Learning from an imprecise teacher: probabilistic and evidential approaches

Christophe Ambroise[1], Thierry Denœux[1], Gérard Govaert[1], and Philippe Smets[2]

[1] Heudiasyc, Université de Technologie de Compiègne,
B.P. 529, 60205 Compiègne Cedex, France
(e-mail: ambroise,tdenoeux,ggovaert@utc.fr)
[2] IRIDIA,
Université Libre de Bruxelles,
50 av. Roosevelt, 1050 Bruxelles, Belgium
(e-mail: psmets@ulb.ac.be)

Abstract. A type of learning problem is considered, in which the class of training examples is only partially specified. Two approaches to such problems are described: the maximum likelihood approach, in which a probabilistic model relating the imprecise label to the true class is postulated, and the Transferable Belief Model approach, which relies on a non probabilistic formalism for representing and manipulating imprecise information. These two methods are compared experimentally using simulated data sets.

1 Introduction

Discriminant analysis (DA) is a classical tool used in statistics for classifying cases into one of several categories, given the values of some measurement variables. Normally, we use a set of data, called the learning set ($\mathcal{L}$). For each case in $\mathcal{L}$, we know the values taken for each measurement variable, and the classification variable that tells the class to which the case belongs. The classes are finite and unordered. Let Ω denote the set of possible classes: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$.

A learning set is composed of N examples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, where $\mathbf{x}_i$ is a d-dimensional measurement vector describing example i, and $y_i \in \Omega$ denotes the class of that example, as provided by a "teacher". The vector $\mathbf{x}$ for a new case is collected, but the class to which $\mathbf{x}$ belongs, denoted y, is unknown. We want to predict the value of y given the observed values of the measurement variables of $\mathbf{x}$. This is the classical supervised learning problem, for which many techniques exist.

Let us now suppose that, instead of the ideal learning set $\mathcal{L}$ as described above, we have a learning set $\widetilde{\mathcal{L}}$ in which the classes, being provided by an "imprecise teacher", are only partially known. For instance suppose we only know that case $\mathbf{x}_1$ belongs either to ω_1 or ω_2, that case $\mathbf{x}_2$ does not belong to class ω_1, case $\mathbf{x}_3$ belongs either to ω_2 or ω_5 or ω_7, etc. Can we adapt DA methods to such "messy data" case? In fact we face a problem of *partially supervised learning*. For some cases, classes are known as in the supervised learning approach, for some cases, the class is completely unknown as in the unsupervised approach, and for some cases, the classes are only partially known.

Probabilistic solutions could be based on:

- a Bayesian approach where we assess for each case a probability function that describes the class to which it belongs. We then allocate every case to a class (and compute the probability to get that learning set), estimate the needed parameters as in a supervised learning approach, and average the results weighted by the probability of the learning sets.
- a maximum likelihood approach where we estimate the unknown parameters, including the probability with which the case belongs to a given class. The EM method proposed here falls in that category.

The transferable belief model (TBM), a non probabilistic interpretation of "Dempster-Shafer" theory (Smets and Kennes (1994)), provides another approach that can handle elegantly and efficiently such a messy case. This method was first proposed by Denoeux (1995), and is called hereafter the TBM classifier. The aim of the work reported in this paper is to compare the EM and TBM approaches on simulated data sets with partially known class labels.

Previous studies on the TBM classifier were published by Denoeux (1995, 1997), Zouhal and Denœux (1997), De Smet (1998) and Smets (1999). A comparison between results obtained with the TBM classifier and an adapted EM algorithm was already performed by Meyer and Laskey (1999, personal communication). However, this small comparison was not conclusive, hence the idea to reproduce such a study to see how the two methods work, and if one "beats" the other.

2 The EM approach

Classical approaches in DA often consider the conditional class densities to be normal, but more flexibility can be gained by modeling conditional class densities with normal mixture densities. This approach has been advocated by Hastie and Tibshirani (1996), among others; it leads to a hierarchical mixture model. The EM algorithm (Dempster et al. (1977) McLachlan (1997)) provides a very simple and efficient iterative estimation for such models.

In this paper, we consider an original implementation of hierarchical mixtures introduced by Ambroise and Govaert (2000), based on a concept of Maximum Likelihood Estimators (MLE) computed from partially known labels, where each observation may be labeled, unlabeled or partially labeled.

Let us consider $\{(\mathbf{x}_1, y_1, \mathbf{z}_1) ..., (\mathbf{x}_N, y_N, \mathbf{z}_N)\}$ an i.i.d sample where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ and $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)$ are the observed data and $\mathbf{Y} = (y_1, ..., y_N)$ are the unobserved labels:

- $\mathbf{x}_i$ is a feature vector taking values in $I\!R^d$ and following a mixture model:
$$f(\mathbf{x}_i|\Phi) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i;\theta_k) \tag{1}$$
where $\Phi = (p_1, ..., p_{K-1}, \theta_1, ..., \theta_K)$ denotes the vector of parameters of the model.
- $y_i \in \Omega$ indicates the true label; it follows a multinomial distribution.
- $\mathbf{z}_i$ takes values in $\{0,1\}^K$. $z_{ik} = 1$ means that $\mathbf{x}_i$ may belong to class k and $z_{ik} = 0$ means that $\mathbf{x}_i$ does not belong to class k. It can be considered as an available expert knowledge which constrains the set of possible labels.

In this paper we assume conditional independence of the z_{ik}, given y, and consider the following distribution for $\mathbf{z}_i|y_i$:

$$P(z_{i\ell} = 1|y_i = k) = \begin{cases} 1 & \text{if } \ell = k \\ \\ \epsilon & \text{otherwise} \end{cases}$$

where $\epsilon \in [0,1]$ is an unknown parameter. This distribution can be interpreted as follows: the expert always indicates a subset of classes which includes the true class, but has some doubts which are identically distributed over the classes.

We want to estimate the posterior distribution of the true label y_i, taking into account all the available information $(\mathbf{x}, \mathbf{z})$. In this general setting, the EM algorithm can be used to maximize the log likelihood according to Φ. Considering the incomplete data to be $(\mathbf{X}, \mathbf{Z})$ and the missing data to be $\mathbf{Y}$, the EM algorithm is an iterative algorithm which starts from an initial value of the parameter vector, Φ^0, and maximizes the expectation of the complete data likelihood at each iteration.

3 The TBM approach

Let L_i denote the subset of Ω that represents what we know about the class to which case $\mathbf{x}_i$ belongs. The learning set $\widetilde{\mathcal{L}}$ is $\{(\mathbf{x}_i, L_i) : i = 1, 2 \ldots, N\}$.

Intuitively, the method can be described by an anthropomorphic model. Each case $\mathbf{x}_i$ in $\mathcal{L}_{pk}$ is considered as an individual. Let y_i denote the true class to which $\mathbf{x}_i$ belongs. All $\mathbf{x}_i$ knows about y_i is that $y_i \in L_i$.Then, $\mathbf{x}_i$ looks at the unknown case and expresses "his" belief bel_i about y. If $\mathbf{x}$ is "close" to $\mathbf{x}_i$, $\mathbf{x}_i$ would defend that $y = y_i$. As $\mathbf{x}_i$ only knows that $y_i \in L_i$, then all what $\mathbf{x}_i$ can express about case $\mathbf{x}$ is that $y_i \in L_i$. If $\mathbf{x}$ is not "close" to $\mathbf{x}_i$, $\mathbf{x}_i$ cannot say anything about y.

This description is formalized as follows. $\mathbf{x}_i$ can only state: "case $\mathbf{x}$ belongs to the same set of classes as myself", which is represented by a belief function with $m^{\Omega}_{i0}(L_i) = 1$. Let $\delta(\mathbf{x}_i, \mathbf{x})$ be the distance (according to some distance measure δ) between $\mathbf{x}_i$ and $\mathbf{x}$. If $\delta(\mathbf{x}_i,\mathbf{x})$ is small, then what $\mathbf{x}_i$ states is reliable, if $\delta(\mathbf{x}_i,\mathbf{x})$ is large, it is not reliable: the largest $(\mathbf{x}_i,\mathbf{x})$, the less reliable. The impact of this reliability is represented by a discounting on m_{i0}^{Ω} into m_i^{Ω}. So $m_i^{\Omega}(L_i) = \varphi(\delta(\mathbf{x}_i,\mathbf{x}))$ and $m_i^{\Omega}(\Omega) = 1 - \varphi(\delta(\mathbf{x}_i,\mathbf{x}))$, where $\varphi(\delta) \in [0,1]$ and is decreasing with δ. Thus, every case $\mathbf{x}_i$ generates such a simple support function bel_i^{Ω} that concerns the value of y.

Consider now what information case $\mathbf{x}$ collects. It receives all these simple support functions bel_i^{Ω}, and combines them by Dempster's rule of combination into a new belief function bel^{Ω} that represents the belief held by case $\mathbf{x}$ about y and induced by the collected belief functions bel_i^{Ω}: $bel^{\Omega} = \bigcap_{i=1,\dots,N} bel_i^{\Omega}$ where $\cap$ denotes the conjunctive combination operator. If a decision must be made regarding the value of y, we build the pignistic probability $BetP^{\Omega}$ on Ω from bel^{Ω} by the application of the pignistic transformation (described and justified in Smets and Kennes (1994)) and use the classical expected utility theory in order to take the optimal decision (Denoeux (1997)).

Details concerning the particular implementation of the method employed in this study are described in a Master Thesis (De Smet, 1998) available from the authors. This implementation uses a local generalized Euclidean distance function based on a covariance matrix S_i that depends on $\mathbf{x}_i$ and whose parameters are computed using the cases in the neighborhood of $\mathbf{x}_i$.

4 Experiments

4.1 Learning tasks

Two artificial learning tasks were used in this study. In each case, the imprecise label L_i is never erroneous, i.e. the true class of $\mathbf{x}_i$ always belongs to L_i.

Case Study 1 (Double isoscele triangles): We have three classes $\omega_1, \omega_2, \omega_3$. Each class is made of data from two subsets. In each subset, data are two-dimensional normally distributed with means (μ_X, μ_Y) and variance matrix σI as given in Table 1. The underlying nature of these data is that the pure cases come from well clustered data ($\sigma = 0.5$) located at the 3 corners of an isoscele triangle, whereas the partially known cases come from largely spread data ($\sigma = 2$) located between the two other pure classes. The partial class label were given randomly for all those cases in 2 subgroups. The total learning set is made of three hundred cases.

The testing set is made of six sets of 50 cases from each of the six subgroups.

Case Study 2 (The nut): We have three classes and two variables:

Table 1. The double isoscele triangles data distribution.

	ω_1 sub 1	ω_1 sub 2	ω_2 sub 1	ω_2 sub 2	ω_3 sub 1	ω_3 sub 2
μ_X	10	15	20	12.5	15	17.5
μ_Y	10	18.6	10	14.3	10	14.3
σ	.5	2.	.5	2.	.5	2.
cases	50 ω_1	25 ω_1, ω_2 25 ω_1, ω_3	50 ω_2	25 ω_1, ω_2 25 ω_2, ω_3	50 ω_3	25 ω_1, ω_3 25 ω_2, ω_3

- The ω_1 cases come from a normal distribution with mean (0,0) and variance matrix I, but all cases with a distance to (0,0) not included in [1, 1.2] were deleted. Hence the ω_1 cases are in a "crown". Fifty cases were labeled ω_1, 25 (ω_1, ω_2), and 25 (ω_1, ω_3).
- The ω_2 cases come from a normal distribution with mean (0,0) and variance matrix $0.7I$, but all cases with $x \leq 0.2$ or a distance to (0,0) larger than 0.9 were deleted. Hence the ω_2 are like a left half nut. Fifty cases were labeled (ω_1, ω_2) and 50 (ω_2, ω_3).
- The ω_3 cases mirror the ω_2 cases, so all cases with $x \geq -0.2$ or a distance to (0,0) larger than .9 were deleted. Hence the ω_3 are like a right half nut. Fifty cases were labeled (ω_1, ω_3) and 50 (ω_2, ω_3).

4.2 Results

The rates of correct classification for the two methods and the two learning tasks are shown in Table 2. For the EM method, each class was modeled by a mixture of 4 Gaussian distributions, with equal proportions and scalar covariance matrices. As shown by Table 2, the performances of the two methods are almost exactly equivalent for these two tasks.

Table 2. Percentage of correct classification for the 2 problems and the two methods. For each problem, the results for 10 independent training sets are given, as well as the mean and standard deviation over the 10 trials.

		1	2	3	4	5	6	7	8	9	10	mean	std
Triangle	EM	85.3	84.3	86.3	88.0	86.7	87.0	83.3	85.7	90.7	88.0	86.5	2.1
	TBM	86.7	85.0	86.7	89.0	87.3	88.0	84.0	86.3	88.3	86.7	86.8	1.5
Nuts	EM	90.7	87.7	91.3	89.7	87.3	93.7	90.7	93.3	91.3	93.7	90.9	2.3
	TBM	90.7	88.3	89.3	92.7	89.7	94.0	88.0	87.3	94.3	87.7	90.2	2.6

5 Discussion

It seems the TBM-classifier and the EM-classifier provide nice tools, but do they fill a real need? The answer is affirmative. Real life hardly complies with the perfect knowledge usually required by classical statistical tools. Real life provides messy data, not idealized data as one hopes for. As an example, consider the clinician who collects during the 1980's the data from three hundred patients suffering from a given disease D. In the 80's such patients were classified as ω_1 or as ω_2. Then as science advances, a new category ω_3 is described for patients with disease D. So, during the 90's, our clinician collects two hundred data classified as ω_1, ω_2 or ω_3. The clinician comes to you and asks for a computerized classifier. How to handle the first three hundred cases, given that the ω_1 cases were in fact ω_1 or ω_3, and the ω_2 cases were ω_2 or ω_3, and their exact classes cannot be re-assessed? Are you going to throw away the three hundred cases as useless? With the classifiers presented here, you can proceed with all the five hundred cases.

Of the two methods described here, which one is the best? As usual, an empirical study such as the one presented in this paper does not provide a final answer to this question. The real conclusion will be that both approaches have merits, and should be included in a discrimination toolbox. Note that the TBM approach has been generalized by Denœux and Zouhal (2001) to the case where knowledge of class membership is represented by a belief function or a possibility function on Ω. Extensions of the EM approach to more complex cases can also be considered and are left for further study.

References

AMBROISE, C. and GOVAERT, G (2000): EM for partially known labels. Proceedings of IFCS'2000, Namur, Belgium, vol. 1.

DE SMET, Y. (1998): Application de la théorie des fonctions de croyance aux problèmes de classification. Master Thesis, Université Libre de Bruxelles.

DEMPSTER, A.P. LAIRD, N. M. and RUBIN, D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1-38.

DENŒUX, T. (1995): A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. SMC*, 25(05):804-813.

DENŒUX, T. (1997): Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095-1107.

DENŒUX, T. and ZOUHAL, L. M. (2001): Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, to appear.

HASTIE, T. and TIBSHIRANI, R. J. (1996): Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society B*, 58:155–176.

MC LACHLAN, G. J. and KRISHNAN, T. (1997): *The EM Algorithm and Extensions.* Wiley.

SMETS, P. and KENNES, R. (1994): The Transferable Belief Model. *Artificial Intelligence*, 66:191–243.

SMETS, P. (1999): Practical uses of belief functions. In *UAI'99*, Stockholm, Sweden.

ZOUHAL, L. M. and DENŒUX, T. (1997): An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. SMC C*, 28(2):263–271.