HAVE I SEEN YOU BEFORE? PRINCIPLES OF BAYESIAN PREDICTIVE CLASSIFICATION REVISITED

Jukka Corander[1,3], *Yaqiong Cui*[1], *Timo Koski*[2] *and Jukka Sirén*[1]

[1]Department of Mathematics and statistics, University of Helsinki, FI-00014, Finland
[2]Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden
[3]Department of Mathematics, Åbo Akademi University, FI-20500 Åbo, Finland

ABSTRACT

A general inductive Bayesian classification framework is introduced for data from multiple finite alphabets using predictive representations based on random urn models and generalized exchangeability. We develop a novel principle of generative supervised and semi-supervised probabilistic classification based on marginalizing simultaneous predictive classification probabilities for all test items jointly given training data. Both the simultaneous and marginalized classifiers have attractive theoretical properties and are illustrated in numerical examples with sparse training data to achieve higher correct classification rates compared to a standard predictive classifier which assigns all test items independently to a source. Optimal simultaneous and marginal supervised predictive classifiers are shown to become equivalent classification rules under generalized exchangeability when the amount of training data increases. The predictive framework can under the form of exchangeability considered here coherently handle missing observations without added computational complexity. We discuss briefly a range of options for generalizing the simultaneous classification framework to allow for different data types and less restrictive modeling assumptions.

1. INTRODUCTION

We consider the generic problem of assigning some items into a discrete set of classes using observed characteristics (features) of the items and assessing the uncertainty related to the assignments conditional on all relevant information available. This task is here referred to as *classification* in a broad sense, which may range from the *supervised* classification, where all the eligible classes are *a priori* given, to the *unsupervised* classification (often referred to as *clustering*), where the classes are only identified by their contents, *i.e.* the items to be classified. An intermediate situation between these two extremes is referred to as *semi-supervised* classification, where a certain set of eligible classes is *a priori* determined, while the items are not forced to solely be allocated to such classes, but can also form previously unknown groups during the classification task. It should be noticed that the concept of semi-supervised classification in the literature is used in various distinct meanings, and apart from the definition used here, it may refer to a situation where partial (*e.g.* pairwise) constraints are imposed on an otherwise unsupervised classification [1]. Semi-supervised learning represents in general a wide range of approaches to solving learning tasks [2].

We focus on developing *inductive learning principles* in the context of supervised and semi-supervised classification in the sense these two were defined above. This means that the inductive principles stem from specific assumptions of probabilistic invariance concerning the joint predictive model for training and future (test) datasets. An operationalization of the inductive learning approach is obtained by combining: (i) a model of *simultaneous classification* or *data labeling* defined by generating random urns and assigning data items to them, (ii) a predictive model for future data given training data, and (iii) an algorithm that can compute the sought predictive probabilities.

As discussed in [2], an inductive approach creates a prediction function which is defined on the entire feature space of future data, whereas *transductive learning* performs predictions only for the test set explicitly considered, i.e. no decision rules are inferred beyond the labels for the unlabeled test data. The semi-supervised methods considered in [2] represent transductive learning. Our approach is conceptually related to the work of Ray Solomonoff [3],[4], who developed a formal theory of inductive inference where predictions need not be sequential.

At the heart of the predictive model considered here lies an assumption of predictive invariance corresponding to a generalization of deFinetti's representation theorem for data from finite alphabets [5],[6]. Although this representation is related to the standard assumption of independently and identically distributed (*i.i.d.*) data, there exists an important conceptual difference between modeling data generating mechanism through an *i.i.d.* assumption and predictive invariance. Namely, the former is by necessity defined conditional on a fixed probability measure, which in reality is nearly always unknown and can only be learned from finite data to some degree of accuracy. In contrast, the predictive approach defines inductively distribution for any future dataset by updating knowledge under axioms of probability using the information from available data, while coherently acknowledging the remaining uncertainty about underlying generating probability measure. Such uncertainty can be considerable in the presence of sparse training data and it should intuitively affect any learning decisions made. The two modeling approaches are therefore united in the predictive sense only at the limit when the amount of training data tends to infinity.

In generative supervised classification it is typical to classify test items or samples individually, independently of the classification decisions made for any other item in the test set. The formal justification of this follows in fact from the *i.i.d.* assumption based approach, where the data for any test item is conditionally independent of the data for any other item given the fixed generative probability measure. However, inductive learning theory based on predictive modeling implies that the test items *should be* classified *simultaneously*, because the item data can be considered conditionally independent only given their *joint labeling* with respect to the *a priori* specified classes. Marginal dependence exists between the test items in the predictive probability distribution, since the underlying generating probability measure is not exactly known. We will show that this dependence vanishes at the limit when the amount of training data tends to infinity, and consequently, the two learning approaches become equivalent.

The simultaneous predictive supervised classification principle was pioneered by Seymour Geisser [7],[8],[9] with a primary focus on a Gaussian modeling context, but it has received limited attention in the statistical or machine learning literature over the years as it was not developed into an explicitly operational approach to solving practical application problems. Here we operationalize this approach by combining a generative model for classification structures with the predictive machinery to derive inductive classifiers for multivariate discrete data. We also generalize the inductive learning principle in two important respects. Firstly, we develop the corresponding simultaneous classifier for a semi-supervised situation, where some or all test items are allowed to be generated from *a priori* unknown models lacking training data, while at the same time training data exists for a finite set of putative sources of the test items.

Secondly, we introduce a novel inductive classification principle which stems from an application of law of total probability to the simultaneous classifier. Such a classifier is referred to as a *marginalized*, in contrast to the standard *marginal* classifier which treats all test items independently of each other. The marginalized classifier follows directly from axioms of probability under the predictive learning framework (supervised or semi-supervised) when asking separately for each test item the question about the posterior probability that it should be labeled with a particular source. Thus, at first sight it may appear as if the marginalized classifier is learning in the same fashion as the marginal one, however, the difference is that the marginalized classifier summarizes inductively the total evidence in the combined training and test data supporting any particular labeling for any particular item.

2. ACKNOWLEDGMENTS

Work of JC and YC was funded by the ERC grant no. 239784 and work of JS was funded by the FICS graduate school. This extended abstract is based on work submitted for publication elsewhere.

3. REFERENCES

[1] Basu, S. (2005). Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments. Ph.D. thesis, Department of Computer Sciences, UT at Austin.

[2] Chapelle, O., Schölkopf, B. and Zien, A. (2006). Introduction to Semi-Supervised Learning. In Chapelle, O., Schölkopf, B. and Zien, A. (Editors). Semi-Supervised Learning. Cambridge (MA), MIT Press, pp. 1-12.

[3] Solomonoff, R.J. (1964). A formal theory of inductive inference. *Inform. Control* **7**, 1-22.

[4] Solomonoff, R.J. (2008). Three kinds of probabilistic induction: universal distributions and convergence theorems. *Comput. J.* **51**, 566-570.

[5] Bernardo, J.M. and Smith, A.F.M. (1994). Bayesian Theory. Chichester: Wiley.

[6] Kallenberg, O. (2005). Probabilistic symmetries and invariance principles. New York: Springer-Verlag.

[7] Geisser, S. (1964). Posterior odds for multivariate normal classifications. *J. Roy. Statist. Soc.* **B 26**, 69-76.

[8] Geisser, S. (1966). Predictive discrimination. In Krishnajah, P.R. (Ed.). Multivariate analysis. New York and London: Academic Press, 149-163.

[9] Geisser, S. (1993). Predictive Inference: An introduction. London: Chapman & Hall.