Deriving TF-IDF as a Fisher Kernel

Charles Elkan

Department of Computer Science and Engineering,
University of California, San Diego
elkan@cs.ucsd.edu
http://www.cs.ucsd.edu/users/elkan/

Abstract. The Dirichlet compound multinomial (DCM) distribution has recently been shown to be a good model for documents because it captures the phenomenon of word burstiness, unlike standard models such as the multinomial distribution. This paper investigates the DCM Fisher kernel, a function for comparing documents derived from the DCM. We show that the DCM Fisher kernel has components that are similar to the term frequency (TF) and inverse document frequency (IDF) factors of the standard TF-IDF method for representing documents. Experiments show that the DCM Fisher kernel performs better than alternative kernels for nearest-neighbor document classification, but that the TF-IDF representation still performs best.

1 Introduction

A fundamental property of text documents, regardless of language, is that if a word occurs once, it is likely that the same word will occur again. This phenomenon is called burstiness [3]. Unfortunately, standard probabilistic models for documents, in particular multinomial distributions, do not allow for burstiness, since they assume that each word in a document is generated independently. These models are therefore incorrect in a fundamental way. In recent research, an alternative distribution has been proposed called the Dirichlet compound multinomial (DCM) [7]. This distribution can capture the phenomenon of burstiness. Experimentally, DCM models lead to significantly better classification accuracy than multinomial models on standard document collections [7].

In this paper, we derive the Fisher kernel for the DCM distribution. A Fisher kernel is a function that measures the similarity of two data items not in isolation, but rather in the context provided by a probability distribution. For documents, a Fisher kernel measures how much two members of a collection are similar taking into account a whole corpus as background information. We show that the Fisher kernel based on the DCM has a mathematical form related to the well-known TF-IDF representation for documents [1]. This demonstration is a new approach towards explaining why the TF-IDF heuristic is justified and why it is so successful experimentally. We provide experimental results for nearest neighbor classification for seven different kernel functions, that is for

M. Consens and G. Navarro (Eds.): SPIRE 2005, LNCS 3772, pp. 296–301, 2005.

seven different document representations: TF-IDF, the DCM Fisher kernel, the multinomial Fisher kernel, the Bhattacharyya kernel [6], and L_0, L_1, and L_2 normalized representations.

2 The Dirichlet Compound Multinomial Distribution

Throughout this paper, we assume the so-called "bag of words" representation for documents. In this representation, a document x is a vector of counts $\langle x_1, \ldots, x_w, \ldots, x_W \rangle$ where x_w is the number of appearances of word w and W is the vocabulary size. The DCM distribution is

$$p(x) = \frac{n!}{\prod_{w=1}^{W} x_w!} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w=1}^{W} \frac{\Gamma(x_w + \alpha_w)}{\Gamma(\alpha_w)}$$

where the length of the document is $n = \sum_w x_w$ and $s = \sum_w \alpha_w$ is the sum of the DCM parameters [2] [7]. Like a multinomial, formally a DCM is a distribution over alternative count vectors of the same length n. Since different lengths give rise to different distributions, but a corpus always contains documents of different lengths, we assume that a corpus is modeled by a family of DCMs that all have the same parameter values α_w.

Given a set D of documents, there is no closed-form expression for the maximum likelihood α_w parameter values. However, these can be approximated closely as

$$\alpha_w = \frac{\sum_{d \in D} I(x_{dw} \geq 1)}{\sum_{d \in D} \Psi(s + n_d) - \Psi(s)}$$

where x_{dw} is the count of word w in document d, n_d is the length of document d, and $\Psi(\cdot)$ is the digamma function (proof to be published elsewhere).

For typical document sets s and n_d are both in the hundreds, so $\Psi(s+n_d) - \Psi(s)$ is around one for each document, and therefore

$$\alpha_w \approx \frac{1}{|D|} \sum_{d \in D} I(x_{dw} \geq 1) \tag{1}$$

where $|D|$ is the number of documents in the collection. Since $x_{dw} = 0$ for most documents d and most words w, $\alpha_w \ll 1$ for most words. For example, for a DCM trained on one class of newsgroup articles, the average α_w is 0.004. Of the 59,826 parameters, 99% are below 0.1, only 17 are above 0.5, and only 5 are above 1.0.

3 The Fisher Kernel for the DCM

In general, a kernel function $k(x, y)$ is a way of measuring the resemblance between two data items x and y. Standard kernel functions are scalar products in some space of alternative representations of data items, that is $k(x, y) = s(x) \cdot s(y)$ where $s(x)$ is a re-representation of x.

For documents, one common approach is to re-represent a count vector by L_2 normalization: $s(x) = x/||x||_2$ where $||x||_2 = \sqrt{\sum_w x_w^2}$. This yields what is called cosine similarity, since $k(x, y) = x \cdot y/||x||_2||y||_2$ is the cosine of the angle between the vectors x and y. Intuitively, this re-representation is unsatisfying for at least two reasons: (a) repeated appearances of one word in the same document are of decreasing informativeness–a consequence of the burstiness phenomenon, and (b) words that appear across a large number of different documents are less informative. TF-IDF (term frequency-inverse document frequency) representations were proposed to address these concerns several decades ago [1] [9]. Most commonly, each term frequency x_w (i.e. each word count) is (a) log-transformed and (b) multiplied by the log of the inverse of the number of documents that word w appears in. This specific version of TF-IDF is

$$\text{TF-IDF}(x_w) = \log(x_w + 1) \cdot \log \frac{|D|}{\sum_{d \in D} I(x_{dw} > 0)}.$$

Typically (and in our experiments below) the TF-IDF representation is then L_2 normalized.

The normalized TF-IDF representation and the corresponding kernel are among the best approaches for retrieving documents relevant to a query, and for categorizing documents into classes. However, no compelling theoretical reason for preferring TF-IDF to other heuristic representations is known [9]. Here, we show that a representation similar to TF-IDF arises naturally from the DCM.

A high-level motivation for TF-IDF is that it incorporates knowledge about the distribution of all documents into the similarity measure for individual documents. Given a probability distribution $p(x)$, the Fisher kernel measures the similarity of x and y in the context of this distribution: $k(x, y) = s(x)^T H s(y)$ where $s(x) = \nabla x$ is the Fisher score vector for x, i.e. the vector of partial derivatives of the log-likelihood $l(x) = \log p(x)$ with respect to the parameters α_w, and H is the Hessian of second partial derivatives of $l(x)$ with respect to the parameters [4] [5]. With this definition, $k(x, y)$ is invariant to changes in the parameterization of p. However, H is usually approximated by the identity matrix, and in this case the Fisher kernel is different for different parameterizations.

For the DCM, the partial derivative of the log-likelihood is

$$\frac{\partial l(x)}{\partial \alpha_w} = \Psi(s) - \Psi(s + n) + \Psi(x_w + \alpha_w) - \Psi(\alpha_w). \tag{2}$$

The Fisher kernel $k(x, y)$ is then the scalar product of the partial derivative vectors for x and y.

Asymptotic values for the digamma function give insight into these score vectors. For $z \geq 1$, $\Psi(z)$ is close to $\log(z - 0.5)$ with the difference tending to zero as z tends to infinity. Similarly, $\Psi(z)$ is close to $-1/z + \Psi(1)$ for $z \ll 1$, where $\Psi(1) \approx -0.577$, with the difference tending to zero as z tends to zero from above. As mentioned in Section 2, for a typical corpus $\alpha_w \ll 1$ for most words w. Therefore, Equation (2) can be approximated as

$$\frac{\partial l(x)}{\partial \alpha_w} \approx \Psi(s) - \Psi(s + n) + I(x_w \geq 1)[\log(x_w - 0.5) + 1/\alpha_w - \Psi(1)].$$

In this form the Fisher score is clearly related to TF-IDF. First, given a document, the term $\Psi(s) - \Psi(s+n)$ is the same for all words w and as explained in Section 2, it is typically around minus one for all documents. Therefore, it has little influence on the ranking of which documents y are closest to a document x, i.e. which y give the smallest $k(x, y)$ values. Second, the term $\log(x_w - 0.5)$ is a log transform of term frequency. Finally, Equation (1) says that $1/\alpha_w \approx |D| / \sum_{d \in D} I(x_{dw} \geq 1)$ which is precisely inverse document frequency.

4 Experiments

In this section, we examine the performance of nine methods for document classification. Two methods are Bayesian classifiers based on training multinomial and DCM models. Seven methods are k-nearest neighbor classifiers. Of these, three use different L_p normalizations of documents $s(x) = x/||x||_p$ for $p = 0, 1, 2$. One nearest neighbor method uses the TF-IDF representation with L_2 normalization. Finally, three nearest neighbor methods use theoretically motivated kernels: the Fisher DCM kernel and two that are representative of those proposed in other recent research. The Bhattacharyya kernel uses the representation $s(x) = \langle \sqrt{x_1/||x||_1}, \ldots, \sqrt{x_W/||x||_1} \rangle$, following the experiments of [6]. The Fisher kernel based on the multinomial distribution uses the representation $s(x) = \langle x_1/\hat{\theta}_1, \ldots, x_W/\hat{\theta}_W \rangle$ where $\hat{\theta}_w = \sum_{d \in D} x_{dw} / \sum_{d \in D} n_d$ is the maximum likelihood parameter value for word w for the multinomial distribution fitted to the given document collection.

Bayesian classification uses Bayes' rule and a different DCM or multinomial model learned from the training documents in each class. However, classification using a Fisher kernel uses just one DCM or multinomial model learned from the entire collection of training documents.

We use two standard document collections called industry sector and 20 newsgroups. Documents are tokenized, stop words removed, and count vectors extracted using the Rainbow toolbox [8]. The industry sector[1] collection contains 9555 documents distributed in 104 classes. It has a vocabulary of 55,055 words, and each document contains on average 606 words. The data are split into halves for training and testing. The 20 newsgroups[2] collection contains 18,828 documents belonging to 20 classes. This collection has a vocabulary of 61,298 words with an average document length of 116 words. The data are split into 80/20 fractions for training and testing.

Table 1 shows classification accuracy averaged over ten splits of each document collection. Nearest neighbor results are for $k = 3$ neighbors. Not surprisingly, the TF-IDF kernel performs best. Both Fisher kernel methods perform well, particularly on the industry sector collection, which has only a small number of documents per class. The DCM Fisher kernel performs slightly better than the multinomial Fisher kernel.

[1] `http://www.cs.umass.edu/~mccallum/code-data.html`
[2] `http://people.csail.mit.edu/people/jrennie/20Newsgroups`

Table 1. Accuracy averaged over ten random splits for different classifiers

Collection	M	DCM	L0	L1	L2	TF-IDF	Bhattacharyya	Fisher-DCM	Fisher-M
20 news	0.843	0.845	0.606	0.675	0.778	0.828	0.744	0.677	0.665
industry	0.791	0.795	0.624	0.017	0.567	0.881	0.254	0.761	0.735

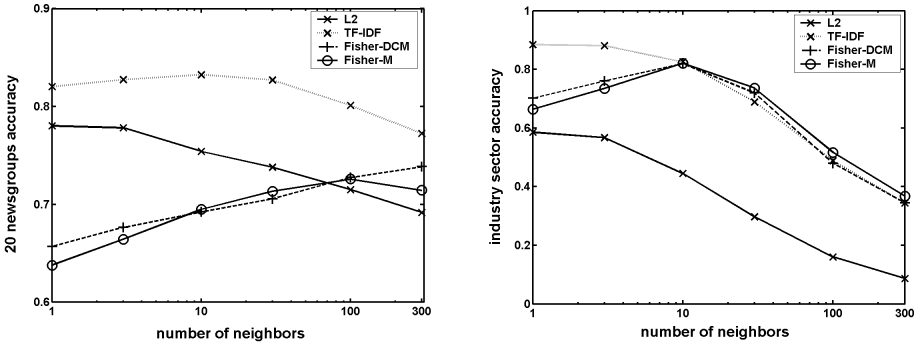

Fig. 1. Average accuracy scores for increasing numbers of nearest neighbors for 20 newsgroups (left) and industry sector (right)

Given the theoretical arguments in favor of the DCM over the standard multinomial model, it is surprising that a Bayesian classifier using multinomial models performs so well. We have three explanations for this. First, our multinomial model uses additive smoothing with constant 0.01 instead of with constant 1.0, which is the standard Laplace smoothing, but performs considerably worse. Second, both the 20 newsgroups and industry sector collections consist of relatively short documents, in which burstiness is less apparent than in longer documents. Third, it is well-known that Bayesian classifiers can be highly accurate even when they use models that produce inaccurate probabilities, since the ordering of the probabilities may still be correct.

Figure 1 shows how four of the k-nearest neighbor methods perform as k varies. Both Fisher kernel methods benefit from using many neighbors on the 20 newsgroups collection, while the performance of cosine similarity decreases as more neighbors are used. This fact possibly indicates that cosine similarity can identify neighbors correctly only if they are very close, whereas the Fisher kernel methods and TF-IDF can pick out not-so-close neighbors well also. For each of the four methods, the optimum value of k is smaller on the industry sector collection. This is perhaps because each class has fewer members in this collection, so each document has fewer genuine neighbors.

5 Discussion

Although the TF-IDF representation for documents is widely used, its origin is heuristic and it does not have a convincing theoretical basis [9]. However, TF-

IDF implicitly contains an important insight: the similarity of two documents (or two data items in general) should be a function not just of the documents themselves, but also of the context of other documents in which they lie.

Fisher kernels are a general implementation of this idea of exploiting background context when computing the degree of similarity of two data items. Above, we have derived and investigated the Fisher kernel induced by the Dirichlet compound multinomial (DCM) distribution. We have shown that the expression for the DCM Fisher kernel contains components similar to the log-term-frequency and inverse-document-frequency components of TF-IDF. We have also shown experimentally that nearest neighbor classifiers based on the DCM Fisher kernel perform well, although not as well as TF-IDF-based classifiers.

We are excited about continuing the research of this paper in three directions. First, we want to experiment with collections of longer documents, where we expect the superiority of the DCM over the multinomial to be greater. Second, we want to use the DCM Fisher kernel in an SVM classifier, since SVMs are generally rather more accurate than nearest neighbor methods. Third, given that TF-IDF remains the best known representation for documents, can we find a new probability distribution whose Fisher kernel is even more similar to TF-IDF?

Acknowledgments. David Kauchak and Rasmus Madsen assisted with the experimental part of this paper.

References

1. Akiko Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1):45–65, 2003.
2. N. Balakrishnan, Norman L. Johnson, and Samuel Kotz. *Discrete Multivariate Distributions*. New York: John Wiley and Sons Inc., 1997.
3. Kenneth W. Church and William A. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.
4. Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Proceedings of NIPS*, pages 914–920, 2000.
5. Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of NIPS*, pages 487–493, 1999.
6. Tony Jebara and Risi Kondor. Bhattacharyya and expected likelihood kernels. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 57–73, 2003.
7. Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. To appear in *Proceedings of ICML*, 2005.
8. Andrew K. McCallum. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.* www.cs.cmu.edu/~mccallum/bow, 1996.
9. Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.