Robust Models in Information Retrieval

Nedim Lipka and Benno Stein
Bauhaus-Universität Weimar
Germany
<first name>.<last name>@uni-weimar.de

***Abstract*—Classification tasks in information retrieval deal with document collections of enormous size, which makes the ratio between the document set underlying the learning process and the set of unseen documents very small. With a ratio close to zero, the evaluation of a model-classifier-combination's generalization ability with leave-n-out-methods or cross-validation becomes unreliable: The generalization error of a complex model (with a more complex hypothesis structure) might be underestimated compared to the generalization error of a simple model (with a less complex hypothesis structure). Given this situation, optimizing the bias-variance-tradeoff to select among these models will lead one astray. To address this problem we introduce the idea of robust models, where one intentionally restricts the hypothesis structure within the model formation process. We observe that—despite the fact that such a robust model entails a higher test error—its efficiency "in the wild" outperforms the model that would have been chosen normally, under the perspective of the best bias-variance-tradeoff. We present two case studies: (1) a categorization task, which demonstrates that robust models are more stable in retrieval situations when training data is scarce, and (2) a genre identification task, which underlines the practical relevance of robust models.**

***Keywords*-retrieval model, bias, overfitting, machine learning**

I. Introduction

Supervised learning means to build a function, called classifier, from labeled training examples in order to predict the labels of unseen examples. The predictive behavior of a classifier is rooted in its generalization ability, i.e., by explaining the observed data under a set of simplifying assumptions. These assumptions are sometimes called *inductive bias* [15]; they are often implicitly introduced, among others by the model that represents the data, by the sample selection process, or by the learning algorithm. Given a classifier in a concrete learning situation the *statistical bias* quantifies the error that is caused by this simplification, while the inductive bias can be considered as the rationale (the logical argument) for this error. Accepting a higher bias will reduce the variance of the learned classifier and may entail a lower generalization error—a connection which is known as bias-variance-tradeoff. If only a very small amount of training data is available, choosing among different complex models in order to determine the best bias-variance-tradeoff becomes a game of chance: all learning methods, which try to build classifier with minimum generalization error, rely on the assumption that the examples are representative.

The investigations of this paper are motivated by the extreme relations in information retrieval. We are working on classification tasks such as genre analysis on the Web or the semi-automatic maintenance of large repositories, where the size ratio υ between the sample S (comprising training and test data) and the set of unseen documents is close to zero. As a consequence even sophisticated learning strategies are misguided by S if the feature vectors $\mathbf{x} \in S$ consist of many and highly variant features. Reason for the misguidance is that the concept of representativeness inevitably gets lost for $\upsilon \ll 1$ and, as a result, it is no longer possible to apply standard model selection or feature selection. However, we argue that even in such extreme learning situations classifiers can be built that generalize well: the basic idea is to *withhold information* contained in S from the learner. Conceptually, such a restriction cannot be left to the learner but must happen intentionally, by means of a task-oriented model formation. By model formation we denote the mapping α from a set of real-world objects O (the real documents) onto a set X of feature vectors.

Contributions. We put the model formation process in the focus of the retrieval performance analysis. In particular, we propose to identify the robustness of a model with the inductive bias that is intentionally introduced within the model formation process.[1] We evaluate these considerations: variants of the vector space model are compared with respect to different model formation functions, and, for the field of genre identification the robustness of the state-of-the-art retrieval models is analyzed. Altogether, the idea of robust models can be considered as a model selection paradigm that suggests to prefer the "inferior" model under certain circumstances.

Existing Research. The existing research can be distinguished into the following areas: theoretical analysis of sample complexity, multiple evaluations of a training sample S, and semi-supervised learning.

1) The sample complexity is related to the question of how many training examples are needed such that a learner converges with high probability to a successful hypothesis [15]. A key factor is the size of the learner's underlying hypothesis space. There are upper bounds linear in $VC(H)$, the Vapnik-Chernovenkis dimension of the hypothesis space [2], [26], and logarithmically in $|H|$, the size of the hypothesis space.
2) A multiple evaluation of training samples can be realized with ensemble classifiers or collaborative filtering techniques [22], [16], [4]. They can be considered as experts, each of which focusing on different aspects of

[1] This form of an inductive bias is sometimes called *restriction bias*.

the training samples, and the combined expertise can alleviate the negative impact of a small set S.

3) Semi-supervised learning approaches like those mentioned in [20], [1] are appropriate if along with a small set of training samples S a large sample of unlabeled, but representative data is given. A promising approach in this regard is the integration of domain knowledge into the learning phase [5].

II. Robust Models

Starting point is a classification task $\langle O, Y\rangle$ (see Figure 1, left), where we are given a set of objects O, the population, which can be classified by a real-world classifier into k classes $Y = \{1, \ldots, k\}$. A real-world classifier should be understood as a decision machine that is unrestricted in every respect. By contrast, computer algorithms work on an abstraction $\mathbf{x}$ of a real-world object o. Without loss of generality $\mathbf{x}$ is a p-dimensional vector, where each dimension i is interpreted as a value of a feature x_i of the real-world object o. The process of deriving $\mathbf{x}$ from o is called model formation, denoted as α, $\alpha : O \to X$. X comprises the feature vectors of the population; it constitutes a multiset, implying the identity $|O| = |X|$ and preserving in X the class distribution of O. The (unknown) function c maps X onto the classes in Y; c is called target concept or ideal classifier. The task of an inductive learner is to build an approximation h of the target concept c, exploiting only information contained in a sample S of training examples $\{(\mathbf{x}, c(\mathbf{x}))\}$. The function h is called a hypothesis for the target concept; it is characterized by its generalization error, $err(h)$, also called prediction error, real error, or true error [7], [24], [27], [19]. $err(h)$ can be defined as the probability of wrong classification:

$$P(h(\mathbf{x}) \neq c(\mathbf{x}))$$

The minimization of this error is the ultimate goal of a classification task. $err_S(h)$ is called test error if S is not used for the construction of h by a learner.

$$err_S(h) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} loss_{0/1}(h(\mathbf{x}), c(\mathbf{x})),$$

where $loss_{0/1}(h(\mathbf{x}), c(\mathbf{x}))$ is 0 if $h(\mathbf{x}) = c(\mathbf{x})$, and 1 otherwise. The learning algorithm selects a hypothesis h from the space H of possible hypotheses, and hence H defines a lower bound for $err(h)$. This lower bound is denoted as $err(h^*)$

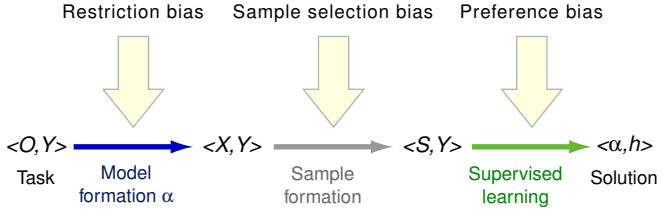

Figure 1. Illustration of a classification task $\langle O, Y\rangle$ and its machine-based solution. The model formation function α associates real-world objects with feature vectors. A restriction bias is introduced at model formation time, and other biases may be introduced within subsequent steps.

and quantifies the expected difference between an optimum hypothesis $h^* \in H$ and the target concept c:

$$err(h^*) := \min_{h \in H} err(h)$$

$err(h^*)$ is called structural bias or model bias [7]. Note that the learner itself can introduce a so-called preference bias, and that a sample selection bias may be introduced during the formation of S (see Figure 1). Choosing between different model formation functions $\alpha_1, \ldots, \alpha_m$ means to choose between different representations $X_1, \ldots, X_m$ along with different hypotheses spaces $H_{\alpha_1}, \ldots, H_{\alpha_m}$, and hence to introduce a more or less rigorous structural bias. If training data is plentiful, the best model can be found by minimizing $err_S(h)$ against the different representations. However, if training data is scarce, we even may prefer α_i over α_j although the former is outperformed under S:

$$err_S(h^*_{\alpha_i}) \;>\; err_S(h^*_{\alpha_j}),$$

where $h^*_{\alpha_i} \in H_{\alpha_i}$, $h^*_{\alpha_j} \in H_{\alpha_j}$, and $i \neq j$. I.e., we introduce a higher restriction bias than suggested by S, accepting a higher error err_S, but still expecting a lower generalization error:

$$err(h^*_{\alpha_i}) \;<\; err(h^*_{\alpha_j})$$

We call the model under α_i to be more robust than the model under α_j, or, to be the robust model for the task $\langle O, Y\rangle$.

III. Case Study I: Text Categorization

The following experiments evaluate the behavior of the generalization error err, the sample error err_S, and the relation between err and err_S. In our study we vary vector space retrieval models by employing different functions α while keeping the inductive learner unchanged. This way, the difference in the retrieval model's robustness is reflected by the classification performance of the obtained solutions. The inductive learner in the setting is an SVM with a linear kernel [8], [25] and $\langle O, Y\rangle$ is a text categorization task on the Reuters Corpus Volume RCV1 [13]. We consider the corpus in its entirety in the role of the population O. The set Y of class labels is defined by the four top-level categories in RCV1: corporate/industrial, economics, government/social, and markets. The corpus contains $|O| = 663768$ uniquely classified documents whose distribution is shown in Table I.

Table I
Document distribution in the top-level categories of RCV1.

Top-level category	Number of documents
corporate/industrial	292 348
economics	51 148
government/social	161 523
markets	158 749

The different model formation functions α_i yield different object representations X_i. Let S be a sample, drawn i.i.d. from X_i, with $|S| = 800$. The extreme ratio of $\upsilon = 0.0012$ between the sizes of S and X_i reflects a typical information retrieval situation as it is encountered in the real world; in fact, $\upsilon = 0.0012$ may still be considered as optimistic.

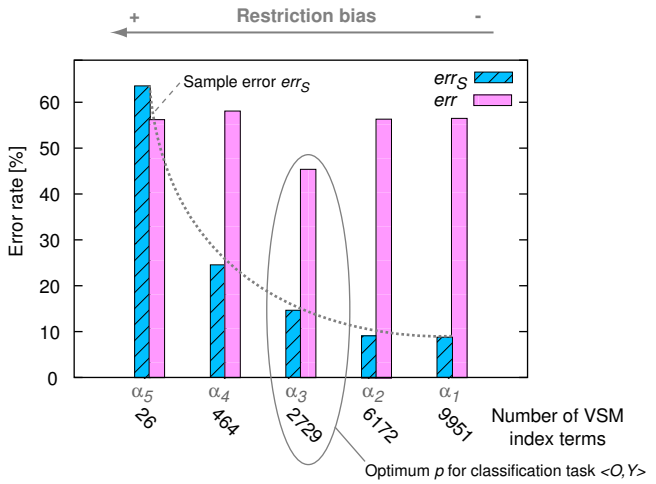

Figure 2. Cross-validated error estimates (hashed bars) and generalization errors (plain bars) for five different solutions $\langle \alpha_i, h \rangle$ of $\langle O, Y \rangle$. The $\alpha_1, \ldots, \alpha_5$ affect H by the employed number of index terms. The learning approach is a SVM with a linear kernel; the training sample S contains 800 examples.

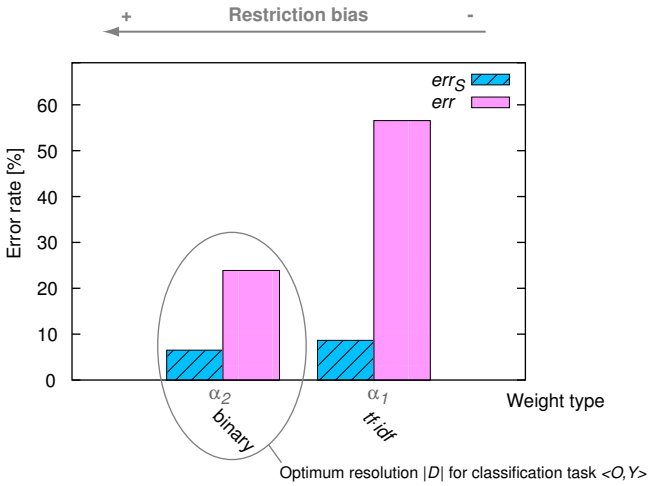

Figure 3. Cross-validated error estimates (hashed bars) and generalization errors (plain bars) for two different solutions $\langle \alpha_i, h \rangle$ of $\langle O, Y \rangle$. α_1 and α_2 affect H by using a different granularity for the feature variable domains. Again, the learning approach is an SVM, and the sample size $|S|$ is 800.

A. Experiment 1

For a document $o \in O$ the function $\alpha_i(o)$ computes a vector space model, where $i = 1, \ldots, 5$, is associated with a certain number p of used index terms (see the x-coordinate in Figure 2 for the actually chosen values for p). The reduction of the feature number p is achieved by introducing prefix equivalence classes for the index terms: the term weights of words that start with the same letter sequence are counted to the $tf{\cdot}idf$-value of the same index term. In our experiments the prefix length is varied between 1 and 10. The plot in Figure 2 reveals, as expected, that the cross-validated error estimates (hashed bars) increase with the impairment of the vector space model. Interestingly, this monotonic behavior cannot be observed for the generalization error: for $p = 2729$ the value becomes minimum, a further reduction of p leads to underfitting. To understand the importance of this result, recall that the generalization error cannot be observed in the information retrieval practice. Put another way, the best solution for $\langle O, Y \rangle$ can be missed easily, since only the analysis results with respect to S are at our disposal.

B. Experiment 2

We now modify α_i by coarsening the feature domain D of the index terms, going from the $tf{\cdot}idf$ retrieval model to the boolean retrieval model. Figure 3 shows the results for the two extremal α_i. Observe that the cross validated errors for both retrieval models are pretty close to each other; in fact, they differ only by one percent. Hence, there is a high risk to select the "wrong" model. This is particularly crucial here since the difference between in the achievable generalization errors is enormous.

That the err_S statistic may lead one astray—even if it relies on cross validation—has been observed and discussed before [18]. We would like to point out that our analyses go beyond these (and similar results): Firstly, we report on realistic information retrieval experiments and the current practice of experiment implementation and experiment evaluation. Secondly, and presumably more important, the focus of our analyses is on the impact of the model formation function α. The above analysis is distantly related to the feature selection problem, which also can eventuate some bias on the estimates of classifier parameters. This kind of bias is also called "feature subset selection bias" or simply "selection bias" [21].

IV. Case Study II: Genre Identification

Web Genre Identification is a prime example for a IR classification task. We begin by explaining how we construct a robust genre retrieval model. Then we report on an experiment for uncovering robustness characteristics when err is incalculable. The genre of a document provides information related to the document's form, purpose, and intended audience. In order to identify a documents genre we need a solution for the classification task $\langle O, Y \rangle$ where O is a set of documents and Y, $Y = \{1, \ldots, k\}$ is a set of genre class labels, also called genre palette. Current Web genre retrieval models achieve a low sample error but do not generalize at Web scale. Though the genre paradigm attracted much interest as positive or negative filter criterion for Web search results, automatic genre identification could not convince in the Web retrieval practice by now.

The development of genre retrieval models is an active research field with several open questions, and only little is known concerning the robustness of a retrieval model. Early work dates back to 1994, where Karlgren and Cutting presented a feasibility study for a genre analysis based on the Brown corpus [9]. Later on several publications followed investigating different corpora, using more intricate or less complex retrieval models, stipulating other concepts of genre, or reporting on new applications. The sizes of existing corpora varies between 200 and 2500 documents sorted into 3 to 16 genres [12], [14], [3], [6]—while there are 20-50 billions of

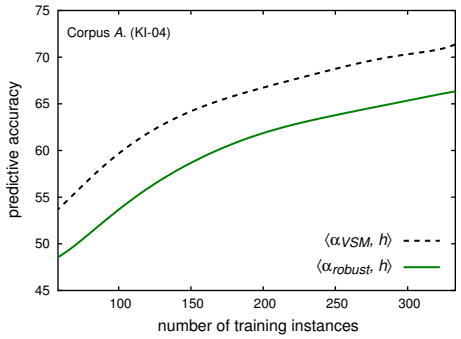

Figure 4. Predictive accuracy of the classification solutions $\langle \alpha_{VSM}, h \rangle$ and $\langle \alpha_{robust}, h \rangle$, depending on the size of the training set, which is drawn from corpus A (KI-04), estimated by a test sample of corpus A.

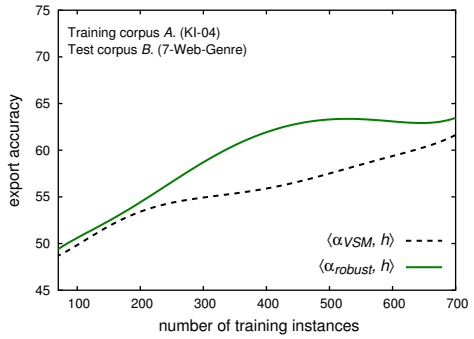

Figure 5. Export accuracy of the classification solutions $\langle \alpha_{VSM}, h \rangle$ and $\langle \alpha_{robust}, h \rangle$, depending on the size of the training set, which is drawn from corpus A (KI-04), estimated by a test sample of corpus B (7-Web-Genre).

indexed Web documents.

A. A Robust Genre Retrieval Model

For our robust genre retrieval model we introduce the features "Maximum Term Concentration" and "Gini Coefficient" based on genre-specific core vocabularies and concentration measures for the model formation. Let T_y denote the core vocabulary specific for the genre $y \in Y$. The terms in T_y should be both predictive and frequent for y. Terms with such characteristics can be identified in Y with approaches from topic identification research, in particular Popescul's method and the weighted centroid covering method [10], [11]. In order to mine genre-specific core vocabulary both methods must be adapted: they do not quantify whether a term is *representative* for y; a deficit, which can be repaired, see [23]. In the simplest case, the relation between T_y and a document o can be quantified by computing the fraction of o's terms from T_y, or by determining the coverage of T_y by o's terms. However, if genre-specific vocabulary tends to be concentrated in certain places on a Web page, this characteristic is not reflected by the mentioned features, and hence it cannot be learned by a classifier h. Examples for Web pages on which genre-specific core vocabulary is unequally distributed: private home pages (e.g. address vocabulary), discussion forums (e.g. terms from mail headers), and non-personal home pages (e.g. terms related to copyright and legal information). The following two statistics quantify two different vocabulary concentration aspects:

1) *Maximum Term Concentration.* Let $o \in O$ be represented as a sequence of terms, $s = \{w_1, \ldots, w_m\}$, and let $W_i \subset s$ be a text window of length l in s starting with term i, say, $W_i = \{w_i, \ldots, w_{i+l-1}\}$. A natural way to measure the concentration of terms from T_y in different places of s is to compute the following function for different W_i:

$$\kappa_{T_y}(W_i) = \frac{|W_i \cap T_y|}{l}, \qquad \kappa_{T_y}(W_i) \in [0,1]$$

The overall concentration is defined as the maximum term concentration:

$$\kappa^*_{T_y} = \max_{W_i \subset d} \kappa_{T_y}(W_i), \qquad \kappa^*_{T_y} \in [0,1]$$

2) *Gini Coefficient.* In contrast to the κ_{T_y} statistic, which quantifies the term concentration strength within a text window, the Gini coefficient can be used to quantify to which extent genre-specific core vocabulary is distributed unequally over a document. Again, let W_i be a text window of size l sliding over s. The number of genre-specific terms from T_y in W_i is $\nu_i = |T_y \cap W_i|$. Let A denote the area between the uniform distribution line and the Lorenz curve of the distribution of ν_i, and let B denote the area between the uniform distribution line and the x-axis. The Gini coefficient is defined as the ratio $g = A/B, g \in [0,1]$. A value of $g = 0$ indicates an equal distribution; the closer g is to 1 the more unequal ν_i is distributed.

B. Analysis

We show the influence of model robustness in this experiment while we test two different robust IR classification solutions on documents sampled from a different corpus. Our analysis is based on the Web genre corpora "KI-04" [14] with the 8 Web genre classes article, discussion, shop, help, personal home page, non-personal home page, link collection and download, denoted as A, and the "7-Web-Genre" [17] with the genres blog, listing, eshop, home page, FAQ, search page and online newspaper front page, denoted as B. We estimated the predictive accuracy ($= 1 - err_S$) of a classification solution by cross validation on corpus A, i.e. all documents for compiling the classification solutions and estimating err_S come from A. Additionally, for these compiled classification solutions we estimated err_S with documents from B and the genres "listing" (mapped to "link collection"), "eshop" (mapped to "shop") and "home page" (mapped to "personal home page") whereas we call $1 - err_S$ export accuracy.

Our assumption is that genre corpora may be representative for the population but are biased because of one or more of the following reasons:

1) The corpus is compiled by a small group of editors who share a similar understanding of genre.
2) The editors introduce subconsciously an implicit correlation between topic and genre.
3) The editors collect their favored documents only.

4) The editors rely on a single search engine whose ranking algorithm is biased towards a certain document type.

A consequence is that the cross-validated error estimate provides no reliable means to prefer one genre classifier over another. This fact is demonstrated in the following, and it is also shown that a robust model (a model with higher restriction bias) may be inferior on a test set S but will do a better job with respect to generalization. The presented effects are not a consequence of overfitting but of the extreme size ratio υ between S and the World Wide Web. The following model formation functions α_{VSM} and α_{robust} are employed:

1) α_{VSM} computes $\mathbf{x}$ with a simple vector space model using $tf \cdot idf$ term weighting scheme, comprising about 3500 features.
2) α_{robust} uses the proposed concentration measures, maximum concentration and Gini coefficient of core vocabulary distributions, impose one feature (= one dimension in $\mathbf{x}$) per genre class $y \in Y$ and measure. We enriched the representation by part-of-speech features. The entire model comprises 98 features.

Again, an SVM determines the hypothesis h in the solutions $\langle \alpha_{VSM}, h \rangle$ and $\langle \alpha_{robust}, h \rangle$. Observe that the solution $\langle \alpha_{VSM}, h \rangle$ achieves a significantly higher predictive accuracy than $\langle \alpha_{robust}, h \rangle$ (see Figure 4); with respect to the sample size both show the same consistency characteristic. We explain the high predictive accuracy of $\langle \alpha_{VSM}, h \rangle$ with its higher training data sensibility, which is beneficial in homogeneous corpora. Even by using a cross validation the predictive accuracy and the export accuracy will considerably diverge.

The impact of model robustness is unveiled when analyzing the export accuracy, which drops significantly (by 21%) for $\langle \alpha_{VSM}, h \rangle$ (see Figure 4 and Figure 5). For $\langle \alpha_{robust}, h \rangle$ the export accuracy drops only by 8%. The better performance of $\langle \alpha_{robust}, h \rangle$ is a consequence of its small number of features, which is more than an order of magnitude smaller compared to $\langle \alpha_{VSM}, h \rangle$.

V. Conclusion

We argue to identify the restriction bias that is introduced within the model formation process with the robustness of the resulting retrieval model. In two case studies we analyze the impact of the restriction bias on the retrieval performance, and we observe that the idea of robust models is highly usable: it captures effects on the generalization error that cannot be attributed to properties of the inductive learner nor to the hypothesis structure. Robust models are a means to reduce the overfitting problem for retrieval tasks where the ratio between the training sample and the set of unseen documents is extremely small.

References

[1] M. Amini and P. Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 105–112, 2002.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the vapnik-chernovenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[3] E. Boese and A. Howe. Effects of web document evolution on genre classification. In *CIKM'05*. ACM Press, Nov. 2005.

[4] R. Caruana, A. Munson, and A. Niculescu-Mizil. Getting the most out of ensemble selection. In *Sixth International Conference on Data Mining*, pages 828–833, 2006. IEEE Computer Society.

[5] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602, 2008.

[6] L. Freund, C. Clarke, and E. Toms. Towards genre classification for IR in the workplace. In *1st international conference on Information interaction in context*, pages 30–36, 2006. ACM.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[8] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

[9] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *15th. International Conference on Computational Linguistics, Coling 94*, pages 1071–1075, Kyoto, Japan, 1994.

[10] D. Lawrie, W. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–357, 2001.

[11] D. J. Lawrie and W. Croft. Generating hierarchical summaries for web searches. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 457–458, 2003.

[12] Y. Lee and S. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. 2002.

[13] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[14] S. Meyer zu Eißen and B. Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. In S. Biundo, T. Frühwirth, and G. Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, Berlin Heidelberg New York, Sept. 2004. Springer.

[15] T. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[16] N. Oza and K. Tumer. Classifier ensembles: Select real-world applications. *Inf. Fusion*, 9(1):4–20, 2008.

[17] M. Santini. Common criteria for genre classification: Annotation and granularity. In *ECAI-Workshop TIR-06*, Riva del Garda, Italy, 2006.

[18] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10(2):153–178, 1993.

[19] C. Schaffer. A conservation law for generalization performance. In *ICML*, pages 259–265, 1994.

[20] V. Sindhwani and S. Keerthi. Large scale semi-supervised linear SVMs. In *29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 477–484, 2006.

[21] S. Singhi and H. Liu. Feature subset selection bias for classification learning. In *23rd international conference on Machine learning*, pages 849–856, New York, NY, USA, 2006. ACM.

[22] K. Sridharan and S. Kakade. An information theoretic framework for multi-view learning. In R. Servedio, T. Zhang, R. Servedio, and T. Zhang, editors, *COLT*, pages 403–414. Omnipress, 2008.

[23] B. Stein and S. Meyer zu Eißen. Retrieval Models for Genre Classification. *Scandinavian Journal of Information Systems (SJIS)*, 20(1):91–117, 2008.

[24] G. Valentini and T. Dietterich. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *The Journal of Machine Learning Research*, 5:725–775, 2004.

[25] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, New York, 1982.

[26] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2000.

[27] D. Wilson and R. Randall. Bias and the probability of generalization. In *International Conference on Intelligent Information Systems (IIS '97)*, page 108, Washington, DC, USA, 1997. IEEE Computer Society.