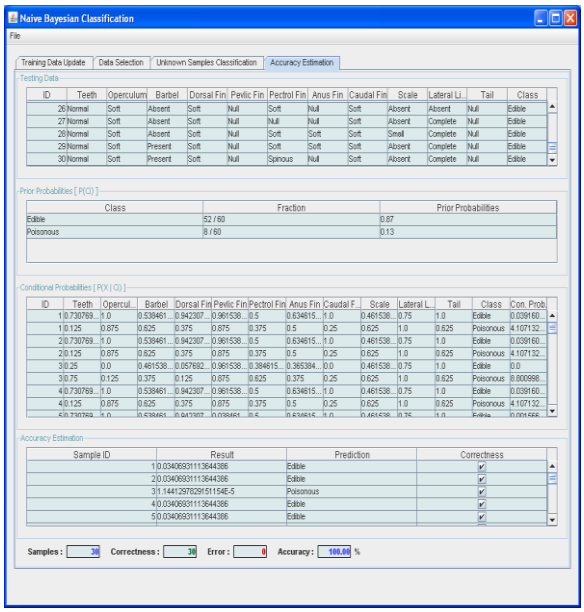

Figure 4 .Estimation Accuracy Form

When the user clicks the Unknown Sample Classification tab, the form appears as shown in figure 5.In this form, the user can see the result of calculation of probabilities by Naive Bayesian Classification method and predicting result. The research who are interested in fish to study that data if maximum probabilities of edible is gather than maximum probabilities of poisons, then Naive Bayesian Classification method put those fish as resulting edible .

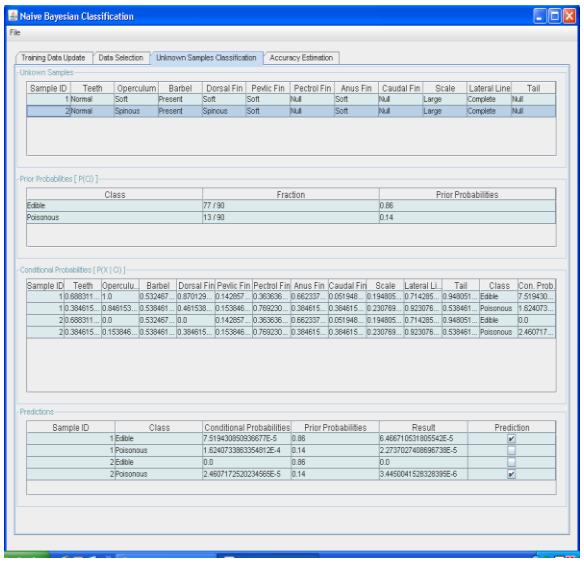

Figure 5. Unknown Sample Classification Form

4.4. Estimate Accuracy of Chat

Estimating the accuracy of a classifier induced by supervised learning algorithms is important not only to predict its future prediction accuracy but also for choosing a classifier from a given set. The holdout estimate is a random number that depends on the division into two third of training set and one third of test set. This system calculates the estimation accuracy of fish testing data as shown in figure 6.

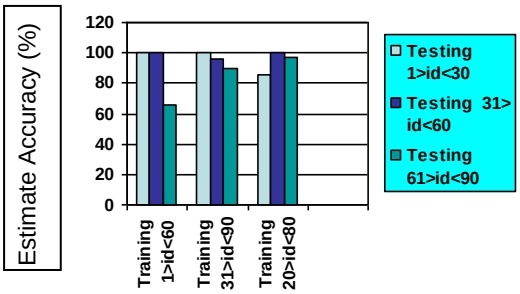

Figure 6. Accuracy Estimation Result

5. Conclusion

Classification is an important technique in data mining. Classification can be used for prediction the class of data objects. The Bayesian classifier uses the naïve Bayesian formula to calculate the probability of each class given the values of all attributes. The Bayesian Classification is based on Byes theorem of posterior probability. This paper will be predicted the class label of fish unknown sample given the training data. This system receiving the user's input testing data of fish. The relative performance of the naïve Bayesian classifier can serve as an estimate of the conditional independence of attributes. The paper has presented generating of classification from large datasets. This approach demonstrates efficiency and effectiveness in dealing with fish data for classification. This paper considered the classification problem of fish by using Naïve Bayesian classifier .The estimate the accuracy of testing data by using holdout method. This system calculation of accuracy result value is high. Many results of fish testing data are hundred percentages. So, this system will be predicted the class label of unknown sample is correct.

References

[1] G.Dimitrois" *Scaling up the Naive Bayesian Classifier:Using Decision Trees for Feature Selection"*,University of California ,Riverside, CA 92521,1-909-787-50-190.

[2] Hamilton*"Correlation used Feature Selection for Machine Learning"* (Ph.D- Thesis) April,1999.

[3] J.Ha and M.Kamber.*"Data Mining Concepts and Techniques"* Morgan Kaufman, 2001.

[4] Kohavi Ron, *"A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection"*.

[5] Kyaw San Win.M.Research*"Classification for Freshwater Fish"*.

[6].Zaiane Osmar, 1999*"Principles of Kno-wledge Discovery in Databases"*

[7] T.Yoshimasa and T.Jun'ichi *"Training a Naive Bayesian Classification via the EM algorithm with a class Distribution Constraint"* Tokyo 113-0033 JAPAN.

[8] U.M.Fayyad, G.Piatetsky-Shaoiro, P.Smyth, and Uthurusamy*"Advances in Knowledge Discovery and Data Mining."* AAAI/MIT Press, 1996.

[9] W.Dr.Dusko *"The poisonous fish of the Croation Adriatic"*ISBN 953-96204-22, DRAGA, 1994.

[10] http://www.math.upatras.gr/~esdlab/en/mem -bers/kotsiantis/5_Kotsiantisogitboost%20of%20simble%20bayesian...No%205.pdf) .