Bayesian Weighting Schemes for PPIs

Muhammad Usman Hafeez

June 6, 2018

Bayesian Weighting Schemes are a probabilistic weighting scheme that leverage Bayes' Theorem to use existing experimental data to calculate a weight that represents the reliability of a specific interaction occurring. Let A and B be two events in our sample space S. Bayes' Theorem is stated as

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

The mathematical derivation is as follows. We know that given two events A and B
$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$. This is because the probability of both A and B occurring is the same as the probability that A occurs first and then given this has happened, B occurs or similarly the probability that B occurs first and given this has happened A occurs. If A and B are independent of each other then $P(B \mid A) = P(B)$ or $P(A \mid B) = P(A)$ but the formula still holds. By virtue of symbol manipulation we have

$$\begin{aligned} P(A, B) &= P(A \mid B)P(B) \\ \implies P(A \mid B) &= \frac{P(A, B)}{P(B)} \\ \implies P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)} \end{aligned}$$

This can also be understood more intuitively with the help of an example. Consider a disease like diabetes. Then there are people who either have the disease or do not have the disease. They are D+(for having diabetes) or D-(for not having diabetes) respectively. However, the lab test for whether a patient has diabetes is not entirely accurate (having some accuracy below 100 percent). Then we might test a patient that is D+ and get back a positive result which could be a false-positive or a true-positive. To ascertain whether the test is worth doing we need to quantify how accurate the test is and we can do this by calculating $P(D+ \mid +)$ or the probability that the patient has the disease given we get a positive result. We can calculate this using Bayes' Theorem.
Say we take 100 people out of which 45 are $D+$. We also happen to know that

for this particular test if a person is $D+$ we get a positive result 43 percent of the time.
Then for $P(D+ \mid +)$ our sample space is the number of people who have tested positive for diabetes (but this does not necessarily mean that they are actually $D+$). Then the probability that a given person from this sample space is actually $D+$ would be $\frac{P(D+,+)}{P(+)}$ or the ratio of people who tested positive and have the disease to the people who tested positive. Now that we have an intuition for the formula and a motivation for its use, we can use the numbers we made up above to show how to use Bayes' Theorem in practice.

$$\begin{aligned} P(D+ \mid +) =& \frac{P(+ \mid D+)P(D+)}{P(+)} \\ =& \frac{0.43.0.45}{P(+ \mid D+)P(D+) + P(+ \mid D-)P(D-)} \\ =& \frac{0.43.0.45}{(0.43.0.45) + (0.57.0.55)} \\ =& 0.38 \end{aligned}$$

We will also need to introduce one more concept that will later help simplify calculations at the risk of introducing some error into our weighting scheme. In particular, it will somewhat inflate the weighting schemes but we will attempt to remedy this by capping all weights at 0.75. This concept is called conditional independence. Consider three events A, B and C. Then A and B are conditionally independant with respect to C iff:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

That is to say, the probability of B hapenning given C has hapenned does not affect the probability of A hapenning given C has hapenned. Compare this to the equation for two independent events and this is more readily clear.

$$P(A, B) = P(A)P(B)$$

We can extend this to many more conditionally independent variables as such

$$P(A_1, A_2, A_3, ...A_k \mid C) = P(A_1 \mid C)P(A_2 \mid C)P(A_3 \mid C)...P(A_k \mid C) = \prod_k P(A_k \mid C)$$

1 Setup

To begin understanding the paper we need to understand the different sets and variables required for a Bayesian Weighting Scheme. We start with an indicator variable I (also known as a binary random variable). This binary random variable records whether the interaction between two proteins p_i and p_j where $i \neq j$ occurs in the interactome. This means that even if the interaction does not occur in a particular signalling pathway or is not an edge in any of the

k-shortest paths, if the two proteins are known to interact, $I = 1$ and $I = 0$ if they do not interact in the same sense. Mathematically,

$$I = \left\{ \begin{array}{l} 1 \text{ if the interaction occurs} \\ 0 \text{ if the interaction does not occur} \end{array} \right\}$$

Then we have an interaction vector E where E can be thought of as a $1 * n$ vector or as a list of n elements. Entries of E are given by E_k's which is an indicator variable that records whether a particular evidence type k suggests the interaction happens or not. Then similarly to I,

$$E_k = \left\{ \begin{array}{l} 1 \text{ if evidence type k suggests the interaction occurs} \\ 0 \text{ if evidence type k suggests the interaction does not occur} \end{array} \right\}$$

So consider a simplified example of what E might look like. We might have four types of evidences for a particular interaction between proteins p_i and p_j. Then if E_1 and E_2 suggest the interaction does happen then and E_3 and E_4 suggest that it does not, then $E_1 = E_2 = 1$ and $E_3 = E_4 = 0$ and $E = [1, 1, 0, 0]$. Notation wise when we write $P(E)$ we mean to say that this is the probability that E equals the one particular interaction vector.
Then we have the necessary set up and can begin to use data to calculate a probability that an interaction happens given the interaction has a particular interaction vector i.e $P(I \mid E)$. We cannot do so directly but Bayes' Theorem will allow us to do so indirectly using experimental data.

2 Calculating $P(I \mid E)$

We begin by constructing a set of true-positives and true-negatives using GO terms to construct a gold-standard positive and negative set. This means that a protein pair is placed in the true positive set P if both are co-annotated by at least one GO term. From the remaining protein pairs, $10. \mid P \mid$ proteins are randomly selected to be N, the true negative gold standard. These sets,P and N are constructed so there is no overlap between them, that is to say there are no elements in common between the two sets. In experimental terms this means we have constructed two sets of terms such that there are no false-positives or false-negatives. We then make the assumption that true positives in the remaining unmarked protein pairs occur at the same rate as true positives occur in the set of true positives and true negatives. This is to say that the sample of true positives and true negatives is held to be representative of the entire interactome considered. Then the probability that a randomly selected pair is a positive or that the interaction occurs is $\frac{|P|}{|P \cup N|}$ the total number of positives over the total number of proteins pairs (both positive and negative). Similarly the probability that a randomly selected pair is negative is $\frac{|N|}{|P \cup N|}$. Using these probabilities let us reconsider our indicator variable I and observe that

$$P(I=i) = \left\{ \begin{array}{l} \frac{|P|}{|P \cup N|} \text{ if } i=1 \\ \\ \frac{|N|}{|P \cup N|} \text{ if } i=0 \end{array} \right\} \tag{1}$$

Then consider a set Z_k which is the set of proteins that have been seen interacting during an experiment or evidence type k. Then similarly we begin to compute $P(E_k = e \mid I = i)$. This is the probability that we have a particular value for E_k given that the interaction occurs or given that the interaction does not occur. For example, we might observe an interaction to occur but evidence type k might suggest that it does not. There are three other possible combinations for e and i. They are illustrated below.

$$P(E_k = e \mid I = i) = \left\{ \begin{array}{l} \frac{|P \cap Z_k|}{|P|} \text{ if } i=1, e=1 \\ \\ \frac{|N \cap Z_k|}{|N|} \text{ if } i=0, e=1 \\ \\ \frac{|P \setminus Z_k|}{|P|} \text{ if } i=1, e=0 \\ \\ \frac{|P \setminus Z_k|}{|P|} \text{ if } i=0, e=0 \end{array} \right\} \tag{2}$$

If $i = 1$ and $e = 1$ for a given protein pair then this protein pair both does interact and has been observed to interact. If $i = 1, e = 0$ then the protein pair does interact but has not been observed to interact under evidence type k. If $i = 0, e = 0$ then the interaction neither occurs nor has been observed to interact under evidence type k. If $i = 0, e = 1$ then the interaction does not occur in the interactome but there is evidence by experiment type k to suggest it could.
Using our probabilities of $P(I = i)$ and $P(E_k = e \mid I = i)$ then we can begin to apply Bayes' Theorem and all of the set up we have performed. Then let the cost of an edge be

$$\mathrm{c}_{uv} = P(I = 1 \mid E)$$

$$= \frac{P(E|I=1)P(I=1)}{P(E)}$$

$$= \frac{P(E|I=1)P(I=1)}{P(E)}$$

$$= \frac{P(E|I=1_P(I=1)}{P(E,I=0+P(E,I=1)}$$

$$= \frac{P(I=1)\prod_k P(E_k|I=1)}{P(I=0)\prod_k P(E_k|I=0)+P(I=1)\prod_k P(E_k|I=1)}$$

Then we can calculate $P(I \mid E)$ in terms of probabilities we already know how to calculate and we can weight the interactome probabilistically using experimental data.