Correcting The Kullback-Leibler Distance for Feature Selection

Frans M. Coetzee

GenuOne, Inc., 2 Copley Square, Boston MA 02216

Abstract

A frequent practice in feature selection is to maximize the Kullback-Leibler (K-L) distance between target classes. In this note we show that this common custom is frequently suboptimal, since it fails to take into account the fact that classification occurs using a finite number of samples. In classification, the variance and higher order moments of the likelihood function should be taken into account to select feature subsets, and the Kullback Leibler distance only relates to the mean separation. We derive appropriate expressions and show that these can lead to major increases in performance.

Key words: Kullback-Leibler distance, feature selection, classification error, Receiver Operating Curve (ROC), Neyman-Pearson classification.

1 Introduction

A common approach to feature selection is to select between feature subsets such that the Kullback-Leibler distance between the resulting class conditional probability densities is maximized. In this paper we show that this approach does not properly account for the finite sample sizes that appear when classifying finite sequences of mixed samples. We present corrected criteria for evaluating feature subsets, which can lead to major performance improvements.

The following notation frames the problem. We assume that we are provided groups of $N < \infty$ samples to be classified, *all* taken from one of two class-dependent distributions $\mathbf{p}(x|H_i), i = 0, 1$ over a finite alphabet X of size M defined by a specific subset. For clarity in notation, we may as appropriate henceforth also denote $\mathbf{p}(x|H_0)$ by q, and $\mathbf{p}(x|H_1)$ by p, respectively.

The Kullback-Leibler distance is a scalar summarizing the dissimilarity of two density functions. For densities p and q it is given by:

$$D\left(p:q\right)=\sum_{i=1}^{M}p(\chi_i)\ln(p(\chi_i)/q(\chi_i)) \tag{1}$$

with the standard assumption of non-zero probability of all alphabet elements for the two class distributions [1] .

Since the K-L divergence is a single deterministic scalar that summarizes the difference of arbitrarily high-dimensional functions, it is a useful construct in algorithms. It has a number of uses in coding theory and arises most naturally in the study of frequentist approaches to probability (e.g. the method of types – Cover and Thomas (1993)). Its use in feature selection appears straightforward: when selecting between two sets of features in building a classifier, select the set on which the two classes have the most widely separated densities, as measured by the K-L divergence. In this paper we investigate this claim more carefully. For optimal classification, a classifier compares the likelihood ratio computed on the data to a threshold. It turns out that the set of outputs of any optimal classifier evaluated on input data from a class, has a *mean* expressed by a K-L divergence of the class-conditional pdfs. However, as a random variable, the likelihood ratio also has higher order moments, which impact its being compared with the optimal threshold for class separation. This paper shows that by accounting for these higher order moments, the performance in classifying samples drawn from two sources can be much improved. This argument is developed in the next few sections; we return to further discussion in Sections 5 and 6.

To formally frame the problem, let us denote two possible feature subsets by α and β, and the class conditional densities on the symbol set defined by the feature product, by $p_\alpha(x)$ and $p_\beta(x)$ under hypothesis H_1, and $q_\alpha(x)$ and $q_\beta(x)$ under hypothesis H_0, respectively. The common procedure is: estimate either the one-sided distance $D\left(p:q\right)$, or more generally the symmetric K-L distance (also known as J-distance) $J\left(p:q\right)=D\left(p:q\right)+D\left(q:p\right)$ for different subsets of features, and select the feature subset that maximizes the chosen distance. In short, the rule is:

$$\text{If}\quad J\left(p_\alpha:q_\alpha\right)>J\left(p_\beta:q_\beta\right)\quad\text{select }\alpha\text{, else select }\beta \tag{2}$$

This approach has long been used, either explicitly, or implicitly as a building block in other approaches (see for example, summaries in Kanal (1974) and

[1] Problems with zero probability of a symbol can be reduced to this type of problem, with aberrant symbols processed independently.

Boekee and Van Der Lubbe (1979)). The most extensive analysis of this approach was performed in the classic paper by Novovicova et al. (1996), where the authors assumed a particular form of parametric density for the classes, and the J-divergence is optimized using an embedded EM algorithm. However, even here the use of the J-divergence as being optimal for finite sample classifications is explicitly assumed, rather than proved (Equation 12 in the reference). Our approach further differs from earlier approaches in that we directly account for the likely desired operating point of the classification step; the optimal features in a classification problem is generally a function of the desired operating point. We will defer further discussion of references until our method has been explained.

2 Classification and the K-L distance

In this section we consider the relationships between the statistics of the likelihood ratio test on a source classification problem, the K-L distribution, and the error that occurs when the optimal Bayes classifier is used for finite sample classification.

The classical Bayesian classification solution minimizes the error of choosing the class from which the samples were obtained, based on the class-conditional densities and the priors of the classes. For linear costs on Type I and Type II errors, the well known theory (Van Trees, 1971) yields a likelihood based test of the form:

$$\text{If } l(x) = \ln\left(\mathbf{p}(x_1, \ldots x_N|H_1)/\mathbf{p}(x_1, \ldots x_N|H_0)\right) > T \text{ select } H_1, \text{ else select } H_0 \tag{3}$$

Here T is a scalar threshold determined by the cost of different errors and relative priors of the two distributions. Note that in the ideal Bayesian design case the discriminator system directly implements the above function. Since all samples in (3) are assumed to arise from one source, the likelihood ratio is a random variable that can be conditioned on the source classes. With the additional assumption of independence of samples, we find:

$$\begin{aligned} l(x|H_i) &= \ln\left(p(x_1, x_2, ..x_N)/q(x_1, x_2, ..x_N)\right)|H_i \\ &= \sum_{j=1}^{N} \ln\left(p(x_j)/q(x_j)\right)|H_i \end{aligned} \tag{4}$$

The Kullback-Leibler distance arises naturally by converting the summation above to a frequency formulation (the continuous analogue being Lebesgue

integration). Denoting the number of times that symbol χ_j occurs in the N samples by $\#(\chi_j)$, rearranging the summation over the alphabet, and normalizing by the number of samples N, we find:

$$l_N(x|H_i) = \left[\sum_{j=1}^{M} (\#(\chi_j)/N) \ln \left(p(\chi_j)/q(\chi_j) \right) | H_i \right] \tag{5}$$

Since $\#(\chi_j)$ is a binomial variable, it follows directly that:

$$\#(\chi_j)/N|H_1 \to p(\chi_j) \tag{6}$$
$$\#(\chi_j)/N|H_0 \to q(\chi_j) \tag{7}$$

as $N \to \infty$ by almost any metric of convergence that is of interest (e.g. expectation, probability, mean square, and uniform all included). As a result:

$$\mathbf{E}\left\{ l_N(x)|H_i \right\} = \begin{cases} -D\left(q:p\right) & i=0 \\ \quad D\left(p:q\right) & i=1 \end{cases} \tag{8}$$

where we ignore that, strictly speaking, equality holds only for those densities where the probability of a symbol is an integer multiple of $1/N$. From this equation we see that the standard approach uses the separation of the means of the two class-conditional likelihood ratio distributions as a proxy for the degree of overlap of the same distributions. However, for finite N, the likelihood ratio $l_N(x)$ is itself a non-trivial random variable, and some metric that more accurately captures the gross outlines of its distribution should be used. For simplicity, later we will use a Gaussian approximation that yields metrics similar to the Mahalanobis distance, although other metrics could be explored.

Using the constraint that the frequency counts have to lie on a simplex, the joint probability of the frequency counts is multinomial. The class-conditional means are given by Equation 8, while application of some algebra shows that the variance is given by:

$$\mathrm{var}\left\{ l_N(x)|H_1 \right\} = \frac{1}{N}[\ \Psi\left(p:q\right) - \ D^2\left(p:q\right)] \tag{9}$$
$$\mathrm{var}\left\{ l_N(x)|H_0 \right\} = \frac{1}{N}[\ \Psi\left(q:p\right) - \ D^2\left(q:p\right)] \tag{10}$$

where we define:

$$\Psi\left(p:q\right) = \sum_{j=1}^{M} p(\chi_j) [\ln\left(p(\chi_j)/q(\chi_j)\right)]^2 \tag{11}$$

We now consider the error in classification that can be expected when the likelihood ratio is used to implement the optimal Bayes classifier for finite N. While for general densities no closed form solution exists, we can make headway by assuming each class-conditioned likelihood ratio is approximately Gaussian. This assumption is usually well motivated since the linear combination of binomial distributions approaches a Gaussian distribution under rather general conditions for even moderate N(Johnson and Kotz, 1969; Feller, 1950).

It is straightforward to show that given two one dimensional Gaussian distributions $G(\mu_0, \sigma_0)$ and $G(\mu_1, \sigma_1)$, where we assume means $\mu_1 > \mu_0$, variances σ_0^2 and σ_1^2, and threshold T, the misclassification error has two major components:

$$P_\epsilon = p(H_0)\Phi\left\{\frac{T-\mu_0}{\sigma_0}\right\} + p(H_1)\Phi\left\{\frac{\mu_1 - T}{\sigma_1}\right\} \tag{12}$$

where $\Phi\{\cdot\} = 1/2\ \mathrm{erfc}(\cdot/\sqrt{2})$ and erfc is the standard complementary error function. Further, the ROC curve defined by false alarm rate $P_F\{T\}$ and detection rate $P_D\{T\}$ is given by:

$$P_F\{T\} = P\{l_N(x|H_0) > T\} \ = \ \Phi\left\{\frac{T-\mu_0}{\sigma_0}\right\} \tag{13}$$

$$P_D\{T\} = P\{l_N(x|H_1) > T\} \ = \ \Phi\left\{\frac{T-\mu_1}{\sigma_1}\right\} \tag{14}$$

We now consider two specific cases of interest.

No Class Preference

The class-conditional errors on each of the two classes are equal when the threshold:

$$T = \frac{\sigma_0\sigma_1}{\sigma_0+\sigma_1}\left[\frac{\mu_0}{\sigma_0} + \frac{\mu_1}{\sigma_1}\right] \tag{15}$$

whence the error is given by:

$$P_\epsilon = \Phi\left\{\frac{\mu_1-\mu_0}{\sigma_1+\sigma_0}\right\} \tag{16}$$

Returning to the original problem, it follows that the limiting classification error when sample classification is performed, is given by:

$$P_\epsilon \simeq \Phi\left\{\sqrt{N}\ \Gamma\left(p:q\right)\right\} \tag{17}$$

where

$$\Gamma(p:q) = \frac{D(p:q) + D(q:p)}{[\Psi(p:q) - D^2(p:q)]^{1/2} + [\Psi(q:p) - D^2(q:p)]^{1/2}} \tag{18}$$

Class Preference: Neyman-Pearson Classification

In this approach the false alarm rate $P_F\{T\}$ is maintained at constant level. Define $\kappa = \Phi^{-1}\{P_F\{\cdot\}\}$. It follows that:

$$T = \mu_0 + \sigma_0 \kappa \tag{19}$$

$$P_D\{P_F\{\cdot\}\} \simeq \Phi\left\{-\sqrt{N}\,\Omega(p:q;\kappa)\right\} \tag{20}$$

where

$$\Omega(p:q;\kappa) = \frac{(\mu_1 - \mu_0) - \sigma_0\kappa}{\sigma_1} \tag{21}$$

Returning to the feature subset selection problem, we see that at a fixed operating point, detection will be improved by optimizing the following criterion:

$$\Omega(p:q;\kappa) = \frac{D(p:q) + D(q:p)}{[\Psi(p:q) - D^2(p:q)]^{1/2}} - \frac{\kappa}{\sqrt{N}}\left[\frac{\Psi(q:p) - D^2(q:p)}{\Psi(p:q) - D^2(p:q)}\right]^{1/2} \tag{22}$$

3 Application to Feature Selection

From the previous section, we can now produce two modified rules for feature selection that account for the finite value of N. When Type I and Type II errors are to be minimized simultaneously:

$$\text{If} \quad \Gamma(p_\alpha : q_\alpha) > \Gamma(p_\beta : q_\beta) \quad \text{select } \alpha\text{, else select } \beta \tag{23}$$

Similarly, when a fixed false alarm rate is to be achieved:

$$\text{If} \quad \Omega(p_\alpha : q_\alpha;\kappa) > \Omega(p_\beta : q_\beta;\kappa) \quad \text{select } \alpha\text{, else select } \beta \tag{24}$$

The careful reader will require that one issue still be addressed: namely that these rules are not coincidental with Equation 2. In short, we have to address

whether the numerator and denominator in Equation 18 are subject to an implicit dependency such that if:

$$J\left(p_\alpha : q_\alpha\right) > \ J\left(p_\beta : q_\beta\right) \tag{25}$$

then it follows that:

$$\Gamma\left(p_\alpha : q_\alpha\right) > \ \Gamma\left(p_\beta : q_\beta\right) \tag{26}$$

and similarly, for Equation 24. We basically have to prove the rules are not simply the same in disguise. We do not expect them to always disagree, but there should be cases where they disagree with each other. This independence is easily proved by example; such an example is discussed in the next section. We note this independence is true in general: the constraints implied by Equations 25 and 26 are regular and impose 2 constraints on a set of dimension $2(M-1)$ spanned by the simplexes from which the class densities are drawn. Hence neither density pairs for which the conditions hold, nor those for which they do not hold, can be considered exceptional. Combined with the fact that there is always at least one multi-dimensional density that exists for a set of marginal distributions, we can expect the disagreement of the selection rules for a large set of distributions.

4 Examples

In this section we present results on a reference data set that show the effect of using the appropriate selection rule for classification problems. The *agaricus-lepiota* or "mushroom" dataset from the UCI Machine Learning repository (Blake and Merz, 1998) is widely used for evaluating discrete feature selection approaches. The dataset has a large number of data points (8124), and 22 features (numbered 1 through 22), all of which are discrete, with each feature having a distinct vocabulary of up to 12 symbols. These features are shown in Table 1. The rules generated on this database are widely used in practice to distinguish edible and poisonous mushrooms – if you buy wild mushrooms, your continued health may in fact depend on the accuracy of this database. The large amount of data relative to feature space allows for accurate evaluations of performance, and bounds on performance for different feature sets and rules are known(Duch et al., 1997).

The data was used to construct histograms for subsets of groups of three features, for which the histograms are highly accurate. An extensive search over all pairs of groups of three features requires significant computation but is possible. We studied randomly selected triplets of features, where the new

1	cap-shape	2	cap-surface	3	cap-color
4	bruises?	5	odor	6	gill-attachment
7	gill-spacing	8	gill-size	9	gill-color
10	stalk-shape	11	stalk-root	12	stalk-surface-above-ring
13	stalk-surface-below-ring	14	stalk-color-above-ring	15	stalk-color-below-ring
16	veil-type	17	veil-color	18	ring-number
19	ring-type	20	spore-print-color	21	population
22	habitat				

Table 1
Features of mushroom data set, numbered 1 through 22.

criteria disagreed with the standard J-divergence when selecting the optimal feature set, until we had a 1000 cases. In each case we calculated the full ROC curve on each feature subset and evaluated the optimal classifiers in full. For clarity and due to the limited space here, we only consider feature subsets where both of the new criteria differed from the J-divergence (i.e. the new criteria indicate that both at the Bayesian operating point, and the $P_f = 0.15$ point, the newly selected feature subset would outperform the commonly selected subset). This rather stringent joint condition occurred for roughly 10% of the feature subset pairings.

Table 2 contains ten examples of feature triplets, the values of the different selection criteria calculated for the class-conditional distributions on the subset, and the Bayesian error rate P^* (equal class errors) as well as the Neyman-Pearson detection error rate $P_d^{0.15}$ at a false alarm rate of $P_f = 0.15$ (the full ROC curve was calculated and interpolated to find these error rates).

Table 3 contains ten examples of feature triplets where the new criteria yielded a different subset selection from the common procedure, but the performance was in fact worse. We ascribe this error to the fact that the new criteria models the likelihood statistics only up to second order; while usually this yields improvement over just using the mean separation (J-divergence), it is not always adequate. We note that such inferior performance is not the norm: Table 4 shows that in most cases (80%+) the new criteria result in an improvement, and on average the improvement in performance at both operating points is higher than when the new criteria result in inferior performance.

Figure 1 shows the receiver operating curves for two representative feature subsets $\alpha = (2, 15, 16)$ and $\beta = (4, 6, 8)$. This case is the first in Table 2.

Set α	Set β	J_α	J_β	Γ_α	Γ_β	Ω_α	Ω_β	P^*_α	P^*_β	$P^{0.15}_\alpha$	$P^{0.15}_\beta$
(2,15,16)	(4,6,8)	3.18	2.29	0.70	0.86	1.32	1.63	73.6	78.1	59.4	67.5
(11,13,16)	(8,12,17)	5.01	4.85	0.88	1.21	1.53	1.90	83	90.8	79.9	92.8
(6,11,19)	(4,11,18)	6.61	6.11	1.15	1.41	2.21	3.76	87.5	92.4	88.1	93.6
(7,13,18)	(11,16,22)	5.63	4.98	0.95	1.00	1.37	1.76	74.2	83.4	67.9	82.6
(10,12,21)	(8,10,21)	6.83	5.84	1.12	1.33	2.05	3.90	80.9	89.6	76.7	93.3
(4,6,15)	(7,16,20)	5.52	4.06	0.89	1.38	1.59	2.35	77.2	88.7	69.7	89.3
(2,4,12)	(16,18,20)	5.61	5.43	0.99	1.55	1.88	3.00	84.2	89.6	84.2	90.3
(7,13,22)	(16,18,20)	6.44	5.43	1.18	1.55	2.05	3.00	86.6	89.6	87.6	90.3
(13,17,18)	(10,12,13)	5.11	4.09	0.96	0.99	1.22	1.41	71.7	78.2	64.5	74.6
(13,16,19)	(4,17,21)	5.89	5.45	1.07	1.25	1.62	3.54	83.4	87.3	82.9	89

Table 2
Selected triplets of features and performance, where selection criteria differ, and new criteria yield an improvement.

Set α	Set β	J_α	J_β	Γ_α	Γ_β	Ω_α	Ω_β	P^*_α	P^*_β	$P^{0.15}_\alpha$	$P^{0.15}_\beta$
(10,18,22)	(7,10,21)	4.13	4.10	0.91	0.96	1.42	3.96	81.5	76.4	80.5	57.8
(1,14,16)	(11,16,18)	2.83	2.42	0.65	0.68	1.20	2.04	71.7	70.3	59	56
(10,12,17)	(6,13,17)	2.83	2.03	0.72	0.73	0.97	1.16	70	69.4	62.7	61.7
(8,10,11)	(4,18,20)	8.01	5.48	1.45	1.59	2.62	2.86	92	90.2	93.3	91.4
(9,10,16)	(3,11,16)	6.43	5.04	1.06	1.17	1.82	3.75	85.7	83.6	85.7	80.7
(4,10,11)	(4,8,11)	8.38	7.28	1.63	1.68	2.89	3.97	92.5	91.6	93.9	93.2
(3,4,16)	(2,8,21)	4.95	3.67	0.86	0.87	1.52	2.03	79	75.7	68.7	68.5
(6,8,22)	(12,13,16)	3.15	2.64	0.78	0.86	1.25	1.39	77.5	75.3	74.2	71.7
(6,7,9)	(3,11,16)	5.80	5.04	1.12	1.17	1.63	3.75	85.2	83.6	85.2	80.7
(10,18,19)	(6,11,18)	3.74	3.45	0.79	0.80	1.10	1.78	79.4	72.9	70.1	60.6

Table 3
Selected triplets of features and performance, where selection criteria differ, and new criteria cause under-performance (fewer cases overall).

5 Feature Space Search

Up to now the paper discussed only the problem of choosing between two subsets of features for later use in a classification scheme. The general feature selection problem involves absolute, or more generally a co-final directed set

Case	% Of Sets	$\overline{\Delta P^*}$ (%)	$\overline{\Delta P^{0.15}}$ (%)
Superior	81.9	6.34	10.2
Inferior	18.1	-2.24	-5.23
Overall	100	4.79	7.37

Table 4
Mean improvement in error on sets where new criteria are used and performance is increased, and percentage where criteria differ, and new criteria cause underperformance.

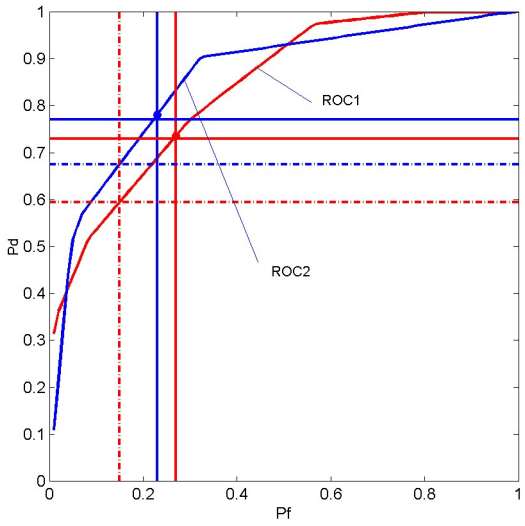

Fig. 1. Receiver operating curves (ROC), ROC1 for features $\alpha = (2, 15, 16)$ and ROC2 for features $\beta = (4, 6, 8)$ on the mushroom data set. The new criteria yields improved performance. The operating points for an optimal Bayesian classifier minimizing Type I and Type II error simultaneously (straight lines), and for a Neyman-Pearson classifier with $P_f = 0.15$ (dashed lines) are shown.

ranking (Munkres, 1975) of all 2^N subsets of a set of N features. Such ranking of the subsets requires, first, a method for traversing the feature subset space, and second, a local method for comparing selected pairs of subsets. Our approach addresses the second part of the problem – how to better rank any two subsets; the global search is still responsible for correctly interpreting the pairwise preference, and being robust against errors.

For any reasonable sizes of the alphabet, the selection problem is computationally insoluble; in addition to the sheer computational complexity, the ranking is further dependent on the desired operating point – a point frequently unacknowledged. Even in our selected examples in Table 2, different operating points favor different features. True feature selection requires the ordering of ROC curves, not points. The brute force solution requires calculating the

ROC curves for all feature subsets, and sorting based on performance at the desired operating point. And if that is not enough of a challenge, in wrapper approaches the additional load of a restricted classifier architecture is also imposed.

The plethora of feature selection procedures all try to reduce the complexities involved (see Blum and Langley (1997)) for a review). Some taxonomies for grouping methods exist; the major separation is that of approaches that explicitly perform selection before classification from those where the feature selection is implicit. Explicit approaches include methods that compare performance of a seed set and the set formed by adding features to (forward approach) or trimming features from (backward approach) the seed set, often with some probabilistic or backtracking annealing components(Koller and Sahami, 1996; Pudil et al., 1994a). Our new criteria are obvious candidates for replacing the J-divergence or KL-divergence in these approaches. In other applications the architecture, estimation and evaluation are intermingled – for example in Attribute Value Taxonomy generation and tree building, split variable selection corresponds to maximizing class divergence (Baker and McCallum, 1998; Kang et al., 2004). Properly accounting for the separation of finite runlength samples is relevant and our new criteria should be considered.

A further issue is performing feature selection with limited data. Little effort has been expended in this area. We note that the author previously proposed using the directed-set structure that exist on the ROC curves to bound the feature subset search to only a statistically valid region of the feature space supported by the finite data(Coetzee et al., 2001). This latter approach also sorts ROC curves, and selects families of subsets based on operating point. The approach was required in our applications where the operating point varies in a wide range. The approach used in the current paper is extremely helpful in pre-sorting large data sets for such computationally complex searches, since we can select for particular operating points, yet both criteria (23) and (24) use scalar statistics that retain the benefit of having a single number that captures performance on the dataset.

6 Conclusion

It is helpful here to discuss why the effects described in this communication are not widely recognized, especially when the standard J-divergence based approaches are so widespread. It appears that a remarkable degree of confusion exists as to the number N in the likelihood ratio. In introductions of the K-L distance such as in Cover(Cover and Thomas, 1993), the problem that is analyzed (correctly) is how many samples are required to be drawn consecutively from a *single* distribution to distinguish between two possible source

distributions. As N becomes large, equi-partition theorems hold, and the K-L distribution appears naturally in various error exponents.

However, in a practical classification problem, samples are obtained from *two distributions* and the problem is more accurately represented by a mixture distribution. Practitioners frequently confuse the two problems. Even if a large number N' of samples are obtained in the latter problem, what is of importance for the classification error bound is the run-length N of batches of samples guaranteed to be pulled sequentially from one of the classes – which is typically small or one in most applications. *The large-number asymptotics captured by the K-L distributions do not apply simply because you have a large number of combined samples from the two distributions.* Further, this effect of finite run-length is not the same as that of estimating the K-L divergence from a finite set of labeled data. The latter is also a difficult problem, but is a secondary effect relative to the focus of this paper.

The other frequent source of confusion results from using K-L divergence implicitly as part of a larger feature selection approach that aims to retain information. In these approaches features are eliminated such that the per-class densities are disturbed minimally by the features being eliminated(Pudil et al., 1994b; Koller and Sahami, 1996). While intuitively these approaches should yield reasonable results, there is no reason to believe that minimizing distortion of per-class distributions as measured by K-L divergence generally minimizes subsequent classification error between the resulting class-conditional distributions. Similarly, ranking bounds on classification performance on feature subsets does not necessarily translate into selecting the optimal feature subset, especially since bounds such as Bhattacharya bounds pay a heavy penalty in discrimination and continuity to achieve their general applicability.

We note that if the features of the samples are not independent, our approach still functions perfectly. The important independence requirement relates to the drawing of data samples. If the samples are drawn dependently, converting the likelihood ratio to a frequentist formulation (the Lebesgue integration in (5)), is not always possible, and the K-L distribution does not naturally arise. Little work exists on analyzing the dependent case, beyond some work on higher order extensions of the method of "types" – notably by Csiszar(see Csiszar (1998) for a review). The independence condition is however not an issue in most problems typically considered as classification problems; the dependence problem usually crops up in areas of stochastic game theory and time series prediction.

We close with one final interesting point: the selection of subsets does not change with N when neither class is favored (Equation 18). In contrast, when the false alarm rate has to be minimized, the run-length may influence the choice of feature subsets (Equation 22). In both cases, though, the overall

rate of error decreases with run-length (Equation 17 and Equation 20), as one would expect.

7 Acknowledgments

The author thanks the editor and two anonymous reviewers for suggestions that markedly clarified and improved the paper.

References

Baker, L. D., McCallum, A. K., 1998. Distributional clustering of words for text classification. In: Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval. Melbourne, AU, pp. 96–103.

Blake, C., Merz, C., 1998. UCI repository of machine learning databases. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97 (1-2), 245–271.

Boekee, D. E., Van Der Lubbe, J. C. A., 1979. Some aspects of error bounds in feature selection. Pattern Recognition 11, 353–360.

Coetzee, F., Glover, E., Lawrence, S., Giles, C. L., July 2001. Feature selection in web applications using roc inflections and power set pruning. In: Symposium on Applications and the Internet - SAINT 2001. San Diego, CA.

Cover, Thomas, 1993. Principles of Information Theory. Wiley and Sons.

Csiszar, I., 1998. The method of types. IEEE Transactions on Information Theory 44 (6), 2502–2523.

Duch, W., Adamczak, R., Grąbczewski, K., Ishikawa, M., Ueda, H., 1997. Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches. In: Proceedings of the European Symposium on Artificial Neural Networks (ESANN'97). pp. 109–114.

Feller, W., 1950. An Introduction to Probability Theory and Its Applications, 3rd Edition. Vol. 1 of Probability and Mathematical Statistics. John Wiley and Sons, iSBN 0-471-25708-7.

Johnson, N. L., Kotz, S., 1969. Discrete Distributions. Probability and Mathematical Statistics. John Wiley and Sons, iSBN 0-471-44360-3.

Kanal, L. N., 1974. Patterns in pattern recognition. IEEE Transactions on Information Theory 20, 697–722.

Kang, D., Silvescu, A., Zhang, J., Honavar, V., 2004. Generation of attribute value taxonomies from data for data-driven construction of accurate and

compact classifiers. In: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04). Brighton, UK, in print.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. In: Proc. 13th International Conference on Machine Learning. Morgan Kaufmann, pp. 284–292.

Munkres, J. R., 1975. Topology: A First Course. Prentice-Hall, iSBN 0-13-925495-1.

Novovicova, J., Pudil, P., Kittler, J., February 1996. Divergence based feature selection for multimodal class densities. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (2), 218–223.

Pudil, P., Novovicova, J., Kittler, J., November 1994a. Floating search methods in feature-selection. Pattern Recognition Letters 15 (11), 1119–1125.

Pudil, P., Novovicova, J., Kittler, J., 1994b. Simultaneous learning of decision rules and important attributes for classification problems in image analysis. Image and Vision Computing 12 (3), 193–198.

Van Trees, H. L., 1971. Detection Estimation and Modulation Theory. Vol. 1-3. Wiley and Sons.