Pattern Recognition Letters 45 (2014) 39–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Sequential pattern recognition by maximum conditional informativity ☆

Jiří Grim *

Institute of Information Theory and Automation, P.O. Box 18, 18208 Prague 8, Czech Republic

ARTICLE INFO

Article history:
Received 25 October 2013
Available online 13 March 2014

Keywords:
Multivariate statistics
Statistical pattern recognition
Sequential decision making
Product mixtures
EM algorithm
Shannon information

ABSTRACT

Sequential pattern recognition assumes the features to be measured successively, one at a time, and therefore the key problem is to choose the next feature optimally. However, the choice of the features may be strongly influenced by the previous feature measurements and therefore the on-line ordering of features is difficult. There are numerous methods to estimate class-conditional probability distributions but it is usually computationally intractable to derive the corresponding conditional marginals. In literature there is no exact method of on-line feature ordering except for the strongly simplifying naive Bayes models. We show that the problem of sequential recognition has an explicit analytical solution which is based on approximation of the class-conditional distributions by mixtures of product components. As the marginal distributions of product mixtures are directly available by omitting superfluous terms in the products, we have a unique non-trivial possibility to evaluate at any decision level the conditional informativity of unobserved features for a general problem of statistical recognition. In this way the most informative feature guarantees, for any given set of preceding measurements, the maximum decrease of decision uncertainty.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Sequential decision-making is an important area of statistical pattern recognition. Unlike the standard scheme considering all features of the classified object at once, the sequential recognition includes the features successively, one at a time. Usually, the goal is to reduce the number of features which are necessary for the final decision. Thus, the classification based on the currently available feature measurements is either terminal or the sequential recognition is continued by choosing the next feature. For this reason the sequential decision scheme should include a stopping rule and a suitable ordering procedure to optimally choose the next feature.

The traditional motivation for sequential recognition assumes that, for a certain reason, the feature measurements are expensive and therefore, if a reliable classification is achievable with a small subset of features, the optimal feature ordering and stopping rule may reduce the total recognition cost. However, in most pattern recognition applications all features are measured simultaneously and with negligible costs. Obviously, there is no need of sequential decision-making when the features can be used simultaneously.

On the other hand, there are problems which are sequential by their nature but the statistical properties of features may differ at different stages of classification. Thus the weak classifiers of [26] can use different feature sets, the recognized patterns in orthotic engineering may develop [31] or the state of the classified object is influenced by control actions [22,4]. In this sense, instead of sequential recognition, we have to solve a sequence of formally different recognition problems.

Practical problems of sequential recognition usually have different specific aspects which may require highly specific solutions. For example, most of the present approaches can be traced back to the theoretical results of Wald [30] which are closely related to the quality control of goods. Wald proposed the sequential probability ratio test to verify the quality of a commodity in a shipment by efficient sampling – with the aim to minimize the costs of the control procedure as a whole. Given a large shipment containing a single type of goods, the test guarantees the optimal trade-off between the number of tested items and the probability of incorrect quality evaluation.

The repetition of identical tests of goods in the Wald's problem naturally implies a sequence of independent, identically distributed measurements, and thus any ordering of measurements is pointless in this case. The generalized sequential probability ratio test provides optimal solutions only for two-class problems and class-conditionally independent features. It can be further extended and modified [8] but, even if we admit different statistical

☆ This paper has been recommended for acceptance by Qian Xiaoning.
* Tel.: +420 266052215; fax: +420 284683031.
E-mail address: grim@utia.cas.cz

http://dx.doi.org/10.1016/j.patrec.2014.02.024
0167-8655/© 2014 Elsevier B.V. All rights reserved.

properties of features in different classes, the independence assumption remains prohibitive because the typical problems of pattern recognition usually involve strongly interrelated features.

In the case of generally dependent features, the key problem of sequential recognition is the optimal on-line ordering of feature measurements. We recall that the off-line (a priori) feature ordering (closely related to the well-known feature selection algorithms [24]), is less efficient because it cannot reflect the values of the previously observed features. As it will be shown later, the optimal choice of the most informative feature at a given stage may be strongly influenced by the values of the preceding feature measurements and, for this reason, the knowledge of the underlying conditional distributions is of basic importance. There are numerous methods to estimate the unknown probability distributions in classes but it is usually computationally intractable to derive on-line the conditional marginals of unobserved features for a given subset of preceding feature measurements.

In this paper we show that, approximating the class-conditional distributions by mixtures of product components, we have a unique possibility to solve exactly the on-line feature ordering problem for a general multi-class problem of statistical recognition. Marginal distributions of product mixtures are directly available by omitting superfluous terms in the products and therefore we can evaluate, for any given set of preceding measurements, the conditional Shannon informativity of the unobserved features. The most informative feature guarantees the maximum decrease of decision uncertainty – with respect to the estimated conditional distributions.

In the following sections we first discuss the related work (Section 2) and briefly describe the product mixture model (Section 3) in application to Bayesian decision-making (Section 4). The information controlled sequential recognition is described in Section 5 and the properties of the method are illustrated by a numerical example in Section 6.

2. Related work

According to our best knowledge, the exact solution of the on-line feature ordering problem is available in the literature only for so-called naive Bayes classifiers based on the strongly simplifying assumption that the features are statistically independent in each class [1,2,21,7]. A more general setup has been considered by Fu [8], who proposed a dynamic programming approach to the on-line ordering of features. However, in order to reduce the arising computational complexity, the features are assumed to be statistically independent or Markov dependent and continuous variables have to be discretized.

Šochman and Matas [26,27] have recently proposed to circumvent the computational difficulties by combining so-called weak classifiers from a large set in the framework of the AdaBoost algorithm. The arising sequence of strong classifiers plays a role of sequential measurements which are not independent. The joint conditional density of all measurements, whose estimation is intractable, is approximated by the class-conditional response of the sequence of strong classifiers. The method called WaldBoost applies the AdaBoost algorithm to selecting and ordering the measurements and to approximation of the sequential probability ratio in the Wald's decision scheme. The WaldBoost algorithm is justified by the asymptotic properties of AdaBoost and yields a nearly optimal trade-off between time and error rate for the underlying two-class recognition problems.

One of the most natural application fields of sequential recognition is that of medical diagnostics [1,2,7]. In the case of computer-aided medical decision-making we assume the final decision to be made by a physician, and therefore the main purpose of the sequential procedure should be to accumulate maximum diagnostically relevant information along with the preliminary evaluation. The number of both possible diagnoses and potentially available features may be very large, and therefore the main advantage of the sequential procedure is the optimal choice of diagnostically relevant questions. There is no need for a stopping rule, the process may continue as long as the user is willing and able to answer the questions. The output of the classifier is given by the Bayes formula in the form of *a posteriori* probabilities of possible diagnoses which may be useful for the physician – in addition to the patient's answers and recommended medical tests.

3. Mixtures of product components

Let $\boldsymbol{x}$ be an N-dimensional vector of discrete features

$$\boldsymbol{x} = (x_1, x_2, \dots, x_N) \in \mathcal{X}, \quad x_n \in \mathcal{X}_n, \quad \mathcal{N} = \{1, 2, \dots, N\}$$

and $\mathcal{N}$ be the related index set of the variables x_n. Approximating unknown discrete probability distributions by product mixtures, we assume the following conditional independence model:

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m), \quad \boldsymbol{x} \in \mathcal{X}, \quad \mathcal{M} = \{1, \dots, M\}, \tag{1}$$

with the component weights

$$\boldsymbol{w} = (w_1, w_2, \dots, w_M), \quad w_m \geqslant 0, \quad \sum_{m \in \mathcal{M}} w_m = 1,$$

and the product distributions

$$F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m), \quad x_n \in \mathcal{X}_n, \; m \in \mathcal{M}. \tag{2}$$

Here $f_n(x_n|m)$ are univariate discrete probability distributions and $\mathcal{M}$ is the component index set.

Since the late 1960s the standard way to compute maximum-likelihood estimates of mixture parameters is to use the EM algorithm [28,6,9]. Formally, given a finite set $\mathcal{S}$ of independent observations of the underlying N-dimensional random vector

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots\}, \quad \boldsymbol{x} = (x_1, x_2, \dots, x_N) \in \mathcal{X}, \tag{3}$$

we maximize the corresponding log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m) \right] \tag{4}$$

by means of the following EM iteration equations:

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad w'_m = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x}), \tag{5}$$

$$f'_n(\xi|m) = \sum_{\boldsymbol{x} \in \mathcal{S}} \frac{\delta(\xi, x_n) q(m|\boldsymbol{x})}{\sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x})}, \quad \xi \in \mathcal{X}_n, \; n \in \mathcal{N}, \tag{6}$$

where $\delta(\xi, x_n)$ is the δ-function notation ($\delta(\xi, x_n) = 1$ for $\xi = x_n$ and zero otherwise) and the apostrophe denotes the new parameter values in each iteration. In the case of high dimensionality ($N \approx 10^2$) the EM algorithm has to be carefully implemented to avoid underflow problems [13].

Let us recall that the number of components in the mixture is a parameter to be specified in advance. One can easily imagine that there are many different possibilities to fit a mixture of many components to a large number of multidimensional feature vectors whereby each possibility may correspond to a local maximum of the related log-likelihood function. For this reason the log-likelihood criterion nearly always has local maxima and therefore the iterative computation depends on the starting-point.

Nevertheless, in the case of large data sets ($|\mathcal{S}| \approx 10^3$) and large number of components ($M \approx 10^2$), possible local maxima usually do not differ very much from each other. With identical number of components the EM algorithm expectedly achieves similar local maxima, the similar values of the log-likelihood criterion imply comparable approximation quality and therefore, according to our experience, large approximating mixtures can be initialized randomly without any relevant risk of unhappy consequences.

4. Decision making based on product mixtures

Considering the framework of statistical pattern recognition we assume that the vector $\boldsymbol{x}$ is to be classified with respect to a finite set of classes $\Omega = \{\omega_1, \ldots, \omega_K\}$. We approximate the class-conditional distributions $P(\boldsymbol{x}|\omega)$ by product mixtures:

$$P(\boldsymbol{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m F(\boldsymbol{x}|m), \quad \bigcup_{\omega \in \Omega} \mathcal{M}_\omega = \mathcal{M}, \tag{7}$$

where w_m are probabilistic weights, $F(\boldsymbol{x}|m)$ are the product components and $\mathcal{M}_\omega, \omega \in \Omega$ are disjoint index sets. Note that, in this way, the component index m simultaneously identifies the value of the class variable $\omega \in \Omega$.

Having estimated the class-conditional distributions $P(\boldsymbol{x}|\omega)$ and *a priori* probabilities $p(\omega), \omega \in \Omega$ we can write, for any given $\boldsymbol{x} \in \mathcal{X}$, the Bayes formula

$$p(\omega|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega)p(\omega)}{P(\boldsymbol{x})}, \quad P(\boldsymbol{x}) = \sum_{\omega \in \Omega} P(\boldsymbol{x}|\omega)p(\omega). \tag{8}$$

The posterior distribution $p(\omega|\boldsymbol{x})$ can be used to define a unique decision by means of Bayes decision function

$$d : \mathcal{X} \to \Omega, \quad d(\boldsymbol{x}) = \arg \max_{\omega \in \Omega} \{p(\omega|\boldsymbol{x})\}, \quad \boldsymbol{x} \in \mathcal{X}, \tag{9}$$

which is known to minimize the probability of classification error.

We recall that the unique classification (9) is accompanied by information loss because of the suppressed *a posteriori* probabilities of classes. Typically, in medical decision-making the probabilities of possible diagnoses are preferable to a unique deterministic decision. For the sake of sequential recognition we consider simultaneously the Bayes formula (8) as the classifier output since the *a posteriori* probabilities provide more subtle description of the resulting classification. By using the Shannon entropy

$$H_{\boldsymbol{x}}(\Omega) = \sum_{\omega \in \Omega} - p(\omega|\boldsymbol{x}) \log p(\omega|\boldsymbol{x})) \tag{10}$$

we have a reliable quantitative measure of the decision uncertainty implied by the Bayes formula (8).

The approximation power of product mixtures has often been underestimated – probably because of their formal similarity with so-called "naive Bayes" model (cf. [1–3,23]). We recall that, in the case of product mixtures, the term "naive Bayes" is incorrectly used because the independence assumption applies to mixture components and not to the approximated class-conditional distributions. Actually, the product mixtures (1) and (2) are suitable for describing complex statistical properties of strongly interrelated features. In the case of discrete variables the product mixture is a universal approximator [12] in the sense that any discrete distribution can be expressed in the form (1). In the case of continuous variables the approximation potential of product mixtures approaches the universality of nonparametric kernel estimates [25] with the increasing number of components.

In recent years we have applied product mixtures successfully to multidimensional pattern recognition [15,16], mammographic screening [19], texture modeling [14], image forgery detection [20], classification of documents [18] and others (cf. also [23]).

5. Sequential recognition

The sequential recognition assumes a successive evaluation of features. Each time, either terminal classification is to be performed or the next feature has to be optimally chosen. The goal of the sequential classification is to assign the observed subset of features to a class $\omega \in \Omega$ or, more exactly, to reduce the uncertainty $H_{\boldsymbol{x}}(\Omega)$ (cf. Eq. (10)) of the related *a posteriori* probability distribution (8). Note that the entropy $H_{\boldsymbol{x}}(\Omega)$ approaches zero when the *a posteriori* probabilities concentrate at a single value $\omega \in \Omega$. If we use the Shannon entropy $H_{\boldsymbol{x}}(\Omega)$ as a measure of decision uncertainty [29] then a natural way to choose the next feature measurement is to maximize the corresponding conditional Shannon information about the class variable ω contained in the considered feature. We show that, in this sense, the most informative feature provides the maximum expected decrease of uncertainty of the *a posteriori* distribution.

Motivated by an earlier idea [11] we use the fact that, in the case of product mixtures, there is a simple possibility to derive any marginal distributions by deleting superfluous terms in the products. In this way, we have at any decision level a unique possibility to evaluate the exact conditional informativity of the remaining features.

In particular, let $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ be a given subset of known feature measurements. Then for the sub-vector $\boldsymbol{x}_C$

$$\boldsymbol{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C, \quad \mathcal{C} = \{i_1, \ldots, i_k\} \subset \mathcal{N}, \tag{11}$$

and a variable $x_n, (n \in \mathcal{N} \setminus \mathcal{C})$, we can directly write the formulae both for the related marginals

$$F_C(\boldsymbol{x}_C|m) = \prod_{i \in C} f_i(x_i|m), \quad m \in \mathcal{M}_\omega, \quad \omega \in \Omega,$$

$$P_{C|\omega}(\boldsymbol{x}_C|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m F_C(\boldsymbol{x}_C|m),$$

$$P_C(\boldsymbol{x}_C) = \sum_{\omega \in \Omega} p(\omega) \sum_{m \in \mathcal{M}_\omega} w_m F_C(\boldsymbol{x}_C|m),$$

$$P_{n,C}(x_n, \boldsymbol{x}_C|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m f_n(x_n|m) F_C(\boldsymbol{x}_C|m),$$

$$P_{n,C}(x_n, \boldsymbol{x}_C) = \sum_{\omega \in \Omega} p(\omega) \sum_{m \in \mathcal{M}_\omega} w_m f_n(x_n|m) F_C(\boldsymbol{x}_C|m),$$

and for the conditional distributions of $x_n \in \mathcal{X}_n$

$$P_{n|C\omega}(x_n|\boldsymbol{x}_C, \omega) = \frac{P_{n,C|\omega}(x_n, \boldsymbol{x}_C|\omega)}{P_{C|\omega}(\boldsymbol{x}_C|\omega)} = \sum_{m \in \mathcal{M}_\omega} W_m^\omega(\boldsymbol{x}_C) f_n(x_n|m), \quad (P_{C|\omega}(\boldsymbol{x}_C|\omega) > 0), \tag{12}$$

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} \bar{W}_m^\omega(\boldsymbol{x}_C) f_n(x_n|m). \tag{13}$$

Here $W_m^\omega(\boldsymbol{x}_C)$ and $\bar{W}_m^\omega(\boldsymbol{x}_C), (\omega \in \Omega)$ are the component weights corresponding to the observed feature vector $\boldsymbol{x}_C \in \mathcal{X}_C$:

$$W_m^\omega(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j \in \mathcal{M}_\omega} w_j F_C(\boldsymbol{x}_C|j)}, \quad m \in \mathcal{M}_\omega, \ \omega \in \Omega,$$

$$\bar{W}_m^\omega(\boldsymbol{x}_C) = \frac{p(\omega) w_m F_C(\boldsymbol{x}_C|m)}{\sum_{\vartheta \in \Omega} p(\vartheta) \sum_{j \in \mathcal{M}_\vartheta} w_j F_C(\boldsymbol{x}_C|j)}.$$

In view of the above equations, the conditional Shannon informativity of the remaining variables x_n can be computed for arbitrary sub-vector $\boldsymbol{x}_C \in \mathcal{X}_C$.

In particular, if $x_n, (n \in \mathcal{N} \setminus C)$ is an unobserved feature, then the conditional information about Ω contained in the related random variable X_n, given a sub-vector $\boldsymbol{x}_C$, can be expressed by means of the Shannon formula

$$I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega) = H_{\boldsymbol{x}_C}(\Omega) - H_{\boldsymbol{x}_C}(\Omega|\mathcal{X}_n). \tag{14}$$

Equivalently, we can write

$$I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega) = H_{\boldsymbol{x}_C}(\mathcal{X}_n) - H_{\boldsymbol{x}_C}(\mathcal{X}_n|\Omega), \tag{15}$$

where $H_{\boldsymbol{x}_C}(\mathcal{X}_n)$, $H_{\boldsymbol{x}_C}(\mathcal{X}_n|\Omega)$ are the respective entropies:

$$H_{\boldsymbol{x}_C}(\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} - P_{n|C}(x_n|\boldsymbol{x}_C) \log P_{n|C}(x_n|\boldsymbol{x}_C), \tag{16}$$

$$H_{\boldsymbol{x}_C}(\mathcal{X}_n|\Omega) = \sum_{\omega \in \Omega} P_{\Omega|C}(\omega|\boldsymbol{x}_C) \times \sum_{x_n \in \mathcal{X}_n} - P_{n|C\omega}(x_n|\boldsymbol{x}_C, \omega) \log P_{n|C\omega}(x_n|\boldsymbol{x}_C, \omega), \tag{17}$$

$$P_{\Omega|C}(\omega|\boldsymbol{x}_C) = \frac{P_{C|\omega}(\boldsymbol{x}_C|\omega)p(\omega)}{P_C(\boldsymbol{x}_C)}. \tag{18}$$

Finally, we can use the statistical information $I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega)$ to define the next most informative feature x_{n_0}, given $\boldsymbol{x}_C$:

$$n_0 = \arg\max_{n \in \mathcal{N} \setminus \mathcal{C}} \{I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega)\}. \tag{19}$$

Let us note that, in view of Eq. (14), the most informative feature x_{n_0} actually minimizes the conditional entropy $H_{\boldsymbol{x}_C}(\Omega|\mathcal{X}_n)$

$$\begin{aligned} H_{\boldsymbol{x}_C}(\Omega|\mathcal{X}_n) &= \sum_{x_n \in \mathcal{X}_n} P_{n|C}(x_n|\boldsymbol{x}_C) H_{x_n, \boldsymbol{x}_C}(\Omega) \\ &= \sum_{x_n \in \mathcal{X}_n} P_{n|C}(x_n|\boldsymbol{x}_C) \sum_{\omega \in \Omega} - P_{\Omega|nC}(\omega|x_n, \boldsymbol{x}_C) \log P_{\Omega|nC}(\omega|x_n, \boldsymbol{x}_C) \end{aligned} \tag{20}$$

which can be viewed as the expected value of the decision uncertainty with respect to the random variable X_n. In other words the most informative feature x_{n_0} guarantees the maximum decrease of the expected decision uncertainty $H_{\boldsymbol{x}_C}(\Omega|\mathcal{X}_n)$.

As mentioned earlier in Section 4, the entropy $H_{\boldsymbol{x}_C}(\Omega)$ of the posterior distribution (18) is a natural measure of decision uncertainty given the sub-vector of observed feature measurements $\boldsymbol{x}_C$. For this reason it is well applicable as a stopping rule. In particular, if we define

$$H_{\boldsymbol{x}_C}(\Omega) = \sum_{\omega \in \Omega} - P_{\Omega|C}(\omega|\boldsymbol{x}_C) \log P_{\Omega|C}(\omega|\boldsymbol{x}_C) \tag{21}$$

$$H(\Omega) = \sum_{\omega \in \Omega} - p(\omega) \log p(\omega), \tag{22}$$

then a reasonable condition to stop sequential recognition and to make final decision is the inequality

$$\frac{H_{\boldsymbol{x}_C}(\Omega)}{H(\Omega)} < \tau, \quad (0 < \tau < 1). \tag{23}$$

Intuitively it would be quite plausible to apply the maximum *a posteriori* probability $P_{\Omega|C}(\omega_0|\boldsymbol{x}_C)$ in the stopping condition in a similar way:

$$P_{\Omega|C}(\omega_0|\boldsymbol{x}_C) > \tau, \quad 0 < \tau < 1. \tag{24}$$

However, the resulting stopping rule (24) would be less sensitive than (23), because it cannot distinguish between different accompanying less probable "noisy" alternatives.

We recall also that the relation (19) "looks" only one step ahead. This widely used strategy has been proposed by Cardillo and Fu [5] to reduce the computational complexity of the on-line feature ordering. In our case we could easily choose the most informative pair of features, in analogy with the formulae above. Obviously, evaluation of the most informative pair of variables is a qualitatively superior strategy; but if only one feature measurement is accepted in each step, the more complex computation would be devaluated in the next step because the knowledge of the new feature measurement x_{n_0} may essentially change the underlying conditional probabilities (cf. Fig.1).

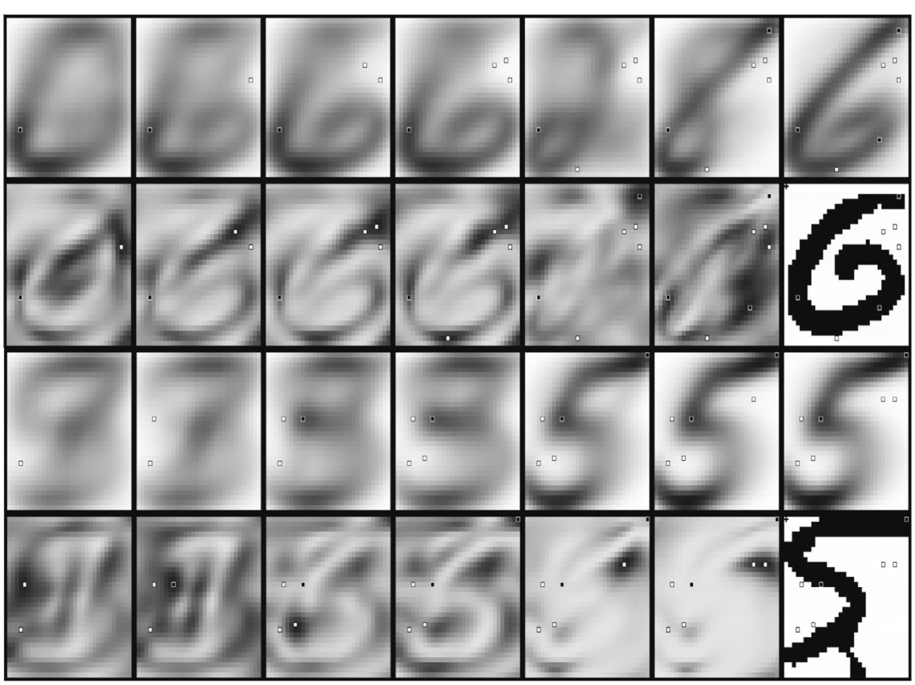

Fig. 1. Sequential recognition of the numerals six and five. The odd rows show the changing expectation of the classifier. The even rows show the informativity of raster fields corresponding to the currently uncovered (white or black) raster fields and finally the input image. Note that the expected images may strongly change in case of unexpected raster field value.

The proposed method of sequential recognition, as described in this section, assumes that the class-conditional distributions $P(\boldsymbol{x}|\omega),\ \omega \in \Omega$ are estimated in the form of product mixtures. The whole procedure can be summarized in three steps:

1. Evaluation of the conditional informativity $I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega)$ for all unobserved features $x_n, n \in \mathcal{N} \setminus \mathcal{C}$, given the subvector $\boldsymbol{x}_C$ of preceding feature measurements.
2. Choice of the most informative feature x_{n_0} by Eq. (19) and inclusion of the new feature measurement x_{n_0} into the subvector $\boldsymbol{x}_C$.
3. Evaluation of the stopping rule (23). A valid condition implies making the final Bayesian classification according to the *a posteriori* distribution (18), otherwise the algorithm continues by point 1.

6. Illustrating example

The proposed sequential recognition controlled by maximum conditional informativity guarantees the best possible strategy in the sense that, for any sub-vector of previously observed features $\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k})$, the next chosen feature measurement minimizes the expected decision uncertainty $H_{\boldsymbol{x}_C}(\Omega|\mathcal{X}_n)$. The result is of theoretical nature. In other words, any other method of sequential recognition based on the same probabilistic description of classes $P(\boldsymbol{x}|\omega)p(\omega), \omega \in \Omega$ can only approach or achieve the same expected decrease of decision uncertainty. In this respect the only application dependent aspect is the quality of the estimated product mixtures.

Obviously, the practical justification of product mixtures by numerical examples is beyond the scope of the present paper, here we refer mainly to our papers published in the last years (cf. Section 4) and to the paper of Lowd and Domingos [23]. The following example of recognition of numerals in the binary raster representation rather illustrates different properties of the proposed method, especially the great variability of decisions in the initial stages of sequential recognition and also the well known trade-off between the error rate and the related number of observed variables. Simultaneously, we make use of the possibility to visualize the changing "expectations" of the classifier with the increasing number of uncovered raster fields.

In recent years we have repeatedly applied multivariate Bernoulli mixtures to recognition of handwritten numerals from the NIST benchmark database, with the goal of verifying different decision-making aspects [15,16]. The considered NIST Special Database 19 (SD19) contains about 400,000 handwritten numerals in binary raster representation (about 40,000 for each numeral). We normalized all digit patterns to a 32×32 binary raster to obtain 1024-dimensional binary data vectors $\boldsymbol{x} \in \{0,1\}^{1024}$. In order to guarantee the same statistical properties of the training- and test-data set, we have used the odd samples of each class for training and the even samples for testing. Also, to increase the variability of the binary patterns, we extended the training data sets four times by making three differently rotated variants of each pattern (by -2, -1 and +1 degrees) with the resulting 80,000 training data vectors for each class.

We approximated the class-conditional distributions of the 1024-dimensional binary patterns by multivariate Bernoulli mixtures

$$P(\boldsymbol{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m), \quad x_n \in \{0,1\}, \tag{25}$$

$$f_n(x_n|m) = \theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \quad 0 \leqslant \theta_{mn} \leqslant 1,\ \omega \in \Omega.$$

In order to estimate the class-conditional distributions (25), we have used the structural modification of the EM algorithm [15,16,10] with the goal of suppressing the noisy parameters of the model. Nevertheless, the resulting components are formally identical with (25), i.e., we have

$$F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} \theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \tag{26}$$

with the only difference being that some of the parameters θ_{mn} are fixed and replaced by their common mean values. The resulting number of components was M = 2007 with the number of parameters θ_{mn} totaling to 1,797,878. The quality of the estimated class-conditional mixtures has been verified by classifying the numerals from the independent test sets (20,000 for each class) with the resulting global error rate of 2.696%. The corresponding classification error matrix is shown in Table 1 in detail.

According to the sequential scheme we assume that the recognized numeral on the raster is not visible and the raster fields become uncovered successively. For this purpose, given a sub-vector of visible raster fields $\boldsymbol{x}_C$, we have to evaluate at each stage the conditional informativity $I_{\boldsymbol{x}_C}(\mathcal{X}_n, \Omega)$ for all the remaining raster fields $x_n, (n \in \mathcal{N} \setminus C)$. In other words, according to (12), (13), we have to compute the marginal distributions

$$P_{C|\omega}(\boldsymbol{x}_C|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m \prod_{i \in \mathcal{C}} \theta_{mi}^{x_i}(1-\theta_{mi})^{1-x_i},$$

$$P_C(\boldsymbol{x}_C) = \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} p(\omega) w_m \prod_{i \in \mathcal{C}} \theta_{mi}^{x_i}(1-\theta_{mi})^{1-x_i},$$

$$P_{nC|\omega}(x_n, \boldsymbol{x}_C|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m \prod_{i \in \mathcal{C} \cup \{n\}} \theta_{mi}^{x_i}(1-\theta_{mi})^{1-x_i},$$

$$P_{nC}(x_n, \boldsymbol{x}_C) = \sum_{\omega \in \Omega} p(\omega) P_{nC|\omega}(x_n, \boldsymbol{x}_C|\omega),$$

in order to evaluate the conditional distributions

Table 1
Classification error matrix obtained by applying the estimated class-conditional mixtures (number of components: $M = 2007$, number of parameters: 1,797,878) to the independent test set. The class-conditional error rates are summarized in the last column and the global percentage of errors is given in the last row.

Class	0	1	2	3	4	5	6	7	8	9	Error rate (%)
0	19,892	5	74	24	36	46	42	4	41	18	1.437
1	7	22,006	40	12	46	16	26	123	67	9	1.548
2	25	55	19,617	65	46	15	20	38	130	25	2.091
3	23	17	103	19,835	3	172	1	29	295	78	3.507
4	41	8	18	1	18,925	13	64	82	75	350	3.330
5	40	27	18	210	12	17,713	52	11	167	53	3.223
6	94	22	40	9	30	172	19,527	2	70	3	2.213
7	9	29	113	30	86	6	0	20,282	46	346	3.175
8	27	40	61	145	25	86	17	34	19,278	77	2.587
9	14	18	19	103	177	39	2	243	180	18,972	4.022
								Mean classification error:			2.696

Table 2
Sequential recognition test for differently chosen thresholds of posterior entropy. The first column contains different threshold values and in the same row follow the related error rates for different numerals. The next row contains the corresponding mean number of observed raster fields. The last column contains the corresponding mean values.

Numerals	0	1	2	3	4	5	6	7	8	9	Mean values (%)
Entropy threshold: 0.05	2.6	1.5	14.0	12.2	4.5	10.5	4.8	7.8	13.4	7.7	7.8%
Mean number of fields:	20.6	37.0	23.7	22.7	18.2	22.2	24.4	36.5	32.7	52.6	29.2
Entropy threshold: 0.10	3.0	1.7	19.0	13.9	5.3	13.4	5.4	8.6	16.1	8.3	9.4%
Mean number of fields:	14.9	20.0	17.7	17.8	13.4	16.5	17.3	30.8	23.9	39.8	21.2
Entropy threshold: 0.15	3.6	2.1	22.4	15.4	6.0	15.4	6.4	8.8	18.8	8.8	10.6%
Mean number of fields:	9.8	16.0	14.1	14.8	11.9	13.7	13.5	24.4	18.9	34.2	17.2
Entropy threshold: 0.20	4.1	2.2	26.4	16.2	7.1	16.5	8.1	10.8	21.0	9.2	12.0%
Mean number of fields:	6.9	13.0	11.4	13.2	9.6	11.9	10.6	19.5	16.0	27.5	14.0
Entropy threshold: 0.25	5.1	2.2	28.9	17.2	8.0	17.9	9.1	12.2	24.7	11.0	13.5%
Mean number of fields:	5.7	12.0	10.3	11.5	8.1	10.5	9.2	18.0	13.1	21.5	12.0
Entropy threshold: 0.30	5.7	2.3	30.8	17.9	8.8	18.6	9.2	14.6	26.4	11.9	14.4%
Mean number of fields:	4.7	10.9	9.6	10.5	7.5	9.1	8.1	16.9	11.6	17.1	10.6
Entropy threshold: 0.35	6.7	8.5	31.9	18.8	10.5	22.7	9.6	13.9	28.3	12.3	16.2%
Mean number of fields:	4.4	9.5	8.5	9.2	6.9	8.2	7.7	13.4	9.4	12.1	8.9

$$P_{n|C\omega}(x_n|\boldsymbol{x}_C,\omega) = \frac{P_{n,C|\omega}(x_n,\boldsymbol{x}_C|\omega)}{P_{C|\omega}(\boldsymbol{x}_C|\omega)} = \sum_{m\in\mathcal{M}_\omega} W_m^\omega(\boldsymbol{x}_C)\theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \quad x_n\in\mathcal{X}_n, \tag{27}$$

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{n,C|\omega}(x_n,\boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{\omega\in\Omega}\sum_{m\in\mathcal{M}_\omega} \bar{W}_m^\omega(\boldsymbol{x}_C)\theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \quad x_n\in\mathcal{X}_n. \tag{28}$$

Finally, by using the probabilities ($n\in\mathcal{N}\setminus\mathcal{C},\omega\in\Omega$):

$$P_{n|C\omega}(1|\boldsymbol{x}_C,\omega) = \sum_{m\in\mathcal{M}_\omega} W_m^\omega(\boldsymbol{x}_C)\theta_{mn}, \tag{29}$$

$$P_{n|C\omega}(0|\boldsymbol{x}_C,\omega) = 1 - P_{n|C\omega}(1|\boldsymbol{x}_C,\omega),$$

$$P_{n|C}(1|\boldsymbol{x}_C) = \sum_{\omega\in\Omega} p(\omega)\sum_{m\in\mathcal{M}_\omega} W_m(\boldsymbol{x}_C)\ \theta_{mn}, \tag{30}$$

$$P_{n|C}(0|\boldsymbol{x}_C) = 1 - P_{n|C}(1|\boldsymbol{x}_C),$$

we can compute the Shannon entropies $H_{\boldsymbol{x}_C}(\mathcal{X}_n), H_{\boldsymbol{x}_C}(\mathcal{X}_n|\Omega)$ to obtain the resulting conditional informativity $I_{\boldsymbol{x}_C}(\mathcal{X}_n,\Omega)$, (cf. (16), (17) and (22)).

Let us remark that the probabilities (30) in the raster arrangement can be interpreted as the conditional expectation of the raster image given the feature measurements $\boldsymbol{x}_C$. In this sense Fig. 1 shows examples of changing "expectation" of the classifier with the increasing number of uncovered raster fields (odd rows). Similarly, for each expected image, we can visualize the corresponding conditional informativity of features by displaying suitably normed informativity values $I_{\boldsymbol{x}_C}(\mathcal{X}_n,\Omega)$ in raster arrangement (even rows). Other examples can be found in the supplementary material. Note that, in case of a surprising raster field value, the expected image may essentially change as it can be seen in Fig. 1. For this reason it would be hardly possible to reduce the number of relevant classes at an early stage of sequential decision-making.

In the experiments the sequential recognition has been stopped by thresholding the normed posterior entropy $H_{\boldsymbol{x}_C}(\Omega)/H(\Omega)$ (cf. (23)). In order to illustrate the trade-off between the classification accuracy and the number of uncovered raster fields, we have tested several thresholds in the stopping rule (23) on the independent test set. Table 2 describes the sequential classification results in detail. The first column contains in even rows the different threshold values τ and in the same row follow the corresponding error rates for different numerals. The next row contains the related mean numbers of uncovered raster fields. The last column contains the global mean values for the underlying stopping rule. It can be seen that, for the threshold $\tau = 0.05$, about 30 raster fields are sufficient in the mean to achieve the classification error of 7.8%. Recall that with all 1024 uncovered raster fields our sequential recognition scheme achieves the same global error 2.696% as the non-sequential classifier from Table 1.

7. Concluding remarks

The sequential problem of statistical pattern recognition can be solved in full generality by approximating the class-conditional distributions using mixtures of product components. In particular, at each stage, given a set of observed measurements, we can compute the conditional informativity of all remaining features and choose the next most informative feature. The most informative feature minimizes the expected decision uncertainty with respect to the estimated product mixtures.

We recall that the product mixtures are suitable to approximate unknown multidimensional and multimodal probability distributions (cf. Section 4). Moreover, we have shown earlier that the mixtures of product components can be used as a knowledge base of the Probabilistic Expert System PES [12]. This system has recently been applied to reproduce the statistical properties of the confidential questionnaire data from the Czech Census 2001 [17]. We recall that by using the final interactive software product [32] the user can derive, with a high degree of accuracy, the marginal distribution of any query variable, possibly conditioned on the values of a set of any evidence variables. In this way the statistical properties of arbitrary subpopulations can be studied in detail.

In case of medical decision-making the sequential classification can be used to design interactive statistical databases - in the sense of the above mentioned census application. The initial database can be designed by medical experts and further developed by means of interactive questioning software. The expert knowledge can be introduced in the database by manually editing the component parameters and, on the other hand, an open access medical expert system can accumulate user-supplied anonymous "questionnaires". We recall that the statistical knowledge base in the form of a product mixture can be estimated from incomplete data [17] and repeatedly upgraded by the increasing data sets. Simultaneously it is possible to identify and remove unreliable questionnaires as data records having low probability. In the case of advanced database the relevance of new manually designed components and features may be automatically verified by means of the EM algorithm in terms of component weights estimated from data.

Acknowledgment

Supported by the Grants of the Czech Science Foundation, No. 14-02652S and 14-10911S.

References

[1] M. Ben-Bassat, Myopic policies in sequential classification, IEEE Trans. Comput. C-27 (1978) 170–174.
[2] M. Ben-Bassat, Pattern-based interactive diagnosis of multiple disorders: the medas system, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-2 (1980) 148–160.
[3] M. Ben-Bassat, D. Teeni, Human-oriented information acquisition in sequential pattern classification: part I - single membership classification, IEEE Trans. Syst. Man Cybern. 14 (1) (1984) 131–138.
[4] A.D. Bimbo, F. Pernici, Towards on-line saccade planning for high-resolution image sensing, Pattern Recognit. Lett. 27 (15) (2006) 1826–1834.
[5] G.P. Cardillo, K.S. Fu, On suboptimal sequential pattern recognition, IEEE Trans. Comput. C-17 (8) (1968) 789–792.
[6] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Statist. Soc. B (39) (1977) 1–38.
[7] B.J. Flehinger, R.L. Engle, HEME: a self-improving computer program for diagnosis-oriented analysis of hematologic diseases, IBM J. Res. Dev. 19 (6) (1975) 557–564.
[8] K.S. Fu, Sequential Methods in Pattern Recognition and Machine Learning, Academic, New York, 1968.
[9] J. Grim, On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions, Kybernetika l8 (3) (1982) 173–190. <http://dml.cz/dmlcz/124132>.
[10] J. Grim, Multivariate statistical pattern recognition with non-reduced dimensionality, Kybernetika 22 (2) (1986) 142–157. <http://dml.cz/dmlcz/125022>.
[11] J. Grim, Sequential decision-making in pattern recognition based on the method of independent subspaces, in: F. Zitek (Ed.), Proceedings, DIANA II Conference on Discriminant Analysis, Mathematical Institute of the AS CR, Prague, 1986, pp. 139–149.
[12] J. Grim, Knowledge representation and uncertainty processing in the probabilistic expert system PES, Int. J. Gen Syst. 22 (2) (1994) 103–111.
[13] J. Grim, Preprocessing of screening mammograms based on local statistical models, in: Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2011), ACM, Barcelona, 2011, pp. 1–5.
[14] J. Grim, M. Haindl, P. Somol, P. Pudil, A subspace approach to texture modeling by using Gaussian mixtures, in: B. Haralick, T.K. Ho (Eds.), Proceedings of the 18th IAPR International Conference on Pattern Recognition ICPR 2006, IEEE Computer Society, Los Alamitos, 2006, pp. 235–238.
[15] J. Grim, J. Hora, Iterative principles of recognition in probabilistic neural networks, Neural Networks 21 (6) (2008) 838–846.
[16] J. Grim, J. Hora, Computational Properties of Probabilistic Neural Networks, in: International Conference on Artificial Neural Networks – ICANN 2010 Part II, Springer, Berlin, 2010, pp. 52–61. LNCS 5164.
[17] J. Grim, J. Hora, P. Boček, P. Somol, P. Pudil, Statistical model of the 2001 Czech census for interactive presentation, J. Off. Stat. 26 (4) (2010) 673–694.
[18] J. Grim, J. Novovičová, P. Somol, Structural Poisson mixtures for classification of documents, in: Proceedings of the 18th IAPR International Conference on Pattern Recognition ICPR 2006, 2006b, pp. 1–4. http://dx.doi.org/10.1109/ICPR.2008.4761669.
[19] J. Grim, P. Somol, M. Haindl, J. Daneš, Computer-aided evaluation of screening mammograms based on local texture models, IEEE Trans. Image Process. 18 (4) (2009) 765–773.
[20] J. Grim, P. Somol, P. Pudil, Digital image forgery detection by local statistical models, in: I. Echizen et al. (Eds.), Proceedings of 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IEEE Computer Society, Los Alamitos, California, 2010, pp. 579–582.
[21] A.B.S. Hussain, R.W. Donaldson, Suboptimal sequential decision schemes with on-line feature ordering, IEEE Trans. Comput. C-23 (1974) 582–590.
[22] M. Kurzynski, A. Zolnierek, Sequential pattern recognition: naive Bayes versus fuzzy relation method, Proceedings International Conference on Computational Intelligence for Modelling, Control and Automation, vol. 1, IEEE, 2005, pp. 1165–1170.
[23] D. Lowd, P. Domingos, Naive Bayes models for probability estimation, in: Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 529–536.
[24] J. Novovičová, P. Pudil, J. Kittler, Divergence based feature selection for multimodal class densities, IEEE Trans. Pattern Anal. Mach. Intell. 18 (2) (2005) 218–223.
[25] E. Parzen, On estimation of a probability density function and its mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[26] J. Šochman, J. Matas, WaldBoost – learning for time constrained sequential detection, in: Computer Vision and Pattern Recognition, (CVPR 2005), IEEE Computer Society Conference on CVPR, 2005, vol. 2, pp. 20–25.
[27] J. Šochman, J. Matas, Learning fast emulators of binary decision processes, Int. J. Comput. Vision 83 (2) (2009) 149–163.
[28] M.I. Schlesinger, Relation between learning and self learning, Kibernetika (Kiev) 2 (1968) 81–88 (in Russian).
[29] I. Vajda, Theory of Statistical Inference and Information, Kluwer Academic Publishers, Dordrecht and Boston, 1989.
[30] A. Wald, Sequential Analysis, John Wiley & Sons, New York, 1947.
[31] R. Willink, A sequential algorithm for recognition of a developing pattern with application in orthotic engineering, Pattern Recognit. 41 (2) (2008) 627–636.
[32] <http://ro.utia.cas.cz/dem.html>.