Learning from partially labelled data—with confidence

Eric Gaussier ERIC.GAUSSIER@XRCE.XEROX.COM
Cyril Goutte CYRIL.GOUTTE@XRCE.XEROX.COM
Xerox Research Centre Europe, 6 ch. de Maupertuis, F-38240 Meylan, France

Abstract

In this paper, we propose a unifying treatment of several strategies for training mixture models from label-deficient data. After a review of different approaches to estimating classification models on partially labelled data using mixture models, we identify a number of problems which lead us to propose a new EM variant. The aim is to better handle unlabelled data and provide a more confident discrimination decision. This is illustrated by an experimental comparison of the different models on the *Leptograpsus* crab data.

1. Overview

Supervised Machine Learning techniques have reached a level of sophistication that allows the efficient automatic training of various linear and non-linear models, even in situations like Natural Language Processing tasks where examples live in very-high dimensional spaces (Joachims, 1998). In many situations, however, labelling data is a costly and time-consuming process. Annotating biological texts, for example, requires the help of educated biologists who, in addition to being expensive, may be reluctant to carry out tedious annotation tasks. On the other hand, unannotated data is often plentiful. In biology, querying PubMed (PubMed, 2005) with few appropriately formed queries can easily return thousands of documents. Unsupervised learning techniques may be applicable to this plethora of unlabelled data. However, they are often less sophisticated than supervised methods, tend to require more data to reach comparable performance, and, most importantly, are intrinsically unable to satisfactorily address typical supervised learning problems, such as discriminant analysis. The development of methods that can learn from a combination of labelled and unlabelled (aka partially-labelled) data is therefore of high practical relevance. In the context of mixture models, such work was initiated by Miller and Uyar (Miller & Uyar, 1997), following early work on the topic (Shashahani & Landgrebe, 1994), and applied for example by (Nigam et al., 2000).

Other techniques have been proposed for learning from partially labelled data. In the context of statistical learning, let us mention for example the transductive inference approach to Support Vector Machines (Gammerman et al., 1998; Joachims, 1999), or the use of Fisher kernels (Jaakkola & Haussler, 1999) obtained from probabilistic models of the unsupervised data. Although these various techniques may in principle be combined, some investigations have revealed that this does not necessarily yield an improvement in performance (Goutte et al., 2004).

In this contribution, we focus on the mixture model approach. We identify crucial issues with respect to their use on partially-labelled data: soft partitioning versus hard partitioning, the *cluster impurity* problem, and the treatment of unlabelled data clusters. We propose a unifying review of different learning strategies, and propose a new EM variant for learning mixture models on partially-labelled data. This is illustrated by experimental results obtained by Gaussian mixtures, similar to Mixture Discriminant Analysis (Hastie & Tibshirani, 1996), on the *Leptograpsus* crab data.

2. Estimating mixture models on partially-labelled data

In situations where few labelled data co-exist with large amounts of unlabelled data, combining labelled and unlabelled data to design statistical classifiers meets the needs for models that are both adequate, ie they correctly model the data to be classified, and discriminative, ie they allow for a clear separation between data from different classes. Unlabelled data plays a crucial role in model adequacy since the

Appearing in *Proc. of the 22st ICML Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, August 2005.

amount of labelled data is typically insufficient to accurately estimate model parameters. Miller and Uyar (Miller & Uyar, 1997) propose two EM variants for designing statistical classifiers from label-deficient data, using mixture models. These two variants are close but differ in the definition of the latent variables used in the EM algorithm (cf. section 2.2). Miller and Uyar also discuss two ways to map mixture components to class labels. In *hard-partitioning*, the mixture components unequivocally determine the class labels, ie each mixture component is associated to one and only one class. In *soft partitioning* the assignment is probabilistic, ie each component may in principle generate labels from all classes.

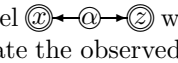
The difference between the two EM variants proposed in (Miller & Uyar, 1997) is mainly formal, since they represent two ways to maximise the same likelihood. However, the difference in partitioning methods leads to models with different levels of flexibility. In all cases, however, class labels for unlabelled data can naturally be derived through maximum a posteriori classification. Test data may be directly integrated as unlabelled data in the learning phase, improving the model estimation and getting directly a class assignment in the process. This may be seen as an instance of *transductive inference* (Vapnik, 1998).

In the following, we assume that we have a dataset $\mathcal{D} = (\mathcal{U}, \mathcal{L})$ with $\mathcal{U} = \{x^{(i)}\}_{i=1,\dots m}$ the unlabelled examples set and $\mathcal{L} = \{(x^{(i)}, z^{(i)})\}_{i=m+1,\dots n}$ the labelled examples set. We model the data using a mixture model with K components α:

$$P(x,z) = \sum_{\alpha=1}^{K} P(\alpha)P(x|\alpha)P(z|\alpha)$$

for labelled data and

$$P(x) = \sum_{z} P(x,z) = \sum_{\alpha} P(\alpha)P(x|\alpha)$$

for unlabelled data. This corresponds to a simple graphical model  where unobserved components α generate the observed data x and corresponding label z independently. Training mixture models is conveniently done using the EM algorithm (Dempster et al., 1977). We review below 3 versions of EM that have been proposed so far for handling partially labelled data. For the last two versions, we then discuss hard and soft partitioning of labels with respect to components. We describe the problem of *cluster impurity* which has adverse effect on the discriminative power of the model when classes are unbalanced and badly aligned with the underlying density. This leads us to introduce a new version of EM to train mixture models on partially-labelled data, which takes into account the proportion of unlabelled data associated with each mixture component.

2.1. EM0

A simple way to bypass the problems raised by partially-labelled data in probabilistic modelling is to strip labelled data of their label, and estimate the mixture parameters for $P(x)$ on all the (now unlabelled) data. Of course, this does not directly yield a classifier. However, the model for $P(x)$ may be used for example to derive a Fisher kernel, which may then be used to obtain a classifier from the labelled or partially labelled data (Hofmann, 2000; Goutte et al., 2004).

The log-likelihood $L_0 = \sum_i \ln P(x^{(i)})$ is maximised using the EM algorithm. We call this version EM0. The E-step equation is:

$$C^{(t)}(\alpha, i) = P^{(t)}(\alpha|x^{(i)}) = \frac{P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)}{\sum_\alpha P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)} \quad (1)$$

For mixing parameters $P(\alpha)$, the M-step equation is:

$$P^{(t+1)}(\alpha) = \frac{1}{N}\left(\sum_i C^{(t)}(\alpha, i)\right) \quad (2)$$

The M-step equations for $P(x|\alpha)$ depend on the form of the component distributions and are given in the appendix for Gaussian mixtures. As this is an EM algorithm, iterating the E- and M-step equations guarantees convergence to a (local) maximum of L_0.

2.2. EM1 and EM2

The above probabilistic modelling, although simple, is somewhat deficient, since it does not make use of all available information, i.e. the labels available for part of the data. Taking labels into account leads to the log-likelihood

$$\begin{aligned} L_1 &= \sum_{i\in\mathcal{U}} \ln P(x^{(i)}) + \sum_{i\in\mathcal{L}} \ln P(x^{(i)}, z^{(i)}) \\ &= L_0 + \sum_{i\in\mathcal{L}} \ln P(z^{(i)}|x^{(i)}) \end{aligned} \quad (3)$$

Miller and Uyar (Miller & Uyar, 1997) propose two versions of the EM algorithm that optimise the same log-likelihood (eq. 3). Although they differ on which latent variables are considered during EM, they both optimise the same likelihood. The should therefore yield very similar results, as exemplified in the experimental section of (Miller & Uyar, 1997).

In **EM1**, the only unobserved variable for labelled *and* unlabelled data is the component α. As this is a latent variable model, the log-likelihood (eq. 3) is again

maximised using the EM algorithm. For unlabelled data, the E-step equation is unchanged (see eq. 1). For labelled data it becomes:

$$\begin{aligned} C^{(t)}(\alpha, i) &= P^{(t)}(\alpha|x^{(i)}, z^{(i)}) \\ &= \frac{P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)P^{(t)}(z^{(i)}|\alpha)}{\sum_\alpha P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)P^{(t)}(z^{(i)}|\alpha)} \end{aligned} \quad (4)$$

Using this notation the M-step equations for $P(\alpha)$ (and $P(x|\alpha)$) are the same as before (eq. 2), using the new expression of $C^{(t)}(\alpha, i)$ for labelled examples. The additional M-step equation needed for re-estimating $P(z|\alpha)$ is:

$$P^{(t+1)}(z|\alpha) = \frac{\sum_{i \in \mathcal{L}, z^{(i)}=z} C^{(t)}(\alpha, i)}{\sum_{i \in \mathcal{L}} C^{(t)}(\alpha, i)} \quad (5)$$

Again, EM convergence proofs guarantee that iterating the E- and M-step equations for EM-1 lead to a (local) maximum of L_1.

In **EM2**, the latent variables are the component α for all data, as well as the label z for all the unlabelled examples. Although the unobserved variable are different from EM-1, this is still a latent variable model. Therefore, we again maximise the likelihood L_1 using EM. In the E-step, we need to estimate the expectation of the joint observation of α and z for all unlabelled examples $i \in \mathcal{U}$:

$$\begin{aligned} P^{(t)}(\alpha, z|x^{(i)}) &= \frac{P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)P^{(t)}(z^{(i)}|\alpha)}{\sum_\alpha P^{(t)}(\alpha)P^{(t)}(x^{(i)}|\alpha)} \\ &= C^{(t)}(\alpha, i)P^{(t)}(z|\alpha) \end{aligned} \quad (6)$$

with $C^{(t)}(\alpha, i)$ defined as in eq. 1, while for all labelled examples $i \in \mathcal{L}$, the E-step equation is still as eq. 4.

Because $\sum_z C^{(t)}(\alpha, i)P^{(t)}(z|\alpha) = C^{(t)}(\alpha, i)$, it turns out that the M-step equations for $P(\alpha)$ and $P(x|\alpha)$ are, again, unchanged. The only modification to the M-step equations in this variant concerns $P(z|\alpha)$, which now takes unlabelled data into account:

$$\begin{aligned} P^{(t+1)}(z|\alpha) &= \frac{1}{\sum_i C^{(t)}(\alpha, i)} \\ &\times \left(\sum_{i \in \mathcal{U}} C^{(t)}(\alpha, i)P^{(t)}(z|\alpha) + \sum_{\substack{i \in \mathcal{L} \\ z^{(i)}=z}} C^{(t)}(\alpha, i) \right) \end{aligned} \quad (7)$$

Models are trained by iterating the E- and M-step equations until convergence. Again, this is an EM algorithm and therefore converges to a (local) maximum of L_1. In fact, because they maximise the same likelihood, EM1 and EM2 should yield similar models, as exemplified in (Miller & Uyar, 1997).

These models may then be used to provide posterior class probabilities for unlabelled data:

$$P(z|x) = \frac{P(z, x)}{P(x)} = \frac{\sum_\alpha P(z|\alpha)P(x|\alpha)P(\alpha)}{\sum_\alpha P(x|\alpha)P(\alpha)} \quad (8)$$

The unlabelled examples from the training data and (when applicable) new examples may then be classified on the basis of $P(z|x)$.

2.3. Hard vs. Soft Partitioning

There are two ways to assign components to classes. In *hard partitioning*, the assignment is binary, ie $P(z|\alpha) = 1$ if component α is associated with class z and $P(z'|\alpha) = 0$ for all other classes. In *soft partitioning*, the assignment is probabilistic, ie each component α potentially generates examples from all classes, $P(z|\alpha) \in [0, 1]$. It has been reported that *soft partitioning* yields better results than *hard partitioning*, at least for balanced classes (Miller & Uyar, 1997). However, it also potentially faces the problem of *cluster impurity*, especially when classes are unbalanced. This happens when all components contain examples from several classes instead of "specialising" to one or few classes. As an illustration, consider the model shown in figure 1. Out of eight components, three are associated almost exclusively with the largest class. The rest generates varying proportions of all classes. As a consequence, better modelling these components using unlabelled data will not help the discrimination task, which is to discriminate each class versus the others. In such cases, the resulting generative model will therefore have poor discriminative power.

One way to impose purity on the model is to use *hard partitioning* of components to classes. It turns out that hard partitioning is easy to implement by initialising $P(z|\alpha)$ to binary (0/1) values and using the regular EM equations (this is apparent by cycling through eq. 4, 6 and 5, 7). Note that for *hard partitioning*, EM-1 and EM-2 are identical as $P(z|\alpha)$ is fixed: the class label is uniquely determined by the component assignment.

Situations where the ratio of labelled to unlabelled examples is very low pose an additional problem of reliability of the class assignments. In that situation, it is likely that some components (which we will call *unlabelled components*) will model *only* unlabelled data. In this case, *hard partitioning* will "arbitrarily" assign a class label to the component.[1] This assignment is

[1]Note that *soft partitioning* will also yield arbitrary class probabilities in that case: eq. (7) becomes, in this situation, $P^{(t+1)}(z|\alpha) \approx P^{(t)}(z|\alpha)$.

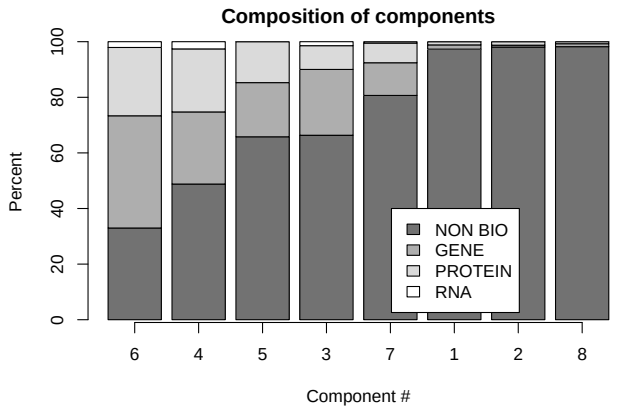


Figure 1. The problem of component impurity. The graph shows the composition of each component of an 8-component mixture modelling four unbalanced classes with soft partitioning. The largest class "takes over" most components, and the smaller classes are not well represented by components. As a consequence, the mixture model is unlikely to yield helpful discriminative information.

arbitrary in the sense that it has no influence on the likelihood. This in turn will lead the model to take arbitrary decisions at classification time, and do so with high confidence. In order to address this problem, we need to take into account the composition of each component in labelled and unlabelled examples. In the next section we propose a new variant of EM that does so in a principled way and addresses both problems of *cluster impurity* and *unlabelled components*.

3. Improving classification confidence with EM3

One way to deal with unbalance between labelled and unlabelled data is to down-weight the unlabelled data by introducing a multiplicative parameter $\lambda \in [0,1]$ in front of the unlabelled contribution in eq. 3 (or 7), as advocated for example in (Nigam et al., 2000). This however does not solve our problem since unlabelled components can still be present in the final solution. In order to take into account the composition of each mixture component in a principled way, we restore the symmetry between labelled and unlabelled data by introducing an additional "label" for unlabelled examples. In binary classification for example, instead of having $z^{(i)} \in \{+,-\}$ for $i \in \mathcal{L}$, we now have $z^{(i)} \in \{+,0,-\}$ for all $i \in \mathcal{U}$. *All* observations are pairs $\left(x^{(i)}, z^{(i)}\right)$, and α is the only latent variable. Table 1 gives an overview of the differences between EM3 and the other EM variants presented earlier (we use square brackets to indicate partially observed data, as z in EM1 or EM2, observed for labelled data only).

Parameters are trained by maximising the log-likelihood taking these new "labels" into account:

$$\begin{aligned} L_3 &= \sum_{i\in\mathcal{D}} \ln P(x^{(i)}, z^{(i)}) \\ &= L_1 + \sum_{i\in\mathcal{U}} \ln P(x^{(i)}, z=0) \end{aligned} \tag{9}$$

Note that this likelihood is partially completed but is still not the complete likelihood (for which α are observed).

The model is still a latent variable model which we again train using the EM algorithm. It turns out that the E-step equation is identical to eq. 4 for all data, and the M-step equations for $P(\alpha)$ and $P(x|\alpha)$ are unchanged. The main difference is the re-estimation formula for $P(z|\alpha)$, which a/ runs over all data and b/ has an additional value $z = 0$:

$$P^{(t+1)}(z|\alpha) = \frac{\sum_{\substack{i\in\mathcal{D} \\ z^{(i)}=z}} C^{(t)}(\alpha, i)}{\sum_i C^{(t)}(\alpha, i)} \tag{10}$$

As before, iterating the E- and M-step equations is an EM algorithm and therefore guarantees convergence to a (local) maximum of L_3.

Note that using EM3 with *soft partitioning* essentially amounts to turning the semi-supervised learning problem into a supervised learning problem with one additional class. The situation is different with *hard partitioning* because of a key difference: the hard partitioning constraint is imposed only on "true" labels, not on the added "fake" label. Unlabelled data may therefore be distributed over all components. Despite this difference, hard partitioning can still be implemented using the appropriate initial conditions.

In EM3, unlabelled components retain the possibility to generate examples with the added "fake" label, rather than be forced to generate an arbitrary label. Once the model is trained using EM3, we use it to classify unlabelled or new examples. The main issue is to distribute the probability mass associated with the "fake" z ($P(z=0|\alpha)$) onto the "real" labels. Using an additional variable ℓ for the "real" labels, the posterior probability of ℓ given an example x is obtained as:

$$\begin{aligned} P(\ell|x) &= \sum_z \sum_\alpha P(\ell|z) P(z|\alpha) P(\alpha|x) \\ &= \sum_\alpha P(\alpha|x) P(z=\ell|\alpha) \\ &\quad + P(\ell|z=0) \sum_\alpha P(\alpha|x) P(z=0|\alpha) \end{aligned} \tag{11}$$

	Observed	Complete	Log-Likelihood to maximise
EM0	$x^{(i)}$	$\left(x^{(i)}, \alpha^{(i)}\right)$	$L_0 = \sum_i \ln P(x^{(i)})$
EM1	$\left(x^{(i)}, [z^{(i)}]\right)$	$\left(x^{(i)}, \alpha^{(i)}, [z^{(i)}]\right)$	$L_1 = L_0 + \sum_{i \in \mathcal{L}} \ln P(z^{(i)}\|x^{(i)})$
EM2	$\left(x^{(i)}, [z^{(i)}]\right)$	$\left(x^{(i)}, \alpha^{(i)}, z^{(i)}\right)$	L_1
EM3	$\left(x^{(i)}, z^{(i)}\right)$	$\left(x^{(i)}, \alpha^{(i)}, z^{(i)}\right)$	$L_3 = L_1 + \sum_{i \in \mathcal{U}} \ln P(z^{(i)}\|x^{(i)})$

Table 1. Observed and completed examples, and associated log-likelihood for the different training methods considered here. Square brackets indicate partial observations.

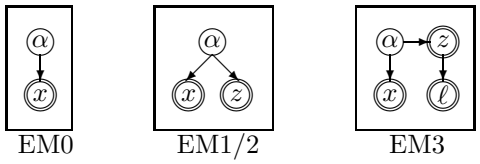


Figure 2. Graphical models for EM0, EM1/2 and EM3.

where we have used the fact that $P(\ell = +|z = -) = P(\ell = -|z = +) = 0$. The first term in eq. 11 may be seen as the contribution from the labelled data, while the second term is the contribution from the unlabelled data. This corresponds to the graphical model in figure 2 (right), where for comparison we also present the graphical models associated with EM0 and EM1/2 (these graphical models are slightly simplified to illustrate their differences).

$P(\ell|z = 0)$ represents a prior probability that an unlabelled component generates examples from class ℓ. One possibility is $P(\ell|z = 0) = 1/2$, a uniform, non-informative prior. Another possibility is to reflect class priors. This may however lead to undesirable results. For example, a component α, associated with the positive class, but containing lots of unlabelled data will typically yield $P(z = 0|\alpha) \gg P(z = +|\alpha)$. If $P(\ell|z = 0)$ is biased towards the negative class, this will lead to the counter-intuitive decision that examples from this positive component are classified as negative examples... We address this problem[2], in a manner similar to (Nigam et al., 2000), but applied at a different level, namely the final categorization decision, by down-weighting the influence of the unlabelled examples in the decision, by a factor $\lambda \in [0, 1]$:

$$\begin{aligned} P(\ell|x) &= \sum_\alpha P(\alpha|x)P(z = \ell|\alpha) \\ &\quad + \lambda P(\ell|z = 0) \sum_\alpha P(\alpha|x)P(z = 0|\alpha) \end{aligned}$$

[2]Alternatively, we can use an improper prior on the label generation, $P(\ell = +|z = 0) = P(\ell = -|z = 0) = \rho \ll 1$. In the experiments, we use $\rho = 0.01$, which corresponds to a uniform $P(\ell|z)$ with $\lambda = 0.02$.

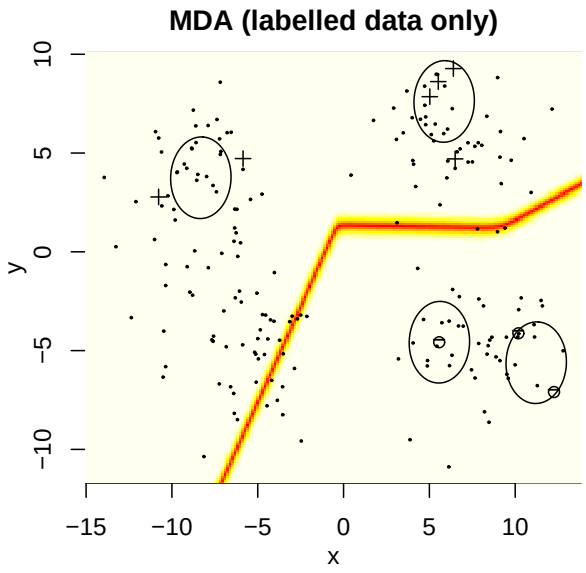


Figure 3. Mixture discriminant analysis using the 9 labelled examples only.

4. Application to Mixture Discriminant Analysis

As an illustration, we use a mixture of Gaussian with hard partitioning to address a discriminant analysis problem with very few labelled data. We use the *Leptograpsus* crab data (See (Ripley, 1996)). It records morphological features of 50 specimens of each sex of each of two species of rock crab (orange or blue). Projected in the first two canonical variates as shown in (Ripley, 1996, p.97), the examples are distributed in 4 groups of roughly equal sizes and shapes. In our experiments, we randomly labelled 9 examples: 6 correspond to male specimens (+ class) and 3 correspond to female specimens (- class). All 3 females turn out to belong to the *orange* species, such that there is no labelled example from the "blue female" group (bottom left on the figures).

Using hard partitioning is quite similar to using *mixture discriminant analysis* (MDA) (Hastie & Tibshirani, 1996). Standard MDA essentially fits a mixture

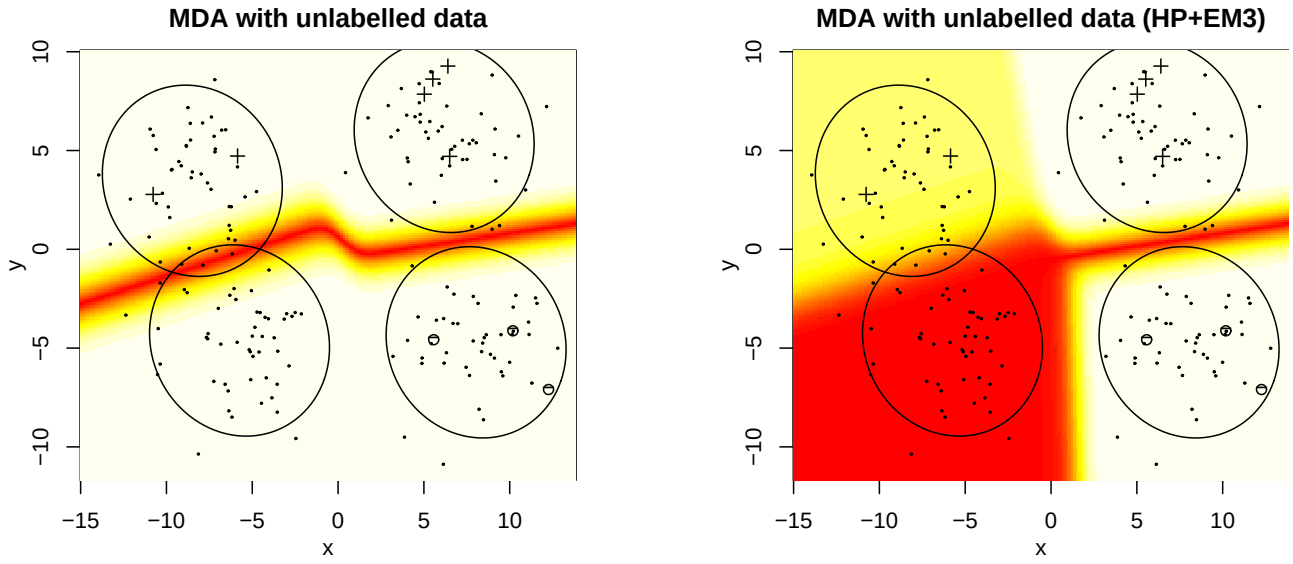


Figure 4. Mixture discriminant analysis results using 2 components per class for EM1 (left) and EM3 (right). Color shading indicates assignment probability: dark/red for intermediate probabilities, light/yellow for extreme probabilities.

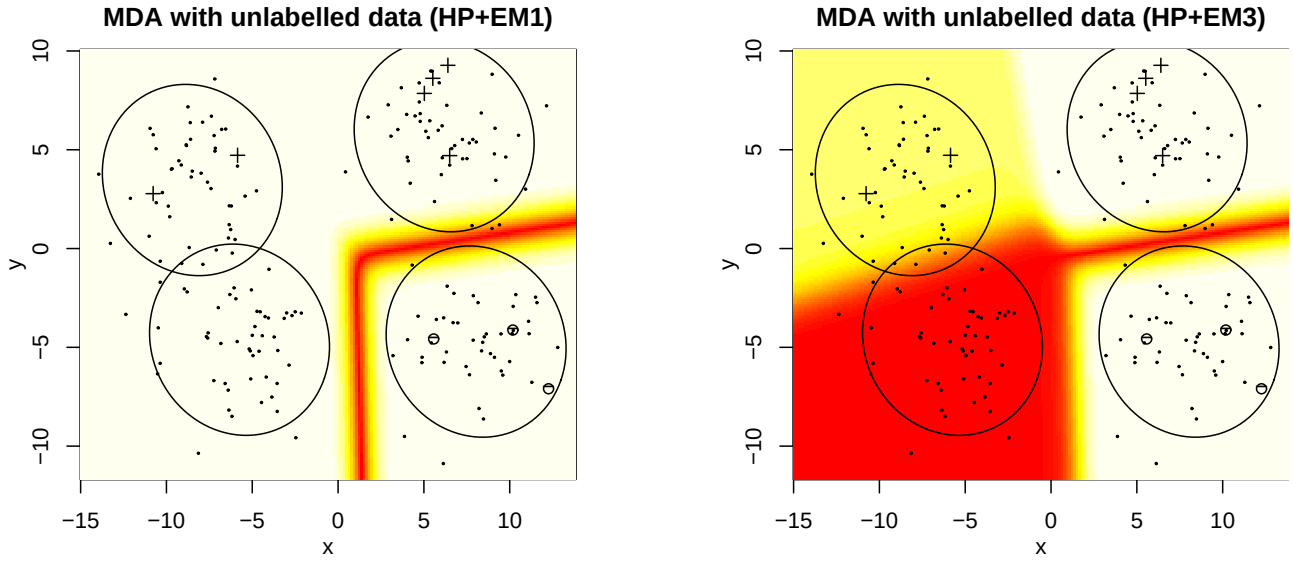


Figure 5. Mixture discriminant analysis results using 3 positive components and 1 negative component class for EM1 (left) and EM3 (right). Color shading indicates assignment probability: dark/red for intermediate probabilities, light/yellow for extreme probabilities.

of Gaussian on examples from each class, and uses the resulting model to discriminate between the classes (within the context of partially labelled data, hard partitioning has the added twist that one estimates a global mixture model and takes unlabelled data into account when estimating the mixture parameters).

In a first experiment, we fit an "ideal" mixture with 2 components per class and common covariance matrices. Figure 3 shows the result of MDA when we use only the 9 labelled examples. Predictably, the model estimates are fairly poor. The left part of the class boundary seems somewhat arbitrary and cuts through the blue female group. Taking unlabelled data into account yields a different picture. In both examples in figure 4 we use hard partitioning, trained either with EM1 (left) or with EM3 (right). In the ideal setting of

2 components per class, both methods correctly identify the boundary between the two classes. The main difference is in the confidence of the two models. Because of the hard assignment of each component to either class, the model obtained by EM1 displays a very crisp boundary. In particular points in the blue female (lower left) group are classified to the '-' class with high confidence, ie $P(z = -|x) \approx 1$. However, there are no labelled examples in this component, so this outcome is essentially an artefact of the ideal setting of 2 component per class, rather than a result of the evidence in the data. The results from EM3 are qualitatively different. For the three components with labelled data, the correct labelling is identified. The component corresponding to blue female is associated to unlabelled data with almost 100% probability, thus, from eq. 11, the assignment probability to either class is about 50%, indicated by a darker (red) zone in figure 4, right. This shows that contrary to EM1, EM3 yields a model that does take into account the intrinsic uncertainty associated with a component with no labelled examples.

This is further illustrated in figure 5, where we trained a misspecified mixture model, with 3 positive components and a single negative component. It is important to notice that because only 3 components have associated labelled examples, the likelihoods for both structure (2+2 and 3+1) are essentially identical. There is therefore no evidence in the data for choosing between the two structures.[3] With this misspecified model, EM1 again produces a crisp decision boundary, this time excluding the blue female group, and does so with high confidence: $P(z = +|x) \approx 1$ in this group. On the other hand, the model obtained by EM3 (right panel in figure 5) yields similar results as before: the three components with associated labelled examples are well identified, while the last component has high uncertainty (50-50 probability). Another way to look at this is that there is no objective way, based on the data, to know whether the discrimination task is "male vs. female" or "orange female vs. rest". The choice between the two structures is therefore arbitrary, but it has large consequences at classification time. In both cases, about 25% of the test data may be misclassified; however, with EM3, the model does identify the fact the the decision for the test examples in the blue female category are intrinsically very uncertain. With a cost that takes uncertainty into account (for example negative log-likelihood on test examples), the EM3 model will perform much better than the EM1 model under misspecification.

[3] In particular, information criteria such as AIC, BIC, etc. (Akaike, 1974; Schwartz, 1978) will not help since both the likelihood and structure penalty are identical for the two models.

5. Discussion

In this work, we have first reviewed several EM variants, including EM1 and EM2 proposed in (Miller & Uyar, 1997), for training mixture models on partially labelled data. This led us to identify two major problems: cluster impurity and unlabelled components. Cluster impurity can be solved by using hard partitioning. Indeed, the performance we obtain with soft partitioning is below what we observe with hard partitioning. This is somewhat in contrast with the findings of (Miller & Uyar, 1997), who report better performance for soft partitioning. We attribute that to a variant of the *cluster impurity* problem: for unbalanced datasets, the resulting soft partitioning models have components for which $P(z|\alpha)$ is biased towards the largest class. In EM-2, this effect is reinforced by eq. 7, in which unlabelled examples dominate. As most unlabelled examples are from the largest class, components tend to be dominated by the largest class.

Unlabelled components, and more generally components which display a low labelled/unlabelled ratio, pose the additional problem of bringing little, and to a certain extent unreliable, information, which is nevertheless used in the final categorization decision. Previous models failed to account for this fact. We showed that, by introducing an additional fake label, it was possible to "model" this uncertainty. As mentioned before, this approach is similar to the down-weighting factor for unlabelled data used in (Nigam et al., 2000). One important difference between the two approaches, however, is that (Nigam et al., 2000) down-weight the unlabelled data at parameter estimation time, while in our case we down-weight the unlabelled components at discrimination time, in order to better model the uncertainty associated with components with few or no labelled examples. That situation is likely to arise whenever there are very few labelled examples. There are situations, however, where the two approaches certainly need be combined, a combination we plan to investigate in the future.

6. Conclusion

In this paper, we provided a unifying treatment of several strategies for training mixture models from partially labelled data. We emphasized two major problems associated with the use of mixture models in this setting: cluster impurity and unlabelled components. This led us to develop a new model and associated EM

variant, EM3. We have showed, in particular, that this new model led to a better treatment of the uncertainty associated with components with few or no labelled examples. These results confirm several experimental studies on the usefulness of combining labelled and unlabelled data for training categorizers.

References

Akaike, H. (1974). A new look at the statistical model identification. *19*, 716–723.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Gammerman, A., Vovk, V., & Vapnik, V. (1998). Learning by transduction. *Uncertainty in Artificial Intelligence, Proceedings of the Fourteenth Conference* (pp. 145–155). Morgan Kaufmann.

Goutte, C., Gaussier, E., Cancedda, N., & Déjean, H. (2004). Generative vs discriminative approaches to entity recognition from label-deficient data. *Le poids des mots—Actes des 7èmes journées internationales d'analyse statistique des données textuelles* (pp. 515–523). Presses Universitaires de Louvain.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, *58*, 155–176.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 289–296). Morgan Kaufmann. http://www2.sis.pitt.edu/ dsl/UAI/uai99.html.

Hofmann, T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. *Advances in Neural Information Processing Systems 12* (p. 914). MIT Press.

Jaakkola, T. S., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems 11* (pp. 487–493).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning (ECML98)* (pp. 137–142). Springer Verlag.

Joachims, T. (1999). Transductive inference for text classification using support vector machine. *Machine Learning—Proceedings of the 16th International Conference (ICML'99)* (pp. 200–209). Morgan Kaufmann.

Miller, D. J., & Uyar, H. S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems (NIPS*9)* (pp. 571–577). MIT Press.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, *39*, 103–134.

PubMed (2005). http://www.pubmed.org. (last visited March 31, 2005).

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shashahani, B., & Landgrebe, D. (1994). The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, *32*, 1087–1095.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley.

A. M-step equations for $P(x|\alpha)$

For a Gaussian mixture:

$$P(x|\alpha) \propto \exp\left(-\frac{1}{2}\left(x-\mu_\alpha\right)^\top \Sigma^{-1}\left(x-\mu_\alpha\right)\right)$$

The M-step re-estimation formulas are:

$$\mu_\alpha^{(t+1)} = \frac{\sum_i C^{(t)}(\alpha,i)x^{(i)}}{\sum_i C^{(t)}(\alpha,i)}$$

$$\Sigma^{(t+1)} = \frac{1}{N}\left(\sum_{\alpha,i} C^{(t)}(\alpha,i)(x^{(i)}-\mu_\alpha^{(t)})(x^{(i)}-\mu_\alpha^{(t)})^\top\right)$$

For a multinomial mixture such as PLSA (Hofmann, 1999), x is a couple (f,e) and $P(x|\alpha) = P(f|\alpha)P(e|\alpha)$:

$$P^{(t+1)}(e|\alpha) = \frac{\sum_{i,e^{(i)}=e} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)}$$

$$P^{(t+1)}(f|\alpha) = \frac{\sum_{i,f^{(i)}=f} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)}$$