

Bayesian and Dempster-Shafer models for combining multiple sources of evidence in a fraud detection system

Fabrice Daniel

Artificial Intelligence Department of Lusion, Paris, France
<http://www.lusion.com>

March 2021

ABSTRACT

Combining evidence from different sources can be achieved with Bayesian or Dempster-Shafer methods. The first requires an estimate of the priors and likelihoods while the second only needs an estimate of the posterior probabilities and enables reasoning with uncertain information due to imprecision of the sources and with the degree of conflict between them. This paper describes the two methods and how they can be applied to the estimation of a global score in the context of fraud detection.

Keywords: Naive Bayes classifier, Dempster-Shafer theory, Dempster's rule of combination, fraud detection

1 INTRODUCTION

Fraud detection mainly relies on expert driven methods that implement a set of rules and data driven approaches implementing machine learning (ML) models. Both provide an estimate (or a score) for a new transaction to be fraudulent.

While each ML model naturally returns a fraud probability, the experts can also attach a probability to each rule. They can also be automatically calculated from the labeled history. Combining them together produces a global score that can be used in a near real time system to rank a set of transactions having the highest probability to be fraudulent. By obtaining this ranking, investigators can concentrate their efforts on the suspect transactions with the highest probability of being true frauds.

The most common approaches for combining scores are summing individual scores or returning the highest score among the triggered rules. This is not entirely satisfactory given that summing scores is equivalent to averaging the probabilities returned by each predictor (rule or model). It also does not take into account the uncertainty of each predictor and the degree of conflict between them.

For the Lusion fraud system, we work on implementing more appropriate approaches.

This paper proposes two ways for addressing this problem. The first is to use **Bayesian** methods [5]; the second is to combine the scores by using **Dempster-Shafer** theory [6].

2 BAYESIAN APPROACH

2.1 Fundamental concepts

- Let E_i be the evidence i that corresponds to a fraud detection rule (ML model) with $1 \leq i \leq n$,
- Let $E = \{E_1, \dots, E_n\}$ be a set of evidence,
- There are two hypotheses, H_f and H_g , corresponding to a fraudulent and genuine transaction respectively,
- Let $P(H_f|E_i)$ be the probability of transaction being fraudulent given E_i ,
- Let $P(H_g|E_i)$ be the probability of transaction being genuine given E_i ,
- Let $P(E_i|H_f)$ be the probability of E_i being triggered given H_f , and is called the likelihood.

In the context of a Fraud Detection System (FDS), $P(H_f|E_i)$ corresponds to the output of a ML model E_i or to the probability attached by an expert to rule E_i .

$$\begin{aligned} P(H_f|E) &= \frac{P(E|H_f)P(H_f)}{P(E)} \\ &= \frac{P(E_1, \dots, E_n|H_f)P(H_f)}{P(E_1, \dots, E_n)} \end{aligned} \quad (1)$$

Because $P(E_1, \dots, E_n|H_f)$ in (1) is intractable we assume that the events E_i are independent, so we can use a naive Bayes model to compute the combined probability $P(H_f|E)$.

$$\begin{aligned} P(H_f|E) &= \frac{P(E_1|H_f)P(E_2|H_f)\dots P(E_n|H_f)P(H_f)}{P(E_1, \dots, E_n)} \\ &= P(H_f) \frac{\prod_{i=1}^n P(E_i|H_f)}{P(E_1, \dots, E_n)} \end{aligned} \quad (2)$$

And because

$$P(E) = P(H_f)P(E|H_f) + P(H_g)P(E|H_g) \quad (3)$$

we find

$$P(H_f|E) = \frac{1}{Z}P(H_f) \prod_{i=1}^n P(E_i|H_f) \quad (4a)$$

$$P(H_g|E) = \frac{1}{Z}P(H_g) \prod_{i=1}^n P(E_i|H_g) \quad (4b)$$

Where Z is a normalization factor:

$$\begin{aligned} Z &= P(E) \\ &= P(H_f)P(E|H_f) + P(H_g)P(E|H_g) \\ &= P(H_f) \prod_{i=1}^n P(E_i|H_f) + P(H_g) \prod_{i=1}^n P(E_i|H_g) \end{aligned} \quad (5)$$

When n becomes large, there is a risk of vanishing precision. To fix this issue we can apply a logarithm to transform the product into a sum.

The main shortcoming of this approach is the independence assumption that is not the case in most of the real problems. In [8] an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the surprising implausible efficacy of naive Bayes classifiers.

2.2 Example

Table 1 shows an example with numerical values from [5].

Let E_1 and E_2 be triggered rules or ML model outputs.

The dataset has 30 transactions including 7 frauds.

	fraud	genuine	$P(E_i H_f)$	$P(E_i H_g)$
E_1	4	6	0.57	0.26
E_2	1	2	0.14	0.09

Table 1: Example with two pieces of evidence E_1 and E_2

This table means that:

- E_1 is triggered by 10 transactions out of 30 in the dataset
- When E_1 is observed, 4 transactions were true frauds, and 6 were not
- $P(E_1|H_f)$ is the likelihood, the probability of observing E_1 , given H_f , meaning that a transaction is fraudulent. Among the 7 frauds observed, 4 of them trigger the rule E_1
- $P(E_1|H_g)$ is the likelihood, the probability of observing E_1 , given H_g , meaning that a transaction is genuine. Among the 23 genuine transactions observed, 6 of them trigger the rule E_1

First we estimate prior probabilities:

$$P(H_f) = \frac{7}{30} = 0.23$$

$$P(H_g) = \frac{23}{30} = 0.77$$

Assume that we observe the set $E = \{E_1, E_2\}$ of evidence. Then let's compute the likelihoods:

$$\begin{aligned} P(E|H_f) &= \prod_{i=1}^n P(E_i|H_f) \\ &= P(E_1|H_f) \cdot P(E_2|H_f) \\ &= 0.57 \times 0.14 \\ &= 0.0798 \end{aligned}$$

$$\begin{aligned} P(E|H_g) &= \prod_{i=1}^n P(E_i|H_g) \\ &= P(E_1|H_g) \cdot P(E_2|H_g) \\ &= 0.26 \times 0.09 \\ &= 0.0234 \end{aligned}$$

This means the likelihood of this transaction being a fraud is 0.0798.

We then compute the marginal likelihood $P(E)$:

$$\begin{aligned} P(E) &= P(H_f)P(E|H_f) + P(H_g)P(E|H_g) \\ &= 0.23 \times 0.0798 + 0.77 \times 0.0234 \\ &= 0.0184 + 0.0180 \\ &= 0.0364 \end{aligned}$$

So the posterior probabilities are:

$$\begin{aligned} P(H_f|E) &= \frac{P(E|H_f) \cdot P(H_f)}{P(E)} \\ &= \frac{0.0798 \times 0.23}{0.0364} \\ &= 0.504 \end{aligned}$$

$$\begin{aligned} P(H_g|E) &= \frac{P(E|H_g) \cdot P(H_g)}{P(E)} \\ &= \frac{0.0234 \times 0.773}{0.0364} \\ &= 0.495 \end{aligned}$$

The probability of this transaction being a fraud is 0.504.

3 DEMPSTER-SHAFER APPROACH

Using Naive Bayes is only possible if we can obtain priors and likelihoods estimates from an expert or from historical data.

In most cases we only have the posterior probabilities attached to rules or models. It can come from an **expert estimate**¹ attached to each rule or from a machine learning model probability prediction.

A way to combine probabilities of fraudulence given individual rules is to use Dempster-Shafer theory.

Dempster-Shafer theory (DST) provides a framework for combining different sources of evidence into a global belief for a given hypothesis [1, p. 36].

3.1 Fundamental concepts

Let Ω be the universe of all the possible states, meaning the set of all the N hypothesis, also called the *frame of discernment*.

$$\Omega = \{H_1, \dots, H_N\} \quad (6)$$

We can define a set 2^Ω , named the *power set*, that contains all the possible subsets of Ω , including the empty set.

$$2^\Omega : \{\emptyset, \{H_1\}, \dots, \{H_N\}, \{H_1, H_2\}, \dots, \Omega\} \quad (7)$$

In our case we assume a universe of two hypotheses H_f and H_g for fraudulent and genuine transactions.

$$2^\Omega : \{\emptyset, \{H_f\}, \{H_g\}, \{H_f, H_g\}\} \quad (8)$$

The theory of evidence assigns a belief mass to each element of the *power set*. Formally the mass function, called the *basic mass assignment* (BMA), *basic belief assignment* (BBA) or *basic probability assignment* (BPA) depending on the source², is defined by:

$$m : 2^\Omega \rightarrow [0, 1] \quad (9)$$

First, the mass of the empty set is zero:

$$m(\emptyset) = 0 \quad (10)$$

Second, the masses of all the members of the power set add up to a total of 1:

$$\sum_{A \in 2^\Omega} m(A) = 1 \quad (11)$$

The mass $m(A)$ in (11) is interpreted as the part of belief placed strictly on A . It expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to A but to no particular subset of A . This quantity differs from a probability since the total mass can be given either to singleton hypotheses H_n or to composite ones A [1, p. 38].

¹When an expert attaches 0.75 to a rule, he estimates that when this rule is triggered it has 75% chances to be a fraud.

²basic probability assignment (BPA) seems to be the most frequent naming

While a probability can only be assigned to the singletons H_f or H_g , a mass $m(A)$ can also be assigned to the composition $\{H_f, H_g\}$.

Belief mass on $m(A)$ where A is a singleton is interpreted as: *A is true*.

Belief mass on $m(A)$ where A is non-atomic is interpreted as: *one of the A components is true, but the source is uncertain about which one of them is true*.

Elements of Ω having $m(A) \neq 0$ are called *focal elements*.

3.2 Belief, Plausibility, uncertainty and probability interval

The belief (also named credibility) $bel(A)$ for a set A is defined as the sum of all the masses of subsets of the set of interest:

$$bel(A) = \sum_{B|B \subseteq A} m(B) \quad \forall A \subseteq \Omega \quad (12)$$

So in our case where $A \in 2^\Omega$ and 2^Ω defined in (8), we find

$$bel(H_f) = m(H_f) \quad (13a)$$

$$bel(H_g) = m(H_g) \quad (13b)$$

The plausibility $pl(A)$ is the sum of all the masses of the sets B that intersect the set of interest A :

$$pl(A) = \sum_{B|B \cap A \neq \emptyset} m(B) \quad \forall A \subseteq \Omega \quad (14)$$

So in our case:

$$pl(H_f) = m(H_f) + m(H_f, H_g) \quad (15a)$$

$$pl(H_g) = m(H_g) + m(H_f, H_g) \quad (15b)$$

It can be shown that

$$bel(A) \leq pl(A) \quad \forall A \subset \Omega \quad (16)$$

This equation can be interpreted as "certain implies plausible" [1, p. 40].

Plausibility and belief are related to each other as follows:

$$pl(A) = 1 - bel(\bar{A}) \quad \forall A \subset \Omega \quad (17)$$

According to [1, p. 41] a probability interval can be defined as the interval with $bel(A)$ and $pl(A)$ as its lower and upper bound respectively

$$bel(A) \leq P(A) \leq pl(A) \quad (18)$$

The difference between $pl(A)$ and $bel(A)$ is the ignorance about a specific hypothesis A .

The author also states that :

“if focal sets are only singletons (i.e. we assign only masses to singleton hypothesis), then the mass distributions, credibility measures, plausibility ones and commonalities are merged and coincide with a probability distribution.”

In our case, we are only assigning a mass to each of the triggered Fraud detection rules. For each of them, an expert has attached a score, meaning a probability to be a fraud given that it has been triggered. Each rule or machine learning model only returns the probability for a transaction to be a fraud. So the focal set are only singletons, meaning we only have the following masses defined : $m_i(H_f)$ and $m_i(H_g)$, where i is the i^{th} triggered rule.

$$m_i(\bar{H}_f) = m_i(H_g) \quad (19a)$$

$$m_i(\bar{H}_g) = m_i(H_f) \quad (19b)$$

$$m_i(H_f, H_g) = 0 \quad (19c)$$

This means in our case:

$$P(A) = bel(A) = pl(A) \quad (20)$$

If we want to also consider the rules not triggered in the model, we should assign the whole mass to the uncertainty with $m_j(H_f, H_g) = 1$ where j is the j^{th} not-triggered rule.

Here, we only consider the masses on the triggered rules.

3.3 Dempster’s rule of combination

When several rules are triggered, we want to calculate the probability for a transaction to be a fraud.

Dempster-Shafer proposes a combination rule for calculating the set of masses $m_{1,2}$ from m_1 and m_2 .

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) \\ = \frac{1}{1 - K} \sum_{B \cap C = A \neq \emptyset} m_1(B)m_2(C) \quad (21)$$

Where K is a measure of the degree of conflict between two mass sets. $1 - K$ is the *normalization factor*.

$$K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \\ = (m_1 \oplus m_2)(\emptyset) \quad K \in [0, 1] \quad (22)$$

Having K near to 0 means there is small conflict between the two mass sets and 1 means that they are in total conflict.

From (21) and (22) we can compute $m_{1,2}(H_f)$ and $m_{1,2}(H_g)$:

$$K = (m_1(H_f).m_2(H_g)) \\ + (m_1(H_g).m_2(H_f)) \quad (23a)$$

$$m_{1,2}(H_f) = \frac{m_1(H_f).m_2(H_f)}{1 - K} \quad (23b)$$

$$m_{1,2}(H_g) = \frac{m_1(H_g).m_2(H_g)}{1 - K} \quad (23c)$$

After getting the combined mass $m_{1,2}$ we need to get the corresponding beliefs. In our case we saw with (13a) and (13b) that they can be directly derived from m , so:

$$bel(H_f) = m_{1,2}(H_f) \quad (24a)$$

$$bel(H_g) = m_{1,2}(H_g) \quad (24b)$$

And as seen with (20): $P(A) = bel(A)$

So we get:

$$P(H_f) = bel(H_f) \quad (25a)$$

$$P(H_g) = bel(H_g) \quad (25b)$$

3.4 Combining more than two sources

Equation (21) is relative to two masses sets only. In a rule engine or more generally in a multiple source decision system (rules + several machine learning models) we need to potentially handle many rules that can be triggered together on the same transaction.

So we need to be able to compute :

$$(m_1 \oplus m_2 \oplus \dots \oplus m_n)(A) \quad (26)$$

According to [3]

$$m_1 \oplus m_2 \oplus m_3 = (m_1 \oplus m_2) \oplus m_3 \\ = m_1 \oplus (m_2 \oplus m_3) \\ = m_2 \oplus (m_1 \oplus m_3) \quad (27)$$

We use this property to combine as many sources as needed.

Now we have a framework to compute the probability of a fraud from several sources of evidence.

3.5 Examples

3.5.1 Without uncertainty

Table 2 shows an example with numerical values from [2].

Assume two sources S_1 and S_2 (rules or models) providing sets of masses m_1 and m_2 respectively.

	S_1	$m_1(H_f) = 0.6$	$m_1(H_g) = 0.4$
S_2			
	$m_2(H_f) = 0.8$	0.48	0.32 (conflict)
	$m_2(H_g) = 0.2$	0.12 (conflict)	0.08

Table 2: Two sources and no mass assignment to uncertainty

Then, as per the Dempster-Shafer combination rule:

$$\begin{aligned}
K &= (m_1(H_f).m_2(H_g)) \\
&+ (m_1(H_g).m_2(H_f)) \\
&= (0.6 \times 0.2) + (0.4 \times 0.8) \\
&= 0.44 \\
(m_1 \oplus m_2)(H_f) &= \frac{1}{1 - 0.44} m_1(H_f).m_2(H_f) \\
&= \frac{0.48}{0.56} \\
&= 0.8571 \\
(m_1 \oplus m_2)(H_g) &= \frac{1}{1 - 0.44} m_1(H_g).m_2(H_g) \\
&= \frac{0.08}{0.56} \\
&= 0.1428
\end{aligned}$$

Since, in our case, focal sets are singletons:

$$P(H_f) = bel(H_f) = m_{1,2}(H_f) = 0.8571$$

The probability of this transaction being a fraud is 0.8571.

3.5.2 With uncertainty

A major advantage of Dempster-Shafer is its capacity to consider uncertainty. It gives the ability to return a probability interval instead of a point estimate.

To illustrate this principle, let us assume the fraud detection rule engine allows the experts to assign an uncertainty mass to any rule.

In table 3 we assume that the expert assigns masses $m_1(H_f, H_g) = 0.2$ and $m_2(H_f, H_g) = 0.5$ to quantify respectively the uncertainty of the sources S_1 and S_2 .

So we compute :

$$\begin{aligned}
K &= (m_1(H_f).m_2(H_g)) \\
&+ (m_1(H_g).m_2(H_f)) \\
&= (0.6 \times 0.2) + (0.4 \times 0.3) \\
&= 0.17 \\
(m_1 \oplus m_2)(H_f) &= \frac{1}{1 - 0.17} m_1(H_f).m_2(H_f) \\
&= \frac{0.21}{0.83} \\
&= 0.253 \\
(m_1 \oplus m_2)(H_g) &= \frac{1}{1 - 0.24} m_1(H_g).m_2(H_g) \\
&= \frac{0.08}{0.83} \\
&= 0.024
\end{aligned}$$

$$\begin{aligned}
(m_1 \oplus m_2)(H_f, H_g) &= \frac{1}{1 - 0.24} (m_1(H_f).m_2(H_f, H_g) \\
&+ m_1(H_g).m_2(H_f, H_g) \\
&+ m_2(H_f).m_2(H_f, H_g) \\
&+ m_2(H_g).m_2(H_f, H_g) \\
&+ m_1(H_f, H_g).m_2(H_f, H_g)) \\
&= \frac{0.6}{0.83} \\
&= 0.723
\end{aligned}$$

$$\begin{aligned}
pl(H_f) &= m(H_f) + m(H_f, H_g) \\
&= 0.253 + 0.723 \\
&= 0.976
\end{aligned}$$

The true probability for this transaction to be a fraud is in this interval:

$$\begin{aligned}
bel(H_f) &\leq P(H_f) \leq pl(H_f) \\
0.25 &\leq P(H_f) \leq 0.98
\end{aligned}$$

Now let's study, with the example from table 4, what happens if after a while, the expert gets more statistics about these rules, such that they can reduce the uncertainty.

In this case we find that the probability interval is reduced, with the minimum probability increasing from 0.25 to 0.40 while the masses $m_1(H_f)$ and $m_2(H_f)$ remain unchanged.

$$0.40 \leq P(H_f) \leq 0.77$$

With this information the fraud scoring engine is able to provide a more informative ranking of the transactions, by including an estimate of the uncertainty.

Some transactions, not detected as fraudulent with a point estimate, become suspect when uncertainty is taken into consideration.

For instance if we modify the previous example by removing uncertainty, allocating the remaining mass to H_g such as $m_1(H_g) = 0.3$ and $m_2(H_g) = 0.7$, we find $P(H_f) = 0.5$; if the threshold for a fraud detection is set to $\tau > 0.5$ then this transaction is not considered fraudulent, while in the previous case it is considered suspicious.

Now if we reduce the uncertainty mass by half, and distribute the remaining mass equally between H_f and H_g by adding 0.025 to each of them, then we find

$$0.50 \leq P(H_f) \leq 0.70$$

The interval is reduced and the credibility increases up to 0.5, so the ranking algorithm can decide to give it a better rank even though the plausibility is lower (0.70 compared to 0.77).

$S_2 \backslash S_1$	$m_1(H_f) = 0.7$	$m_1(H_g) = 0.1$	$m_1(H_f, H_g) = 0.2$
$m_2(H_f) = 0.3$	0.21	0.03 (conflict)	0.06
$m_2(H_g) = 0.2$	0.14 (conflict)	0.02	0.04
$m_2(H_f, H_g) = 0.5$	0.35	0.05	0.10

Table 3: Two sources and masses assignment to uncertainty

$S_2 \backslash S_1$	$m_1(H_f) = 0.7$	$m_1(H_g) = 0.2$	$m_1(H_f, H_g) = 0.1$
$m_2(H_f) = 0.3$	0.21	0.06 (conflict)	0.03
$m_2(H_g) = 0.6$	0.42 (conflict)	0.12	0.06
$m_2(H_f, H_g) = 0.1$	0.07	0.02	0.01

Table 4: Two sources and reduced masses assignment to uncertainty

4 CONCLUSION

Combining probability estimates of fraud detection rules and ML models predictions by using Dempster-Shafer has two advantages compared to a pure Bayesian approach. First it's applicable to any situation where knowledges or historical data are not available to estimate the prior probabilities. Second it can represents the level of uncertainty, providing an interval instead of only a point estimate for the true probability.

A future improvement study could be focused on how to weight ML models or some specific rules. Such weighting can be useful to reflect the business impact in term of cost³ for a given rule or model.

Another future study could be focused on ranking the combined probability estimates of suspect transactions not only from the point estimates but also by using their respective uncertainty.

REFERENCES

- [1] A. Bellenger. "Semantic Decision Support for Information Fusion Applications". en. In: (2013), p. 222.
- [2] Q. Chen et al. "Data Classification Using the Dempster-Shafer Method". en. In: *Journal of Experimental & Theoretical Artificial Intelligence* 26.4 (Oct. 2014), pp. 493–517. ISSN: 0952-813X, 1362-3079. DOI: 10.1080/0952813X.2014.886301.
- [3] J. Dezert, A. Tchamova, and F. Dambreville. "On the Mathematical Theory of Evidence and Dempster's Rule of Combination". en. In: (2011), p. 12.
- [4] J. Frery. "Ensemble Learning for Extremely Imbalanced Data Flows". en. In: (2019), p. 150.
- [5] J. C. Moso and J. K. Kenei. "Credit Card Fraud Detection Using Bayes Theorem". en. In: 07.04 (2018), p. 6.
- [6] S. Panigrahi et al. "Credit Card Fraud Detection: A Fusion Approach Using Dempster-Shafer Theory and Bayesian Learning". en. In: *Information Fusion* 10.4 (Oct. 2009), pp. 354–363. ISSN: 15662535. DOI: 10.1016/j.inffus.2008.04.001.
- [7] A. D. Pozzolo. "Adaptive Machine Learning for Credit Card Fraud Detection". en. In: (2015), p. 199.
- [8] H. Zhang. "The Optimality of Naive Bayes". en. In: (2004), p. 6.

³the cost concept depends on the context, it can be monetary but it can also be reputation, time or anything else. According to literature such as [4] and [7] this concept is known to be very hard to quantify in the context of payments fraud