Statistical Models in Data Mining: A Bayesian Classification

Ravindra Changala, Annapurna Gummadi, T Janardhan Rao
Guru Nanak Institutions Technical Campus, Hyderabad.

Abstract—The concept of conditional probability is introduced in Elementary Statistics. The conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. In this section we extend the discussion of conditional probability to include applications of Bayes' theorem, which we use for revising a probability value based on additional information that is later obtained. One key to understanding the essence of Bayes’ theorem is to recognize that we are dealing with sequential events, whereby new additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event. The paper presents how bayes theorem used in data mining classification and prediction of tuple of class labels. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes’ theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. In this context, the terms prior probability and posterior probability are commonly used.

Keywords—Bayes’ theorem, Data mining, Classification, Prediction, Naïve Bayesian classifiers.

I. INTRODUCTION

In the words of anonymous saying there are two problems in modern science: too many people using different terminology to solve the same problems and even more people using the same terminology to address completely different issues. This is particularly relevant to the relationship between traditional statistics and the new emerging field of knowledge data discovery (KDD) and data mining (DM). Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and drawing conclusions from data.

II. STATISTICAL MODELS

The aim of this chapter is to present the main statistical issues in Data mining (DM) and Knowledge Data Discovery (KDD) and to examine whether traditional statistics approach and methods substantially differ from the new trend of KDD and DM. We address and emphasize some central issues of statistics which are highly relevant to DM and have much to offer to DM.

Traditional statistics emphasizes the mathematical formulation and validation of a methodology, and views simulations and empirical or practical evidence as a less form of validation. The emphasis on rigor has required proof that a proposed method will work prior to its use. In contrast, computer science and machine learning use experimental validation methods. In many cases mathematical analysis of the performance of a statistical algorithm is not feasible in a specific setting, but becomes so when analyzed asymptotically. At the same time, when size becomes extremely large, studying performance by simulations is also not feasible. It is therefore in settings typical of DM problems that asymptotic analysis becomes both feasible and appropriate.

III. DATA MINING

Databases today can range in size into the terabytes. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest? The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous. The first and simplest analytical step in data mining is to describe the data summarize its statistical attributes, visually review it using charts and graphs, and look for potentially meaningful links among variables.

As emphasized in the section on the data mining process, collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. You must build a predictive model based on patterns determined from known results, then test that model on results outside the original sample. A good model should never be confused with reality (you know a road map isn’t a perfect representation of the actual road), but it can be a useful guide to understanding your business. The final step is to empirically verify the model. For
example, from a database of customers who have already responded to a particular offer, you’ve built a model predicting which prospects are likeliest to respond to the same offer. Can you rely on this prediction? Send a mailing to a portion of the new list and see what results you get identifying exceptions, or finding interactions.

Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions. The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional.

IV. CLASSIFICATION

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. For example, you may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation, vulnerable to switching over to a competing long-distance phone service, or a good candidate for a surgical procedure. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long-distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database.

V. BAYES’ THEOREM

The concept of conditional probability is introduced in Elementary Statistics. We noted that the conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. We used $P(B|A)$ to denote the conditional probability of event B occurring, given that event A has already occurred. The following formula was provided for finding $P(B|A)$:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

In addition to the above formal rule, the textbook also included this “intuitive approach for finding a conditional probability”:

The conditional probability of B given A can be found by assuming that event A has occurred and, working under that assumption, calculating the probability that event B will occur. In this section we extend the discussion of conditional probability to include applications of Bayes’ theorem (or Bayes’ rule), which we use for revising a probability value based on additional information that is later obtained. One key to understanding the essence of Bayes’ theorem is to recognize that we are dealing with sequential events, whereby new additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event. In this context, the terms prior probability and posterior probability are commonly used.

A prior probability is an initial probability value originally obtained before any additional information is obtained.

A posterior probability is a probability value that has been revised by using additional information that is later obtained.

VI. BAYESIAN CLASSIFICATION

“What are Bayesian classifiers?” Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes’ theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this
Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers, allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification.

VII. BAYES THEOREM IN DATA MINING

Bayes’ theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. Let $X$ be a data tuple. In Bayesian terms, $X$ is considered “evidence.” As usual, it is described by measurements made on a set of $n$ attributes. Let $H$ be some hypothesis, such as that the data tuple $X$ belongs to a specified class $C$. For classification problems, we want to determine $P(H|X)$, the probability that the hypothesis $H$ holds given the “evidence” or observed data tuple $X$. In other words, we are looking for the probability that tuple $X$ belongs to class $C$, given that we know the attribute description of $X$. $P(H|X)$ is the posterior probability, or a posteriori probability, of $H$ conditioned on $X$.

For example, suppose our world of data tuples is confined to customers described by the attributes age and income, respectively, and that $X$ is a 35-year-old customer with an income of $40,000. Suppose that $H$ is the hypothesis that our customer will buy a computer. Then $P(H|X)$ reflects the probability that customer $X$ will buy a computer given that we know the customer’s age and income. In contrast, $P(H)$ is the prior probability, or a priori probability, of $H$ conditioned on $X$.

For our example, this is the probability that any given customer will buy a computer, regardless of age, income, or any other information, for that matter. The posterior probability, $P(H|X)$, is based on more information (e.g., customer information) than the prior probability, $P(H)$, which is independent of $X$. Similarly, $P(X|H)$ is the posterior probability of $X$ conditioned on $H$. That is, it is the probability that a customer, $X$, is 35 years old and earns $40,000, given that we know the customer will buy a computer. $P(X)$ is the prior probability of $X$. Using our example, it is the probability that a person from our set of customers is 35 years old and earns $40,000. “How are these probabilities estimated?” $P(H)$, $P(X|H)$, and $P(X)$ may be estimated from the given data, as we shall see below. Bayes’ theorem is useful in that it provides a way of calculating the posterior probability, $P(H|X)$, from $P(H)$, $P(X|H)$, and $P(X)$. Bayes’ theorem is

VIII. PREDICTING A CLASS LABEL USING BAYESIAN CLASSIFICATION

We wish to predict the class label of a tuple using naïve Bayesian classification, given the same training data. The data tuples are described by the attributes age, income, student, and credit rating. The class label attribute, buys computer, has two distinct values (namely, yes and no). Let $C_1$ correspond to the class buys computer = yes and $C_2$ correspond to buys computer = no. The tuple we wish to classify is $X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit rating} = \text{fair})$.

We need to maximize $P(X|C_i)P(C_i)$, for $i = 1, 2$. $P(C_i)$, the prior probability of each class, can be computed based on the training tuples:

- $P(\text{buys computer} = \text{yes}) = \frac{9}{14} = 0.643$
- $P(\text{buys computer} = \text{no}) = \frac{5}{14} = 0.357$

To compute $P(X|C_i)$, for $i = 1, 2$, we compute the following conditional probabilities:

- $P(\text{age} = \text{youth} | \text{buys computer} = \text{yes}) = \frac{2}{9} = 0.222$
- $P(\text{age} = \text{youth} | \text{buys computer} = \text{no}) = \frac{3}{5} = 0.600$
- $P(\text{income} = \text{medium} | \text{buys computer} = \text{yes}) = \frac{4}{9} = 0.444$
- $P(\text{income} = \text{medium} | \text{buys computer} = \text{no}) = \frac{2}{5} = 0.400$
- $P(\text{student} = \text{yes} | \text{buys computer} = \text{yes}) = \frac{6}{9} = 0.667$
- $P(\text{student} = \text{yes} | \text{buys computer} = \text{no}) = \frac{1}{5} = 0.200$
- $P(\text{credit rating} = \text{fair} | \text{buys computer} = \text{yes}) = \frac{6}{9} = 0.667$
- $P(\text{credit rating} = \text{fair} | \text{buys computer} = \text{no}) = \frac{2}{5} = 0.400$

Using the above probabilities, we obtain

- $P(X|\text{buys computer} = \text{yes}) = P(\text{age} = \text{youth} | \text{buys computer} = \text{yes}) X P(\text{income} = \text{medium} | \text{buys computer} = \text{yes}) X P(\text{student} = \text{yes} | \text{buys computer} = \text{yes}) X P(\text{credit rating} = \text{fair} | \text{buys computer} = \text{yes}) = 0.044 X 0.444 X 0.667 X 0.667 = 0.044$. Similarly,

- $P(X|\text{buys computer} = \text{no}) = 0.222 X 0.444 X 0.667 X 0.400 = 0.019$. To find the class, $C_i$, that maximizes $P(X|C_i)P(C_i)$, we compute

- $P(X|\text{buys computer} = \text{yes})P(\text{buys computer} = \text{yes}) = 0.044 X 0.643 = 0.028$
- $P(X|\text{buys computer} = \text{no})P(\text{buys computer} = \text{no}) = 0.019 X 0.357 = 0.007$

Therefore, the naïve Bayesian classifier predicts buys computer = yes for tuple $X$. 
“What if I encounter probability values of zero?”
Recall that in Equation, we estimate \( P(X|C_i) \) as the product of the probabilities \( P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i) \), based on the assumption of class conditional independence. These probabilities can be estimated from the training tuples (step 4). We need to compute \( P(X|C_i) \) for each class (\( i = 1, 2, \ldots, m \)) in order to find the class \( C_i \) for which \( P(X|C_i)P(C_i) \) is the maximum (step 5). Let’s consider this calculation.

For each attribute-value pair (i.e., \( A_k = x_k \), for \( k = 1, 2, \ldots, n \)) in tuple \( X \), we need to count the number of tuples having that attribute-value pair, per class (i.e., per \( C_i \), for \( i = 1, \ldots, m \)). In Example, we have two classes (\( m = 2 \)), namely buys computer = yes and buys computer = no.

Therefore, for the attribute-value pair student = yes of \( X \), say, we need two counts the number of customers who are students and for which buys computer = yes (which contributes to \( P(X|\text{buys computer} = \text{yes}) \)) and the number of customers who are students and for which buys computer = no (which contributes to \( P(X|\text{buys computer} = \text{no}) \)). But what if, say, there are no training tuples representing students for the class buys computer = no, resulting in \( P(\text{student} = \text{yes}|\text{buys computer} = \text{no}) = 0 \)? In other words, what happens if we should end up with a probability value of zero for some \( P(x_k|C_i) \)? Plugging this zero value into Equation would return a zero probability for \( P(X|C_i) \), even though, without the zero probability, we may have ended up with a high probability, suggesting that \( X \) belonged to class \( C_i \)!

A zero probability cancels the effects of all of the other (posteriori) probabilities (on \( C_i \)) involved in the product. There is a simple trick to avoid this problem. We can assume that our training database, \( D \), is so large that adding one to each count that we need would only make a negligible difference in the estimated probability value, yet would conveniently avoid the case of probability values of zero. This technique for probability estimation is known as the Laplacian correction or Laplace estimator, named after Pierre Laplace, a French mathematician who lived from 1749 to 1827. If we have, say, \( q \) counts to which we each add one, then we must remember to add \( q \) to the corresponding denominator used in the probability calculation. We illustrate this technique in the following example.

Using the Laplacian correction to avoid computing probability values of zero. Suppose that in some training database, \( D \), containing 1,000 tuples, we have 0 tuples with income = low, 990 tuples with income = medium, and 10 tuples with income = high. The probabilities of these events, without the Laplacian correction, are 0, 0.990 (from 999/1000), and 0.010 (from 10/1000), respectively. Using the Laplacian correction for the three quantities, we pretend that we have 1 more tuple for each income-value pair. In this way, we instead obtain the following probabilities (rounded up to three decimal places):

\[
\frac{1}{1,001} = 0.001, \quad \frac{991}{1,001} = 0.988, \quad \text{and} \quad \frac{11}{1,001} = 0.011.
\]

The “corrected” probability estimates are close to their “uncorrected” counterparts, yet the zero probability value is avoided.

IX. CONCLUSION
This paper presents the role of statistical models in the data mining to classify and predict class tuples in an efficient manner. The Bayes theorem performs well as per the above example and respective classifiers all data mining tasks in Weka software with accurate results. Byes theorem effectively drafted uncertain data and in predicting next event to be happened. We integrate the uncertain data model with Bayes theorem and propose new techniques to calculate conditional probabilities. Besides laying the theoretical foundations for enhancing naive Bayesian classification to process uncertain data sets, we show how to put these concepts into practice. Our experimental evaluation demonstrates that the classifiers for uncertain data can be efficiently constructed and effectively classify and predict even highly uncertain data. Further, the proposed classification is more stable and more suitable for mining uncertain data than the previous work.

REFERENCES
[1] Data Mining: Concepts and Techniques Second Edition, Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber.Elsevier Publications.