A Statistical Learning Model of Text Classification for Support Vector Machines

Thorsten Joachims
GMD Forschungszentrum IT, AIS.KD
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Thorsten.Joachims@gmd.de

ABSTRACT

This paper develops a theoretical learning model of text classification for Support Vector Machines (SVMs). It connects the statistical properties of text-classification tasks with the generalization performance of a SVM in a quantitative way. Unlike conventional approaches to learning text classifiers, which rely primarily on empirical evidence, this model explains why and when SVMs perform well for text classification. In particular, it addresses the following questions: Why can support vector machines handle the large feature spaces in text classification effectively? How is this related to the statistical properties of text? What are sufficient conditions for applying SVMs to text-classification problems successfully?

1. INTRODUCTION

There are at least two ways to motivate why a particular learning method is suitable for a particular learning task. Since ultimately one is interested in the performance of the method, one way is through comparative studies. Previous work [11, 4] presents such studies showing that Support Vector Machines (SVMs) deliver state-of-the-art classification performance. However, success on benchmarks is a brittle justification for a learning algorithm and gives only limited insight. Therefore, this paper analyzes the suitability of SVMs for learning text classifiers from a theoretical perspective.

In particular, this paper presents an abstract model of text-classification tasks. This model is based on statistical properties of text-classification problems that are both observable and intuitive. Using this model, it is possible to prove what types of text-classification problems are efficiently learnable with SVMs. The central result is an upper bound connecting the expected generalization error of an SVM with the statistical properties of text-classification tasks.

This paper is structured as follows. After a short introduction to SVMs, it will identify the key properties of text-classification tasks. They motivate the model formally defined in Section 4. In addition to verifying the assumptions of the model against real data, this section proves the learnability results. Section 5 further validates the model using experiments, before Section 6 analyzes the complexity of text-classification tasks and identifies sufficient conditions for goods generalization performance.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

2. SUPPORT VECTOR MACHINES

SVMs [18] were developed by V. Vapnik et al. based on the structural risk minimization principle from statistical learning theory. In their basic form, SVMs learn linear decision rules $h(\vec{x}) = sign\{\vec{w}\cdot\vec{x}+b\}$ described by a weight vector $\vec{w}$ and a threshold b. Input is a sample of n training examples $S_n = ((\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n))$, $\vec{x}_i \in \Re^N$, $y_i \in \{-1, +1\}$. For a linearly separable S_n, the SVM finds the hyperplane with maximum Euclidean distance to the closest training examples. This distance is called the margin δ, as depicted in Figure 1. For non-separable training sets, the amount of training error is measured using slack variables ξ_i. Computing the hyperplane is equivalent to solving the following primal optimization problem [18].

Optimization Problem 1 (SVM (primal)).

$$\textit{minimize:} \quad V(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\textit{subj. to:} \quad \forall_{i=1}^{n} : y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i \tag{2}$$

$$\forall_{i=1}^{n} : \xi_i > 0 \tag{3}$$

The constraints (2) require that all training examples are classified correctly up to some slack ξ_i. If a training example lies on the "wrong" side of the hyperplane, the corresponding ξ_i is greater or equal to 1. Therefore, $\sum_{i=1}^{n}\xi_i$ is an upper bound on the number of training errors. The factor C in (1) is a parameter that allows trading off training error vs. model complexity. Note that the margin of the resulting hyperplane is $\delta = 1/||\vec{w}||$.

Instead of solving OP1 directly, one can also consider the follwing dual program.

Optimization Problem 2 (SVM (dual)).

$$\textit{maximize:}\ W(\vec{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j \alpha_i \alpha_j (\vec{x}_i\cdot\vec{x}_j) \tag{4}$$

$$\textit{subj. to:} \sum_{i=1}^{n} y_i\alpha_i = 0 \tag{5}$$

$$\forall i \in [1..n] : 0 \leq \alpha_i \leq C \tag{6}$$

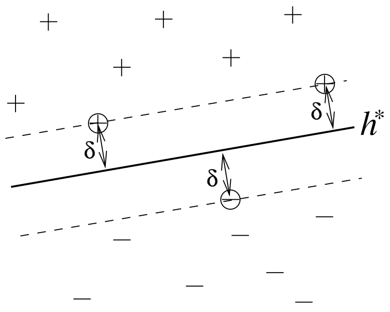

Figure 1: A binary classification problem (+ vs. −) in two dimensions. The hyperplane h^* separates positive and negative training examples with maximum margin δ. The examples closest to the hyperplane are called *support vectors* (marked with circles).

Duality implies that $W(\vec{\alpha}^*) = V(\vec{w}^*, b^*, \vec{\xi}^*)$ at the respective solutions of both programs, and that $W(\vec{\alpha}) \leq V(\vec{w}, b, \vec{\xi})$ for any feasible point. From the solution of the dual, the primal solution can be constructed as

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \quad \text{and} \quad b = y_{usv} - \vec{w}\cdot\vec{x}_{usv} \tag{7}$$

where $(\vec{x}_{usv}, y_{usv})$ is some training example with $0 < \alpha_{usv} < C$. For all but degenerate cases, such training examples exist and the hyperplane is called *stable*. One special family of hyperlanes considered in the following are called *unbiased* hyperplanes. Such hyperplanes are forced to pass through the origin, either by adding the constraint $b = 0$ in OP1, or equivalently by removing the constraint (5) in OP2. From a practical perspective for text classification, SVM restricted to unbiased hyperplane achieve a performance similar to general (i.e. biased) hyperplanes. For the experiments in this paper, SVM^{Light} [12] is used for solving the dual optimization problem[1]. More detailed introductions to SVMs can be found in [2, 18].

3. PROPERTIES OF TEXT-CLASSIFICATION TASKS

To make useful statements about why a particular learning methods should work well for text classification, it is necessary to identify key properties of text-classification tasks. Given a bag-of-words representation, the following properties hold:

High-Dimensional Feature Space. Independent of the particular choice of terms, text-classification problems involve high-dimensional feature spaces. If each word occurring in the training documents is used as a feature, text-classification problems with a few thousand training examples can lead to 30,000 and more attributes.

Sparse Document Vectors. While there is a large space of potential features, each document contains only a small number of distinct words. This implies that document vectors are very sparse.

[1] http://www-ai.informatik.uni-dortmund.de/svm_light

MODULAIRE BUYS BOISE HOMES PROPERTY

Modulaire Industries said it acquired the design library and manufacturing rights of privately-owned Boise Homes for an undisclosed amount of cash. Boise Homes sold commercial and residential prefabricated structures, Modulaire said.

USX, CONSOLIDATED NATURAL END TALKS

USX Corp's Texas Oil and Gas Corp subsidiary and Consolidated Natural Gas Co have mutually agreed not to pursue further their talks on Consolidated's possible purchase of Apollo Gas Co from Texas Oil. No details were given.

JUSTICE ASKS U.S. DISMISSAL OF TWA FILING

The Justice Department told the Transportation Department it supported a request by USAir Group that the DOT dismiss an application by Trans World Airlines Inc for approval to take control of USAir. "Our rationale is that we reviewed the application for control filed by TWA with the DOT and ascertained that it did not contain sufficient information upon which to base a competitive review," James Weiss, an official in Justice's Antitrust Division, told Reuters.

E.D. And F. MAN TO BUY INTO HONG KONG FIRM

The U.K. Based commodity house E.D. And F. Man Ltd and Singapore's Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo's 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

Figure 2: Four documents from the Reuters-21578 category "corporate acquisitions" that do not share any content words.

Heterogeneous Use of Terms. Consider the 4 documents shown in Figure 2. All documents are Reuters-21578 articles from the category "corporate acquisitions". Nevertheless, the overlap between their document vectors is very small. In this extreme case, the documents do not share any content words. The only words that occur in at least two documents are "*it*", "*the*", "*and*", "*of*", "*for*", "*an*", "*a*", "*not*", "*that*", and "*in*". All these words are stopwords and it is unlikely that they help discriminate between documents about corporate acquisitions and other documents. However, each document contains good keywords indicating a "corporate acquisition", just that they are different.

High Level of Redundancy. While there are generally many different features relevant to the classification task, often several such cues occur in one document. These cues are partly redundant. Table 1 [11] shows the results of an experiment on the Reuters "corporate acquisitions" category. All features (after stemming and stopword removal) are ranked according to their (binary) empirical mutual information (EMI) with the class label (cf. e.g. [14]). Then a naive

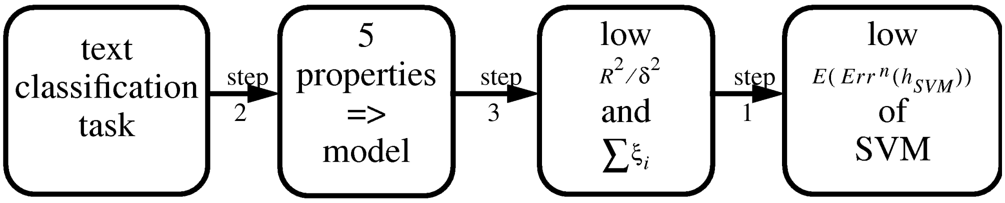

Figure 3: Structure of the argument.

used features by EMI rank	PRBE
1-200	89.6
201-500	71.3
501-1000	63.3
1001-2000	58.0
2001-4000	55.4
4001-9947	47.5
random (no learning)	21.8

Table 1: Learning without using the "best" features.

Bayes classifier is trained using only those features ranked 1-200, 201-500, 501-1000, 1001-2000, 2001-4000, 4001-9947. The results in Table 1 show that even features ranked lowest still contain considerable information and are somewhat relevant. A classifier using only those "worst" features has a precision/recall break-even point (PRBE) (e.g. [11]) much better than random.

Frequency Distribution of Words and Zipf's Law. The occurrence frequencies of words in natural-language follow Zipf's law [19]. Zipf's law states that if one ranks words by their term frequency, the r-th most frequent words occurs roughly $\frac{1}{r}$ times the term frequency of the most frequent words. This implies that there is a small number of words that occurs very frequently, while most words occur very infrequently.

4. A DISCRIMINATIVE MODEL OF TEXT CLASSIFICATION

The goal of this section is a statistical learning model of text-classification tasks. Using a three step approach as illustrated in Figure 3, it provides the relationship between the properties of text-classification tasks identified above and the expected error rate of an SVM. The first step shows that large margin combined with low training error is a sufficient condition for good generalization accuracy. The second step abstracts the properties of text-classification tasks into a model, which the third step connects to large-margin separation.

4.1 Step 1: Bounding the Expected Error Based on the Margin

The following bound [14, 18] shows that large margin combined with low training error leads to high generalization accuracy. It uses results limiting the number of leave-one-out errors [10, 13]. The key quantities are the margin δ as defined in Section 2, the maximum Euclidean length R of the document vectors $\vec{x}$, and the training loss $\sum \xi_i$.

THEOREM 1 (BOUND ON EXPECTED ERROR OF SVM). *The expected error rate $\mathcal{E}(Err^n(h_{SVM}))$ of a SVM based on n training examples with $0 \leq ||\vec{x}_i|| \leq R$ for all points with non-zero probability and some constant C, is bounded by*

$$\mathcal{E}(Err^n(h_{SVM})) \leq \frac{\rho\, \mathcal{E}\left(\frac{R^2}{\delta^2}\right) + \rho\, C'\, \mathcal{E}\left(\sum_{i=1}^{n+1} \xi_i\right)}{n+1}$$

with $C' = C\,R^2$ if $C \geq 1/(\rho\,R^2)$, and $C' = C\,R^2 + 1$ otherwise. For unbiased hyperplanes ρ equals 1, and for general stable hyperplanes ρ equals 2. The expectations on the right are over training sets of size $n+1$.

The proof can be found in [14]. Note the R acts as a scaling constant for the margin δ, as it can easily be seen in Optimization Problem 1. For example, the squared margin δ^2 can always be doubled by scaling the document vectors $\vec{x}$ to twice their length. The bound in Theorem 1 accounts for such scaling.

4.2 Step 2: TCat-Concepts as a Model of Text-Classification Tasks

Unfortunately, it is not possible to simply look at a new text-classification task and immediately have a good idea of whether it has a large margin. The margin property is observable only after training data becomes available and requires training the SVM. To overcome this problem, this second step lays the basis for connecting the large-margin property with more intuitive and more meaningful properties of text-classification tasks.

Consider the following stereotypical text classification task. While this task is artificial and hypothetical, it will serve as a motivation for the model developed in this section. For this example task, the following describes how documents from the two classes differ in terms of the frequency with which certain types of words occur in them. Figure 4 graphically illustrates the corresponding "word-frequency histogram".

Stopwords Independently of whether a document is from the positive or the negative class, each document contains 20 word occurrences from a set of 100 words (i.e. lexicon entries). These high-frequency words are typically considered stopwords. Note that this does not specify the individual word frequences, i.e. it is open whether one word occurs 20 times, or 20 different words each occur once, or something in between.

Medium Frequency There are 1,000 medium-frequency words in the lexicon. From a subset of 600 such entries, again each positive and negative document contains (any bag of) 5 occurrences. But there are also

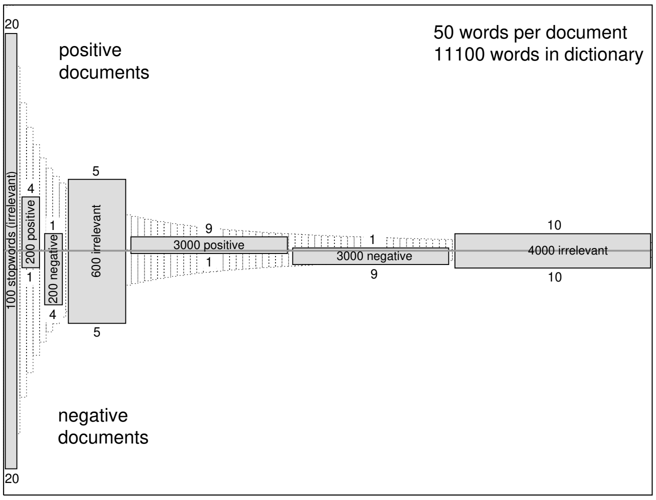

Figure 4: A simple example of a TCat-concept.

two groups of 200 entries each that occur primarily in positive or negative documents, respectively. In particular, from one group there are 4 occurrences in each positive document and only 1 in each negative document. Respectively, from the other group there are 4 occurrences in each negative document while there is only one in each positive document.

Low Frequency Similarly, for the remaining 10,000 entries in the low-frequency part of the lexion, there is a subset of 4,000 entries of which there are 10 occurrences in both positive and negative documents. But there are two sets of 3,000 entries each that occur primarily in positive or negative documents with a frequency of 9 versus 1.

In how far does this example resemble the properties of text-classification tasks identified in Section 3?

High-Dimensional Input Space: There are 11,100 features, which is on the same order of magnitude as real text-classification tasks.

Sparse Document Vectors: Each document is only 50 words long, which means there are at least 11,050 zero entries in each document vector.

High Level of Redundancy: In each document there are 4 medium-frequency words and 9 low-frequency words that indicate the class of the document. Considering the document length of 50 words, this is a fairly high level of redundancy.

Heterogeneous Use of Terms: Both the positive and the negative documents each have a group of 200 medium-frequency words and a group of 3,000 low-frequency words. From each group there can be an arbitrary subset of 4 for the medium-frequency words and 9 for the low-frequency words in each document. Considering only the medium-frequency words, this implies that there can be 50 documents in the same class that do not share a single medium-frequency term from this group. This mimics the property of text classification tasks identified in Section 3.

Zipf's Law: There is a small number of words (100 stopwords) that occur very frequently, a set of 1,000 words of medium frequency, and a large set of 10,000 low-frequency words. This does resemble Zipf's law.

To abstract from this particular example, the following definition introduces a parameterized model that can describe text-classification tasks more generally.

DEFINITION 1 (HOMOGENEOUS TCAT-CONCEPTS). *The TCat-concept*

$$TCat([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s]) \tag{8}$$

describes a binary classification task with s disjoint sets of features (i.e. words). The i-th set includes f_i features. Each positive example contains p_i occurrences of features from the respective set, and each negative example contains n_i occurrences. The same feature can occur multiple times in one document.

This definition does not include noise (e.g. violations of the occurrence frequencies prescribed by the TCat-concept). However, the model can be extended to handle noise in a straightforward way [14]. Applying the definition to the example in Figure 4, it is easy to verify that the example can be described as a

$$\begin{array}{ll} TCat([20\!:\!20\!:\!100], & \#\text{ high freq.} \\ \quad [4\!:\!1\!:\!200], [1\!:\!4\!:\!200], [5\!:\!5\!:\!600], & \#\text{ medium freq.} \\ \quad [9\!:\!1\!:\!3000], [1\!:\!9\!:\!3000], [10\!:\!10\!:\!4000] & \#\text{ low freq.} \\ \quad) & \end{array}$$

concept. While this is an artificial example, is it possible to model real text-classification tasks as TCat-concepts?

Empirical Validation. Consider text-classification tasks from the Reuters-21578[2], the WebKB[3], and the Ohsumed[4] collection. The following analysis shows how they can be modeled as TCat-concepts.

Let us start with the category "course" from the WebKB collection. First, we partition the feature space into disjoint sets of positive indicators, negative indicators, and irrelevant features. Using the simple strategy [14] of selecting features by their odds ratio, there are 98 high-frequency words that indicate positive documents (odds ratio greater than 2) and 52 high-frequency words indicating negative documents (odds ratio less than 0.5). An excerpt of these words is given in Figure 5. Similarly, there are 431 (341) medium-frequency words that indicate positive (negative) documents with an odds ratio greater than 5 (less than 0.2). In the low-frequency spectrum there are 5,045 positive indicators (odds ratio greater than 10) and 24,276 negative indicators (odds ratio less than 0.1). All other words in the vocabulary are assumed to carry no information.

To abstract from the details of particular documents, it is useful to analyse what a typical document for this task looks like. In some sense, an "average" document captures what is typical. An average WebKB document is 277 words long. For positive examples of the category "course", on average 27.7% of the 277 occurrences come from the set of 98 high-frequency positive indicators while these words account

[2]http://www.research.att.com/~lewis/reuters21578.html
[3]http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data
[4]ftp://medir.ohsu.edu/pub/ohsumed

	high frequency	medium frequency	low frequency
pos	**98 words** all any assignment assignments available be book c chapter class code course cse description discussion document due each eecs exam exams fall final ... section set should solution solutions spring structures students syllabus ta text textbook there thursday topics tuesday unix use wednesday week will you your	**431 words** account acrobat adapted addison adt ahead aho allowed alternate announced announcement announcements answers appointment approximately ... tuesdays turing turn turned tuth txt uidaho uiowa ullman understand ungraded units unless upenn usr vectors vi walter weaver wed wednesdays weekly weeks weights wesley yurttas	**5045 words** 002cc 009a 00a 00om 01oct 01pm 02pm 03oct 03pm 03sep 04dec ... gradable gradebook gradebooks gradefreq1 gradefreq2 gradefreq3 graders gradesheet gradients grafica grafik ... zimmermann zinc zipi zipser zj zlocate znol zoran zp zwatch zwhere zwiener zyda
neg	acm address am austin ca california center college computational conference contact current currently d department dr faculty fax graduate group he ... me member my our parallel performance ph pp proceedings professor publications recent research sciences support technical technology university vision was working **52 words**	aaai academy accesses accurate adaptation advisor advisory affiliated affiliations agent agents alberta album alumni amanda america amherst annual ... victoria virginia visiting visitors visualization vita vitae voice wa watson weather webster went west wi wife wireless wisconsin worked workshop workshops wrote yale york **341 words**	0a 0b 0b1 0e 0f 0r 0software 0x82d4ff 100k 100mhz 100th 1020x620 102k 103k ... lunar lunches lunchtime lund lundberg lunedi lung luniewski luo luong lupin lupton lure lurker lus ... zuo zuowei zurich zvi zw zwaenepoel zwarico zwickau zwilling zygmunt zzhen00 **24276 words**
	high frequency	medium frequency	low frequency

Figure 5: Indicative words for the WebKB category "course" partitioned by occurrence frequency.

for only 10.4% of the occurrences in an average negative document. Assessing the percentages analogously also for the other word groups, they can be directly translated into the following TCat-concept.

$$
\begin{array}{lll}
TCat_{course}(& [77:29:98], [4:21:52], & \#\text{ high freq.} \\
 & [16:2:431], [1:12:341], & \#\text{ medium freq.} \\
 & [9:1:5045], [1:21:24276], & \#\text{ low freq.} \\
 & [169:191:8116] & \#\text{ rest} \\
 &)
\end{array}
$$

This shows that the text-classification task connected with the WebKB category "course" can be modeled as a TCat-concept, if one assumes that documents are of homogeneous length and composition. It can be shown that this assumption of homogeneity can be relaxed [14].

Similar TCat-concepts can also be found for other tasks. For the Reuters-21578 category "earn" the same procedure leads to the TCat-concept

$$
\begin{array}{lll}
TCat_{earn}(& [33:2:65], [32:65:152], & \#\text{ high freq.} \\
 & [2:1:171], [3:21:974], & \#\text{ medium freq.} \\
 & [3:1:3455], [1:10:17020], & \#\text{ low freq.} \\
 & [78:52:5821] & \#\text{ rest} \\
 &)
\end{array}
$$

as an average case model. The model for the Ohsumed category "pathology" is

$$
\begin{array}{lll}
TCat_{pathology}(& [2:1:10], [1:4:22], & \#\text{ high freq.} \\
 & [2:1:92], [1:2:94], & \#\text{ medium freq.} \\
 & [5:1:4080], [1:10:20922], & \#\text{ low freq.} \\
 & [197:190:13459] & \#\text{ rest} \\
 &).
\end{array}
$$

Note that in particular the model for "pathology" is substantially different from the other two. This verifies that TCat-concepts can capture some properties of real text-classification tasks that have the potential to differentiate between tasks. The following studies their relevance for generalization performance.

4.2.1 Step 3: Learnability of TCat-Concepts

This final step provides the connection between TCat-concepts and the bound for the generalization performance of an SVM. The first lemma shows that homogeneous TCat-concepts are generally separable with a certain margin. Using the fact that term frequencies obey Zipf's law, a second lemma shows that the Euclidean length of document vectors is small for text-classification tasks. These two results lead to the main learnability result for TCat-concepts.

LEMMA 1 (MARGIN OF NOISE-FREE TCAT-CONCEPTS).

For $TCat([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s])$-concepts, there is always a hyperplane passing through the origin that has a margin δ bounded by

$$
\delta^2 \geq \frac{ac - b^2}{a + 2b + c} \qquad \textit{with} \qquad
\begin{array}{l}
a = \sum_{i=1}^{s} \frac{p_i^2}{f_i} \\
b = \sum_{i=1}^{s} \frac{p_i n_i}{f_i} \\
c = \sum_{i=1}^{s} \frac{n_i^2}{f_i}
\end{array}
\qquad (9)
$$

PROOF. Define $\vec{p}^T = (p_1, \ldots, p_s)^T$ and $\vec{n}^T = (n_1, \ldots, n_s)^T$, as well as the diagonal matrix F with $f_1, \ldots, f_s$ on the diagonal.

The margin of the maximum-margin hyperplane that separates a given training sample $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ and that passes through the origin can be derived from the solution of the following optimization problem.

$$\begin{aligned} W(\vec{w}) &= min \frac{1}{2} \vec{w}^T \vec{w} && (10) \\ s.t. \quad & y_1[\vec{x}_1^T \vec{w}] \geq 1 && \\ & \vdots && (11) \\ & y_n[\vec{x}_n^T \vec{w}] \geq 1 && \end{aligned}$$

The hyperplane corresponding to the solution vector $\vec{w}^*$ has a margin $\delta = (2W(\vec{w}^*))^{-0.5}$. By adding constraints to this optimization problem, it is possible to simplify its solution and get a lower bound on the margin. Let us add the additional constraint that within each group of f_i features the weights are required to be identical. Then $\vec{w}^T \vec{w} = \vec{v}^T F \vec{v}$ for a vector $\vec{v}$ of dimensionality s. The constraints (11) can also be simplified. By definition, each example contains a certain number of features from each group. This means that all constraints for positive examples are equivalent to $\vec{p}^T \vec{v} \geq 1$ and, respectively, $\vec{n}^T \vec{v} \leq -1$ for the negative examples. This leads to the following simplified optimization problem.

$$\begin{aligned} W'(\vec{v}) &= min \frac{1}{2} \vec{v}^T F \vec{v} && (12) \\ s.t. \quad & \vec{p}^T \vec{v} \geq 1 && (13) \\ & \vec{n}^T \vec{v} \leq -1 && (14) \end{aligned}$$

Let $\vec{v}^*$ be the solution. Since $W'(\vec{v}^*) \geq W(\vec{w}^*)$, it follows that $\delta \geq (2W'(\vec{v}^*))^{-0.5}$ is a lower bound for the margin. It remains to find an upper bound for $W'(\vec{v}^*)$ that can be computed in closed form. Introducing Lagrange multipliers, the solution $W'(\vec{v}^*)$ equals the value $L(\vec{v}, \alpha_+, \alpha_-)^*$ of

$$L(\vec{v}, \alpha_+, \alpha_-) = \frac{1}{2} \vec{v}^T F \vec{v} - \alpha_+ (\vec{p}^T \vec{v} - 1) + \alpha_- (\vec{n}^T \vec{v} + 1) \quad (15)$$

at its saddle-point. $\alpha_+ \geq 0$ and $\alpha_- \geq 0$ are the Lagrange multipliers for the two constraints (13) and (14). Using the fact that

$$\frac{dL(\vec{v}, \alpha_+, \alpha_-)}{d\vec{v}} = 0 \quad (16)$$

at the saddle point one gets a closed form solution for $\vec{v}$.

$$\vec{v} = F^{-1} [\alpha_+ \vec{p} - \alpha_- \vec{n}] \quad (17)$$

For ease of notation one can equivalently write

$$\vec{v} = F^{-1} X Y \vec{\alpha} \quad (18)$$

with $X = (\vec{p}, \vec{n})$, $Y = diag(1, -1)$, and $\vec{\alpha}^T = (\alpha_+, \alpha_-)$ appropriately defined. Substituting into the Lagrangian results in

$$L(\vec{\alpha}) = 1^T \vec{\alpha} - \frac{1}{2} \vec{\alpha}^T Y X^T F^{-1} X Y \vec{\alpha} \quad (19)$$

To find the saddle points one has to maximize this function over $\vec{\alpha}^T = (\alpha_+, \alpha_-)^T$ subject to $\alpha_+ \geq 0$ and $\alpha_- \geq 0$. Since only a lower bound on the margin is needed, it is possible to drop the constraints $\alpha_+ \geq 0$ and $\alpha_- \geq 0$. Removing the constraints can only increase the objective function at the solution. So the unconstrained maximum $L'(\vec{\alpha})^*$ is greater or equal to $L(\vec{\alpha})^*$. Setting the derivative of (19) to 0

$$\frac{dL'(\vec{\alpha})}{d\vec{\alpha}} = 0 \quad \Leftrightarrow \quad \vec{\alpha} = (Y X^T F^{-1} X Y)^{-1} 1 \quad (20)$$

and substituting into (19) yields the unconstrained maximum:

$$L'(\vec{v}, \vec{\alpha})^* = \frac{1}{2} 1^T (Y X^T F^{-1} X Y)^{-1} 1 \quad (21)$$

The special form of $(Y X^T F^{-1} X Y)$ makes it possible to compute its inverse in closed form.

$$\begin{aligned} (Y X^T F^{-1} X Y)^{-1} &= \begin{pmatrix} \vec{p}^T F^{-1} \vec{p} & -\vec{p}^T F^{-1} \vec{n} \\ -\vec{n}^T F^{-1} \vec{p} & \vec{n}^T F^{-1} \vec{n} \end{pmatrix}^{-1} && (22) \\ &= \begin{pmatrix} a & -b \\ -b & c \end{pmatrix}^{-1} && (23) \\ &= \frac{1}{ac - b^2} \begin{pmatrix} a & b \\ b & c \end{pmatrix} && (24) \end{aligned}$$

Substituting into (21) completes the proof. $\square$

The lemma shows that any set of documents, where each document is fully consistent with the specified TCat-concept, is always linearly separable with a certain minimum margin. Note that separability implies that the training loss $\sum \xi_i$ is zero. While this paper considers only the case of full consistency and zero noise, [14] shows how these assumtions can be relaxed.

It remains to bound the maximum Euclidean length R of document vectors before it is possible to apply Theorem 1. Clearly, the document vector of a document with l words cannot have a Euclidean length greater than l. Nevertheless, this bound is very loose for real document vectors. To bound the quantity R more tightly it is possible to make use of Zipf's law.

Assume that the term frequencies in every document follow the generalized Zipf's law [15]

$$TF_r = \frac{c}{(r+k)^\phi} \quad (25)$$

with typical parameter values $k \approx 5$, $\phi \approx 1.3$, and c scaling with document length. This assumption about Zipf's law does not imply that a particular word occurs with a certain frequency in every document. It is much weaker; it merely implies that the r-th most frequent word in a document occurs with a particular frequency. In slight abuse of Zipf's law for short documents, the following lemma connects the length of the document vectors to Zipf's law. Intuitively, it states that many words in a document occur with low frequency, leading to document vectors of relatively short Euclidean length.

Lemma 2 (Length of Document Vectors). *If the ranked term frequencies TF_r in a document with l terms have the form of the generalized Zipf's law*

$$TF_r = \frac{c}{(r+k)^\phi} \quad (26)$$

based on their frequency rank r, then the squared Euclidean length of the document vector $\vec{x}$ of term frequencies is bounded by

$$||\vec{x}|| \leq \sqrt{\sum_{r=1}^{d} \left(\frac{c}{(r+k)^\phi} \right)^2} \quad \text{with } d \text{ such that } \sum_{r=1}^{d} \frac{c}{(r+k)^\phi} = l$$

PROOF. From the connection between the frequency rank of a term and its absolute frequency it follows that the r-th most frequent term occurs

$$TF_r = \frac{c}{(r+k)^{\phi}} \tag{27}$$

times. The document vector $\vec{x}$ has d non-zero entries which are the values $TF_1, ..., TF_d$. Therefore, the Euclidian length of the document vector $\vec{x}$ is

$$\vec{x}^T\vec{x} = \sum_{r=1}^{d}\left(\frac{c}{(r+k)^{\phi}}\right)^2 \tag{28}$$

□

Combining Lemma 1 and Lemma 2 with Theorem 1 leads to the following main result.

THEOREM 2 (LEARNABILITY OF TCAT-CONCEPTS). *For $TCat([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s])$-concepts and documents with l terms distributed according to the generalized Zipf's law $TF_r = \frac{c}{(r+k)^{\phi}}$, the expected generalization error of an (unbiased) SVM after training on n examples is bounded by*

$$\mathcal{E}(Err^n(h_{SVM})) \leq \rho\frac{R^2}{n+1}\,\frac{a+2b+c}{ac-b^2} \quad \text{with} \quad \begin{array}{l} a = \sum_{i=1}^{s}\frac{p_i^2}{f_i} \\ b = \sum_{i=1}^{s}\frac{p_i n_i}{f_i} \\ c = \sum_{i=1}^{s}\frac{n_i^2}{f_i} \\ R^2 = \sum_{r=1}^{d}\left(\frac{c}{(r+k)^{\phi}}\right)^2 \end{array}$$

unless $\forall_{i=1}^{s} : p_i = n_i$. d is chosen so that $\sum_{r=1}^{d}\frac{c}{(r+k)^{\phi}} = l$. For unbiased SVMs ρ equals 1, and for biased SVMs ρ equals 2.

PROOF. Using the fact that TCat-concepts are separable (and therefore stable), if at least for one i the value of p_i is different from n_i, the result from Theorem 1 reduces to

$$\mathcal{E}(Err^n(h_{SVM})) \leq \frac{1}{n+1}\,\rho\, E\!\left(\frac{R^2}{\delta^2}\right) \tag{29}$$

since all ξ_i are zero for a sufficiently large value of C. Lemma 1 gives a lower bound for δ^2 which can be used to bound the expectation

$$E\!\left(\frac{R^2}{\delta^2}\right) \leq \rho\frac{a+2b+c}{ac-b^2}\, E(R^2) \tag{30}$$

It remains for us to give an upper bound for $E(R^2)$. R^2 is the maximum Euclidian length of any feature vector in the training data. Since the term frequencies in each example follow the generalized Zipf's law $TF_r = \frac{c}{(r+k)^{\phi}}$, it is possible to use Lemma 2 to bound R^2 and therefore $E(R^2)$. □

Empirical Validation. The TCat-model and the lemmata leading to the main result suggest that text classification tasks are generally linearly separable (i.e. $\sum \xi_i = 0$), and that the normalized inverse margin R^2/δ^2 is small. This prediction can be tested against real data.

Reuters	$\frac{R^2}{\delta^2}$	$\sum_{i=1}^{n}\xi_i$
earn	1143	0
acq	1848	0
money-fx	1489	27
grain	585	0
crude	810	4
trade	869	9
interest	2082	33
ship	458	0
wheat	405	2
corn	378	0

WebKB	$\frac{R^2}{\delta^2}$	$\sum_{i=1}^{n}\xi_i$
course	519	0
faculty	1636	0
project	741	0
student	1588	0

Ohsumed	$\frac{R^2}{\delta^2}$	$\sum_{i=1}^{n}\xi_i$
Pathology	11614	0
Cardiovasc.	4387	0
Neoplasms	2868	0
Nervous Sys.	3303	0
Immunologic	2556	0

Table 2: Normalized inverse margin and training loss for the Reuters (27,658 features), the WebKB (38,359 features), and the Ohsumed data (38,679 features) for $C = 50$. As suggested by model-selection experiments, TFIDF-weighting is used for Reuters and Ohsumed, while the representation for WebKB is binary. No stemming is performed and stopword removal is used only on the Ohsumed data.

First, Table 2 indicates that all Ohsumed categories, all WebKB tasks, and most Reuters-21578 categories are linearly separable (i.e. $\sum \xi_i = 0$). This means that there is a hyperplane so that all positive examples are on one side of the hyperplane, while all negative examples are on the other. Inseparability on some Reuters categories is often due to dubious documents (consisting only of a headline) or obvious misclassifications of the human indexers.

Second, separability is possible with a large margin. Table 2 shows the size of the normalized inverse margin for the ten most frequent Reuters categories, the WebKB categories, and the five most frequent Ohsumed categories. Intuitively, R^2/δ^2 can be treated as an "effective" number of parameters due to it's link to VC-dimension [18]. Compared to the dimensionality of the feature space, the normalized inverse margin is typically small.

These experimental findings in connection with the theoretical results from above validate that TCat-concepts do capture an important and widely present property of text classification tasks.

5. COMPARING THE THEORETICAL MODEL WITH EXPERIMENTAL RESULTS

The previous sections formally describes that a large expected margin with low training error leads to a low expected prediction error. Furthermore, they indicate how margin is related to the properties of TCat-concepts, and experimentally verify that real text-classification tasks can be modeled with TCat-concepts. This section verifies not only that the individual steps are well justified, but also that their conjunction produces meaningful results. To show this, this section compares the generalization performance as predicted by the model with the generalization performance found in experiments.

In Section 4.2 a TCat-model for the WebKB category

	model $\mathcal{E}(Err^n(h_{SVM}))$	experiment $Err^n_{test}(h_{SVM})$
WebKB "course"	11.2%	4.4%
Reuters "earn"	1.5%	1.3%
Ohsumed "pathology"	94.5%	23.1%

Table 3: Comparing the expected error predicted by the model with the error rate and the precision/recall breakeven point on the test set for the WebKB category "course", the Reuters category "earn", and the Ohsumed category "pathology" with TF weighting and $C = 1000$. No stopword removal and no stemming are used.

"course" was estimated. Furthermore, the parameters of Zipf's law for the full WebKB collection are $c = 470000$, $k = 5$, and $\phi = 1.25$. Subject to the assumptions of Theorem 2, substituting the estimated values into the bound leads to the following characterization of the expected error.

$$\mathcal{E}(Err^n(h_{SVM})) \leq \frac{0.2331 \cdot 1899.7}{n+1} \leq \frac{443}{n+1} \tag{31}$$

n denotes the number of training examples. Consequently, after training on 3957 examples the model predicts an expected generalization error of less than 11.2%.

An analog procedure for the Reuters category "earn" leads to the bound

$$\mathcal{E}(Err^n(h_{SVM})) \leq \frac{0.1802 \cdot 762.9}{n+1} \leq \frac{138}{n+1} \tag{32}$$

so that the expected generalization error after 9603 training examples is less than 1.5%. Similarly, the bound for the Ohsumed category "pathology" is

$$\mathcal{E}(Err^n(h_{SVM})) \leq \frac{7.4123 \cdot 1275.8}{n+1} \leq \frac{9457}{n+1}, \tag{33}$$

leading to an expected generalization error of less than 94.5% after 10,000 training examples.

Table 3 compares the expected generalization error predicted by the estimated models with the generalization performance observed in experiments. While it is unreasonable to expect that the model precisely predicts the exact performance observed on the test set, Table 3 shows that the model captures which classification tasks are more difficult than others. In particular, it does correctly predict that "earn" is the easiest task, "course" is the second easiest task, and that "pathology" is the most difficult one. While the TCat model is probably not detailed enough to be suitable for performance estimation in most application settings (e.g. [13]), this gives some validation that TCat-concepts can formalize the key properties of text-classification tasks relevant for learnability with SVMs. More can be found in [14].

6. SENSITIVITY ANALYSIS: DIFFICULT AND EASY LEARNING TASKS

The previous section revealed that the bound on the expected generalization error can be large for some TCat-concepts while it is small for others. Going through different scenarios, it is now possible to identify the key properties that make a text-classification task "easy" or "difficult" for an SVM to learn [14].

Occurrence Frequency Given that the other parameters stay constant, the bound on the error rate decreases, if the frequency of the discriminative features is increased.

Discriminative Power of Term Sets The extent to which vocabulary differs between classes makes a difference for learnability. The value of the bound decreases, if the difference in class conditional word frequencies increases.

Level of Redundancy The higher the redundancy, the lower the bound on the generalization error. This implies that it is desirable to have many clues in each document.

Similarly, the model can be used to analyse the effect of TFIDF weighting on the effectiveness of SVMs depending on the properties of the task [14].

7. LIMITATIONS OF THE MODEL AND OPEN QUESTIONS

Every model abstracts from reality in some sense and it is important to clearly point the assumptions out.

First, each document is assumed to exactly follow the same generalized Zipf's law, neglecting variance and discretization inaccuracies that occur especially for short documents. In particular, this implies that all documents are of equal length.

Second, the model fixes the number of occurrences from each word set in the TCat-model. While the degree of violation of this assumption can be captured in terms of attribute noise, it might be useful and possible not to specify the exact number of occurrences per word set, but only upper and lower bounds. This could make the model more accurate. However, it comes with the cost of an increased number of parameters, making the model less understandable. While the formal analysis of noise in [14] demonstrates that the model does not break in the presence of noise, the bounds could be tightened. Along the same lines, parametric noise models could be incorporated to model the types of noise in text-classification problems.

Finally, the general approach taken in this paper is to model only upper bounds on the error rate. While these are important to derive sufficient conditions for the learnability of text-classification tasks, lower bounds may be of interest as well. They could answer the question of which text-classification tasks cannot be learned with SVMs.

8. RELATED WORK

While other learning algorithms can also be analyzed in terms of formal models, these models typically make assumptions unjustified for text.

The most popular such algorithm is naive Bayes. Naive Bayes is commonly justified using assumptions of conditional independence or linked dependence [3]. However, these assumptions are generally accepted to be false for text. While more complex dependence models can somewhat remove the degree of violation [17], a principal problem with using generative models for text remains. Finding a generative models for natural language appears much more difficult than solving a text classification task. Therefore, this paper presented a discriminative model of text classification. It does not model language, but merely constrains the distribution of words enough to describe classification accuracy. This way it is possible to avoid false independence assumptions.

Another model used to describe the properties of text is the 2-Poisson model [1]. However, like the Bernoulli model it is rejected by tests [8, 9]. Description oriented approaches [7] [6] [5] provide powerful modeling tool and can avoid high-dimensional feature spaces, but require implicit assumptions in the way description vectors are generated.

While different in its motivation and its goal, the work of Papadimitriou et. al is most similar in spirit to the approach presented here [16]. They show that latent semantic indexing leads to a suitable low-dimensional representation, given assumptions about the distribution of words. These assumptions are similar in how they exploit the difference of word distributions. However, they do not show how their assumptions relate to the statistical properties of text and they do not derive generalization-error bounds.

9. SUMMARY AND CONCLUSIONS

This paper develops the first model of learning text classifiers from examples that makes it possible to quantitatively connect the statistical properties of text with the generalization performance of the learner. The model is the result of taking a discriminative approach. Unlike conventional generative models, it does not involve independence assumptions. The discriminative model focuses on those properties of the text classification tasks that are sufficient for good generalization performance, avoiding much of the complexity of natural language.

Based on this discriminative model, the paper explains how SVMs can achieve good classification performance despite the high-dimensional feature spaces in text classification. The resulting bounds on the expected generalization error give a formal understanding of what kind of text-classification task can be solved with SVMs. This makes it possible to identify that – intuitively – high redundancy, high discriminative power of term sets, and discriminative features in the high-frequency range are sufficient conditions for good generalization. Finally, the model provides a formal basis for developing new algorithms that are most appropriate in specific scenarios.

10. REFERENCES

[1] A. Bookstein and D. R. Swanson. Probabilistic models for automated indexing. *Journal of the American Society for Information Science*, 25(5):312–318, 1974.

[2] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] W. Cooper. Some inconsistencies and misnomers in probabilistic information retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–61, 1991.

[4] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of ACM-CIKM98*, November 1998.

[5] N. Fuhr, S. Hartmann, G. Lustig, M. Schwantner, K. Tzeras, and G. Knorz. Air/x - a rule-based multistage indexing system for large subject fields. In *RIAO*, pages 606–623, 1991.

[6] N. Fuhr and G. Knorz. Retrieval test evaluation of a rule based automatic indexing (air/phys). In C. van Rijsbergen, editor, *Research and Development in Information Retrieval: Proceedings of the Third Joint BCS and ACM Symposium*, pages 391–408. Cambridge University Press, July 1984.

[7] N. Gövert, M. Lalmas, and N. Fuhr. A probabilistic description-oriented approach for categorising Web documents. In *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management*, pages 475–482, Kansas City, US, 1999. ACM Press, New York, US.

[8] S. P. Harter. A probabilistic approach to automated keyword indexing. Part I: on the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 26(4):197–206, 1975.

[9] S. P. Harter. A probabilistic approach to automated keyword indexing. Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, 26(5):280–289, 1975.

[10] T. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Conference on AI and Statistics*, 1999.

[11] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137 – 142, Berlin, 1998. Springer.

[12] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11. MIT Press, Cambridge, MA, 1999.

[13] T. Joachims. Estimating the generalization performance of a SVM efficiently. In *Proceedings of the International Conference on Machine Learning*, San Francisco, 2000. Morgan Kaufman.

[14] T. Joachims. *The Maximum-Margin Approach to Learning Text Classifiers: Methods, Theory, and Algorithms.* PhD thesis, Universität Dortmund, 2001. Kluwer, to appear.

[15] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon. *Information and Control*, 2(1):90–99, Apr. 1959.

[16] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In ACM, editor, *PODS '98. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June, 1998, Seattle, Washington*, pages 159–168, New York, NY 10036, USA, 1998. ACM Press.

[17] M. Sahami. *Using Machine Learning to Improve Information Access.* PhD thesis, Stanford University, 1998.

[18] V. Vapnik. *Statistical Learning Theory.* Wiley, Chichester, GB, 1998.

[19] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley, Cambridge, MA, USA, 1949.