Software Artifact Research and Development Laboratory Technical Report SARD04-11, September 1, 2004

Combining Evidence: the Naïve Bayes Model Vs. Semi-Naïve Evidence Combination

Daniel Berleant
Dept. of Electrical and Computer Engineering, Iowa State University, 3215 Coover Hall,
Ames, Iowa 50011, USA
berleant@iastate.edu
(515)-294-3959

Combining different items of evidence is important because it results in a single composite likelihood. For example, a sentence may have a number of features, each associated with some probability that the sentence describes an interaction between biomolecules. What, then, is the composite likelihood that the sentence describes an interaction implied by combining the various sources of evidence provided by various features of the sentence? By determining such a composite likelihood that a given sentence describes an interaction, important applications can be supported. Putative interactions in automatically generated biomolecular interaction network simulators can be rated, sentences can be ranked for curation, etc. In the following paragraphs we give a theoretical comparison of two methods for evidence combination. One is the well-known Naïve Bayes model. The other is semi-naïve evidence combination.

Naïve Bayes and semi-naïve evidence combination both have a similar scalability advantage over full Bayesian analysis using Bayes Theorem to account for whatever dependencies may exist. That scalability is why they are useful. However when used to estimate probabilities that an item (e.g. a sentence) is in some category (e.g. describes a biomolecular interaction) or in the complementary category, semi-naïve evidence combination makes fewer assumptions. Next we explain each method. Following that they are compared and discussed.

Evidence combination with the Naïve Bayes model. This standard method produces probability estimates which are often used for categorization (Lewis 1998). It may be explained as follows.

1) We wish to determine $p(h|f_1,...,f_n)$ where h represents is a hit – i.e., the sentence in question contains an interaction, and $f_1...f_n$ represent features 1 through n (a feature can be any property of the sentence such as, for example, an attribute-value pair).
2) By Bayes' Theorem we have $p(h|f_1,...,f_n)=p(h)p(f_1,...f_n|h)/p(f_1,...,f_n)$.
3) From standard probability we know that $p(a,b)=p(a)p(b|a)$. (This can be visualized using a Venn diagram.) This fact can be extended to 3 variables: $p(a,c,d)=p(a)p(c,d|a)$ where c,d (that is, the simultaneous occurrence of c and d) acts like a single variable – and can even be named such: let b stand for "c and d." The same fact can also be extended by limiting the domain of discourse (technically, the "reference set") to cases where some property e holds: $p(a,b|e)=p(a|e)p(b|a,e)$.
4) The fact and its extensions laid out in the preceding step may be applied to the term $p(f_1,..f_n|h)$ to give $p(f_1,...f_n|h)$
$=p(f_1|h)p(f_2,...,f_n|f_1,h)$
$=p(f_1|h)\underline{p}(f_2|f_1,h)p(f_3,...f_n|f_2,f_1,h)$
$=p(f_1|h)p(f_2|f_1,h)p(f_3|f_2,f_1,h)p(f_4,...f_n|f_3,f_2,f_1,h)$

=... .

This process is described nicely in, for example, Wikipedia (see refs.).

5) To make the process of calculation tractable, we now start making independence assumptions. Assuming that the features are independent of one another given that the sentence being a hit, implies that $p(f_a|h,f_b,f_c,...)=p(f_a|h)$. In other words, the result of the preceding step becomes: $p(f_1,..f_n|h)=p(f_1|h)p(f_2|h)p(f_3|h)...p(f_n|h)$.
6) By making the additional assumption that the features are *unconditionally* independent (i.e. ignoring whether the sentence is a hit or not), we can simplify the computation of the denominator in step 2 as follows: $p(f_1,...,f_n)=p(f_1)p(f_2)p(f_3)...p(f_n)$. This simplifies the calculations because if there are n attributes each with j possible values, then there are j^n possible $p(f_1,...,f_n)$'s to be determined, and for even relatively modest n and j, j^n is a prohibitive number of probabilities to find.
7) Having made both the numerator and denominator in step 2 tractable, the Naïve Bayes model estimates

$$p(h|f_1,...f_n)=\frac{p(h)p(f_1,...f_n|h)}{p(f_1)p(f_2)p(f_3)...p(f_n)}$$
$$\approx\frac{p(h)p(f_1|h)p(f_2|h)p(f_3|h)...p(f_n|h)]}{p(f_1)p(f_2)p(f_3)...p(f_n)}$$

The Naïve Bayes model is tractable to compute but provides only an estimate of the probability that a given sentence is a hit, because of the two sets of assumptions, which are that the features occur independently of one another both when conditioned by h, and when not conditioned by h.

Semi-naïve evidence combination. This method is scalable in the number of features, like Naïve Bayes, but has the advantage of making fewer independence assumptions under certain important conditions. Unlike the Naïve Bayes model, it does not assume that the features are unconditionally independent (e.g., regardless of whether sentences are hits or not). The most parsimonious formula for semi-naïve evidence combination is

$$O(h|f_1,...,f_n)=O_1...O_n/(O_0)^{n-1}$$

where the odds that a sentence describes an *i*nteraction if it has features $f_1,...,f_n$ are $O(h|f_1,...,f_n)$, the odds that a sentence with feature k is a hit are O_k, and the prior odds (i.e. over all sentences in the test set irrespective of their features) that a sentence is a hit are O_0. The equation $O(h|f_1,...,f_n)=O_1...O_n/(O_0)^{n-1}$ just given is in terms of odds, which are ratios of hits to misses. Thus, for example, the odds of flipping a head are 1/1=1 (1 expected success per failure), while the odds of rolling a six are 1/5 (one success expected per five failures). Odds are easily converted to the more familiar probabilities by applying $p=O/(O+1)$. Similarly, $O=p/(1-p)$.

Let us now derive the preceding equation for $O(h|f_1,...,f_n)$, while referring to a concrete example as appropriate to help develop the intuitions behind it. We start by looking at the odds without reference to any features, then account for features one by one.

1) Prior to considering any features, we must rely on the overall number of sentences that are hits compared to those that are not. The ratio of hits to non-hits determines the prior odds, O_0, of a sentence being a hit. For example, consider a sample, quite small for purposes of explanation, of 8 sentences, 4 that are hits and 4 that are not. In this pool, odds O_0 are 4/4=1 (corresponding to a probability of ½).

2) Now consider the influence of feature f_1, possessed by 4 sentences that are hits and 2 that are not. The odds that a sentence possessing feature f_1 is a hit, O_1, now override the prior odds. For the example, these new odds are 4/2, so $O_1=2$. Note that this is consistent with the equation $O(h|f_1,...,f_n)=O_1...O_n/(O_0)^{n-1}$ given earlier because it implies $O(h|f_1)=O_1/(O_0)^0=2/1=2$. The rationale for this equation, however, becomes evident only when additional features are also considered.

3) Next consider feature f_2, possessed by 4 sentences that are hits and 2 that are not, one of which also possesses feature f_1. What are the odds that a sentence possessing both features f_1 and f_2 is a hit? The odds are determined from the pool of sentences with both features.

   The value is the number that are hits divided by the number that are not. For the example, this is 4/1=4. While the number of sentences to count is small in the example, the number may also be very large, as in the case of all sentences on the Web. Regardless, it is not necessary to actually count the relevant sentences as we are able to do in the small example. The process for finding the odds is described next.

   1. Let F_1 be the number of hits possessing feature f_1.
   2. Let F_2 be the number of hits possessing feature f_2.
   3. Let H be the number of sentences that are hits, and let S_H be the set of such sentences.
   4. Then the fraction of hits with feature f_2 is F_2/H.
   5. Here some assumptions are needed: we must assume f_1 and f_2 occur independently of each other within S_H.
   6. Then, the fraction of hits with feature f_1 that also possess feature f_2, is the same as the fraction of hits with feature f_2 regardless of f_1; that is, F_2/H.
   7. Thus, the number of hits possessing features f_1 and f_2 is $F_1(F_2/H)$.
   8. Similarly, we determine the number of *non*-hits possessing features f_1 and f_2. Let F_1' be the number of non-hits possessing feature f_1.
   9. Let F_2' be the number of non-hits possessing feature f_2.
   10. Let H' be the number of sentences that are non-hits and let S_H' be the set of such sentences.
   11. Then the fraction of non-hits with feature f_2 is F_2'/H'.
   12. Here more assumptions are needed: assume f_1 and f_2 occur independently of each other within S_H'.
   13. Then, the fraction of non-hits with feature f_1 that also possess feature f_2, is the same as the fraction of non-hits with feature f_2 regardless of f_1; that is, F_2'/H'.
   14. Thus, the number of non-hits possessing features f_1 and f_2 is $F_1'F_2'/H'$.
   15. The odds of a sentence being a hit if it has features f_1 and f_2 are therefore $O(h|f_1,f_2)=(F_1F_2/H)/(F_1'F_2'/H')=F_1F_2H'/F_1'F_2'H$.
   16. O_1, the odds that a sentence with feature f_1 is a hit, is F_1/F_1', by the definition of odds. Similarly, $O_2=F_2/F_2'$. Similarly, O_0, the prior odds a sentence is a hit without considering its features, is H/H'. Therefore, $O(h|f_1,f_2)=F_1F_2H'/F_1'F_2'H=(F_1/F_1')(F_2/F_2')/(H/H')=O_1O_2/O_0$.
   17. To account for feature f_3, the same reasoning by which odds O_1, implied by feature f_1, are multiplied by the factor O_2/O_0 to account for feature f_2 is applied again. Thus O_1O_2/O_0 is multiplied by the factor O_3/O_0 to account for feature f_3. In symbols, $O(h|f_1,f_2)=O_1O_2/O_0$ is modified to $O(h|f_1,f_2,f_3)=O_1O_2O_3/(O_0)^2$.

18. The same process is repeated. Each feature f_i causes the odds due to previously considered features to be multiplied by the factor O_i/O_0, ultimately giving the formula $O(h|f_1,...,f_n)=O_1...O_n/(O_0)^{n-1}$ if there are n features considered.

Comparison of the Naïve Bayes and semi-naïve evidence combination models. Naïve Bayes is often used for category assignment. The item to classified is put into the category for which Naïve Bayes gives the highest likelihood. In the present context there are two categories, one of hits and one of non-hits, but in general there can be N categories. In either case, the denominator of the Naïve Bayes formula is the same for each category, so it can be ignored. This is fortunate because it is this denominator whose computation assumes that features are unconditionally (i.e. disregarding the category of the sample sentences) independent of one another. However, when the Naïve Bayes formula is used for estimating the probability that a sentence is in a particular category, the denominator must be evaluated. This is problematic because the assumption of unconditional independence is not only unsupported, but most likely wrong. The reason is that the presence of features that provide evidence that the sentence belongs in a particular category are probably correlated. For example, two features that both predict a sentence will be a hit do not occur independently of one another. Rather, if a sentence has one of them (and thus is likely to be a hit), then it will have an enhanced likelihood of possessing the other (because it is also particularly associated with hits).

For this problem of estimating the probability that a particular sentence is a hit or, more generally, belongs to a particular category, semi-naïve evidence combination appears more suitable because it estimates odds (which are easily converted to probabilities) without the problematic assumption that features are unconditionally independent in their occurrence.

Both methods require assuming that features occur independently given (i.e. within) the category. Semi-naïve evidence combination additionally requires assuming features are independent within the complement of the category. In the example, the category is sentences that are hits and the complement of the category is sentences that are not hits. However, if Naïve Bayes is to be used to determine assignment to the hits category or the non-hits category, then it will have to be applied separately to each of those categories, thus requiring the same assumptions as semi-naïve evidence combination. On the other hand, then the denominator of the Naïve Bayes formula would become superfluous, so that the two methods would be making the same independence assumptions.

For cases where there are multiple categories, semi-naïve evidence combination requires independence assumptions for the complement of each category, whereas Naïve Bayes does not. It may be, however, that the effect of assuming features are independent within the complement of a category will tend to counteract any inaccuracy resulting from assuming features are independent within the category. For some problems, the decision of which method to use may ultimately depend on for which formula the required figures are most easily obtained.

Exercise 1. Suppose there is a set of 8 sentences, 4 of which are hits and 4 of which are not. Feature 1 is present in all 4 hits and in 2 non-hits. Feature 2 also occurs in 4 hits and 2 non-hits. There is 1 non-hit with both features. What is the probability estimated by the Naïve Bayes formula that a sentence with both features is a hit? What are the odds for this estimated by the formula for semi-naïve evidence combination? What is the probability implied by these odds? What is the true probability? Repeat this process for the non-hit category. Discuss the results.

References

Lewis, D. *Naïve Bayes at forty: the independence assumption in information retrieval*. in *Conf. Proc. European Conference on Machine Learning*. 1998. Chemnitz, Germany.

Wikipedia, free on-line encyclopedia, Naïve Bayesian classification, http://en.wikipedia.org/wiki/Naive_Bayes.