The Potential Distortion of Bayesian Reasoning

Feng-Jen Yang
Computer Science Department
Prairie View A&M University
Prairie View, TX 77446

Abstract

The Bayesian theorem was formulated in the 18th century and has been adopted as the theoretical basis of uncertainty management as well as the stochastic foundation for forecast-oriented expert systems. Through Bayesian reasoning, people accumulate evidences and draw hypothetical conclusions according to the evidences being observed within the problem domain. Mathematically, the reasoning steps can be represented by a sequence of probabilistic computations. However, without a good mapping to human mental models, Bayesian reasoning is *neither* nature *nor* intuitive. To reduce the mathematical complexity and make it workable with human mental models, an assumption is usually made. This assumption does simplify the probabilistic computation and make it more manageable from the perspective of human cognition. Nevertheless, the simplification also somehow introduces errors to the computation and makes it distorted from the real probabilistic result. The range of this potential error depends on how well is the assumption fitted to the real case being dealt with.

In this paper, I present cases of probabilistic computations to elicit the range of the potential probabilistic distortion caused by Bayesian assumption. Instead of involving the detail but tedious probabilistic computations, I use graphical approach to visualize the distortion being introduced to the result by showing cases from best-fitted, partial-fitted to worst-fitted.

Keywords: *Bayesian Reasoning*, *Bayesian Rule*, *Bayesian Theorem*, *Uncertainty Management*.

1.0 Introduction

Human languages are ambiguous by nature. Imprecise terms such as *usually*, *often*, *sometimes*, *seldom*, and *rarely* are locutions bringing uncertainties to our daily conversations. Although people can easily narrow down the scope of their dialogues and make the subject being discussed more specific, the intrinsic ambiguity of human languages will still introduce a certain extent of uncertainty to the conclusion. Uncertainty management and stochastic conclusion, thus, become inevitable issues in our daily lives. A good way to take ambiguous locutions into consideration while producing computational conclusions is to quantify those locutions from the perspective of probabilities. The idea of quantifying imprecise locutions for probabilistic computations was first done by Simpson in 1944 [1] and repeated by Hakel in 1968 [2]. The resulted quantifications are still in use on contemporary expert systems and surviving the test of time.

Bayesian reasoning, also known as Bayesian theorem or Bayesian rule, is a mathematical approach of deriving probabilistic conclusions based on uncertain evidences. After being formulated in the 18th century, this theorem was soon adopted as the theoretical basis of managing uncertainties and triggered the development of forecast-oriented expert systems since the late 1970s.

Nowadays, Bayesian theorem has offered a stochastic foundation for expert systems to deal with forecast and classification problems, ranging from pattern recognition, medical diagnostic, weather forecast, to natural language processing [3].

2.0 The Inverse Probability

An unnatural property making Bayesian theorem hard to understand is that it turns a probability around. This is also the property that transforms the mathematical equations to meet the predictive purpose required by most of the stochastic systems. The idea can be illustrated by the mathematical notation of conditional probability in which $P(A|B)$ represents the conditional probability of event A given event B, i.e., if event B has happened the probability that event A will also happen.

The way Bayesian theorem turns a conditional probability around is that it provides a method of computing $P(B|A)$ by using $P(A|B)$. The algebra required for this computation is not difficult but *counter-intuitive* and *unnatural*.

2.1 The Transformation of Equations

Mathematically, the conditional probability of H given E is represented by:

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}$$

In the literature of expert systems, E denotes an *evidence* and H denotes a *hypothesis*. This equation is actually representing a prediction of:

> *If the evidence E is observed, how much likely is the hypothesis H true?*

While developing an forecast-oriented system, if the domain expert can provide the values of $P(H \cap E)$ and $P(E)$ then the prediction, $P(H|E)$, is just a simple division. However, $P(H \cap E)$ and $P(E)$ are not intuitive values that human experts can easily come up with disregarding how experienced the domain expert might be. Most of the knowledge engineers can not precisely extract these two values from their interviews with the domain expert. Indeed, getting these two values directly is a kind of information processing against human mental model. This explains why most of the knowledge engineers try the other way around to find alternative values which can have an equal computation of $P(H|E)$ mathematically.

Since human experts can easily come up with $P(E|H)$, $P(E|\overline{H})$, $P(H)$ and $P(\overline{H})$ based on their experiences, the prediction can be computed indirectly as [4]:

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}$$

$$= \frac{P(E \cap H)}{P(E \cap H) + P(E \cap \overline{H})}$$

$$= \frac{P(E \mid H) \times P(H)}{P(E \mid H) \times P(H) + P(E \mid \overline{H}) \times P(\overline{H})}$$

2.2 The Probabilistic Reasoning

The steps of Bayesian reasoning are a series of mathematical computations that transform conditional probabilistic equations into a form that makes sense to the mental models of domain experts.

Considering a given problem domain with n evidences and m hypotheses, the predictive reasoning for:

> *If evidences E_1, E_2, ... and E_n are observed, how much likely is the hypothesis H_i true, where $1 \leq i \leq m$?*

can be performed by the following computation:

$$P(H_i \mid (E_1 \cap E_2 \cap \ldots \cap E_n))$$

$$= \frac{P(H_i \cap (E_1 \cap E_2 \cap \ldots \cap E_n))}{P(E_1 \cap E_2 \cap \ldots \cap E_n)}$$

$$= \frac{P((E_1 \cap E_2 \ldots \cap E_n) \cap H_i)}{\sum_{j=1}^{m} P((E_1 \cap E_2 \cap \ldots \cap E_n) \cap H_j)}$$

$$= \frac{P((E_1 \cap E_2 \cap \ldots \cap E_n) \mid H_i) \times P(H_i)}{\sum_{j=1}^{m} P((E_1 \cap E_2 \cap \ldots \cap E_n) \mid H_j) \times P(H_j)}$$

where $1 \le i \le m$

2.3 The Bayesian Assumption

The previous computation requires the conditional probabilities involving all possible combinations of evidences and hypotheses which are indeed far beyond any domain expert can handle. Thus, in terms of implementation, conditional independence among these evidences is usually assumed so that:

$$P((E_1 \cap E_2 \cap \ldots \cap E_n) \mid H_i))$$

$$= P(E_1 \mid H_i) \times P(E_2 \mid H_i) \times \ldots \times P(E_n \mid H_i)$$

and thus

$$P(H_i \mid (E_1 \cap E_2 \cap \ldots \cap E_n))$$

$$= \frac{P(E_1 \mid H_i) \times P(E_2 \mid H_i) \times \ldots \times P(E_n \mid H_i) \times P(H_i)}{\sum_{j=1}^{m} P(E_1 \mid H_j) \times P(E_2 \mid H_j) \times \ldots \times P(E_n \mid H_j) \times P(H_j)}$$

where $1 \le i \le m$

Based on this assumption, the computation is much more simplified and become cognitively workable by domain experts. Instead of considering all evidences at once and providing probabilities of $P((E_1 \cap E_2 \cap \ldots \cap E_2) \mid H_i)$, *where* $1 \le i \le m$, the experts are now considering only one evidence at a time and providing probabilities of $P(E_i \mid H_j)$, *where* $1 \le i \le n$ *and* $1 \le j \le m$.

However, the assumption that all evidences are conditional independent may not be true in the really life case being dealt with and, as a result, introduces distortions to the final results.

3.0 The Probabilistic Distortion

The potential probabilistic distortion of Bayesian reasoning is related to how well the Bayesian assumption is fitted to real life cases being dealt with. Mathematically, the fitness can be represented by the difference between:

$P(E_1 \cap E_2 \cap \ldots \cap E_n)$ *and* $P(E_1) \times P(E_2) \times \ldots \times P(E_n)$

The less difference, the better fitness.

The range of this distortion is not easy to estimate mentally but can be illustrated by showing the probabilistic computation of cases from *best-fitted*, *partial-fitted* to *worst-fitted*.

3.1 The Best-Fitted Case

The case fitting Bayesian reasoning best happens when there is no conditional dependence among evidences.

Mathematically, this guarantees that there is no difference between:

$P(E_1 \cap E_2 \cap \ldots \cap E_n)$ *and* $P(E_1) \times P(E_2) \times \ldots \times P(E_n)$

So, the computation will produce correct results.

An instance of the best-fitted case can be illustrated as:

Given a universe space U *in which there are three evidences* E_1, E_2 *and* E_3, *and three hypotheses* H_1, H_2, *and* H_3, *where:*

$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$E_1 = \{1, 2, 3\}$, $H_1 = \{1, 4, 7\}$

$E_2 = \{4, 5, 6\}$, $H_2 = \{2, 5, 8\}$

$E_3 = \{7, 8, 9\}, \quad H_3=\{3, 6, 9\}$

In this instance E_1, E_2 *and* E_3 are *mutually independent* so that:

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) \times P(E_2) \times P(E_3)$$

Graphically, it is illustrated as Figure 1.

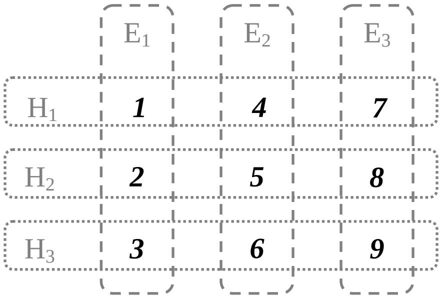

Figure 1. A Best-Fitted Instance

The post probabilities of:

$P(H_i|E_1 \cap E_2 \cap E_3)$, *where* $1 \leq i \leq 3$

are:

$$P(H_i \mid (E_1 \cap E_2 \cap E_3))$$

$$= \frac{P(H_i \cap (E_1 \cap E_2 \cap E_3))}{P(E_1 \cap E_2 \cap E_3)}$$

$$= \frac{P((H_i \cap E_1) \cap (H_i \cap E_2) \cap (H_i \cap E_3))}{P(E_1 \cap E_2 \cap E_3)}$$

$$= \frac{\frac{1}{9} \times \frac{1}{9} \times \frac{1}{9}}{\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}} = \frac{1}{27}, \text{ where } 1 \leq i \leq 3$$

3.2 The Partial-Fitted Case

The case fitting Bayesian reasoning partially happens while some evidences are conditional dependent and others are not.

Mathematically, there is a difference between:

$P(E_1 \cap E_2 \cap \ldots \cap E_n)$ *and* $P(E_1) \times P(E_2) \times \ldots \times P(E_n)$

So, the computation will produce distorted results.

An instance of the partial-fitted case can be illustrated as:

Given a universe space U in which there are three evidences E_1, E_2 *and* E_3, *and three hypotheses* H_1, H_2, *and* H_3, *where:*

$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$E_1 = \{1, 2, 3\}, \quad H_1=\{1, 4, 7\}$

$E_2 = \{1, 2, 3, 4, 5, 6\}, \quad H_2=\{2, 5, 8\}$

$E_3 = \{7, 8, 9\}, \quad H_3=\{3, 6, 9\}$

In this instance $(E_1 \cap E_2)$ *and* E_3 *are mutually independent* and $P(E_1 \cap E_2) = P(E_1)$ so that:

$$P((E_1 \cap E_2) \cap E_3) = P(E_1 \cap E_2) \times P(E_3)$$

$$= P(E_1) \times P(E_3)$$

Graphically, it is illustrated as Figure 2.

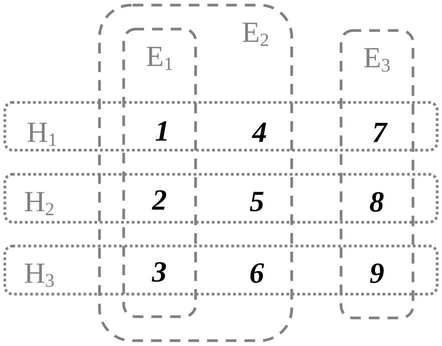

Figure 2. A Partial-Fitted Instance

The post probabilities of:

$P(H_i|E_1 \cap E_2 \cap E_3)$, *where* $1 \leq i \leq 3$

are:

$$P(H_i \mid (E_1 \cap E_2 \cap E_3))$$

$$=\frac{P(H_i \cap ((E_1 \cap E_2) \cap E_3))}{P((E_1 \cap E_2) \cap E_3)}$$

$$=\frac{P(H_i \cap (E_1 \cap E_3))}{P(E_1 \cap E_3)}$$

$$=\frac{P((H_i \cap E_1) \cap (H_i \cap E_3))}{P(E_1 \cap E_3)}$$

$$=\frac{P(H_i \cap E_1) \times P(H_i \cap E_3)}{P(E_i) \times P(E_3)}$$

$$=\frac{\frac{1}{9} \times \frac{1}{9}}{\frac{1}{3} \times \frac{1}{3}} = \frac{1}{9}, \textit{ where } 1 \le i \le 3$$

3.3 The Worst-Fitted Case

The case fitting Bayesian reasoning worst happens while *E_1 is a proper subset of E_2*, *E_2 is a proper subset of E_3,…*, and *E_{n-1} is a proper subset of E_n*.

Mathematically, there is a big difference between:

$P(E_1 \cap E_2 \cap \ldots \cap E_n)$ *and* $P(E_1) \times P(E_2) \times \ldots \times P(E_n)$

In this case, $P(E_1 \cap E_2 \cap \ldots \cap E_2)$ is equal to $P(E_1)$ only, without multiplying any other term. So, the computation may result in very distorted results.

An instance of the worst-fitted case can be illustrated as follows:

Given a universe space U in which there are three evidences E_1, E_2 and E_3, and three hypotheses H_1, H_2, and H_3, where:

$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$E_1 = \{1, 2, 3\}$, $H_1=\{1, 4, 7\}$

$E_2 = \{1, 2, 3, 4, 5, 6\}$, $H_2=\{2, 5, 8\}$

$E_3 = \{\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $H_3=\{3, 6, 9\}$

In this instance $(E_1 \cap E_2 \cap E_3)$ is equal to E_1 so that:

$P(E_1 \cap E_2 \cap E_3) = P(E_1)$

Graphically, it is illustrated as Figure 3.

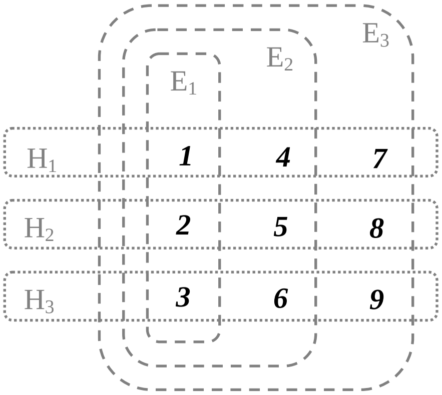

Figure 3. A Worst-Fitted Instance

The post probabilities of:

$P(H_i | E_1 \cap E_2 \cap E_3)$, *where* $1 \le i \le 3$

are:

$$P(H_i \mid (E_1 \cap E_2 \cap E_3))$$

$$=\frac{P(H_i \cap (E_1 \cap E_2 \cap E_3))}{P(E_1 \cap E_2 \cap E_3)} = \frac{P(H_i \cap E_1)}{P(E_1)}$$

$$=\frac{\frac{1}{9}}{\frac{1}{3}} = \frac{1}{3}, \textit{ where } 1 \le i \le 3$$

3.3 The Summary of Results

From the above computational cases, the range of potential probabilistic distortions caused by Bayesian assumption, as summarized in Table 1, becomes obvious. Even though the *best-fitted* instance has the correct result as *1/27*, the *partial-fitted* instance should have the correct result as *1/9* which is *3 times* of the distorted result as *1/27*. Similarly, the *worst-fitted* instance should have the correct result as *1/3* which is *9 times* of the distorted result as *1/27*. The range of potential probabilistic distortions caused by Bayesian assumption could be significant and should be further taken cared while applying Bayesian reasoning.

Table 1. The Potential Probabilistic Distortions

	Computed Result	Correct Result
Best-Fitted Instance	*1/27*	*1/27*
Partial-Fitted Instance	*1/27*	*1/9*
Worst-Fitted Instance	*1/27*	*1/3*

4.0 About Thomas Bayes

Thomas Bayes (1702 – 1761) was a British minister and mathematician. Being a minister, he kept a low profile throughout his life but interested in communications with other mathematicians. The well known Bayesian Theorem was not published by himself but through another minister two years after he died [5]. After Bayes past away, his family asked another minister, Richard Price, to examine his unpublished mathematical papers and found the special one known by the title *An Essay towards Solving a Problem in the Doctrine of Chances.* Price recognized the importance of this work and had it published in the *Philosophical Transactions of the Royal Society* [6].

5.0 References

[1] Simpson, R. 1944. The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency. *The Quartely Journal of Speech*. No. 30, Pages 328 – 330.

[2] Hakel, M.D. 1968. How Often is Often. *American Psychologist.* No. 23, Pages 533 – 534.

[3] D'Agnostini, G. 2003. *Bayesian Reasoning in Data Analysis: A Critical Introduction*. River Edge, NJ: World Scientific Publishing Co. Pte. Ltd.

[4] Firebaugh, M.W. 1989. *Artificial Intelligence: A Knowledge-Based Approach.* Boston, MA: PWS-KENT Publishing Co.

[5] Tabak, J. 2004. *Probability and Statistics: The Science of Uncertainty*. New York City, NY: Facts On File, Inc., Pages. 46 – 50.

[6] Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*. Vol. 53, Pages 370 – 418.