classification

Christopher D. Manning

Prabhakar Raghavan

Hinrich Schütze

search

precision

crawler

links

spam

Introduction to Information Retrieval

recall

query

svm

clustering

index

web

xml

language model

CAMBRIDGE

ranking

CAMBRIDGE

www.cambridge.org/9780521865715

```
SelectFeatures($\mathbb{D}, c, k$)
$V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
$L \leftarrow []$
for each $t \in V$
do $A(t, c) \leftarrow$ ComputeFeatureUtility($\mathbb{D}, t, c$)
    Append($L, \langle A(t, c), t \rangle$)
return FeaturesWithLargestValues($L, k$)
```

Figure 13.6 Basic feature selection algorithm for selecting the k best features.

Exercise 13.4 Table 13.3 gives Bernoulli and multinomial estimates for the word the. Explain the difference.

13.5 Feature selection

FEATURE SELECTION

NOISE FEATURE

OVERFITTING

Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers that, unlike NB, are expensive to train. Second, feature selection often increases classification accuracy by eliminating noise features. A *noise feature* is one that, when added to the document representation, increases the classification error on new data. Suppose a rare term, say arachnocentric, has no information about a class, say *China*, but all instances of arachnocentric happen to occur in *China* documents in our training set. Then the learning method might produce a classifier that misassigns test documents containing arachnocentric to *China*. Such an incorrect generalization from an accidental property of the training set is called *overfitting*.

We can view feature selection as a method for replacing a complex classifier (using all features) with a simpler one (using a subset of the features). It may appear counterintuitive at first that a seemingly weaker classifier is advantageous in statistical text classification, but when discussing the bias-variance tradeoff in Section 14.6 (page 284), we will see that weaker models are often preferable when limited training data are available.

The basic feature selection algorithm is shown in Figure 13.6. For a given class c, we compute a utility measure $A(t, c)$ for each term of the vocabulary and select the k terms that have the highest values of $A(t, c)$. All other terms are discarded and not used in classification. We will introduce three different utility measures in this section: mutual information, $A(t, c) = I(U_t; C_c)$; the χ^2 test, $A(t, c) = X^2(t, c)$; and frequency, $A(t, c) = N(t, c)$.

Of the two NB models, the Bernoulli model is particularly sensitive to noise features. A Bernoulli NB classifier requires some form of feature selection or else its accuracy will be low.

This section mainly addresses feature selection for two-class classification tasks like *China* versus *not-China*. Section 13.5.5 briefly discusses optimizations for systems with more than two classes.

13.5.1 Mutual information

MUTUAL INFORMATION

A common feature selection method is to compute $A(t, c)$ as the expected *mutual information* (MI) of term t and class c.[5] MI measures how much information the presence/absence of a term contributes to making the correct classification decision on c. Formally:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}, \tag{13.16}$$

where U is a random variable that takes values $e_t = 1$ (the document contains term t) and $e_t = 0$ (the document does not contain t), as defined on page 246, and C is a random variable that takes values $e_c = 1$ (the document is in class c) and $e_c = 0$ (the document is not in class c). We write U_t and C_c if it is not clear from context which term t and class c we are referring to.

For MLEs of the probabilities, Equation (13.16) is equivalent to Equation (13.17):

$$\begin{aligned} I(U;C) = {} & \frac{N_{11}}{N} \log_2 \frac{N N_{11}}{N_{1.} N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{N N_{01}}{N_{0.} N_{.1}} \\ & + \frac{N_{10}}{N} \log_2 \frac{N N_{10}}{N_{1.} N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{N N_{00}}{N_{0.} N_{.0}} \end{aligned} \tag{13.17}$$

where the Ns are counts of documents that have the values of e_t and e_c that are indicated by the two subscripts. For example, N_{10} is the number of documents that contain t ($e_t = 1$) and are not in c ($e_c = 0$). $N_{1.} = N_{10} + N_{11}$ is the number of documents that contain t ($e_t = 1$) and we count documents independent of class membership ($e_c \in \{0, 1\}$). $N = N_{00} + N_{01} + N_{10} + N_{11}$ is the total number of documents. An example of one of the MLE estimates that transform Equation (13.16) into Equation (13.17) is $P(U = 1, C = 1) = N_{11}/N$.

Example 13.3: Consider the class *poultry* and the term export in Reuters-RCV1. The counts of the number of documents with the four possible combinations of indicator values are as follows:

	$e_c = e_{\text{poultry}} = 1$	$e_c = e_{\text{poultry}} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 141$
$e_t = e_{\text{export}} = 0$	$N_{01} = 27{,}652$	$N_{00} = 774{,}106$

[5] Take care not to confuse expected mutual information with *pointwise mutual information*, which is defined as $\log N_{11}/E_{11}$ where N_{11} and E_{11} are defined as in Equation (13.17). The two measures have different properties. See Section 13.7.

After plugging these values into Equation (13.17) we get:

$$\begin{aligned} I(U;C) = &\frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49{+}27{,}652)(49{+}141)}\\ &+\frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141{+}774{,}106)(49{+}141)}\\ &+\frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49{+}27{,}652)(27{,}652{+}774{,}106)}\\ &+\frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141{+}774{,}106)(27{,}652{+}774{,}106)}\\ \approx\ &0.000105 \end{aligned}$$

To select k terms $t_1, \ldots, t_k$ for a given class, we use the feature selection algorithm in Figure 13.6: We compute the utility measure as $A(t, c) = I(U_t, C_c)$ and select the k terms with the largest values.

Mutual information measures how much information – in the information-theoretic sense – a term contains about the class. If a term's distribution is the same in the class as it is in the collection as a whole, then $I(U;C) = 0$. MI reaches its maximum value if the term is a perfect indicator for class membership, that is, if the term is present in a document if and only if the document is in the class.

Figure 13.7 shows terms with high mutual information scores for the six classes in Figure 13.1.[6] The selected terms (e.g., london, uk, british for the class *UK*) are of obvious utility for making classification decisions for their respective classes. At the bottom of the list for *UK* we find terms like peripherals and tonight (not shown in the figure) that are clearly not helpful in deciding whether the document is in the class. As you might expect, keeping the informative terms and eliminating the non-informative ones tends to reduce noise and improve the classifier's accuracy.

Such an accuracy increase can be observed in Figure 13.8, which shows F_1 as a function of vocabulary size after feature selection for Reuters-RCV1.[7] Comparing F_1 at 132,776 features (corresponding to selection of all features) and at 10–100 features, we see that MI feature selection increases F_1 by about 0.1 for the multinomial model and by more than 0.2 for the Bernoulli model. For the Bernoulli model, F_1 peaks early, at ten features selected. At that point, the Bernoulli model is better than the multinomial model. When basing a classification decision on only a few features, it is more robust to consider binary occurrence only. For the multinomial model (MI feature selection), the peak occurs later, at 100 features, and its effectiveness recovers somewhat at the end when we use all features. The reason is that the multinomial takes

[6] Feature scores were computed on the first 100,000 documents, except for *poultry*, a rare class, for which 800,000 documents were used. We have omitted numbers and other special words from the top ten lists.

[7] We trained the classifiers on the first 100,000 documents and computed F_1 on the next 100,000. The graphs are averages over five classes.

UK

london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China

china	0.0997
chinese	0.0523
beiing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

elections

election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

sports

soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

Figure 13.7 Features with high mutual information scores for six Reuters-RCV1 classes.

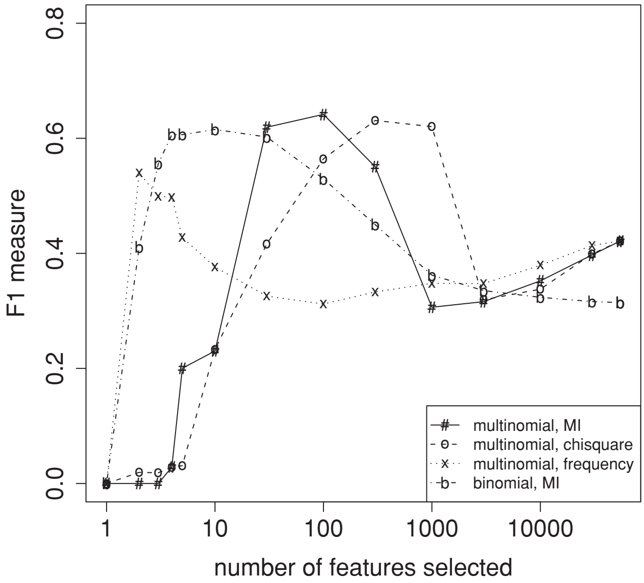

Figure 13.8 Effect of feature set size on accuracy for multinomial and Bernoulli models.

the number of occurrences into account in parameter estimation and classification and therefore better exploits a larger number of features than the Bernoulli model. Regardless of the differences between the two methods, using a carefully selected subset of the features results in better effectiveness than using all features.

13.5.2 χ^2 Feature selection

χ^2 FEATURE SELECTION

INDEPENDENCE

Another popular feature selection method is χ^2. In statistics, the χ^2 test is applied to test the independence of two events, where two events A and B are defined to be *independent* if $P(AB) = P(A)P(B)$ or, equivalently, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In feature selection, the two events are occurrence of the term and occurrence of the class. We then rank terms with respect to the following quantity:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_te_c} - E_{e_te_c})^2}{E_{e_te_c}} \tag{13.18}$$

where e_t and e_c are defined as in Equation (13.16). N is the *observed* frequency in $\mathbb{D}$ and E the *expected* frequency. For example, E_{11} is the expected frequency of t and c occurring together in a document assuming that term and class are independent.

Example 13.4: We first compute E_{11} for the data in Example 13.3:

$$\begin{aligned} E_{11} &= N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N} \\ &= N \times \frac{49 + 141}{N} \times \frac{49 + 27652}{N} \approx 6.6 \end{aligned}$$

where N is the total number of documents as before.

We compute the other $E_{e_te_c}$ in the same way:

	$e_{poultry} = 1$		$e_{poultry} = 0$	
$e_{\textsf{export}} = 1$	$N_{11} = 49$	$E_{11} \approx 6.6$	$N_{10} = 141$	$E_{10} \approx 183.4$
$e_{\textsf{export}} = 0$	$N_{01} = 27{,}652$	$E_{01} \approx 27{,}694.4$	$N_{00} = 774{,}106$	$E_{00} \approx 774{,}063.6$

Plugging these values into Equation (13.18), we get a X^2 value of 284:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_te_c} - E_{e_te_c})^2}{E_{e_te_c}} \approx 284$$

X^2 is a measure of how much expected counts E and observed counts N deviate from each other. A high value of X^2 indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect. In our example, $X^2 \approx 284 > 10.83$. Based on Table 13.6, we can reject the hypothesis that *poultry* and export are independent with

Table 13.6 Critical values of the χ^2 distribution with one degree of freedom. For example, if the two events are independent, then $P(X^2 > 6.63) < 0.01$. So for $X^2 > 6.63$ the assumption of independence can be rejected with 99% confidence.

p	χ^2 critical value
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

only a 0.001 chance of being wrong.[8] Equivalently, we say that the outcome $X^2 \approx 284 > 10.83$ is *statistically significant* at the 0.001 level. If the two events are dependent, then the occurrence of the term makes the occurrence of the class more likely (or less likely), so it should be helpful as a feature. This is the rationale of χ^2 feature selection.

STATISTICAL SIGNIFICANCE

An arithmetically simpler way of computing X^2 is the following:

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})} \quad (13.19)$$

This is equivalent to Equation (13.18) (Exercise 13.14).

Assessing χ^2 as a feature selection method

From a statistical point of view, χ^2 feature selection is problematic. For a test with one degree of freedom, the so-called Yates correction should be used (see Section 13.7), which makes it harder to reach statistical significance. Also, whenever a statistical test is used multiple times, then the probability of getting at least one error increases. If 1,000 hypotheses are rejected, each with 0.05 error probability, then $0.05 \times 1000 = 50$ calls of the test will be wrong on average. However, in text classification it rarely matters whether a few additional terms are added to the feature set or removed from it. Rather, the *relative* importance of features is important. As long as χ^2 feature selection only ranks features with respect to their usefulness and is not used to make statements about statistical dependence or independence of variables, we need not be overly concerned that it does not adhere strictly to statistical theory.

[8] We can make this inference because, if the two events are independent, then $X^2 \sim \chi^2$, where χ^2 is the χ^2 distribution. See, for example, Rice (2006).

13.5.3 Frequency-based feature selection

A third feature selection method is *frequency-based feature selection*, that is, selecting the terms that are most common in the class. Frequency can be either defined as document frequency (the number of documents in the class c that contain the term t) or as collection frequency (the number of tokens of t that occur in documents in c). Document frequency is more appropriate for the Bernoulli model, collection frequency for the multinomial model.

Frequency-based feature selection selects some frequent terms that have no specific information about the class, for example, the days of the week (Monday, Tuesday, . . .), which are frequent across classes in newswire text. When many thousands of features are selected, then frequency-based feature selection often does well. Thus, if somewhat suboptimal accuracy is acceptable, then frequency-based feature selection can be a good alternative to more complex methods. However, Figure 13.8 is a case where frequency-based feature selection performs a lot worse than MI and χ^2 and should not be used.

13.5.4 Feature selection for multiple classifiers

In an operational system with a large number of classifiers, it is desirable to select a single set of features instead of a different one for each classifier. One way of doing this is to compute the X^2 statistic for an $n \times 2$ table where the columns are occurrence and nonoccurrence of the term and each row corresponds to one of the classes. We can then select the k terms with the highest X^2 statistic as before.

More commonly, feature selection statistics are first computed separately for each class on the two-class classification task c versus $\overline{c}$ and then combined. One combination method computes a single figure of merit for each feature, for example, by averaging the values $A(t, c)$ for feature t, and then selects the k features with highest figures of merit. Another frequently used combination method selects the top k/n features for each of n classifiers and then combines these n sets into one global feature set.

Classification accuracy often decreases when selecting k common features for a system with n classifiers as opposed to n different sets of size k. But even if it does, the gain in efficiency owing to a common document representation may be worth the loss in accuracy.

13.5.5 Comparison of feature selection methods

Mutual information and χ^2 represent rather different feature selection methods. The independence of term t and class c can sometimes be rejected with high confidence even if t carries little information about membership of a document in c. This is particularly true for rare terms. If a term occurs once in a large collection and that one occurrence is in the *poultry* class, then this is statistically significant. But a single occurrence is not very informative according to the information-theoretic definition of information. Because its criterion is significance, χ^2 selects more rare terms (which are often less reliable indicators) than mutual information. But the selection criterion of mutual information also does not necessarily select the terms that maximize classification accuracy.

Despite the differences between the two methods, the classification accuracy of feature sets selected with χ^2 and MI does not seem to differ systematically. In most text classification problems, there are a few strong indicators and many weak indicators. As long as all strong indicators and a large number of weak indicators are selected, accuracy is expected to be good. Both methods do this.

Figure 13.8 compares MI and χ^2 feature selection for the multinomial model. Peak effectiveness is virtually the same for both methods. χ^2 reaches this peak later, at 300 features, probably because the rare, but highly significant features it selects initially do not cover all documents in the class. However, features selected later (in the range of 100–300) are of better quality than those selected by MI.

GREEDY FEATURE SELECTION

All three methods – MI, χ^2 and frequency based – are *greedy* methods. They may select features that contribute no incremental information over previously selected features. In Figure 13.7, `kong` is selected as the seventh term even though it is highly correlated with previously selected `hong` and therefore redundant. Although such redundancy can negatively impact accuracy, non-greedy methods (see Section 13.7 for references) are rarely used in text classification due to their computational cost.

Exercise 13.5 Consider the following frequencies for the class *coffee* for four terms in the first 100,000 documents of Reuters-RCV1:

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

Select two of these four terms based on (i) χ^2, (ii) mutual information, (iii) frequency.

13.6 Evaluation of text classification

Historically, the classic Reuters-21578 collection was the main benchmark for text classification evaluation. This is a collection of 21,578 newswire articles, originally collected and labeled by Carnegie Group, Inc. and Reuters,