Short Note:
Naive Bayes Classifiers and Permanence of Ratios

Julián M. Ortiz (jmo1@ualberta.ca)
Department of Civil & Environmental Engineering
University of Alberta

Abstract

The assumption of permanence of ratios allows integration of data from different sources, and with different types, and precision. It originates from the assumption of conditional independence between the different sources of information, given the event of interest.

This concept is known as the Naive Bayes Classifier in statistics. Under the same assumption of conditional independence, it aims at classifying multivariate events given the conditioning information. The conditional probabilities are inferred from training data, which is similar to using a training image for inferring multiple-point statistics used to condition geostatistical simulations.

We recall the theory behind these two concepts, and comment on some future avenues of research.

Introduction

Journel introduced the concept of permanence of ratios as an alternative to the hypothesis of data independence in geostatistical applications, when integrating information from multiple sources [5].

Consider the estimation of the probability of occurrence of an unknown event $\mathbf{A}$, given information from several sources. For now, let us consider only two sources of information, $\mathbf{B}$ and $\mathbf{C}$. Bayes' law permits the calculation of the conditional probability $P(\mathbf{A}|\mathbf{B},\mathbf{C})$:

$$P(\mathbf{A}|\mathbf{B},\mathbf{C}) = \frac{P(\mathbf{A},\mathbf{B},\mathbf{C})}{P(\mathbf{B},\mathbf{C})} \tag{1}$$

with

$$\begin{aligned} P(\mathbf{A},\mathbf{B},\mathbf{C}) &= P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A},\mathbf{B}) \\ &= P(\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A},\mathbf{C}) \end{aligned}$$

However, the calculation of this probability requires the knowledge of the joint probability between $\mathbf{B}$ and $\mathbf{C}$, and the conditional probability of $\mathbf{C}$ given $\mathbf{A}$ and $\mathbf{B}$ (or equivalently, this last requirement can be replaced by the knowledge of the conditional probability of $\mathbf{B}$ given $\mathbf{A}$ and $\mathbf{C}$).

Data Independence

The assumption of data independence states that $\mathbf{B}$ and $\mathbf{C}$ are independent, therefore:

$$P(\mathbf{B}, \mathbf{C}) = P(\mathbf{B}) \cdot P(\mathbf{C})$$

Contrary to what is stated in the paper by Journel [5], this condition by itself does not resolve the problem. An additional assumption is required to simplify Expression 1: conditional independence of $\mathbf{B}$ and $\mathbf{C}$ given the event $\mathbf{A}$ [2], that is:

$$\begin{aligned} P(\mathbf{C}|\mathbf{A}, \mathbf{B}) &= P(\mathbf{C}|\mathbf{A}) \\ P(\mathbf{B}|\mathbf{A}, \mathbf{C}) &= P(\mathbf{B}|\mathbf{A}) \end{aligned}$$

These two assumptions entail:

$$P(\mathbf{A}|\mathbf{B}, \mathbf{C}) = \frac{P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A})}{P(\mathbf{B}) \cdot P(\mathbf{C})} \tag{2}$$

Recall that:

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{B})}{P(\mathbf{A})} = \frac{P(\mathbf{A}|\mathbf{B}) \cdot P(\mathbf{B})}{P(\mathbf{A})} \tag{3}$$

Hence, Expression 2 can also be written:

$$P(\mathbf{A}|\mathbf{B}, \mathbf{C}) = \frac{P(\mathbf{A}|\mathbf{B})}{P(\mathbf{A})} \cdot \frac{P(\mathbf{A}|\mathbf{C})}{P(\mathbf{A})} \cdot P(\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B}) \cdot P(\mathbf{A}|\mathbf{C})}{P(\mathbf{A})}$$

Although correct, this expression is not robust against departures from the assumption of independence. Consider the case when $P(\mathbf{A}) = 0.1$, $P(\mathbf{A}|\mathbf{B}) = 0.8$, and $P(\mathbf{A}|\mathbf{C}) = 0.8$. The conditional probability $P(\mathbf{A}|\mathbf{B}, \mathbf{C})$ is 6.4. Since a probability is being estimated, a good property of the estimators would be to generate results that always fall in the interval $[0, 1]$. The previous assumptions do not provide an estimate (by Bayes' law) that satisfies with this property. This motivates the search for alternative assumptions.

Conditional Independence

A less constraining and, as we will see, more robust approach is to assume the data are conditionally independent given the event $\mathbf{A}$. There is no need to assume they are independent. The expression for the conditional probability of $\mathbf{A}$ given $\mathbf{B}$ and $\mathbf{C}$ is:

$$P(\mathbf{A}|\mathbf{B}, \mathbf{C}) = \frac{P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A})}{P(\mathbf{B}, \mathbf{C})} \tag{4}$$

The joint probability between $\mathbf{B}$ and $\mathbf{C}$ is required in order to compute the conditional probability; however, as Journel pointed out, there is a simple way around this problem that gives the expression for the permanence of ratios assumption [5]. We can consider the expression for $P(\mathbf{A}|\mathbf{B}, \mathbf{C})$ and the expression for $P(\bar{\mathbf{A}}|\mathbf{B}, \mathbf{C})$, where the event $\bar{\mathbf{A}}$ represents

the complement of event $\mathbf{A}$, that is, the event of $\mathbf{A}$ *not* occurring, which is equal to $1 - P(\mathbf{A}|\mathbf{B},\mathbf{C})$. The two expressions and the ratio between them are:

$$P(\mathbf{A}|\mathbf{B},\mathbf{C}) = \frac{P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A})}{P(\mathbf{B},\mathbf{C})} \quad (5)$$

$$P(\bar{\mathbf{A}}|\mathbf{B},\mathbf{C}) = \frac{P(\bar{\mathbf{A}}) \cdot P(\mathbf{B}|\bar{\mathbf{A}}) \cdot P(\mathbf{C}|\bar{\mathbf{A}})}{P(\mathbf{B},\mathbf{C})} \quad (6)$$

$$\frac{P(\bar{\mathbf{A}}|\mathbf{B},\mathbf{C})}{P(\mathbf{A}|\mathbf{B},\mathbf{C})} = \frac{P(\bar{\mathbf{A}}) \cdot P(\mathbf{B}|\bar{\mathbf{A}}) \cdot P(\mathbf{C}|\bar{\mathbf{A}})}{P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A})} \quad (7)$$

Using Expression 3, the terms in Expression 7 can be rearranged to end up with the permanence of ratios expression:

$$\frac{\frac{P(\bar{\mathbf{A}}|\mathbf{B},\mathbf{C})}{P(\mathbf{A}|\mathbf{B},\mathbf{C})}}{\frac{P(\bar{\mathbf{A}}|\mathbf{B})}{P(\mathbf{A}|\mathbf{B})}} = \frac{\frac{P(\bar{\mathbf{A}}|\mathbf{C})}{P(\mathbf{A}|\mathbf{C})}}{\frac{P(\bar{\mathbf{A}})}{P(\mathbf{A})}} \quad (8)$$

This expression states that the incremental information provided by one event $\mathbf{C}$ before and after knowing another event $\mathbf{B}$ is constant (and vice-versa).

Naive Bayes Classifier

Statistical classification helps in the decision to assign a given case into a class, given a set of attributes. The problem is similar to the one of integrating several sources of information (the "attributes") to decide the probability of occurrence of an unknown event (the "case"). In geostatistical applications, the event $\mathbf{A}$ is in general an indicator that represents the presence or absence of an attribute at a given location. For continuous variables, it can be defined as the probability of being below a given threshold. By defining the conditional probability of the event $\mathbf{A}$ occurring given the information provided by the attributes $\mathbf{B}$ and $\mathbf{C}$, in the case with two additional sources of information, the problem of combining knowledge from diverse sources is reduced to a statistical classification problem.

The Naive Bayes classifier was introduced more than forty years ago [1, 6]. It assumes the attributes are conditionally independent, given the case being classified. The assumption is typically displayed as a Bayesian network, as depicted in **Figure 1**. This assumption, as seen before, entails the permanence of ratios. It is then reasonable to explore what forty years of research in statistical classification can offer to further the application of the permanence of ratios assumption in geostatistics.

Comments

The objective of many of the statistical and geostatistical methods is to predict an unknown or unsampled attribute [4]. This prediction can consider a continuous attribute or a categorical one. Unlike regression methods, classification techniques aim to locate the unknowns into one of several (disjoint) classes. Both methods require some rule to assign a value or a class to the unknown event, given the information available. The rule can be based on almost anything. In order to obtain the probability of an unknown event, Bayesian

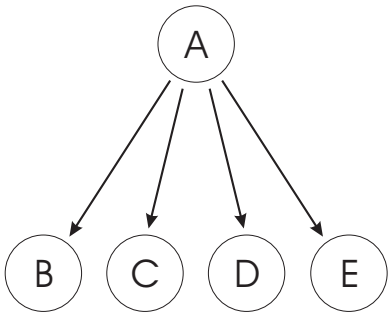

Figure 1: Bayesian network representing the Naive Bayes classifier with attributes **B**, **C**, **D**, and **E**. The conditional independence assumption is shown as no connectors between the attributes.

statistics offers a framework to determine these rules. Classification is then simply made by assigning the event to the class with highest probability.

Notwithstanding their simplicity, Naive Bayes classifiers are extremely successful in terms of accuracy, that is, when the number of incorrect versus correct assignments to classes are computed [3, 4]. This success is mainly due to the fact that they do not need to be precise in estimating the probability, but only have to discriminate correctly between the classes. Thus, the class with highest probability is selected and the case is assigned to it. Hence, there is no requirement for the estimation of the probability to be accurate, but the ranking of the probabilities has to be correct to have good classification.

One of the main advantages of the assumption of conditional independence is the great simplification that it entails in the expression for the conditional probability of **A** given **B** and **C**. However, in a spatial context, this assumption appears as unrealistic. This problem also occurs in statistical applications, where it is common to find that many of the attributes are strongly correlated.

Research in statistics has shown that the Naive Bayes assumption performs well because, despite generating a considerable bias in the estimation of the conditional probability, it generates a low inter-class error, which results in good classification performance. This has been shown in the binary case, although extensions to cases with multiple categories are likely similar [4].

Another interesting application found in the statistical literature is the use of incomplete training data sets to infer the conditional probabilities used in the Naive Bayes model. Depending on the reason for these missing data events, it is possible to infer from the available data intervals for the missing conditional probabilities. These intervals can then be used to improve classification.

A simple extension to the Naive Bayes model consists on considering an additional parameter in the expression of the permanence of ratios. This parameter τ should account for the correlation between the sources of information. This implies a departure from the assumption that the incremental information of **C** remains constant regardless of the additional information provided by **B** to the estimation of the probability of **A** occurring:

$$\frac{\frac{P(\bar{\mathbf{A}}|\mathbf{B},\mathbf{C})}{P(\mathbf{A}|\mathbf{B},\mathbf{C})}}{\frac{P(\bar{\mathbf{A}}|\mathbf{B})}{P(\mathbf{A}|\mathbf{B})}} = \left(\frac{\frac{P(\bar{\mathbf{A}}|\mathbf{C})}{P(\mathbf{A}|\mathbf{C})}}{\frac{P(\bar{\mathbf{A}})}{P(\mathbf{A})}} \right)^{\tau} \tag{9}$$

This can be written equivalently as:

$$P(\mathbf{A}|\mathbf{B},\mathbf{C}) = \frac{P(\mathbf{A}) \cdot P(\mathbf{B}|\mathbf{A}) \cdot P(\mathbf{C}|\mathbf{A})^{\tau}}{P(\mathbf{B},\mathbf{C})}$$

Expression 9 accounts for the redundancy between **B** and **C** regarding event **A**. We are looking for the value of τ, such that:

$$P(\mathbf{C}|\mathbf{A})^{\tau} \approx P(\mathbf{C}|\mathbf{A},\mathbf{B})$$

The question is now how to infer τ. A possible solution would be to look at experimental data and try to understand how τ changes with the nature of the dependence. If **B** and **C** are fully redundant, then $\tau = 0$. If they are independent, then $\tau = 1$. The behavior between these extremes could be inferred from other synthetic cases.

The assumption of conditional independence opens a new avenue of research in integration of information from multiple sources. We can build from the experience in statistics. There are many potential applications and several unsolved issues that must be addressed in future research.

References

[1] J. A. Anderson. Diagnosis by logistic discriminant function: Further practical problems and results. *Applied Statistics*, 23(3):397–404, 1974.

[2] P. Dawid. Conditional independence in statistical theory (With discussion). *J. Roy. Statist. Soc. B*, 41:1–31, 1979.

[3] E. Frank, L. Trigg, G. Holmes, and I. H. Witten. Technical note: Naive bayes for regression. *Machine Learning*, 41:5–25, 2000.

[4] J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[5] A. G. Journel. Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, 34(5):573–596, July 2002.

[6] H. Warner, H. Toronto, L. Veesey, and R. Stephenson. A mathematical approach to medical diagnosis. *Journal of the American Medical Association*, 177:177–183, 1961.