This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

Upper and Lower Tight Error Bounds for Feature Omission with an Extension to Context Reduction

Ralf Schlüter, Eugen Beck, Hermann Ney
Human Language Technology and Pattern Recognition, Computer Science Department
RWTH Aachen University, Ahornstr. 55, 52056 Aachen, Germany
{schlueter,beck,ney}@cs.rwth-aachen.de

Abstract—In this work, fundamental analytic results in the form of error bounds are presented that quantify the effect of feature omission and selection for pattern classification in general, as well as the effect of context reduction in string classification, like automatic speech recognition, printed/handwritten character recognition, or statistical machine translation. A general simulation framework is introduced that supports discovery and proof of error bounds, which lead to the error bounds presented here. Initially derived tight lower and upper bounds for feature omission are generalized to feature selection, followed by another extension to context reduction of string class priors (aka *language models*) in string classification. For string classification, the quantitative effect of string class prior context reduction on symbol-level *Bayes* error is presented. The tightness of the original feature omission bounds seems lost in this case, as further simulations indicate. However, combining both feature omission *and* context reduction, the tightness of the bounds is retained. A central result of this work is the proof of the existence, and the amount of a statistical threshold w.r.t. the introduction of additional features in general pattern classification, or the increase of context in string classification beyond which a decrease in *Bayes* error is *guaranteed*.

1 Introduction

Due to the discontinuity of the local *Bayes* error probability, the computation of the *Bayes* error analytically often remains intractable, even if the underlying distributions are available. This motivates investigations into error bounds that do not suffer from this problem. In information theory, many bounds exist for the *Bayes* error with classification error (zero/one) cost function. Examples for this are the *Chernoff* bound [4], the *Lainiotis* bound [15], or the nearest neighbor bound [8]. However, these general bounds do not cover more specific modelling issues, like feature omission/selection [21], or more complex problems like the effect of context in sequence classification.

For statistical classification, it is well known that the addition of further features potentially improves the classification accuracy. The partial motivation of this work is to provide a better fundamental understanding of feature omission/selection by quantifying the effect of the inclusion of further features on the accuracy of the underlying statistical classifier by way of upper and lower error bounds. Corresponding analyses usually are done empirically on specific classification tasks. However, empirical findings cannot readily be generalized to new tasks and domains, as they usually are subject to specific conditions and modelling choices. In contrast, error bounds provide a general means to quantify the effect of systematic modelling decisions like feature omission/selection, or the choice of symbol prior context length.

In string classification applications like automatic speech recognition [7], statistical machine translation [13], or printed or handwritten character recognition [22], classification refers to string classes, where each class represents a string (or sequence) of symbols (words, characters, phonemes, etc.). Error bounds derived for single symbol classification can be applied to symbol string classification, also. However, this implies error measures based on the symbol *string* error, instead of a symbol-wise error definition.

Traditionally, symbol string prior probabilities are modelled using *Markov*-chains. The corresponding so-called *language models*, providing symbol probability distributions in symbol sequence context of varying length, are an important aspect of many natural language processing tasks. Language modelling paradigms may be based on smoothed n-gram counts [12], or on multi-layer perceptrons [2]. Empirically, using longer context improves perplexity and, up to some extent, also the symbol-level accuracy [20] of string classifiers. For automatic speech recognition (ASR), perplexity and symbol-level (e.g. word) error rate empirically are expected to be connected by a power law [16, p. 186], [11]. Currently unpublished results obtained by the authors' working group recently confirmed these findings for a large variety of perplexities, language modelling approaches and corresponding context lengths in ASR [19, pp. 47–48]. Nevertheless, to the best of the authors' knowledge, currently no *formal* relation is known between the order of the context used in the language model and the accuracy of a resulting classifier. The same applies for handwriting recognition, as it differs from automatic speech recognition only in the form of the input and its preprocessing. Also, other highly rele-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

vant string classification tasks like machine translation [13] can be expected to benefit from a better understanding of the context dependency involved. For example, even for strong neural machine translation systems, the additional use of language models still provides further improvements in translation performance [9].

In [5], an upper bound on the Bayes error of a string classifier using two classes is described. The bound is a function of the class prior and requires a restriction on the class conditional observation distribution. In [17], two bounds on the accuracy difference between a Bayes single symbol classifier and a model classifier (e.g. one learned from data) are presented. These bounds are based on the squared distance and the Kullback-Leibler divergence [14]. The Kullback-Leibler based bound was later tightened and extended to the general class of f-divergences [6] in [18]. The simulation techniques described here are an advancement of the techniques applied in [18].

To find corresponding bounds, an empirical statistical simulation approach was used, to judge, if a measure is a potential candidate for an error bound. This approach has the advantage that the most demanding task, i.e. finding a formal proof, is attempted only if one can be reasonably certain about the existence of a bound for a respective statistical measure. It is also made easier by having a hypothesis for the functional form of the bound. Furthermore, this general approach can lead to tight bounds, as the simulations may provide those distributions, for which the bounds are satisfied with equality.

Using the proposed simulation approach, tight upper and lower bounds on the accuracy difference for the case of feature omission are presented in this work, together with corresponding analytical proofs. It might be of interest that the nearest-neighbor bound does resemble a part of the lower bound presented here. Subsequently, these proofs are extended to general feature selection, which is relevant for any pattern classification task, like e.g. document classification, detection of manufacturing errors in industrial automation, biometric recognition for access control, etc. [10]. Also, the bounds derived for feature omission/selection are transferred to the seemingly unrelated problem of modeling context dependence in string classification, for which also upper and lower error bounds for the symbol-wise *Bayes* error are derived.

The error bounds presented in this work quantitatively reveal a general property of the *Bayes* error. Considering the maximum operation that is needed to compute the *Bayes* error, it can be expected that small changes in the underlying distributions do not necessarily have an impact on the *Bayes* error, provided that the maxima of the class posterior distribution are not affected. The error bounds presented here do not only confirm a corresponding threshold on a statistical measure of the underlying distributions in form of the *Gini* difference introduced here. Also a generally effective threshold is quantified, beyond which changes in the underlying distributions under feature omission/selection and/or context reduction induce changes in *Bayes* error.

The remainder of this work is organized as follows. Sec. 2 gives an overview of the simulation approach applied here to discover error bounds and support their proof. Sec. 3 introduces feature omission, and Sec. 4 shows corresponding exemplary simulations. In Sec. 5, proofs of error bounds for feature omission are presented, which are generalized to feature selection in Secs. 6 and 7. In Sec. 8 string classification is introduced, and Sec. 9 extends the bounds derived in Sec. 5 to context reduction in string classification, followed by corresponding simulations in Sec. 10. Finally, in Sec. 11, error bounds for the combined case of feature omission and context reduction are derived, including simulations for this case. Sec. 12 concludes this work with a final discussion and outlook on further research in this direction. Parts of this work, excluding proofs, and excluding both the more general case of feature selection, and the combined case of feature omission and context reduction were presented earlier in [1].

2 Simulation Approach to Error Bound Discovery

To find corresponding bounds, an empirical statistical simulation approach was used, as described in Fig. 1. To judge, if a measure is a potential candidate for an error bound, millions of distributions were simulated. The simulation here does not require to sample from some distribution, but aims at generating distributions, itself, without assuming further statistical constraints. Rather, the aim is to generate random walks through the space of all possible distributions with the aim to fill the two dimensional space spanned by an accuracy measure and a corresponding statistical measure that is tested for its potential to bound the accuracy measure in question. If an investigated statistical measure did not exhibit a suitable bounding behavior on the accuracy measure, it was discarded. If simulations indicated a potential bound, those distributions which occurred on a (hypothesized) bound were parametrized, which also helped to conjecture

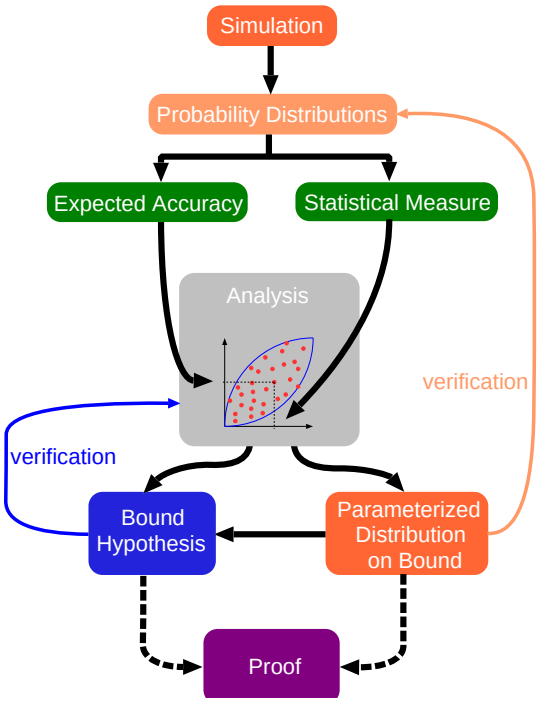

Fig. 1. Sketch of the simulation approach proposed here to support the discovery of novel error bounds.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

the functional form of a bound. These in turn also were used to verify conjectured bounds, i.e. to check, if by perturbing distributions on the bound it is not violated, i.e. falsified. If simulations clearly indicated bounds, the simulations were followed by attempts to find formal proofs. This simulation approach also contributed to the discovery of further error bounds [18].

3 Feature Omission: Definitions

Let $\mathcal{C}$ be the finite set of classes and $\mathcal{X}$ be the set of observations. For simplicity $\mathcal{X}$ is assumed to be finite. Then a classification task maps an observation $x \in \mathcal{X}$ to a class $c \in \mathcal{C}$. Let $pr(c, x)$ be the probability mass function of the true joint distribution. Then the accuracy of a Bayes classifier is:

$$A^* = \sum_x \max_c \left\{ pr(c) pr(x|c) \right\}$$

In contrast to this, omission of feature x leads to the static, prior-only classifier, whose accuracy is defined by:

$$\tilde{A} = \max_c pr(c).$$

To measure the effect of feature omission, the accuracy difference between these classifiers is considered:

$$\Delta A = A^* - \tilde{A} = \sum_x \max_c \left\{ pr(c) pr(x|c) \right\} - \max_c pr(c) \quad (1)$$

Now, the following statistical measure is defined, which will be used in the following. The *Gini* difference is defined as follows:

$$\begin{aligned} \Delta G &= \sum_x pr(x) \sum_c pr(c|x)^2 - \sum_c pr(c)^2 \\ &= \sum_x pr(x) \sum_c \left[pr(c|x) - pr(c) \right]^2 \end{aligned} \quad (2)$$

The term *Gini* difference is chosen here, as it is similar to the *Gini* criterion, as, e.g. used in decision tree learning. In [8], the minuend and subtrahend of the *Gini* difference are known as Bayesian distance.

4 Simulations for Feature Omission

In order to determine the exact relation between the *Gini* difference and the accuracy difference, millions of distributions were simulated to calculate their values of the *Gini* and the accuracy difference. In Fig. 2, the results of such a simulation for $C = |\mathcal{C}| = 8$ classes and a set of $|\mathcal{X}| = 16$ different discrete observations is presented. Note that the axes are using normalized versions of the accuracy difference and the *Gini* difference. An upper and a lower bound for the accuracy difference as a function of the *Gini* difference is visible. This type of simulation also was performed for other combinations of $C = |\mathcal{C}|$ and $X = |\mathcal{X}|$ and from these results the following upper/lower bounds were hypothesized empirically by extensive analysis of the simulations:

$$\frac{C}{C-1}\Delta A \leq \sqrt{\frac{C}{C-1}\Delta G}$$

$$\frac{C}{C-1}\Delta A \geq \begin{cases} 0 & \text{if } 0 \leq \frac{C}{C-1}\Delta G \leq \frac{1}{4} \\ \frac{C}{C-1}\Delta G - \frac{1}{4} & \text{if } \frac{1}{4} \leq \frac{C}{C-1}\Delta G \leq \frac{3}{4} \\ 1 - \sqrt{1 - \frac{C}{C-1}\Delta G} & \text{if } \frac{3}{4} \leq \frac{C}{C-1}\Delta G \leq 1 \end{cases}$$

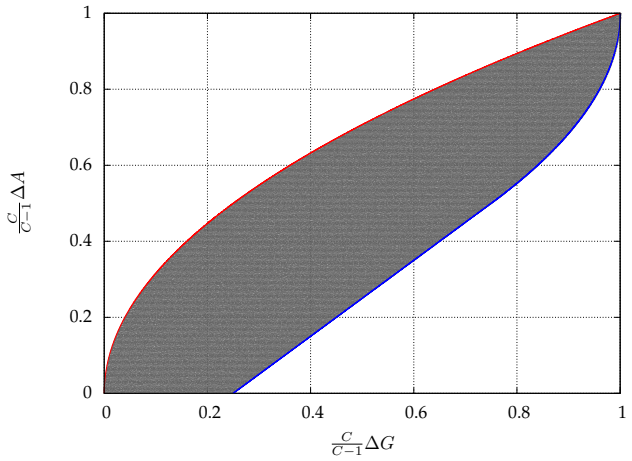

Fig. 2. Simulation results for feature reduction with $C = 8$ classes and $|X| = 16$ observations. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

The last segment of the lower bound (case $\frac{C}{C-1}\Delta G \geq \frac{3}{4}$) can also be written as:

$$\Delta G \leq 2\Delta A - \frac{C}{C-1}\Delta A^2$$

In Subsec. 5.3 we will show that the proof of the tightness requires certain relations between the cardinalities of the set of classes and the set of discrete observations, C and $|X|$ are implied. For the tightness of the linear mid-section of the lower bound, $|X| > C$ is required. On the other hand, for the tightness of the upper bound and the right section of the lower bound only $|X| \geq C$ is required. A simulation shows that for the simplest case of $|X| = 2$ and $C = 3$ indeed both the upper bound, and the mid- and right section of the lower bound are not tight, cf. Fig. 3. In contrast to this, Fig. 4 shows a simulation of the case $|X| = C = 2$, for which only the mid-section of the lower bound is not tight, as suggested by the analytic results presented in Sec. 5.

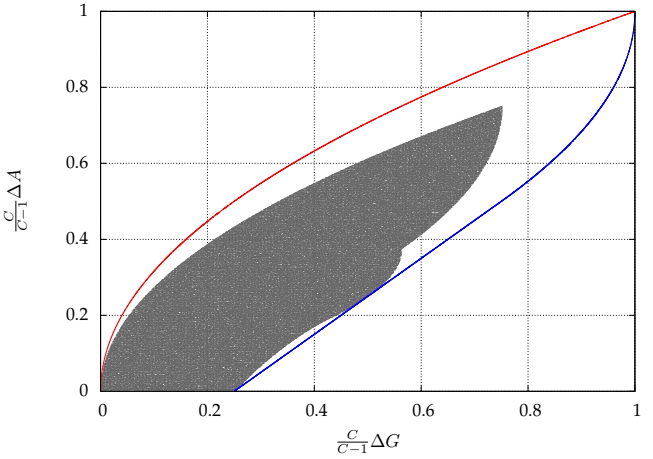

Fig. 3. Simulation results for feature reduction with $C = 3$ classes and $|X| = 2$ observations. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively. Note that in this case we have $|X| < C$, which does not fulfil the requirements for the tightness of the upper, and mid- and right section of the lower bound. This is confirmed by the simulation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

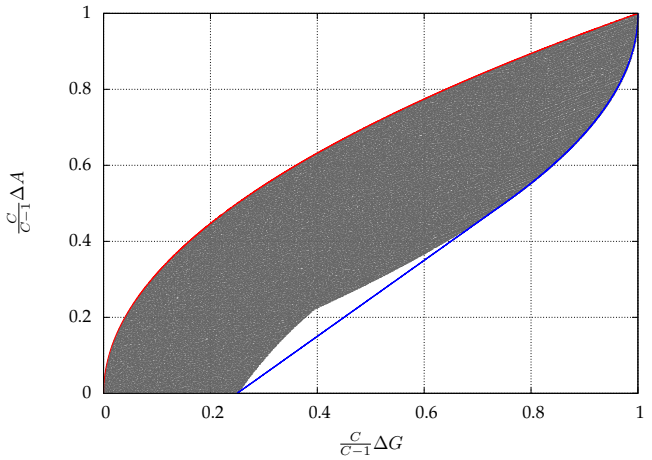

Fig. 4. Simulation results for feature reduction with $C = 2$ classes and $|X| = 2$ observations. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively. Note that in this case we have $|X| = C$, which does not fulfil the requirements for the tightness of the mid-section of the lower bound. This is confirmed by the simulation.

5 Feature Omission: Error Bounds and Proofs

In this section, sketches of the proofs for the conjectured tight lower and upper bounds are presented. Where applicable, also a generalization to the case of a continuous random variable x has been done. First, a number of lemmas are introduced. The bounds presented also are proven to be tight. The proofs for the tightness for each respective bound mainly follow the equality conditions for each derivation step. The proof of the tightness of the derived bounds finally is given by providing a set of parametrized distributions that fulfil equality for the upper bound and each segment of the lower bound that cover the complete domain.

Lemma 1. *Let $p(c|x)$ and $p(c)$ be discrete probability distributions. Then for any index c_0 the following holds:*

$$\sum_c [p(c|x) - p(c)]^2 \geq \frac{C}{C-1} \left(p(c_0|x) - p(c_0)\right)^2$$

Equality here is obtained, iff $p(c|x) - p(c) = \text{const}(x)\ \forall\ c \neq c_0$.

Proof: Apply the *Cauchy-Schwarz* inequality to the sum on the left-hand side of the inequality excluding index c_0:

$$\begin{aligned}\sum_{c\neq c_0} [p(c|x) - p(c)]^2 \geq & \frac{1}{C-1} \left(\sum_{c\neq c_0} [p(c|x) - p(c)]\right)^2 \\ & \text{(using the Cauchy-Schwarz inequality)} \\ = & \frac{1}{C-1} \left(\sum_{c\neq c_0} p(c|x) - \sum_{c\neq c_0} p(c)\right)^2 \\ = & \frac{1}{C-1} \left(p(c_0|x) - p(c_0)\right)^2 \\ & \text{(using normalization)}\end{aligned}$$

□

Lemma 2. *Let $p(c) \geq 0$, $c \in \hat{\mathcal{C}} \subseteq \mathcal{C}$ and $\sum_{c\in\hat{\mathcal{C}}} p(c) = m$. Then it holds:*

$$\sum_{c\in\hat{\mathcal{C}}} p(c)^2 \geq \frac{m^2}{|\hat{\mathcal{C}}|}$$

Proof:

$$\begin{aligned}\sum_{c\in\hat{\mathcal{C}}} p(c)^2 \geq & \left(\sum_{c\in\hat{\mathcal{C}}} p(c)\right)^2 \Big/ \left(\sum_{c\in\hat{\mathcal{C}}} 1^2\right) \\ & \text{(using Cauchy-Schwarz)} \\ = & \frac{m^2}{|\hat{\mathcal{C}}|} \quad \text{(using definition of } m\text{).}\end{aligned}$$

Equality is obtained, iff: $p(c) = \text{const}(c)\ \forall\ c \in \hat{\mathcal{C}}$. □

Lemma 3. *Let $p(c)$ be a discrete probability distribution. Then it holds:*

$$\max_c\{p(c)\} - \sum_c p(c)^2 \leq \frac{1}{4}\frac{C-1}{C}$$

Proof: Define $\lambda := \max_c\{p(c)\}$ and $c_* := \arg\max_c\{p(c)\}$. Then we have:

$$\begin{aligned}\max_c\{p(c)\} - \sum_c p(c)^2 = & \lambda - \lambda^2 - \sum_{c\neq c_*} p(c)^2 \\ \leq & \lambda - \lambda^2 - \frac{(1-\lambda)^2}{C-1} \\ & \text{(using Lemma 2, equality,} \\ & \text{iff: } p(c) = \text{const}(c)\ \forall\ c \neq c_*\ ) \\ = & \frac{1}{4}\frac{C-1}{C} - \frac{C}{C-1}\left(\lambda - \frac{C+1}{2C}\right)^2 \\ & \text{(quadratic complement)} \\ \leq & \frac{1}{4}\frac{C-1}{C}\end{aligned}$$

□

Lemma 4. *Let $p(c)$ be a discrete probability distribution. Then it holds:*

$$\sum_c p(c)^2 \leq \max_c\{p(c)\}$$

Equality is obtained, iff a non-empty subset $\mathcal{C}_0 \subseteq \mathcal{C}$ exists, for which the prior is constant $p(c) = \frac{1}{|\mathcal{C}_0|}$ for $c \in \mathcal{C}_0$ and zero otherwise: $p(c) = 0\ \forall\ c \notin \mathcal{C}_0$.

Proof:

$$\sum_c p(c)^2 \leq \sum_{c'} p(c') \max_c\{p(c)\} = \max_c\{p(c)\}$$

using the ineq. $p(c) \leq \max_{c'} p(c')$. The bound is tight, iff, for each $c \in \mathcal{C}$, either $p(c) = \max_{c'} p(c') = \text{const}(c)$ or $p(c) = 0$. □

In the following, upper and lower tight bounds are given for the *Bayes* accuracy difference in terms of the *Gini* difference. The true distributions needed to define the *Bayes* accuracy (*Bayes* error) will be denoted by pr. Tightness is proven by providing corresponding exemplary distributions. These

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

exemplary distributions will be given in terms of matrices P and vectors b, such that

$$pr(c_i|x \in \mathcal{X}_j) = P_{ij}$$
$$pr(x \in \mathcal{X}_j) = b_j$$

for proper choice of disjoint observation subsets $\mathcal{X}_j \subset \mathcal{X}$, i.e. $\mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset \; \forall \; i \neq j$, with $\bigcup_{j=1}^{J} \mathcal{X}_j \subseteq \mathcal{X}$, and classes c_i with $\{c_i|i = 1, \ldots, C\} \subseteq \mathcal{C}$. Therefore, P is a $C \times J$ probability matrix, and b a J-dimensional probability vector. Cases for given C and J also are valid for larger set sizes, assuming that the probabilities of elements of potential additional subsets of classes and/or observations are zero.

5.1 Upper Bound for Feature Omission

Theorem 1. *Given a discrete random variable c and a continuous random variable x, let $pr(c, x)$ be their true joint probability density function. Then the following tight bound holds for the accuracy difference defined in Eq. (1) and the Gini difference defined in Eq. (2):*

$$\Delta A \leq \sqrt{\frac{C-1}{C}\Delta G} = g_C(\Delta G) \quad with \quad g_C(y) = \sqrt{\frac{C-1}{C}y}$$

Proof:

$$\begin{aligned}
\Delta A &= \int_x pr(x) \left[\max_c\{pr(c|x)\} - \max_c\{pr(c)\}\right] dx \\
&\leq \int_x pr(x) \max_c\{pr(c|x) - pr(c)\} dx \\
&\quad \textit{(equality for } pr(c) = 1/C \; \forall \; c \in \mathcal{C}\textit{)} \\
&\leq \sqrt{\int_x pr(x) \big[\max_c\{pr(c|x) - pr(c)\}\big]^2 dx} \\
&\quad \textit{(using Jensen's inequality, equality iff:} \\
&\qquad \max_c[pr(c|x) - pr(c)] = \text{const}(x) \\
&\qquad\qquad \forall \; x \in \{x'|pr(x') \neq 0\}\textit{)} \\
&\leq \sqrt{\int_x pr(x) \big[\max_c\{|pr(c|x) - pr(c)|\}\big]^2 dx} \\
&\quad \textit{(absolute value of } \max \textit{ argument)} \\
&= \sqrt{\int_x pr(x) \max_c\{[pr(c|x) - pr(c)]^2\} dx} \\
&\quad \textit{(square function is monotonous increasing for positive} \\
&\quad \textit{arguments and can be drawn into maximization)} \\
&\leq \sqrt{\frac{C-1}{C} \int_x pr(x) \sum_c [pr(c|x) - pr(c)]^2 dx} \\
&\quad \textit{(using Lemma 1 with } c_0 = \arg\max_c(pr(c|x) - pr(c)) \\
&\quad \textit{equality, iff: } pr(c|x) - pr(c) = \text{const}(x) \; \forall \; c \neq c_0 \textit{)} \\
&= \sqrt{\frac{C-1}{C}\Delta G}
\end{aligned}$$

Equality is obtained for the following set of distributions:

$$pr(c_i|x \in \mathcal{X}_j) = P_{ij}$$
$$pr(x \in \mathcal{X}_j) = b_j$$

with a $C \times C$ matrix P and a C-dimensional vector b:

$$P = \begin{bmatrix} 1-\lambda & \frac{\lambda}{C-1} & \cdots & \cdots & \frac{\lambda}{C-1} \\ \frac{\lambda}{C-1} & 1-\lambda & \ddots & & \vdots \\ \vdots & \frac{\lambda}{C-1} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1-\lambda & \frac{\lambda}{C-1} \\ \frac{\lambda}{C-1} & \frac{\lambda}{C-1} & \cdots & \frac{\lambda}{C-1} & 1-\lambda \end{bmatrix} \quad \wedge \quad b = \begin{bmatrix} \frac{1}{C} \\ \frac{1}{C} \\ \vdots \\ \frac{1}{C} \end{bmatrix} \tag{3}$$

Using this set of distributions parametrized by $\lambda \in \left[0, \frac{C-1}{C}\right]$ we obtain:

$$\Delta A(\lambda) = \frac{C-1}{C} - \lambda$$
$$\Delta G(\lambda) = (1-\lambda)^2 + \frac{\lambda^2}{C-1} - \frac{1}{C}$$

with

$$\Delta A(\lambda) = \sqrt{\frac{C-1}{C} \cdot \Delta G(\lambda)}.$$

As we also have $\Delta G(\lambda = \frac{C-1}{C}) = 0$ and $\Delta G(\lambda = 0) = \frac{C-1}{C}$, $\Delta G(\lambda)$ covers the complete domain of the Gini difference, as it is a continuous function w.r.t. λ, and the intermediate value theorem applies. Therefore, this proves the tightness of the bound for the complete domain $\Delta G \in \left[0, \frac{C-1}{C}\right]$.

□

5.2 Lower Bound for Feature Omission

As discussed in Sec. 4, the lower bound of the *Gini* difference consists of three different segments. The proofs for each segment are presented individually.

5.2.1 First Segment of the Lower Bound

Theorem 2. *Assume a discrete random variable c and a continuous random variable x, and let $pr(c, x)$ be their true joint probability density function. Then the accuracy difference is non-negative in general:*

$$\Delta A = A^* - \tilde{A} \geq 0, \tag{4}$$

Furthermore, this bound is tight for

$$0 \leq \Delta G \leq \frac{C-1}{4C}.$$

Proof:

$$\begin{aligned}
A^* &= \int_x pr(x) \max_c\{pr(c|x)\} dx \\
&\geq \max_c\{\int_x pr(x) pr(c|x) dx\} \\
&= \max_c\{pr(c)\} = \tilde{A} \\
&\quad (\textit{equality, if: } \arg\max_c pr(c|x) = \text{const}(x) \;)
\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

A necessary condition for equality, i.e. $\Delta A = 0$ can be derived as follows:

$$\begin{aligned}\Delta G \leq & \Delta A + \frac{C-1}{4C} \quad \textit{(using Theorem 3)} \\ = & \frac{C-1}{4C} \quad \textit{(using } \Delta A = 0\textit{)}.\end{aligned}$$

Equality is obtained for the following set of distributions:

$$\begin{aligned} pr(c_i|x \in \mathcal{X}_j) &= P_{ij} \\ pr(x \in \mathcal{X}_j) &= b_j \end{aligned}$$

with

$$P = \begin{bmatrix} 1 & \frac{1}{C} \\ 0 & \frac{1}{C} \\ \vdots & \vdots \\ 0 & \frac{1}{C} \end{bmatrix} \qquad \wedge \quad b = \begin{bmatrix} \frac{1}{2} + \frac{\lambda}{2} \\ \frac{1}{2} - \frac{\lambda}{2} \end{bmatrix} \tag{5}$$

Using this set of distributions parametrized by $\lambda \in [0,1]$ we obtain:

$$\begin{aligned}\Delta A(\lambda) &= 0 \\ \Delta G(\lambda) &= \frac{C-1}{4C}(1-\lambda^2),\end{aligned}$$

with $\Delta G(\lambda = 1) = 0$ and $\Delta G(\lambda = 0) = \frac{C-1}{4C}$. Thus, the domain $\Delta G \in \left[0, \frac{C-1}{4C}\right]$ is completely covered, which can be concluded using the intermediate value theorem. Therefore, the bound for the first segment defined by this interval is tight. □

5.2.2 Second Segment of the Lower Bound

Theorem 3. *Given a discrete random variable c and a continuous random variable x, let $pr(c,x)$ be their true joint probability density function. Then the following holds:*

$$\Delta A \geq \Delta G - \frac{C-1}{4C}$$

This bound is tight for $\frac{1}{4}\frac{C-1}{C} \leq \Delta G \leq \frac{3}{4}\frac{C-1}{C}$.

Proof:

$$\begin{aligned}\Delta G - \Delta A = & \underbrace{\left(\max_c pr(c) - \sum_c pr(c)^2\right)}_{\leq \frac{1}{4}\frac{C-1}{C} \quad \textit{cf. Lemma 3}} \\ & - \int_x pr(x) \underbrace{\left(\max_c pr(c|x) - \sum_c pr(c|x)^2\right)}_{\geq 0 \quad \textit{cf. Lemma 4}} dx \\ \leq & \frac{1}{4}\frac{C-1}{C}\end{aligned}$$

Equality is obtained for the following set of distributions:

$$\begin{aligned} pr(c_i|x \in \mathcal{X}_j) &= P_{ij} \\ pr(x \in \mathcal{X}_j) &= b_j \end{aligned}$$

with

$$P = \begin{array}{c} \\ i=1 \\ 2 \\ \vdots \\ m \\ m+1 \\ \vdots \\ C \end{array} \begin{array}{c} \begin{array}{cccccc} j=1 & 2 & \dots & m & m+1 & m+2 \end{array} \\ \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \frac{1}{C-m+1} \\ 0 & 1 & \ddots & \vdots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & \vdots \\ 0 & \vdots & \ddots & 1 & 0 & 0 \\ 0 & \dots & \dots & 0 & \frac{1}{C-m} & \frac{1}{C-m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & \frac{1}{C-m} & \frac{1}{C-m+1} \end{bmatrix} \end{array} \tag{6}$$

$$\wedge \quad b = \begin{array}{c} \\ j=1 \\ 2 \\ \vdots \\ m \\ m+1 \\ m+2 \end{array} \begin{bmatrix} \frac{C+\lambda}{2C} \\ \frac{1}{2C} \\ \vdots \\ \frac{1}{2C} \\ \lambda\frac{C-m}{2C} \\ (1-\lambda)\frac{C-m+1}{2C} \end{bmatrix}$$

with $m \in \{1,2,\dots,C-1\}$. Using this set of distributions parametrized by $\lambda \in [0,1]$ we obtain:

$$\begin{aligned}\Delta A(\lambda, m) &= \frac{m-1+\lambda}{2C} \\ \Delta G(\lambda, m) &= \frac{m-1+\lambda}{2C} + \frac{1}{4}\frac{C-1}{C} = \Delta A(\lambda, m) + \frac{1}{4}\frac{C-1}{C},\end{aligned}$$

i.e., the distributions of this form are on the bound. Also:

$$\Delta G(\lambda = 0, \quad m = 1) = \frac{1}{4}\frac{C-1}{C} \tag{7}$$

$$\Delta G(\lambda = 1, \ \ m = C-1) = \frac{3}{4}\frac{C-1}{C} \tag{8}$$

$$\Delta G(\lambda = 0, \quad m) = \Delta G(\lambda = 1, m-1). \tag{9}$$

Hence, for $m \in \{1,2,\dots,C-1\}$ and $\lambda \in [0,1]$, the domain $\Delta G \in \left[\frac{1}{4}\frac{C-1}{C}, \frac{3}{4}\frac{C-1}{C}\right]$ is completely covered, which can be concluded using the intermediate value theorem w.r.t. λ and continuity w.r.t. m as shown in Eqs. (7-9). Therefore, the bound for the second segment defined by this interval is tight. □

5.2.3 Third Segment of the Lower Bound

Theorem 4. *Let c and x be discrete random variables and $pr(x,c)$ the associated true joint probability density function, and require*

$$\Delta G \geq \frac{3}{4}\frac{C-1}{C}. \tag{10}$$

Then the following inequality holds:

$$\begin{aligned}\Delta A \geq & \frac{(C-1) - \sqrt{(C-1)(C-1-C\cdot\Delta G)}}{C} \\ \Leftrightarrow \Delta G \leq & 2\Delta A - \frac{C}{C-1}\Delta A^2\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

Proof: First define:

$$
\begin{aligned}
g_{pr} &:= \sum_x \frac{\sum_c pr(x,c)^2}{\sum_c pr(x,c)}, \\
a_{pr} &:= \sum_x \max_c pr(c,x), \\
\lambda &:= \max_c pr(c),
\end{aligned}
$$

such that accuracy difference ΔA and Gini difference ΔG can be represented as

$$\Delta A = a_{pr} - \lambda, \tag{11}$$

$$\Delta G = g_{pr} - \sum_c pr^2(c) \tag{12}$$

Using Theorem 3 and Condition (10) then results in:

$$\Delta A \geq \frac{1}{2}\frac{C-1}{C}. \tag{13}$$

First, a lower bound on the Gini-term of the prior is provided:

$$
\begin{aligned}
\sum_c pr(c)^2 &\geq \lambda^2 + \frac{(1-\lambda)^2}{C-1} \qquad \textit{(using Lemma 2)} \\
&= \frac{C}{C-1}\left(\lambda - \frac{1}{C}\right)^2 + \frac{1}{C} \quad \textit{(quadr. compl.)}
\end{aligned}
$$

Substituting this into the definition of the Gini difference and using Eq. (11) the following inequality is obtained:

$$
\begin{aligned}
\Delta G &\leq g_{pr} - \frac{C}{C-1}\left(\lambda - \frac{1}{C}\right)^2 - \frac{1}{C} \\
&= g_{pr} - \frac{C}{C-1}\left(a_{pr} - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C}
\end{aligned} \tag{14}
$$

Now, the distribution $pr(x,c)$ is modified by redistributing probability mass for a specific $x = x_0$ from all classes to a single class c_0:

$$
q(x,c) := \begin{cases} pr(x,c) & x \neq x_0 \\ \sum_{c'} pr(x_0,c') & x = x_0 \wedge c = c_0 \\ 0 & \textit{otherwise} \end{cases} \tag{15}
$$

This allows to rewrite g_{pr} and a_{pr}:

$$
\begin{aligned}
g_{pr} &= \sum_x \frac{\sum_c q(x,c)^2}{\sum_c q(x,c)} - \sum_c pr(x_0,c) + \frac{\sum_c pr(x_0,c)^2}{\sum_c pr(x_0,c)}, \\
a_{pr} &= \sum_x \max_c q(x,c) - \sum_c pr(x_0,c) + \max_c pr(x_0,c),
\end{aligned}
$$

which can be substituted into Ineq. (14):

$$
\begin{aligned}
\Delta G \leq & g_{pr} - \frac{C}{C-1}\left(a_{pr} - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C} \\
= & g_q - \sum_c pr(x_0,c) + \frac{\sum_c pr(x_0,c)^2}{\sum_c pr(x_0,c)} \\
& - \frac{C}{C-1}\Bigg(a_q - \sum_c pr(x_0,c) + \max_c pr(x_0,c) \\
& \qquad - \Delta A - \frac{1}{C}\Bigg)^2 - \frac{1}{C} \\
= & g_q - \frac{C}{C-1}\left(a_q - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C} \\
& - \sum_c pr(x_0,c) + \frac{\sum_c pr(x_0,c)^2}{\sum_c pr(x_0,c)} \\
& \underbrace{- \frac{C}{C-1}\left(\max_c pr(x_0,c) - \sum_c pr(x_0,c)\right)^2}_{\leq 0} \\
& + \frac{2C}{C-1}\underbrace{\left(a_q - \Delta A - \frac{1}{C}\right)}_{\leq \frac{C-1}{2C},\ \textit{cf. Ineq. (13)}} \\
& \qquad \cdot \underbrace{\left(\sum_c pr(x_0,c) - \max_c pr(x_0,c)\right)}_{\geq 0} \\
\leq & g_q - \frac{C}{C-1}\left(a_q - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C} \\
& - \sum_c pr(x_0,c) + \frac{\sum_c pr(x_0,c)^2}{\sum_c pr(x_0,c)} \\
& + \sum_c pr(x_0,c) - \max_c pr(x_0,c) \\
\leq & g_q - \frac{C}{C-1}\left(a_q - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C} \\
& + \frac{\sum_c pr(x_0,c)\max_{c'} pr(x_0,c')}{\sum_c pr(x_0,c)} - \max_c pr(x_0,c) \\
= & g_q - \frac{C}{C-1}\left(a_q - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C}
\end{aligned}
$$

Note that the form of the inequality remains unchanged by replacing distribution pr by the modified distribution q. In the following, this process is repeated by successively performing the modification presented in Eq. (15) again on each modified distribution q with different pivot observations x_0', until finally

$$
\overline{q}(x,c) = \begin{cases} \sum_{c'} pr(x,c') & c = c_0(x) \\ 0 & \textit{else} \end{cases}
$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

is obtained. The values of $g_{\overline{q}}$ and $a_{\overline{q}}$ then become:

$$\begin{aligned} g_{\overline{q}} &= \sum_x \frac{\sum_c \overline{q}(x,c)^2}{\sum_c \overline{q}(x,c)} = \sum_x \frac{\left(\sum_c pr(x,c)\right)^2}{\sum_c pr(x,c)} \\ &= \sum_{x,c} pr(x,c) = 1 \\ a_{\overline{q}} &= \sum_x \max_c \overline{q}(x,c) = \sum_x \sum_c pr(x,c) = 1. \end{aligned}$$

This results in the final inequality intended for the third segment:

$$\begin{aligned} \Delta G &\leq 1 - \frac{C}{C-1}\left(1 - \Delta A - \frac{1}{C}\right)^2 - \frac{1}{C} \\ &= -\frac{C}{C-1}\Delta A^2 + 2\Delta A \end{aligned}$$

Equality is obtained for the following set of distributions:

$$\begin{aligned} pr(c_i|x \in \mathcal{X}_j) &= \delta_{ij} \\ pr(x \in \mathcal{X}_j) &= b_j \end{aligned} \tag{16}$$

with the Kronecker delta function $\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases}$ *, and*

$$b = \begin{bmatrix} 1-\lambda \\ \frac{\lambda}{C-1} \\ \vdots \\ \frac{\lambda}{C-1} \end{bmatrix}.$$

Using this set of distributions parametrized by $\lambda \in \left[\frac{1}{2}\frac{C-1}{C}, \frac{C-1}{C}\right]$ *we obtain:*

$$\begin{aligned} \Delta A(\lambda) &= \lambda \\ \Delta G(\lambda) &= -\frac{C}{C-1}\lambda^2 + 2\lambda = -\frac{C}{C-1}\Delta A(\lambda)^2 + 2\Delta A(\lambda) \end{aligned}$$

i.e., the distributions of this form are on the bound. Also, with

$$\begin{aligned} \Delta G\big(\lambda = \frac{1}{2}\frac{C-1}{C}\big) &= \frac{3}{4}\frac{C-1}{C}, \\ \Delta G\big(\lambda = \quad \frac{C-1}{C}\big) &= \quad \frac{C-1}{C}, \end{aligned}$$

the proposed domain of the third segment of the lower bound, i.e. $\Delta G \in \left[\frac{3}{4}\frac{C-1}{C}, \frac{C-1}{C}\right]$ *is completely covered, which follows from the intermediate value theorem.* □

Overall, combining the three segments, the lower bound can then be represented as:

$$\Delta A \geq f_C(\Delta G)$$

with

$$f_C(y) = \begin{cases} 0 & \text{for } 0 \leq y \leq \frac{C-1}{4C} \\ y - \frac{C-1}{4C} & \text{for } \frac{C-1}{4C} \leq y \leq \frac{3}{4}\frac{C-1}{C} \\ \frac{(C-1) - \sqrt{(C-1)(C-1-C\cdot y)}}{C} & \text{for } y \geq \frac{3}{4}\frac{C-1}{C} \end{cases}$$

5.3 Tightness of the Derived Error Bounds

The existence of the distributions on the bound provides a sufficient condition for the tightness of the derived bounds. For the linear mid-section of the lower bound, this requires $|X| > C$, cf. Eq. (6). On the other hand, for the upper bound and the right section of the lower bound only $|X| \geq C$ is required, cf. Eqs. (3) and (16). The left section of the lower bound does introduce non-trivial requirements other than $|X| \geq 2$, cf. Eq. (5), which needs to be fulfilled for a non-trivial classification problem. However, it can be expected that the cardinality of the observation set usually far exceeds the number of classes of a classification problem for realistic tasks, therefore these exceptions from the derived bounds' tightness are not considered further here.

6 Conditional to Global Bound

Assume that all distributions used to define accuracy and *Gini* difference are conditioned on some random variable $z \in \mathcal{Z}$, i.e.:

$$\begin{aligned} \Delta\hat{A}(z) &= \sum_x \max_c \left\{pr(c|z)pr(x|c,z)\right\} - \max_c pr(c|z) \\ \Delta\hat{G}(z) &= \sum_x pr(x|z) \sum_c \left[pr(c|x,z) - pr(c|z)\right]^2 \end{aligned}$$

The additional condition z does not change the proofs for the feature omission error bounds derived in Subsecs. 5.1, and 5.2. These upper and lower bounds therefore remain valid for this case, i.e.:

$$g_C\big(\Delta\hat{G}(z)\big) \geq \Delta\hat{A}(z) \geq f_C\big(\Delta\hat{G}(z)\big) \tag{17}$$

Note that the upper bounding function g_C and the lower bounding function f_C are concave and convex, respectively. Then *Jensen*'s inequality [3, p. 182] can be applied to obtain the same bounds for the global case, where condition z is marginalized:

$$\begin{aligned} \Delta\hat{A} &= \sum_z pr(z)\Delta\hat{A}(z) \\ &\leq \sum_z pr(z) g_C\big(\Delta\hat{G}(z)\big) \quad \text{(Eq.(17))} \\ &\leq g_C\big(\sum_z pr(z)\Delta\hat{G}(z)\big) \quad \text{(\textit{Jensen}'s ineq., concave case)} \\ &\leq g_C\big(\Delta\hat{G}\big) \\ \Delta\hat{A} &= \sum_z pr(z)\Delta\hat{A}(z) \\ &\geq \sum_z pr(z) f_C\big(\Delta\hat{G}(z)\big) \quad \text{(Eq.(19))} \\ &\geq f_C\big(\sum_z pr(z)\Delta\hat{G}(z)\big) \quad \text{(\textit{Jensen}'s ineq., convex case)} \\ &\geq f_C\big(\Delta\hat{G}\big) \end{aligned}$$

Nevertheless, it should be mentioned that these global bounds are not necessarily tight, depending on the definition of z.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

7 Feature Selection

The general derivation from the previous Sec. 6 can directly be applied to generalize feature omission to feature selection, i.e. the case when classification is based on a set of D features $x_1^D = x_1, \ldots, x_D$, of which a subset x_1^d with $d < D$ is selected. The accuracy difference for this case then is defined by:

$$\begin{aligned}\Delta\overline{A} &= \sum_{x_1^D} \max_c \left\{pr(c)pr(x_1^D|c)\right\} - \sum_{x_1^d} \max_c \left\{pr(c)pr(x_1^d|c)\right\} \\ &= \sum_{x_1^d} pr(x_1^d) \cdot \Delta\overline{A}(x_1^d),\end{aligned}$$

with the local accuracy difference:

$$\Delta\overline{A}(x_1^d) = \sum_{x_{d+1}^D} \max_c \left\{pr(c)pr(x_{d+1}^D|c, x_1^d)\right\} - \max_c pr(c|x_1^d).$$

Similarly, the *Gini* difference for this case is defined by:

$$\begin{aligned}\Delta\overline{G} &= \sum_{x_1^D} pr(x_1^D) \sum_c \left[pr(c|x_1^D) - pr(c|x_1^d)\right]^2 \\ &= \sum_{x_1^d} pr(x_1^d) \cdot \Delta\overline{G}(x_1^d)\end{aligned}$$

with the local *Gini* difference:

$$\Delta\overline{G}(x_1^d) = \sum_{x_{d+1}^D} pr(x_{d+1}^D) \sum_c \left[pr(c|x_{d+1}^D, x_1^d) - pr(c|x_1^d)\right]^2$$

Identifying x_{d+1}^D with x, and x_1^d with z in Sec. 6, we conclude the corresponding bounds:

$$g_C\left(\Delta\overline{G}\right) \geq \Delta\overline{A} \geq f_C\left(\Delta\overline{G}\right)$$

from the proofs given in Secs. 5 and 6.

Also this derived bound for feature selection has been confirmed by simulations. Fig. 5 shows a simulation of a classification problem with $C = 3$ classes which is reduced from two observations $x_1, x_2 \in X$ with $|X| = 4$ down to the single observation x_1. Clearly, the derived bounds are fulfilled and tightness is obtained.

For the tightness to be fulfilled, also the local bounds on the accuracy and *Gini* difference for the unselected feature $x_2 \in \mathcal{X}_2$ should be tight, which is not fulfilled for e.g. the case $C = 3$ and $|X_2| = 2$, as shown in the simulation in Fig. 6. However, this is not relevant for the cases of (quasi-)continuous features with $|X_d| \gg C \;\forall\; d = 1, \ldots, D$. The case of discrete features taking only a limited number of observations is not further investigated here, but might be treated in further work.

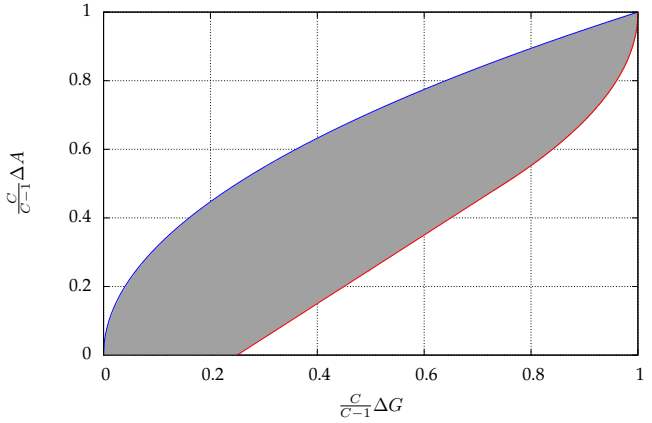

Fig. 5. Simulation results for feature selection with $C = 3$ classes and a selection of one out of two features $x_1 \in X_1$ and $x_2 \in X_2$ with $|X_{1,2}| = 4$ observations each. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

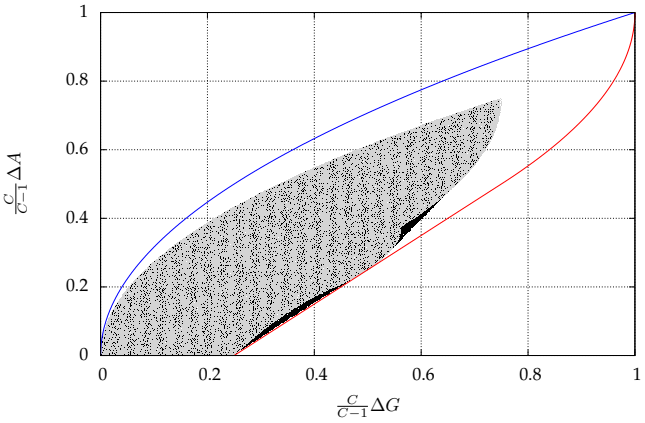

Fig. 6. Simulation results for feature selection with $C = 3$ classes and a selection of a feature $x_1 \in X_1$ with $|X_1| = 6$ observations out of two features, with the second, not selected feature $x_2 \in X_2$ being binary, i.e. $|X_2| = 2 < C$ observations. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively. Note that this case is not tight. The simulations only fill the convex hull of the area reached by the simulations of the local case, which is eqivalent to feature omission with $C = 3$ and $|X| = 2$, as shown in Fig. 3. To illustrate this, the simulation of this case from Fig. 3 is plotted on top here using a brighter gray. The narrow darkened sections show how the convex hull confines the simulations for the feature selection case when dropping the single binary feature x_2.

8 Context Reduction: Definitions

The task of string classification is to map a sequence of observations $x_1^N \in \mathcal{X}^N$ to a sequence of classes $c_1^N \in \mathcal{C}^N$. Note that here the sequence of classes and observations have the same length and no alignment problem is assumed, like in automatic speech recognition. An exemplary task, which would be represented by this model is part-of-speech tagging. Let $pr(c_1^N, x_1^N)$ be the probability mass function of the true joint distribution. Then the accuracy of a Bayes classifier at position i in the string of classes is:

$$A_i^* = \sum_{x_1^N} \max_c \left\{ \sum_{c_1^N : c_i = c} pr(c_1^N) pr(x_1^N|c_1^N) \right\}$$

For the observation model $pr(x_1^N|c_1^N)$, only local dependence is assumed:

$$pr(x_1^N|c_1^N) = \prod_{n=1}^N pr(x_n|c_n).$$

At this point, no specific assumption as to the structure of the class sequence prior distribution (language model) is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

applied. From this general class sequence prior, a position dependent class unigram can be derived:

$$pr_i(c) = \sum_{c_1^N : c_i = c} pr(c_1^N) = \sum_{c_1^i : c_i = c} pr(c_1^i)$$

Note that all positions higher than i can be marginalized. This derived unigram can be used to define a reduced context classifier, whose accuracy is defined by:

$$\tilde{A}_i = \sum_x \max_c pr_i(c, x),$$

with

$$pr_i(c, x) = pr(x|c) \cdot pr_i(c)$$

To measure the effect of the prior context, the accuracy difference between the full, bigram-based classifier A_i^*, and the reduced context classifier $\tilde{A}_i$ that is based on the derived unigram prior, is considered.

$$\Delta A_i = A_i^* - \tilde{A}_i = \sum_{x_1^N} \max_c \{pr_i(c, x_1^N)\} - \sum_x \max_c pr_i(c, x).$$

To see the connection to single symbols, the last equation is rewritten as follows:

$$\begin{aligned}\Delta A_i = & \sum_{x_i} pr_i(x_i) \sum_{y = x_1^N \setminus x_i} pr_i(y|x_i) \max_c \{pr_i(c|x_i, y)\} \\ & - \sum_{x_i} pr_i(x_i) \max_c pr_i(c|x_i) \\ = & \sum_{x_i} pr_i(x_i) \Delta A_i(x_i),\end{aligned}$$

with the local accuracy difference:

$$\Delta A_i(x_i) = \sum_{y = x_1^N \setminus x_i} pr_i(y|x_i) \max_c \{pr_i(c|x_i, y)\} - \max_c pr_i(c|x_i),$$

and the marginals in symbol position i, with $y = x_1^N \setminus x_i$:

$$\begin{aligned}pr_i(x) &= \sum_{c_1^N, x_1^N : x_i = x} pr(c_1^N) pr(x_1^N|c_1^N) \\ &= \sum_c pr_i(c) \cdot pr(x|c), \\ pr_i(c|x) &= \frac{pr_i(c) pr(x|c)}{pr_i(x)}, \\ pr_i(c, x_1^N) &= \sum_{c_1^N : c_i = c} pr(x_1^N|c_1^N) \cdot pr(c_1^N) \\ pr_i(c|x_i, y) &= pr_i(c|x_1^N) = \frac{pr_i(c, x_1^N)}{\sum_{c'} pr_i(c', x_1^N)}, \\ pr_i(c, y = x_1^N \setminus x_i | x_i) &= pr_i(c, x_1^N \setminus x_i | x_i) = \frac{pr_i(c, x_1^N)}{pr_i(x_i)}, \\ pr_i(y|x_i) &= \sum_c pr_i(c, y|x_i),\end{aligned}$$

This is very similar to the case of a single symbol classifier that maps a single (compound) observation $y \in Y$ to a single class $c \in \mathcal{C}$ compared with a classifier that only uses the prior (mapping every observation to the same class). This relation is derived formally in the following section.

9 Extension to Context Reduction in String Classification

For the case of symbol string classification, the *Gini* difference for a specific symbol position i can be rewritten as follows:

$$\Delta G_i := \sum_{x_i} pr_i(x_i) \Delta G_i(x_i)$$

with the local *Gini* difference:

$$\Delta G_i(x_i) := \sum_{y = x_1^N \setminus x_i} pr_i(y|x_i) \sum_c pr_i(c|y, x_i)^2 - \sum_c pr_i(c|x_i)^2$$

Apart from the additional condition on x_i, both the local accuracy difference $\Delta A_i(x_i)$, and the local *Gini* difference $\Delta G_i(x_i)$ effectively can be identified as single symbol cases (conditioned by x_i), such that the same upper and lower bounds apply, as derived for the feature omission case in Subsecs. 5.1, and 5.2:

$$\Delta A_i(x_i) \leq g_C\big(\Delta G_i(x_i)\big) \tag{18}$$

$$\Delta A_i(x_i) \geq f_C\big(\Delta G_i(x_i)\big) \tag{19}$$

Further identifying $y = x_1^N \setminus x_i$ with x, and x_i with z in Sec. 6, we conclude the corresponding bounds:

$$g\big(\Delta G_i\big) \geq \Delta A_i \geq f\big(\Delta G_i\big).$$

from the proofs given in Secs. 5 and 6. Nevertheless, it should be mentioned that these global bounds for the symbol string case are not necessarily tight anymore, as also pointed out for the general case covered in Sec. 6. This is confirmed by the simulations shown in the following section.

Note that this derivation can be generalized to context reduction to n-grams higher than a unigram, by assuming the classes c in the corresponding position to cover more than a single word.

10 Simulations for Context Reduction

Simulation experiments similar to the case of feature omission were performed for symbol string classification. The upper and lower bounds from the symbol case (feature omission) do hold for the string case as shown in Section 9, but the simulations suggest that in this case the bounds are not tight any more, i.e. the simulations do not reach the bound in general, as shown in Fig. 7. Note that the accuracy difference and *Gini* difference is normalized in the plots shown in this section, cf. axis labels.

Although this does not present a formal proof, it should be noted that the simulations were designed expressly to concentrate on areas not yet filled or with low density of points in the corresponding plot. We therefore safely assume that the simulations cover the complete area which can be reached.

In the following Fig. 8, the number of classes C and observations $|X|$ were proportionally reduced, upon which the space between the analytical bounds is much less filled. This might be due to the dependency between the individual position's distributions, which might be stronger for a lower number of classes and observations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

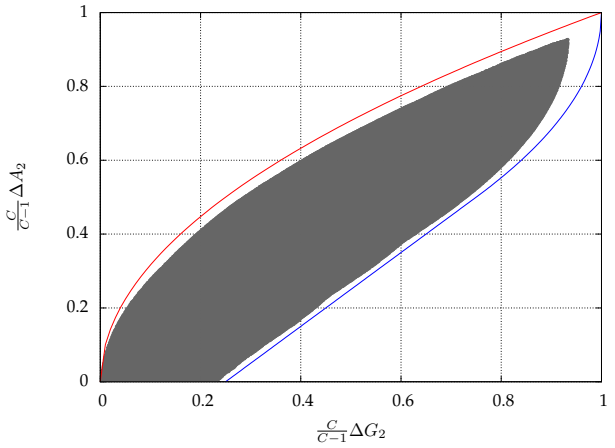

Fig. 7. Simulation results for a string classifier and context reduction with $C = 5$ classes, $|X| = 10$ observations, and sequence length $N = 3$. The accuracy and *Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

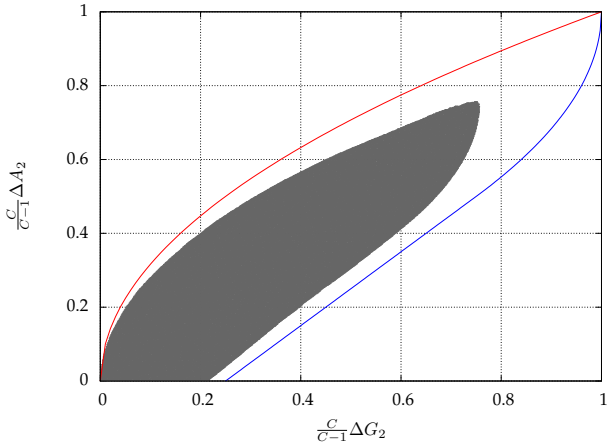

Fig. 8. Simulation results for a string classifier and context reduction with $C = 3$ classes, $|X| = 6$ observations, and sequence length $N = 3$. The accuracy and *Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

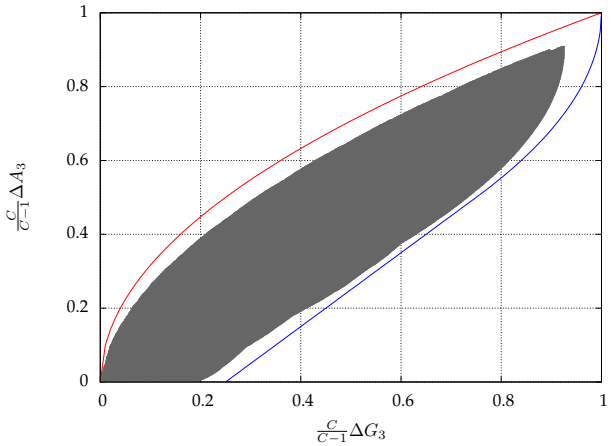

Fig. 9. Simulation results for a string classifier and context reduction with $C = 8$ classes, $|X| = 9$ observations, and sequence length $N = 5$. The accuracy and *Gini* difference was calculated at position $i = 3$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

When (slightly) increasing the length N, apparently no strong difference can be observed, as shown in Fig. 9. The number of observations here was reduced somewhat relative to the number of classes, as the complexity of the simulations apparently is exponential and the number of simulations required to obtain good filling of the space between the bounds increases strongly with the sequence length N.

11 Combined Context & Feature Reduction

The simulations for the case of context reduction in Sec. 10 show that the bound holds in case of sequences, but the tightness of the bounds seems lost. To further analyze this, we also check the case of both feature and context reduction. When both context dependency and feature dependence are dropped, then the corresponding classifier for a symbol in position i is reduced to optimizing the position dependent unigram distribution (cf. Eq. (8)), only. The accuracy for classifying a symbol in position i ignoring context and feature dependence then reduces to

$$\Delta\overline{\mathcal{A}}_i = \max_c pr_i(c).$$

The accuracy difference for this case then reduces to:

$$\Delta\mathcal{A}_i = \sum_{x_1^N} pr(x_1^N) \max_c pr_i(c|x_1^N) - \max_c pr_i(c).$$

Similarly, we define the corresponding reduced *Gini* difference:

$$\Delta\mathcal{G}_i := \sum_{x_1^N} pr(x_1^N) \sum_c pr_i(c|x_1^N)^2 - \sum_c pr_i(c)^2.$$

Similar to the local accuracy and *Gini* difference defined for context reduction only in Sec. 8, in this case the accuracy difference $\Delta\mathcal{A}_i(x_i)$, and the *Gini* difference $\Delta\mathcal{G}_i(x_i)$ effectively can be identified as single symbol cases (conditioned on symbol position i), such that the upper and lower bounds derived for feature omission in Subsecs. 5.1, and 5.2 also apply in this case. In contrast to context reduction alone, here even the tightness of the derived bounds is retained, which can be observed in the simulation results presented in Figs. 10-12. The remaining empty area between the derived tight bounds and the simulations for context reduction and feature omission in Fig. 12 can be attributed to the combinatorial complexity of the corresponding simulation using 8 classes, 9 observations, and sequences length 5.

In the simplest case, tightness is obtained for the case of the bigram distribution degenerating to a unigram. This in turn results in statistical independence of the symbol positions, and the problem falls back to feature reduction in each symbol position independently. Consequently, tightness derives from feature reduction, as proven in Sec. 5, i.e. the bounding cases from feature reduction can be substituted here per symbol position to obtain the bounding cases for applying feature reduction *and* feature omission.

The loss of tightness in case of context reduction only in Sec. 9 might be attributed to an interaction of the averaging steps bounded by applying *Jensen*'s inequality in Ineqs. (18) and (19), and constraints introduced by the language model, which occurs in each symbol position.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2788434, IEEE Transactions on Pattern Analysis and Machine Intelligence

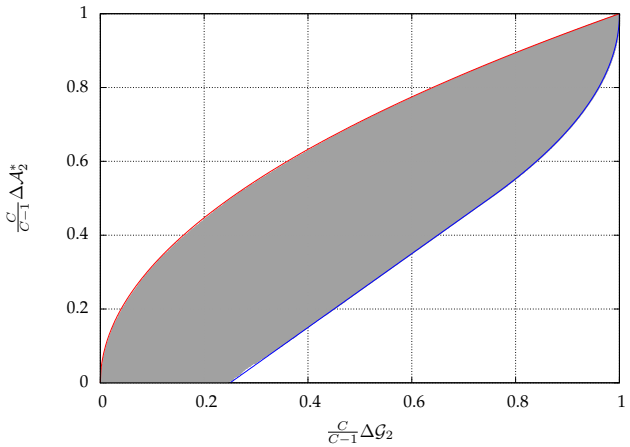

Fig. 10. Simulation results for a string classifier and both feature and context reduction with $C = 5$ classes, $|X| = 10$ observations, and sequence length $N = 3$. The accuracy and *Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

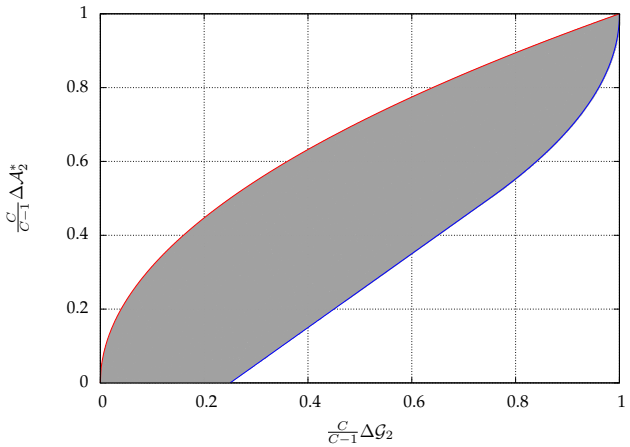

Fig. 11. Simulation results for a string classifier and both feature and context reduction with $C = 3$ classes, $|X| = 6$ observations, and sequence length $N = 3$. The accuracy and *Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

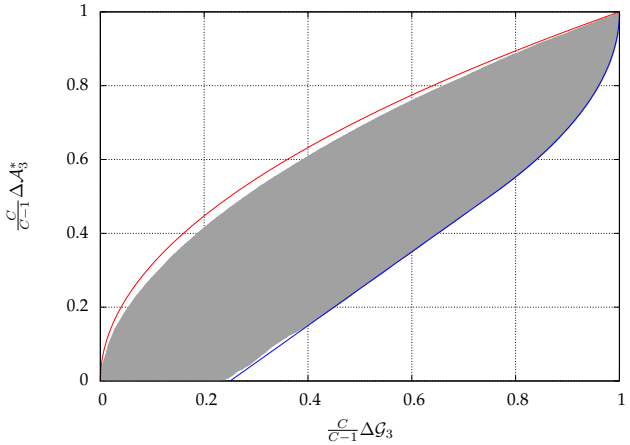

Fig. 12. Simulation results for a string classifier and both feature and context reduction with $C = 8$ classes, $|X| = 9$ observations, and sequence length $N = 5$. The accuracy and *Gini* difference was calculated at position $i = 3$. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown as lines, respectively.

12 Conclusions & Outlook

In this work, novel upper and lower tight bounds for the case of feature omission and feature selection in single symbol classification were derived. Analytical proofs for the corresponding bounds and their tightness are presented. Statistical simulations played an important role in the discovery, as well as in finding formal proofs of these bounds. Furthermore, the bounds derived for the case of feature omission were extended to the case of context-reduction of symbol string priors in symbol string classification. As further simulations suggest, the derived bounds, although being tight for the single symbol case, do not seem to be tight in general for the symbol string case. Although only a limited margin remains if the number of classes C is large enough, the simulations hint at the existence of tighter bounds for the string case, which will be investigated in further work. However, in the combined case of both feature omission *and* context reduction, the tightness of the bound is retained. When considering the normalized accuracy difference and normalized *Gini* difference, the derived bounds have the same shape for all cases. Specifically, the lower bound becomes non-zero, once the normalized *Gini* difference exceeds 0.25, i.e. a normalized *Gini* difference greater than 0.25 induces a non-zero accuracy difference. In other words, exceeding this statistical threshold of 0.25 in (normalized) *Gini* difference *guarantees* a reduction in the *Bayes* error for adding further features, or context. Even though this result might be expected qualitatively from the maximum operation in the computation of the *Bayes* error, to the best of the knowledge of the authors, the explicit threshold generally implied by the error bounds derived here for feature omission/selection in pattern classification and context reduction in string classification has not been reported before.

Acknowledgments

The authors would like to thank Tamer Alkhouli and Malte Nuhn for many insightful conversations on this topic. This work has been supported by a compute time grant on the RWTH ITC cluster. This work was partly funded under the project EU-Bridge (FP7-287658). H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694537). The work reflects only the author's view and the European Research Council Executive Agency is not responsible for any use that may be made of the information it contains.

References

[1] E. Beck, R. Schlüter, H. Ney: "Error Bounds for Context Reduction and Feature Omission," *Interspeech*, pp. 1280–1284, Dresden, Germany, Sept. 2015.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155, 2003.

[3] G. Casella, R.L. Berger, *Statistical Inference*, Duxbury Press, Belmont, California, 1990, 650 pages.

[4] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *The Annals of Mathematical Statistics*, Vol. 23, No. 4, pp. 493–507, 1952.

[5] J. Chu, "Error Bounds for a Contextual Recognition Procedure," *IEEE Transactions on Computers*, Vol. C-20, No. 10, pp. 1203–1207, Oct 1971.

[6] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten," *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, Vol. 8, pp. 85–108, 1963.

[7] Y. Dong, L. Deng: *Automatic Speech Recognition: A Deep Learning Approach*, Springer, Nov. 2014, 320 pages.

[8] P. A. Devijver, "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition," *IEEE Transactions on Computers*, Vol. C-23, No. 1, pp. 70–80, Jan. 1974.

[9] C. Gulcehre, O. Firat, K. Xua, K. Choca, Y. Bengio: "On integrating a language model into neural machine translation," *Computer Speech & Language*, Vol. 45, pp. 137–148, Sept. 2017.

[10] A. K. Jain, R. P. W. Duin, J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 4–37, Jan. 2000.

[11] D. Klakow, J. Peters, "Testing the Correlation of Word Error Rate and Perplexity," *Speech Communication*, Vol. 38, No. 4, pp. 19–28, Sept. 2002.

[12] R. Kneser and H. Ney, "Improved Backing-Off for m-gram Language Modeling," in *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 181–184, Detroit, MI, May 1995.

[13] P. Koehn *Statistical Machine Translation*, Cambridge University Press, New York, NY, 2010.

[14] S. Kullback and R. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86, 1951.

[15] D. Lainiotis, "A class of upper bounds on probability of error for multihypotheses pattern recognition (corresp.)," *IEEE Transactions on Information Theory*, Vol. 15, No. 6, pp. 730–731, Nov. 1969.

[16] J. Makhoul, R. Schwartz, "State of the Art in Continuous Speech Recognition," pp. 165–198, in D. B. Roe, J. G. Wilpon (Eds.): *Voice Communication Between Humans and Machines*, The National Academies Press, Washington, DC, 1994, 560 pages.

[17] H. Ney, "On the Relationship Between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition," in *Proc. Iberian Conference on Pattern Recognition and Image Analysis*, Springer-Verlag Berlin Heidelberg, LNCS 2652, pp. 636–645, Puerto de Andratx, Spain, Jun. 2003.

[18] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, and H. Ney, "Novel Tight Classification Error Bounds under Mismatch Conditions Based on f-Divergence," in *Proc. IEEE Information Theory Workshop*, pp. 432–436, Sevilla, Spain, Sep. 2013.

[19] R. Schlüter: "Automatic Speech Recognition based on Neural Networks," Keynote speech, Intern. Conf. on Speech and Computer (SPECOM), Budapest, Hungary, Aug. 2016. Slides available at http://www.specom2016.hte.hu/documents/1695342/3102158/02_01_Schlueter_SPECOM2016.pdf

[20] H. Schwenk, "Continuous Space Language Models," *Computer Speech & Language*, Vol. 21, No. 3, pp. 492–518, 2007.

[21] J. Tang, S. Alelyani, H. Liu: "Feature Selection for Classification: A Review," in C. Aggarwal (Ed.): *Data Classification: Algorithms and Applications*, pp. 37–64, CRC Press, Boca Raton, FL, 2014.

[22] Q. Ye, D. Doermann: "Text Detection and Recognition in Imagery: A Survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 7, pp. 1480–1500, Jul. 2015.

Ralf Schlüter studied physics at RWTH Aachen University, Germany, and Edinburgh University, UK. He received the Dipl. degree with honors in physics in 1995 and the Dr.rer.nat. degree with honors in computer science in 2000, from RWTH Aachen University. From November 1995 to April 1996 Ralf Schlüter was with the Institute for Theoretical Physics B at RWTH Aachen, where he worked on statistical physics and stochastic simulation techniques. Since May 1996 Ralf Schlüter is with the Computer Science Department at RWTH Aachen University, where he currently is academic director and leads the automatic speech recognition group at the Human Language Technology and Pattern Recognition chair. His research interests cover automatic speech recognition, discriminative training, decision theory, stochastic modeling, and signal analysis.

Eugen Beck studied computer science at RWTH Aachen University, Germany. He received the M.Sc. degree with honors in 2014 from RWTH Aachen University. Since January 2015 Eugen Beck is with the Human Language Technology and Pattern Recognition Group at RWTH Aachen University, where he currently works as reasearch assistant. His research interests include speech recognition, decision theory and keyword search.

Hermann Ney received a master degree in physics in 1977 from the University of Goettingen, Germany and a Dr.-Ing. degree in electrical engineering in 1982 from the Braunschweig University of Technology, Braunschweig, Germany. From 1977-1993, he was with Philips Research Laboratories, Hamburg and Aachen, Germany. From 1988-1989, he was a visiting scientist at ATT Bell Labs, Murray Hilly, NJ. Since 1993, he has been a Professor of Computer Science at RWTH Aachen University in Aachen, Germany.

His research interests lie in the area of machine learning and human language technology including automatic speech recognition and machine translation of text and speech. In automatic speech recognition, he and his team worked on dynamic programming for large vocabulary search, discriminative training and on language modelling. In machine translation, he and his team introduced the alignment tool GIZA++, the method of phrase-based translation, the use of dynamic programming based beam search for decoding, the log-linear model combination and system combination.

His work has resulted in more than 700 conference and journal papers with an H-index of 83 and 36000 citations (based on Google scholar). He is a fellow of both IEEE and ISCA (Int. Speech Communication Association). In 2005, he was the recipient of the Technical Achievement Award of the IEEE Signal Processing Society. In 2010, he was awarded a senior DIGITEO chair at LIMSI/CNRS in Paris, France. In 2012-2013, he was a Distingusihed Lecturer of ISCA. In 2013, he received the IAMT award of honour (IAMT: Int. Association of Machine Translation). In 2016, he received an advanced grant from the European Research Council (ERC).