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.

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