Classification using Pattern Probability Estimators

Jayadev Acharya, ECE, UCSD, jacharya@ucsd.edu
Hirakendu Das, ECE, UCSD, hdas@ucsd.edu
Alon Orlitsky, ECE and CSE, UCSD, alon@ucsd.edu
Shengjun Pan, CSE, UCSD, s1pan@ucsd.edu
Narayana P. Santhanam, EE, U. Hawaii, nsanthan@hawaii.edu

***Abstract*—We consider the problem of classification, where the data of the classes are generated *i.i.d.* according to unknown probability distributions. The goal is to classify test data with minimum error probability, based on the training data available for the classes. The Likelihood Ratio Test (LRT) is the optimal decision rule when the distributions are known. Hence, a popular approach for classification is to estimate the likelihoods using well known probability estimators, *e.g.,* the Laplace and Good-Turing estimators, and use them in a LRT. We are primarily interested in situations where the alphabet of the underlying distributions is large compared to the training data available, which is indeed the case in most practical applications. We motivate and propose LRT's based on pattern probability estimators that are known to achieve low redundancy for universal compression of large alphabet sources. While a complete proof for optimality of these decision rules is warranted, we demonstrate their performance and compare it with other well-known classifiers by various experiments on synthetic data and real data for text classification.**

I. Introduction

Classification is one of the most important problems in the areas of machine learning and information theory. It involves designing decision rules for classifying *test* data into one among several classes characterized by *training* data belonging to them. It has a wide range of important applications like text classification, optical character recognition (OCR), bio-informatics, credit rating and many more. It has been studied extensively and is a well understood problem. In many practical applications of classification, a reasonable model is to assume that the data of each class is generated *i.i.d.* according to a probability distribution [4].

A. Notation and background

The following are some preliminaries about classification and hypothesis testing [2, 11.7] [3]. Let $\mathcal{I}_k^*$ be the collection of all *i.i.d.* distributions over the alphabet $\mathcal{A} \stackrel{\text{def}}{=} \{a_1, a_2, \ldots, a_k\}$ of size k. Let $p^{(1)}, p^{(2)} \in \mathcal{I}_k^*$ be the distributions that generate the data, *i.e.,* sequences, from the classes 1 and 2 respectively. For simplicity, we limit ourselves to classification with two classes, although the arguments presented here extend to more than two classes as well. Let $\overline{X}^{(1)}, \overline{X}^{(2)} \in \mathcal{A}^N$ be training sequences from the two classes, generated randomly according to $p^{(1)}, p^{(2)}$ respectively. A test sequence $\overline{Y}$ is generated according to $p^{(1)}$ with probability π_1, or $p^{(2)}$ with probability $\pi_2 = 1 - \pi_1$ respectively. For simplicity, we assume this prior $\bar{\pi} \stackrel{\text{def}}{=} (\pi_1, \pi_2)$ to be $(1/2, 1/2)$. The goal is to assign a label 1 or 2 to $\overline{Y}$, *i.e.,* classify it as being generated by $p^{(1)}$ or $p^{(2)}$.

A (randomized) decision rule

$$\Omega \stackrel{\text{def}}{=} \{(\omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}), \omega(2|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})) : (\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) \in \mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^n\}$$

assigns label 1 to the test sequence $\overline{Y} = \overline{y}$ with probability $\omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})$ and label 2 with probability $\omega(2|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) = 1 - \omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})$ when the training sequences are $\overline{X}^{(1)} = \overline{x}^{(1)}$ and $\overline{X}^{(2)} = \overline{x}^{(2)}$.

The probability of classifying correctly is therefore

$$P_c = \sum_{\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}} p^{(1)}(\overline{x}^{(1)}) p^{(2)}(\overline{x}^{(2)}) \Big(\frac{1}{2} p^{(1)}(\overline{y}) \omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) + \frac{1}{2} p^{(2)}(\overline{y}) \omega(2|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) \Big)$$

and the error probability is $P_e(\Omega, p^{(1)}, p^{(2)}) \stackrel{\text{def}}{=} 1 - P_c$. Hence, when the distributions $p^{(1)}, p^{(2)}$ are known, the optimal decision rule $\Omega^*(p^{(1)}, p^{(2)}) \stackrel{\text{def}}{=} \arg\min_\Omega P_e(\Omega, p^{(1)}, p^{(2)})$ which minimizes the error probability, i.e., maximizes P_c, is given by the Likelihood Ratio Test (LRT)

$$\omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) = \begin{cases} 1 & \text{if } p^{(1)}(\overline{x}^{(1)}) p^{(2)}(\overline{x}^{(2)}) p^{(1)}(\overline{y}) > \\ & p^{(1)}(\overline{x}^{(1)}) p^{(2)}(\overline{x}^{(2)}) p^{(2)}(\overline{y}), \\ & \textit{i.e., } p^{(1)}(\overline{y}) > p^{(2)}(\overline{y}), \\ 0 & \text{otherwise,} \end{cases}$$

and $\omega(2|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) = 1 - \omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})$, $\forall (\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})$.

Likewise, if $p^{(1)}, p^{(2)}$ are not known, but instead a prior $\mu(dp'^{(1)}, dp'^{(2)})$ over the $(p'^{(1)}, p'^{(2)}) \in \mathcal{I}_k^* \times \mathcal{I}_k^*$ is known, the error probability is

$$P_e(\Omega, \mu) = 1 - \frac{1}{2} \sum_{(\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y})} \Big(P_\mu(\overline{x}^{(1)} \overline{y}; \overline{x}^{(2)}) \omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) + P_\mu(\overline{x}^{(1)}; \overline{x}^{(2)} \overline{y}) \omega(2|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) \Big),$$

where $P_\mu(\overline{z}^{(1)}; \overline{z}^{(2)})$, the probability of $(\overline{z}^{(1)}, \overline{z}^{(2)}) \in \mathcal{A}^* \times \mathcal{A}^*$ under the prior μ, is

$$P_\mu(\overline{z}^{(1)}; \overline{z}^{(2)}) \stackrel{\text{def}}{=} \int_{\mathcal{I}_k^* \times \mathcal{I}_k^*} \mu(dp'^{(1)}, dp'^{(2)}) p'^{(1)}(\overline{z}^{(1)}) p'^{(2)}(\overline{z}^{(2)}).$$

Hence, the optimal decision rule $\Omega^*(\mu) \stackrel{\text{def}}{=} \arg\min_\Omega P_e(\Omega, \mu)$ is the LRT

$$\omega(1|\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}) = \begin{cases} 1 & \text{if } P_\mu(\overline{x}^{(1)} \overline{y}; \overline{x}^{(2)}) > P_\mu(\overline{x}^{(1)}; \overline{x}^{(2)} \overline{y}), \\ & \textit{i.e., } \frac{P_\mu(\overline{x}^{(1)} \overline{y}; \overline{x}^{(2)})}{P_\mu(\overline{x}^{(1)}; \overline{x}^{(2)})} > \frac{P_\mu(\overline{x}^{(1)}; \overline{x}^{(2)} \overline{y})}{P_\mu(\overline{x}^{(1)}; \overline{x}^{(2)})}, \\ 0 & \text{otherwise,} \end{cases}$$

and $\omega(2|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}) = 1-\omega(1|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}), \forall(\overline{x}^{(1)},\overline{x}^{(2)},\overline{y})$. For example, when the prior μ is the uniform prior, the optimal rule is a LRT based on Laplace estimator, while Jeffrey's prior ($\mathrm{Dir}(\frac{1}{2})$) leads to KT estimator based LRT.

B. Classification under large alphabet distributions

As the optimal decision rule when the distributions $p^{(1)}, p^{(2)}$ are known is the LRT, a reasonable approach for classification is to use various probability estimators to estimate the likelihoods and *plug* them into the LRT. Despite the simplicity of this approach, it is known to work very well in practice [4].

Most of the well known probability estimators, *e.g.,* Laplace and KT, estimate the probability of *sequences* and try to assign each sequence a probability that is close to its maximum likelihood among all possible *i.i.d.* distributions over a finite alphabet. However, in most practical applications, the alphabet is large compared to the length of the sequences, or possibly infinite. For example, as we will see in Section III, in a typical text classification experiment with words considered as alphabet symbols, the data available for each class consists of about 2000 documents, each containing about 100 words and the alphabet size k is at least 10000. With a 60:40 training-test split, *i.e.,* 60% of documents set aside for training and remaining 40% used for test, we have $N \approx 120000$ (and $n \approx 100$), which is not so large compared to the alphabet size. In such large alphabet scenarios, there can be a large gap between the estimated probability of a sequence and its maximum likelihood, potentially the actual probability. Indeed no estimator can assign a probability that is close to maximum likelihood for all sequences [6].

It is relevant to use sequence probability estimators when we want to perform almost as good as the optimal classifier that knows $p^{(1)}$, $p^{(2)}$. However, if we can perform close to an optimal classifier that has any information about $p^{(1)}$, $p^{(2)}$, but does not know $p^{(1)}$, $p^{(2)}$ completely, for example their support sets or collections/families to which they belong, then we essentially have a near optimal classifier. To this end, we consider the optimal classifiers that know about the probability *multisets* of $p^{(1)}$, $p^{(2)}$, but do not know anything about the associations between the multisets and the alphabet. As we show, and is intuitive, such classifiers are LRT's based on *pattern* probabilities instead of sequence probabilities, assuming all mappings of the alphabet to the probability multiset are equally likely.

The pattern $\Psi(\overline{z})$ of a sequence $\overline{z}$ conveys the order of appearances of symbols in $\overline{z}$. For example, $\Psi(\texttt{abracadabra}) = 12314151231$. Clearly, the probability of a pattern depends only on the probability multiset of the underlying distribution. In order to compete with classifiers which know the probability multisets of the distributions, we use LRT's that use pattern probability estimators instead of the actual pattern probabilties. In the context of universal compression, it was previously shown in [8] that patterns can be compressed with diminishing per symbol redundancy regardless of alphabet size of the underlying distribution, demonstrating several good pattern probability estimators in the process. Such estimators assign each pattern a probability that is close to its maximum likelihood, and in a way estimate the pattern probabilities accurately. The use of pattern probability estimators for classification was introduced in [9] and preliminary empirical results on text classification were encouraging. In this paper, we further explore these techniques.

In Section II, we show in detail the role of pattern probabilities in classification and consider classifiers based on good pattern probability estimators. Finally, we show several experimental results in Section III for the specific application of text classification, involving both synthetic and actual data sets.

II. Classifiers based on pattern probabilities

A. Single pattern classifiers

The following lemma characterizes the optimal classifier $\Omega^*(\{p^{(1)}\},\{p^{(2)}\})$ when the probability multisets $\{p^{(1)}\}$ of $p^{(1)}$ and $\{p^{(2)}\}$ of $p^{(2)}$ are known. Let $k_1 \stackrel{\text{def}}{=} |\{p^{(1)}\}|$ and $k_2 \stackrel{\text{def}}{=} |\{p^{(2)}\}|$ be the size of the nonzero support sets of $p^{(1)}$ and $p^{(2)}$. For a sequence $\overline{z} \in \mathcal{A}^*$, let $\mathcal{A}(\overline{z})$ be the set of symbols and $m(\overline{z}) = |\mathcal{A}(\overline{z})|$ be number of distinct symbols that have appeared in $\overline{z}$. For brevity, let $m_1 \stackrel{\text{def}}{=} m(\overline{x}^{(1)})$, $m_2 \stackrel{\text{def}}{=} m(\overline{x}^{(2)})$, $\Delta m_1 \stackrel{\text{def}}{=} m(\overline{x}^{(1)}\overline{y}) - m(\overline{x}^{(1)})$ and $\Delta m_2 \stackrel{\text{def}}{=} m(\overline{x}^{(2)}\overline{y}) - m(\overline{x}^{(2)})$ whenever $(\overline{x}^{(1)},\overline{x}^{(2)},\overline{y})$ are clear from the context. (Δm_j is the number of *new* symbols in $\overline{y}$ that have not appeared in $\overline{x}^{(j)}$, for $j = 1, 2.$) Let $u^{\underline{v}} \stackrel{\text{def}}{=} u(u-1)\cdots(u-v+1)$ denote the falling power for integers $u \geq v \geq 0$. Also, $0^{\underline{v}} = 1$ if $v = 0$ and 0 if $v > 0$.

Lemma 1: The decision rule $\Omega^*(\{p^{(1)}\},\{p^{(2)}\})$ is given by

$$\omega(1|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}) = \begin{cases} 1 & \text{if } \frac{1}{(k-m_1)^{\underline{\Delta m_1}}} \frac{p^{(1)}(\Psi(\overline{x}^{(1)}\overline{y}))}{p^{(1)}(\Psi(\overline{x}^{(1)}))} > \\ & \frac{1}{(k-m_2)^{\underline{\Delta m_2}}} \frac{p^{(2)}(\Psi(\overline{x}^{(2)}\overline{y}))}{p^{(2)}(\Psi(\overline{x}^{(2)}))}, \\ 0 & \text{otherwise,} \end{cases}$$

and $\omega(2|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}) = 1-\omega(1|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}), \forall(\overline{x}^{(1)},\overline{x}^{(2)},\overline{y})$.

Proof: Since $\{p^{(1)}\}$,$\{p^{(2)}\}$ are known and all the $k^{\underline{k_1}} \cdot k^{\underline{k_2}}$ associations of $\{p^{(1)}\}$ with $\mathcal{A}$ and $\{p^{(2)}\}$ with $\mathcal{A}$ are equally likely, it is equivalent to a uniform prior μ over $(p'^{(1)}, p'^{(2)}) \in \mathcal{A}^{\underline{\{p^{(1)}\}}} \times \mathcal{A}^{\underline{\{p^{(2)}\}}}$. Then,

$$\begin{aligned} P_\mu(\overline{z}^{(1)};\overline{z}^{(2)}) &= \sum_{\substack{(p'^{(1)},p'^{(2)}) \in \\ \mathcal{A}^{\underline{\{p^{(1)}\}}} \times \mathcal{A}^{\underline{\{p^{(2)}\}}}}} \frac{1}{k^{\underline{k_1}} \cdot k^{\underline{k_2}}} p'^{(1)}(\overline{z}^{(1)}) p'^{(2)}(\overline{z}^{(2)}) \\ &= \prod_{j=1,2} \sum_{p'^{(j)} \in \mathcal{A}^{\underline{\{p^{(j)}\}}}} \frac{1}{k^{\underline{k_j}}} p'^{(j)}(\overline{z}^{(j)}) \\ &\stackrel{(a)}{=} \prod_{j=1,2} \frac{(k-m(\overline{z}^{(j)}))^{\underline{k_j - m(\overline{z}^{(j)})}}}{k^{\underline{k_j}}} p^{(j)}(\Psi(\overline{z}^{(j)})) \\ &= \prod_{j=1,2} \frac{1}{k^{\underline{m(\overline{z}^{(j)})}}} p^{(j)}(\Psi(\overline{z}^{(j)})), \end{aligned}$$

where Equality (a) can be shown as follows. For $j = 1, 2$, $p'^{(j)}(\overline{z}^{(j)}) \neq 0$ when each symbol in $\mathcal{A}(\overline{z}^{(j)})$ is assigned a probability from $\{p^{(j)}\}$. The sum of $p'^{(j)}(\overline{z}^{(j)})$ over all such assignments is the pattern probability $p^{(j)}(\Psi(\overline{z}^{(j)}))$. And once such an assignment is made, mapping the remaining $k_j - m(\overline{z}^{(j)})$ probabilities in $\{p^{(j)}\}$ to $k - m(\overline{z}^{(j)})$ remaining symbols of $\mathcal{A}$ in all the $(k-m(\overline{z}^{(j)}))^{\underline{k_j - m(\overline{z}^{(j)})}}$ different ways lead to the same $p'^{(j)}(\overline{z}^{(j)})$.

The lemma follows by substituting the above $P_\mu(\overline{z}^{(1)};\overline{z}^{(2)})$ in the decision rule $\Omega^*(\mu)$. ∎

In order to match the performance of $\Omega^*(\{p^{(1)}\},\{p^{(2)}\})$, we can consider classifiers that use one of the several pattern probability estimators shown in [8], in place of the actual pattern probabilities. We complete this subsection with a typicality result similar to [7] which shows that estimated pattern probabilities are close to the underlying pattern probabilities with high probability. We consider the *block estimator* $q_{\rm spb}$ for probability of patterns $\overline{\psi}$ of length ℓ, shown in [8, Thm. 11], given by

$$q_{\rm spb}(\overline{\psi}) \stackrel{\rm def}{=} \frac{1}{|\Phi^\ell|}\frac{1}{N(\varphi(\overline{\psi}))}.$$

Here, $\varphi(\overline{\psi}) \stackrel{\rm def}{=} (\varphi_1,\varphi_2,\ldots,\varphi_\ell)$ is the *profile* of the pattern $\overline{\psi}$, where φ_μ is the number of distinct symbols that have each appeared μ times in $\overline{\psi}$, for $\mu = 1,2,\ldots,\ell$. For example, $\varphi(\Psi(\texttt{abracadabra})) = \varphi(12314151231) = (2,2,0,0,1)$. $N(\varphi)$ is the number of patterns with the same profile φ, which is shown in [8, Lem. 3] to be

$$N(\varphi) = \frac{\ell!}{\prod_{\mu=1}^{\ell}(\mu!)^{\varphi_\mu}\varphi_\mu!}.$$

Φ^ℓ is the collection of all profiles of length ℓ, and the number of such profiles, $|\Phi^\ell|$, is same as the number of unordered partitions of ℓ, i.e., the partition number $p(\ell)$ [10, Thm 15.7][8], given by

$$\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{\ell}(1-o(1))\right) \le |\Phi^\ell| \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{\ell}\right).$$

The estimator $q_{\rm spb}$ assigns equal probability to all profiles and equal probability to all patterns with the same profile.

Lemma 2: Let $\overline{X}$ be a sequence of length ℓ generated by an *i.i.d.* distribution p over an arbitrary alphabet. Then, for all $l > \pi\sqrt{\frac{2}{3}}\sqrt{\ell}$,

$$\Pr\left\{\left|\log\frac{q_{\rm spb}(\Psi(\overline{X}))}{p(\Psi(\overline{X}))}\right| \ge l\right\} \le \exp\left(-l+\pi\sqrt{\frac{2}{3}}\sqrt{\ell}\right).$$

Proof: The proof is along the lines of [7, Lem. 15]. We observe that

$$\log\frac{q_{\rm spb}(\Psi(\overline{x}))}{p(\Psi(\overline{x}))} = \log\frac{q_{\rm spb}(\varphi(\Psi(\overline{x})))}{p(\varphi(\Psi(\overline{x})))},$$

since both $q_{\rm spb}$ and p assign equal probabilities to patterns with the same profile. Also,

$$\log(q_{\rm spb}(\varphi(\Psi(\overline{x})))) = \log\frac{1}{|\Phi^\ell|} \ge -\pi\sqrt{\frac{2}{3}}\sqrt{\ell} > -l.$$

Hence, as $p(\varphi(\Psi(\overline{x}))) \le 1$ and $q_{\rm spb}(\varphi(\Psi(\overline{x}))) \le 1$,

$$\left|\log\frac{q_{\rm spb}(\varphi(\Psi(\overline{x})))}{p(\varphi(\Psi(\overline{x})))}\right| \ge l \Rightarrow \log\frac{1}{p(\varphi(\Psi(\overline{x})))} \ge l.$$

Therefore,

$$\begin{aligned}\Pr\left\{\left|\log\frac{q_{\rm spb}(\Psi(\overline{X}))}{p(\Psi(\overline{X}))}\right| \ge l\right\} &= \Pr\left\{\log\frac{q_{\rm spb}(\varphi(\Psi(\overline{X})))}{p(\varphi(\Psi(\overline{X})))} \ge l\right\}\\ &\le \Pr\left\{\log\frac{1}{p(\varphi(\Psi(\overline{X})))} \ge l\right\}\\ &\le |\Phi^\ell|\exp(-l)\\ &\le \exp\left(-l+\pi\sqrt{\frac{2}{3}}\sqrt{\ell}\right).\end{aligned}$$

∎

The bound is uniform for all *i.i.d.* distributions. In particular, substituting $l = 2\pi\sqrt{\frac{2}{3}}\sqrt{\ell}$,

$$\Pr\left\{\left|\log\frac{q_{\rm spb}(\Psi(\overline{X}))}{p(\Psi(\overline{X}))}\right| \ge 2\pi\sqrt{\frac{2}{3}}\sqrt{\ell}\right\} \le \exp\left(-\pi\sqrt{\frac{2}{3}}\sqrt{\ell}\right).$$

Thus, probabilities of long patterns can be estimated correctly to first order in the exponent with high probability. This is a positive result, since we are estimating probabilities of patterns of length $\ell \ge N$. However, it still does not imply that the classifier based on $q_{\rm spb}$ is good in terms of worst case discrepancy between $P_e(\Omega_{\rm spb},p^{(1)},p^{(2)})$ and $P_e(\Omega^*(\{p^{(1)}\},\{p^{(2)}\}),p^{(1)},p^{(2)})$. This requires a stronger result than the above lemma, *i.e.,* typicality in terms of conditional probabilties of patterns, and is an ongoing work.

We observe that *single pattern classifiers* require the alphabet size k to supplied, which may not be known and may therefore needed to be estimated. As we will see in the next subsection, classifiers based on *joint pattern* probabilities do not have this requirement.

B. Joint pattern classifiers

In this subsection, we look at classifiers that attempt to perform as good as the optimal classifier that not only knows $\{p^{(1)}\},\{p^{(2)}\}$, but also knows the (relative) associations between them. We denote the multiset of pairs of probabilities for different symbols by $\{p^{(1)},p^{(2)}\}$. For example, the if $p^{(1)} = (0.7,0.3,0)$ and $p^{(2)} = (0.2,0.6,0.2)$, then $\{p^{(1)},p^{(2)}\} = \{(0.3,0.6),(0.7,0.2),(0,0.2)\}$. Before we proceed to characterize the optimal classifier $\Omega^*(\{p^{(1)},p^{(2)}\})$, we introduce the notion of *joint pattern* of two (or more) sequences, which apart from conveying the patterns of the individual sequences, also conveys the association between their actual symbols. For a pair of sequences $\overline{z}^{(1)},\overline{z}^{(2)} \in \mathcal{A}^*$, the joint pattern $\Psi(\overline{z}^{(1)},\overline{z}^{(2)}) \stackrel{\rm def}{=} (\overline{\psi}^{(1)},\overline{\psi}^{(2)})$, where $\overline{\psi}^{(1)} = \Psi(\overline{z}^{(1)})$ and $\overline{\psi}^{(1)}\overline{\psi}^{(2)} = \Psi(\overline{z}^{(1)}\overline{z}^{(2)})$. For example, $\Psi(\texttt{bab},\texttt{abca}) = (121,2132)$.

Lemma 3: The decision rule $\Omega^*(\{p^{(1)},p^{(2)}\})$ is given by

$$\omega(1|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}) = \begin{cases}1 & \text{if } p(\Psi(\overline{x}^{(1)}\overline{y},\overline{x}^{(2)})) > \\ & p(\Psi(\overline{x}^{(1)},\overline{x}^{(2)}\overline{y})),\\ & \textit{i.e.,}\ \frac{p(\Psi(\overline{x}^{(1)}\overline{y},\overline{x}^{(2)}))}{p(\Psi(\overline{x}^{(1)},\overline{x}^{(2)}))} > \frac{p(\Psi(\overline{x}^{(1)},\overline{x}^{(2)}\overline{y}))}{p(\Psi(\overline{x}^{(1)},\overline{x}^{(2)}))},\\ 0 & \text{otherwise,}\end{cases}$$

where $p = (p^{(1)},p^{(2)})$, and $\omega(2|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y}) = 1 - \omega(1|\overline{x}^{(1)},\overline{x}^{(2)},\overline{y})$, for all $(\overline{x}^{(1)},\overline{x}^{(2)},\overline{y})$.

Proof: Let $k_\vee$ be the nonzero support of $\{p^{(1)}, p^{(2)}\}$, i.e., the number of symbols that are assigned non-zero probabilities by one of $p^{(1)}$ or $p^{(2)}$. All the $k^{\underline{k_\vee}}$ associations of $\{p^{(1)}, p^{(2)}\}$ with $\mathcal{A}$ are equally likely, which is equivalent to the uniform prior μ over $(p'^{(1)}, p'^{(2)}) \in \mathcal{A}^{\underline{\{p^{(1)},p^{(2)}\}}}$. Then,

$$\begin{aligned} P_\mu(\overline{z}^{(1)}; \overline{z}^{(2)}) &= \sum_{\substack{(p'^{(1)},p'^{(2)})\in \\ \mathcal{A}^{\underline{\{p^{(1)},p^{(2)}\}}}}} \frac{1}{k^{\underline{k_\vee}}} p'^{(1)}(\overline{z}^{(1)}) p'^{(2)}(\overline{z}^{(2)}) \\ &= \frac{(k-m)^{\underline{k_\vee - m}}}{k^{\underline{k_\vee}}} p(\Psi(\overline{z}^{(1)}, \overline{z}^{(2)})) \\ &= \frac{1}{k^{\underline{m}}} p(\Psi(\overline{z}^{(1)}, \overline{z}^{(2)})), \end{aligned}$$

where $m = m(\overline{z}^{(1)}\overline{z}^{(2)})$. Substituting the above $P_\mu(\overline{z}^{(1)}; \overline{z}^{(2)})$ in the decision rule $\Omega^*(\mu)$ leads to the desired result. ∎

We now consider a block estimator q_{jpb} for probabilities of joint patterns, that can be used in place of the actual pattern probabilities in $\Omega^*(\{p^{(1)}, p^{(2)}\})$. It is analogous to the estimator q_{spb} seen in last subsection. Let $(\overline{\psi}^{(1)}, \overline{\psi}^{(2)})$ be a joint pattern of *length* (ℓ_1, ℓ_2), *i.e.,* the length of $\overline{\psi}^{(1)}$ is ℓ_1 and of $\overline{\psi}^{(2)}$ is ℓ_2. The profile of this pattern is $\varphi(\overline{\psi}^{(1)}, \overline{\psi}^{(2)}) \stackrel{\text{def}}{=} [\varphi_{\mu_1,\mu_2}]$, a $(\ell_1+1)\times(\ell_2+1)$ integer matrix, where φ_{μ_1,μ_2} is the number of symbols, *i.e.,* labels, that have appeared μ_1 times in $\overline{\psi}^{(1)}$ and μ_2 times in $\overline{\psi}^{(2)}$, for $\mu_1 = 0, 1, \ldots, \ell_1$ and $\mu_2 = 0, 1, \ldots, \ell_2$. By convention, $\varphi_{0,0} \equiv 0$. Like q_{spb}, the estimator q_{jpb} assigns equal probability to all profiles and equal probability to all patterns with the same profile. In other words,

$$q_{\text{jpb}}(\overline{\psi}^{(1)}, \overline{\psi}^{(2)}) \stackrel{\text{def}}{=} \frac{1}{|\Phi^{\ell_1,\ell_2}|} \frac{1}{N(\varphi(\overline{\psi}^{(1)}, \overline{\psi}^{(2)}))},$$

where Φ^{ℓ_1,ℓ_2} is the collection of all distinct profiles of patterns of length (ℓ_1, ℓ_2), and $N(\varphi)$ is the number of joint patterns with the same profile φ. We state without proof the following two lemmas that calculate $N(\varphi)$ and $|\Phi^{\ell_1,\ell_2}|$.

Lemma 4: For all $\ell_1, \ell_2 \geq 0$ and $\varphi \in \Phi^{\ell_1,\ell_2}$,

$$N(\varphi) = \frac{\ell_1!\ell_2!}{\prod_{\mu_1=0}^{\ell_1}\prod_{\mu_2=0}^{\ell_2}(\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}}\varphi_{\mu_1,\mu_2}!}.$$

Proof: A proof is along the lines of [8, Lem. 3]. ∎

$|\Phi^{\ell_1,\ell_2}|$ is same as *joint* partition number $p(\ell_1, \ell_2)$, the number of unordered partitions of (ℓ_1, ℓ_2) where each partition consists of parts that are $2-$tuples of non-negative integers and component wise sums of all parts add to ℓ_1 and ℓ_2 respectively. For example, $p(2,1) = 4$, since $(2,1) = (2,0) + (0,1) = (1,1) + (1,0) = 2 \cdot (1,0) + (0,1)$.

Lemma 5: For all integers $\ell_1, \ell_2 \geq 8$,

$$p(\ell_1, \ell_2) < \exp\left(\ell_1^{2/3} + \ell_2^{2/3} + 2(\ell_1\ell_2)^{1/3} + \ell_1^{1/3} + \ell_2^{1/3}\right).$$

Proof: A simple proof is similar to [10, Thm 15.7]. ∎

While it immediately follows that the per-symbol pattern redundancy of q_{jpb} goes to zero, a typicality result similar to Lemma 2 is stated without proof below.

Lemma 6: Let $\overline{X}^{(1)}, \overline{X}^{(2)}$ be sequences of length $\ell_1, \ell_2 > 8$ generated by *i.i.d.* distributions $p^{(1)}, p^{(2)}$ over an arbitrary alphabet. Then,

$$\Pr\left\{\left|\log\frac{q_{\text{jpb}}(\Psi(\overline{X}^{(1)}, \overline{X}^{(2)}))}{p(\Psi(\overline{X}^{(1)}, \overline{X}^{(2)}))}\right| \geq 2l\right\} \leq e^{-l},$$

where $l = \ell_1^{2/3} + \ell_2^{2/3} + (\ell_1\ell_2)^{1/3} + \ell_1^{1/3} + \ell_2^{1/3}$. ∎

The estimator q_{jpb} and the results can be extended to more than two classes but is computationally intensive owing to profiles of higher dimension. Also, the redundancy of joint patterns increases with the number of classes, *i.e.,* distributions.

III. Experimental Results

We show experimental results for text classification to demonstrate the performance of pattern based classifiers. In this application, one is given a data set consisting of documents, for example, electronic messages from newsgroups, along with their pre-assigned labels, for example, their topic, and the task is to label new documents.

One of the techniques that works reasonably well in practice is *Naive Bayes* [4], which assumes a *Bag of Words* model, *i.e.,* the words in each document are generated *i.i.d.* according to the distribution of the class to which it belongs. Naive Bayes classifiers are LRT's that use one of the several well known probability estimators, for example, Laplace or Good-Turing estimators, to estimate the underlying distributions of the classes from the training documents. Our experiments show that pattern based classifiers, which are essentially Naive Bayes classifiers that use pattern probability estimators, can perform as good as the state-of-the-art techniques like Support Vector Machine (SVM).

In addition to the classifiers based on block estimators for pattern probability, *i.e.,* q_{spb} and q_{jpb}, we also consider q_{sps}, a *sequential* estimator for single pattern probability developed in [8, Thm. 18] and its analogue q_{jps} for joint patterns.

The conditional probability assigned by the estimator q_{sps} to a symbol $a \in \mathcal{A}$ given a sequence $\overline{z} = z_1 z_2 \cdots z_\ell \in \mathcal{A}^\ell$, with the symbol a appearing μ times in $\overline{z}$, is given by

$$q_{\text{sps}}(z_{\ell+1} = a|\overline{z}) \stackrel{\text{def}}{=} \frac{1}{S}\begin{cases} \frac{f_\ell(\varphi_1+1)}{\varphi_0} & \text{if } \mu = 0, \\ (\mu+1)\frac{f_\ell(\varphi_{\mu+1}+1)}{f_\ell(\varphi_\mu)} & \text{if } 1 \leq \mu \leq \ell, \end{cases}$$

where $S \stackrel{\text{def}}{=} f_\ell(\varphi_1+1) + \sum_{\mu'=1}^{\ell} \varphi_{\mu'} \frac{f_\ell(\varphi_{\mu'+1}+1)}{f_\ell(\varphi_{\mu'})}$ is the normalization factor and $f_\ell(\varphi) \stackrel{\text{def}}{=} \max\{\varphi, \lceil \ell^{1/3} \rceil\}$ is a *smoothing function* and $\varphi_0 \stackrel{\text{def}}{=} k - m(\overline{z})$ is the number of *unseen* symbols.

Likewise, the probabilities assigned by the estimator q_{jps} to a symbol $a \in \mathcal{A}$ given sequences $\overline{z}^{(1)} \in \mathcal{A}^{\ell_1}$ and $\overline{z}^{(2)} \in \mathcal{A}^{\ell_2}$, with the symbol a appearing μ_1 and μ_2 times in $\overline{z}^{(1)}$ and $\overline{z}^{(2)}$ respectively, is given by

$$q_{\text{sps}}(z^{(1)}_{\ell_1+1} = a|\overline{z}^{(1)}, \overline{z}^{(2)}) \stackrel{\text{def}}{=}$$

$$\frac{1}{S^{(1)}}\begin{cases} f_{\ell_1,\ell_2}(\varphi_{1,0}+1) & \text{if } (\mu_1,\mu_2) = (0,0), \\ (\mu_1+1)\frac{f_{\ell_1,\ell_2}(\varphi_{\mu_1+1,\mu_2}+1)}{f_{\ell_1,\ell_2}(\varphi_{\mu_1,\mu_2})} & \text{if } 0 \leq \mu_1 \leq \ell_1, 0 \leq \mu_2 \leq \ell_2 \\ & \text{and } (\mu_1,\mu_2) \neq (0,0), \end{cases}$$

where

$$S^{(1)} \stackrel{\text{def}}{=} f_{\ell_1,\ell_2}(\varphi_{1,0}+1) + \sum_{\substack{0\leq\mu_1\leq\ell_1, 0\leq\mu_2\leq\ell_2 \\ (\mu_1,\mu_2)\neq(0,0)}} \varphi_{\mu_1,\mu_2} \frac{f_{\ell_1,\ell_2}(\varphi_{\mu_1+1,\mu_2}+1)}{f_{\ell_1,\ell_2}(\varphi_{\mu_1,\mu_2})}$$

and $f_{\ell_1,\ell_2}(\varphi) \stackrel{\text{def}}{=} \max\{\varphi, \lceil (\ell_1\ell_2)^{1/8} \rceil\}$. The estimate $q_{\text{sps}}(z^{(2)}_{\ell_2+1} = a|\overline{z}^{(1)}, \overline{z}^{(2)})$ is defined similarly.

We use the `rainbow` toolkit [5] for classification, with addional support for pattern based classifiers and optimal classifiers that use actual distributions for synthetic data sets.

A. Synthetic data sets

These experiments are intended to demonstrate that pattern based classifiers work well when the data sets indeed confirm to the Bag of words model. The data sets, which try to resemble actual data sets, were generated as follows. All classes have the same monotone distribution, which is a Zipf distribution [11]. The actual distribution for each class is obtained by permuting the monotone distribution, and ensuring that the distributions of different distributions are not too far apart, and thus being *easily* classifiable unlike real data sets. This is achieved in two ways: permuting the probabilities randomly such that the final index of each probability is within (1) a fixed range of the original index (*i.e.,* rank in monotone) and (2) within a variable range that is proportionally large as the original index, *i.e.,* the smaller probabilities are permuted within a farther range. From the results shown in Table I, it is seen that pattern based classifiers perform favorably. In particular, JPS performs consistently well. Also observed is the generally better performance of sequential estimators than their block counterparts. As we will see in the case of real data sets in the next subsection, the benefits of sequential estimators are more prominent in *skewed* data sets, *i.e.,* data sets with variable number of documents per class and hence non-uniform prior $\bar{\pi}$.

Data set		Classification method						
Exp	k	lap	svm	sps	spb	jps	jpb	best
0.7	7,500	81.4	**85.0**	83.5	77.1	81.9	75.8	90.4
0.7	25,000	68.6	**79.9**	73.7	69.7	77.3	72.8	88.9
0.7	75,000	60.5	**76.0**	62.4	61.6	72.6	71.3	83.1
1.0	7,500	98.6	**99.1**	98.9	95.0	99.0	98.0	99.5
1.0	25,000	94.6	**97.8**	95.8	91.6	**97.8**	96.6	98.6
1.0	75,000	90.3	**96.7**	87.2	84.1	96.0	94.9	98.0
0.7	7,500	89.5	87.1	**90.3**	83.5	88.5	79.6	96.4
0.7	25,000	82.1	83.3	**87.0**	82.2	86.0	77.1	96.8
0.7	75,000	75.7	**80.4**	79.5	77.1	79.6	73.1	95.5
1.0	7,500	98.3	96.6	**98.6**	95.0	98.3	97.0	99.4
1.0	25,000	95.9	95.3	96.8	93.7	**97.1**	95.6	99.2
1.0	75,000	95.2	93.3	94.3	92.3	**96.1**	93.7	99.3

TABLE I
ACCURACY OF DIFFERENT CLASSIFIERS - LAPLACE, SVM, SPS, SPB, JPS, JPB ON TWO-CLASS DATA SETS CONTAINING 2000 DOCUMENTS PER CLASS AND 100 WORDS PER DOCUMENT, AND SPLIT 60-40 FOR TRAINING AND TEST. ZIPF DISTRIBUTIONS ARE GENERATED WITH DIFFERENT EXPONENTS AND SUPPORT SIZES. THE TOP HALF OF THE ROWS CORRESPOND TO DISTRIBUTIONS GENERATED BY 'FIXED RANGE' AND LATER HALF BY 'PROPORTIONAL RANGE' INDEX PERTURBATION.

B. Real world data sets

These experiments demonstrate the favorable performance of pattern based classifiers on some of the well known actual data sets. The collection *Newsgroups*, *i.e.,* `20ng`, is a list of 1000 articles collected from 20 newsgroups. It contains several closely related subgroups, for example, `comp.*`, `sci.*` and `talk.*`. The *Reuters 21758* data sets, *i.e.,* `r52` and a subset `r8`, have 52 and 8 classes respectively and the number of documents per class vary sharply between few thousands to just one or two. The *CADE* dataset, *i.e.,* `cade`, is a collection of Portugese web documents consisitng of 12 classes. It is a fairly large and uneven data set with documents per class ranging between few hundreds to few thousands and is generally difficult to classify. The data set *World Wide Knowledge Base*, *i.e.,* `webkb`, is a small data set of 4 classes of variable number of documents per class. These data sets, along with their training-test split can be obtained from [1]. The results are shown in Table II. As mentioned earlier, we observe the generally lower performance of block estimators, with the JPB estimator faring especially poorly with the skewed data set `r52` and `r8`. While results in general are in favor of SVM, they also show the favorable performance of pattern based classifiers. In particular, although JPS is not the best classifier for any of them, it is the second best in all data sets except for `webkb`, and performs well consistently.

Data set	Classification method					
	laplace	svm	sps	spb	jps	jpb
20ng	80.76	80.80	82.68	**83.05**	83.01	81.92
cade	53.10	52.09	**57.01**	51.22	54.86	49.77
r52	80.53	**92.04**	85.59	83.30	87.32	32.52
r8	90.61	**94.49**	90.36	90.17	91.69	67.38
webkb	83.30	**87.94**	83.37	83.13	83.30	83.02

TABLE II
ACCURACY OF DIFFERENT CLASSIFIERS FOR REAL DATA SETS.

REFERENCES

[1] Ana Cardoso-Cachopo, "Datasets for single-label text categorization," http://web.ist.utl.pt/~acardoso/datasets/.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*, second edition, New York: Wiley-Interscience, 2006.

[3] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitive minimax approach," *IEEE Trans. Inform. Theory*, 48(6): 1504-1517, June 2002.

[4] David D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pp. 4–15, 1998.

[5] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/mccallum/bow, 1996.

[6] A. Orlitsky, N.P. Santhanam, "Speaking of infinity," *IEEE Trans. on Inform. Theory*, 50(10): 2215-2230, Oct. 2004.

[7] A. Orlitsky, N.P. Santhanam, K. Vishwanathan and J. Zhang, "Limit Results on Pattern Entropy," *IEEE Trans. on Inform. Theory*, 52(7): 2954-2964, July 2006.

[8] A. Orlitsky, N.P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. on Inform. Theory*, 50(7): 1469-1481, July 2004.

[9] N. P. Santhanam, A. Orlitsky, and K. Viswanathan, "New tricks for old dogs: Large alphabet probability estimation," *ITW 2007*, 638-643, Sep. 2007.

[10] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge University Press, Cambridge, 1992.

[11] G. K. Zipf, *Human behavior and the principle of least effort: An Introduction to Human Ecology*, Addison-Wesley, Cambridge, USA, 1949.