

Why Chi?

Motivations for the Use of Fisher's Inverse Chi-Square Procedure in Spam Classification

Version .91
May 4, 2004

Gary Robinson
CEO, Transpose, LLC
grobinson@transpose.com

Introduction

This article will attempt to elucidate the theoretical underpinnings of the chi-square-based spam classification technique, and discuss its possibly useful features for spam classification. It assumes familiarity with Gary Robinson's article "Handling Redundancy in Email Token Probabilities" [Robinson] and will refer directly to equation numbers from that article.

The first part of the present article will focus on the basic Fisher calculation given in Equations 2 and 3 of [Robinson] for calculating S and H , our measures of "spamminess" and "hamminess". For simplicity of discussion, we will focus on using just one of those calculations for classification purposes. Combining them as is done in [Robinson] works better, but they can be used separately as well, and doing so is convenient for our discussion. The second part of the article will discuss the question of combining those indicators into an overall indicator for making a classification.

Basis for the Use of Fisher's Inverse Chi-Square Calculation

The key to the chi-square-based spam filter is the fact that Grahamian word probabilities, which represent the "spamminess" of a token, are (approximations to) true probabilities. This allows us to use the Fisher inverse chi-square calculation even though the conditions do not exist where that calculation would traditionally be used.

[Robinson] discusses a null hypothesis that says that the underlying distribution is "drawn from a population of random variables that is uniformly distributed and statistically independent." That article suggests that that is the basis for the use of the inverse chi-square method in its spam classification algorithm. However, that is really a convenient simplification. It is not without merit, but it doesn't engage directly enough with the actual underlying realities to explain the strong performance of the algorithm in testing.

Traditionally, the Fisher calculation is used in situations where the null hypothesis is something that could at least be imagined to exist in the physical world. That is, it there is a possibility that a situation exists such that a p-values can be associated with it. (A p-

value p associated with a value v , on a continuous distribution, says that the probability of a result more extreme than v is p .)

Now consider Grahamian token probabilities. In the real world, some token probabilities appear more often than others do. In fact, when one creates a histogram representing the token probabilities of all tokens extracted from a typical large collection of email, they are not remotely in a uniform distribution. If the token probabilities could be directly used as p-values, they would form a uniform distribution. And then, one could frame a null hypothesis which said essentially that the tokens in an email were in the same distribution as in the general population of tokens. An actual spam or a ham would have the different distribution that applied to spams or hams, resulting in extreme values as the output from the inverse chi-square calculation, giving us a confidence level with which we could reject that null hypothesis and assume instead that the email was a spam or a ham.

This would be a formulation which would be in line with traditional application of the Fisher calculation. However we can't use that formulation, because the token probabilities do not appear in a uniform distribution.

An alternative approach might be to derive a uniform distribution from the token probabilities. For instance, one could create a table from the real-world data in a large population of emails that would map any Grahamian token probability to the probability that a Grahamian probability even more extreme would be selected if a token were randomly chosen from the overall collection of tokens.

That would be a true p-value, derived from the Grahamian token probabilities. Indeed this is what is usually done when Fisher's inverse chi-square calculation is employed, but rather than using a table derived from real-world data, a cumulative distribution function based on a known underlying distribution is used. For instance, if the distribution underlying the null hypothesis is normal, one does not have to build a table because p-values can be mathematically generated that are the same as the ones that would be generated by such a table. (That can't be done with the token probabilities because they don't form a simple, known distribution with a known cumulative distribution function.)

Now, consider this table-driven p-value based on the token probabilities. A null hypothesis could be stated that the tokens in the test email were as if those tokens were chosen randomly from the entire population of tokens. A p-value would be associated with each token, and these p-values could be fed into the Fisher calculation, and we'd arrive at a p-value representing the confidence with which we can reject the null hypothesis, and all would be well... or would it?

The answer is no, it wouldn't. The reason is that the original Grahamian probabilities, which are themselves very meaningful, would be completely lost, replaced by the table-driven p-values. To see why this is a problem, consider the following example.

First we note that there is no reason to assume that there are equal number of spammy vs. hammy words. For example, even today, many people whose email addresses are not

publicly available on the Web receive hams than spams. A moment's reflection will show that in such inboxes, a randomly chosen word is probably going to be more associated with hams than spams.

If we did the table-driven transformation outlined above, a Grahamian token probability of .5, which actually represents neither spammy nor hammy tendencies, would be more associated with spams than the average email for that user. Thus the table-driven p-value would be greater than .5 (assuming the table was laid out so that spammier words were assigned the higher positions).

If we consider an email that contains nothing but neutral tokens, the Fisher calculation would be combing a number of tokens greater than .5, and would therefore tend to output a combined p-value that is close to 1, and closer to 1 as similarly neutral emails were considered that happened to have more tokens.

This would make it very difficult to choose a ham/spam cutoff point based on the output from the calculation such that the cutoff point has a similar meaning for emails of different sizes, making the Fisher calculation not very useful for spam classification.

Note that this problem does not occur in traditional use of the Fisher calculation because if, for example, the null hypothesis involves an underlying normal distribution, a p-value of .5 corresponds to the middle bell-shaped curve and truly is neutral.

So, we want to combine Grahamian token probabilities, but we are not aware of a way to start with a reasonable null hypothesis that represents the actual, "physical" world of emails and arrive at a p-value that is useful to us. So, if we used the traditional determinants of whether the inverse chi-square method can be applied, we would not use it.

However, if we broaden our understanding of the real meaning of the Fisher calculation a bit, we have the opportunity to arrive at a different conclusion. In fact we will see that it makes sense to feed the Grahamian token probabilities directly into the calculation -- and this will also provide a framework for understanding the strong performance of chi-square-based classifiers to date.

While Grahamian token probabilities aren't p-values in the real world due to the non-uniform distribution of those probabilities, they do have a very concrete probabilistic meaning. If one picks a word w and randomly chooses one of the emails containing that word, the Grahamian token probability $f(w)$ is the probability that that email will be a spam. (Note, we're assuming "idealized" Grahamian token probabilities, which strictly represent the underlying actual probabilities in association with each token, which could only be calculated if we had an infinite amount of data. But for practical purposes the $f(w)$'s we calculate from real-world collections of email have been found to be close enough to "work" for our purposes.)

This concrete meaning of $f(w)$ has interesting properties. If a token has $f(w) = .9$, it is not very associated with spams; it only appears in a ham 10% of the time it appears. If $f(w)$ is

.95, it is half as frequently associated with hams. In a very real sense, this second token is half as "good" as the first. That is, all other things being equal, it's associated with half as many good emails. So there is a concept of "goodness" that is directly represented by $f(w)$ and that has a very concrete meaning to users who care about the amount of spam they receive. (For readers who are noting that the .9 vs. .95 duality does not represent a halving of goodness from the perspective of doubling the number of associated hams, recall that the Fisher test is one-sided; we only care about the values near 0. But we do care about the values near 1 too, which is why we calculate $f(w)$ separately from the spammy and hammy perspective, resulting in S and H as discussed in [Robinson].)

Now suppose we randomly choose an amount of goodness, according to a uniform distribution. Then given a goodness value g , the probability of randomly choosing one more extreme is g . In other words, the g 's are p-values in that sense. Note also that .5 is a neutral goodness, as it should be.

Note also that goodness is the probability of "an event of interest" occurring. This has a very concrete meaning in our context, but perhaps unexpectedly, it also has one in traditional use of Fisher's calculation.

Suppose we decide that obtaining a result as extreme or more extreme than the one associated with goodness g is "an event of interest." Then the p-values used in traditional applications of the inverse chi-square method would be goodnesses according to our use of that term.

The underlying principle in both cases is that the goodness g describes the probability of "an event of interest." The cases are different in that the traditional case maps to a p-value and the usage we're describing here doesn't. But the underlying principle and meaning is the same: we're talking about "an event of interest".

It is important to discuss one more difference between the traditional use of the inverse chi-square calculation and the one we are promoting here. In the traditional use, if the result of the inverse chi-square calculation is P , then P is a p-value. That is, let us consider the product G of all the g 's associated with the random variables we are combining. Let us calculate the inverse chi-square result with respect to G . The result is P . Then the following is true if the null hypothesis is true: the probability of a product of g 's smaller than G is P .

In our case, that is not true. That is because our g 's are not p-values in the physical world. Instead, they are "merely" probabilities with respect to "an event of interest". For the purposes we really care about, that is what a p-value is too; but it happens that p-values also have the characteristic, friendly to human intuition, that they cause the inverse chi-square calculation to result in something that is also p-value. That makes it very easy for people to associate a meaning with the inverse chi-square calculation, resulting in its popular usage in contexts where the inputs are p-values.

But, for decisionmaking purposes, it does not actually matter whether or not one can associate that intuitively-friendly meaning, because the deeper probabilistic meaning is the same either way.

That is, if one considers P to represent an amount of evidence, then one can say the following. P stands for amount of evidence that, if the probabilities represented by the g 's happen to be p-values, would result in the probability of a product of g 's smaller than G being P . Whether or not the g 's happen to be p-values in a given case does not change the meaning of P as a measure of evidence; such a case only enables us to associate it with a very intuitively friendly interpretation. Its utility for decisionmaking purposes is unchanged because the deeper probabilistic meaning relating to the probability of "an event of interest" is unchanged.

As another way of looking at this, we can formulate an appropriate null hypothesis in an imaginary world. Imagine a world which is governed by a strange law of nature. Spam emails have $f(w)$'s that are in a uniform distribution, whereas, ham emails have the same distribution they have in the real world; in particular, the distribution of $f(w)$'s in ham emails tends to be skewed toward 0.

In such a world, we could generate a spam filter based on a null hypothesis that the email we're testing is a spam. If the $f(w)$'s in that email have enough of a tendency to be near 0, we can reject that null hypothesis in favor of the alternative hypothesis that the email is a ham. We can use the inverse chi-square calculation to generate a p-value which represents the confidence with which we can reject the null hypothesis.

Perhaps we find that a p-value of .01 is the cut-off point at which we are comfortable rejecting the null hypothesis and accepting the email as ham.

No statistician living in the imaginary world would have a problem with this use of the inverse chi-square calculation. The null hypothesis would represent a realistic possibility in that world, and it would make sense to reject it if the p-value was close enough to 0.

Now, suppose we move to another world, our actual one, where Grahamian token probabilities are not in a uniform distribution in either spams or in hams.

Now, the fact is that this classifier will perform even better in the real world than in the imaginary one, due to the fact that the $f(w)$'s for spams tend towards 1 in the real world. So the output of the inverse chi-square calculation will have a value of less than .01 with less frequency than in the imaginary one. So the classifier will make fewer errors.

(Note: recall that in the discussion here we are only talking about the instance of the inverse chi-square calculation that leads to H , which is Eq. 2 in [Robinson]. We are using H alone for spam classification. When we apply this argument to S , which is the spam-sensitive counterpart to H , we invoke a different imaginary world where it's the hams that have their $f(w)$'s in a uniform distribution.)

According to the traditional use of the inverse chi-square calculation, that calculation would not be applicable to such a situation because there is no obvious null hypothesis that makes sense in the real world and under which the input values are p-values. In the traditional use of the calculation, we would want the output to be a p-value, and that is only true when the inputs are p-values.

However, for our classification purposes, we don't care whether or not the output is a p-value. We only care about whether it helps us classify emails. And it does -- even better than in the imaginary world where the use of the calculation is consistent with tradition.

Clearly, it makes little sense to require p-values as inputs to the calculation, when in the real world, which doesn't have them in our example, the calculation works even better for our purposes than in the imaginary world, which does.

In the real world, the useful information is encoded in the $f(w)$'s associated with the tokens in our database. Those represent real probabilities of interest to us. Whether or not the distribution of these $f(w)$'s allows us to construct a null hypothesis to test that corresponds to an actual physical possibility does not help us in our classification task. It just isn't relevant.

So, we want a way to broaden the prerequisites for using the inverse chi-square calculation when we want to use it for classification. When we really need it to output a p-value (for instance because we want to use it to test whether a new medical treatment is statistically significant), then we need to require the inputs to be p-values. But when we are using it for classification purposes, the requirements are less strict.

One way of expressing that broadening is to say that the Fisher inverse chi-square calculation is a way of combining evidence where that evidence takes the form the occurrence of events of interest which have known probabilities.

It may be interesting to note that we no longer rely on a null hypothesis at all for our classification purposes. The null hypothesis is useful for the traditional use of Fisher because p-values are calculated relative to that null hypothesis. Without p-values, there is no need for a null hypothesis to disprove. Instead, we are combining evidence that is encoded in real-world probabilities.

The assertions made above could and should be studied in more depth and rigor. Our purpose here is not to provide the ultimate guide to when and where to use Fisher's inverse chi-square calculation, but rather to discuss the theoretical basis for its current use in spam classification.

Note also that in some applications it may be convenient to use $L = -\log P$, which is a measure of information. It would be an information-theoretic measure of the amount of evidence at hand.

Now that we have discussed the theoretical basis for using the inverse chi-square calculation for spam filtering, the next question might be to wonder why it would be considered likely to have particularly good performance for that purpose.

First of all, it's a one-sided test, which has certain advantages which will be described in the next section of this article.

Secondly, there is an optimality theorem [Littell & Folks] which we think probably applies for our use, although that has yet to be proven. The theorem essentially says that the Fisher calculation is optimal among all combining procedures for p-values, given certain weak conditions. We think that the fact that we don't have p-values shouldn't change the result, if we formulate the conditions carefully. For instance we have already seen by example that if we imagine a universe in which Grahamian token probabilities are p-values, for classification purposes we can get the same results if we move the calculation to a universe where they are not p-values. Conditions will apply, of course, and those have not been rigorously defined here. But it seems likely to us that the results will translate.

Other methodologies for spam/ham classification also have optimality theorems associated with them, so the existence of one for our purposes would not, in any case, be a deciding factor. Rather the possible optimality would serve to make the procedure worthy of consideration; beyond that we should consider the implications of the one-sided nature of the test, as will be discussed below.

Of course, the final arbiter should be rigorous comparative testing.

[A personal note from the author: I am not sure whether this broadened concept of when to use Fisher's inverse chi-square calculation may help people notice other applications where it could be used, but for which it is not now being used. I have read fairly substantially in the literature about this calculation and I have not noticed this aspect explicitly formulated previously. In addition, people who are experienced in statistics have frequently had a problem with our use of the inverse chi-square calculation because of the fact that there isn't a null hypothesis that 1) is completely imaginable as a circumstance in actual physical world and 2) cleanly maps to our usage. So it seems that it could be difficult for people having that mindset to envision a use like the current one, if they were presented with a problem where it would nevertheless be an appropriate solution. Therefore, I think it is possible that this conscious broadening of the requirements for using the inverse chi-square might be generally useful. I think a more rigorous mathematical formulation would be a good thing, but at the time of this writing I don't have one. However I think that the intuitive understanding of this concept is enough for many practical purposes.]

Advantages of One-sidedness

There are two known advantages to using a one-sided test, when the two sides are combined together as in [Robinson]. There is no reason to assume that there aren't other approaches with one or both of these advantages. But the author of this article is unaware of them, and they aren't in widespread use in the antispam community. As other techniques emerge into common knowledge that have these advantages, they will hopefully be tested against the present technique for reliability in spam/ham classification.

Here are the two advantages:

1) As Fisher is used in [Robinson] the evidence that the email is spam and the evidence that the email is ham are treated separately and only combined at the end. The fact that Fisher is a one-sided test is what makes it very easy to do this. In that paper, these values are represented by S and H for ham and spam.

As mentioned in [Robinson] Tim Peters of SpamBayes pointed out that when both S and H are extreme, that means the data is "confused". There is a lot of evidence pointing in both directions. Even though there may be a couple of orders of magnitude difference between S and H , it has been found in practical use that if both values are "extreme", no reliable classification can be done.

So, the present technique, due to its separate handling of the evidence in favor of hamminess and in favor of spamminess, enables us to sense the existence of this state of confusion where both S and H are extreme, and classify such emails into a middle-range "Unsure" classification. In other known classification schemes such as naïve Bayes, the tendencies toward spamminess and hamminess are considered together, so that when there is a several-order-of-magnitude difference in one compared to the other, one dominates, even though both are extreme. So breaking the two values out as we do is helpful.

2) This second advantage is also due to the separate handling of S and H . [Robinson] describes an approach for taking advantage of the different levels of redundancy in token probability data in spam vs. hams. Due to the one-sided nature of the Fisher test, we are able to treat these different redundancy levels differently in each case and use it for greater classification accuracy. Again, we do not know of competitive approaches that allow for this. Obviously they may exist; but they are not in common use in the antispam community. If/when other such approaches emerge, comparison tests will hopefully be run.

Summary

The purpose of this article has been to discuss the reasons the inverse chi-square test defined by R.A. Fisher has been found to be useful in the task of spam detection. Further research opportunities include:

- 1) providing more rigorous "broadened" guidelines for when and when not to use that test,
- 2) exploring the applicability of the [Littell & Folks] theorem,
- 3) conducting more comparative testing involving competitive approaches and other ways of combining S and H , and
- 4) considering possible applications beyond spam classification.

[Robinson] Robinson, Gary (2004). Handling Redundancy in Email Token Probabilities. http://www.garyrobinson.net/2004/04/improved_chi.html.

[Littell & Folks] Littell, R. C., & Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association*, **68**, 193-194.