Learning from a Test Set

Piotr Juszczak and Robert P. W. Duin

Information and Communication Theory Group,
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, The Netherlands
p.juszczak@ewi.tudelft.nl, r.p.w.duin@ewi.tudelft.nl

Classification of partially labeled data requires linking the unlabeled input distribution $P(\mathbf{x})$ with the conditional distribution $P(y|\mathbf{x})$ obtained from the labeled data. The latter should, for example, vary little in high density regions. The key problem is to articulate a general principle behind this and other such reasonable assumptions. In this paper we provide a new approach to semi-supervised learning based on the stability of estimated labels for the unlabeled dataset, e.g a large test set, and the maximization of the mutual label relation. No clustering assumptions are required and the approach remains tractable even for continuous marginal class densities. We demonstrate the approach on synthetic examples and UCI repository datasets.

1 Introduction

In many classification problems there is an easy access to unlabeled objects and a specified cost, in time or money, to label them. Therefore, usually we label a small number of objects and hope, that they are sufficiently representative for the classification problem. However, to benefit from remaining unlabeled objects, one must exploit implicitly or explicitly the link between density $P(\mathbf{x})$ over objects $\mathbf{x}$ and the conditional $P(y|\mathbf{x})$ representing the posterior probability of the labels y.

Most classification methods do not attempt to explicitly model or incorporate information from the density $P(\mathbf{x})$. However, some classification algorithms such as density based algorithms as the Parzen classifier [1] or transductive SVM [2] have a possibility to relate $P(\mathbf{x})$ to $P(y|\mathbf{x})$; the decision boundary is biased to fall preferentially in low density regions of $P(\mathbf{x})$.

In such algorithms, the unlabeled objects, e.g a large test set to be classified, provide additional information about the structure of the domain while the few labeled objects identify the classification task expressed in this structure. A tacit assumption in this context is to associate high-density clusters in data with pure classes. When this assumption is appropriate, it is only required to label a single object per cluster to classify the whole dataset.

The presented problem is in broad terms related to a number of other problems like maximum entropy discrimination [3], data clustering by information bottleneck [4], and minimum-entropy data partitioning [5].

In this paper we investigate label propagation from a small labeled set over a large unlabeled set for density based classifiers in the semi-supervised learning framework, using as an example the Parzen classifier. The main difference between the various semi-supervised learning algorithms proposed in literature, such as spectral methods [6], random walks [7], graph mincuts [8] and transductive SVM [2], lies in the way of realizing the assumption of the labels consistency. However, the following three assumptions are often made about the representation space where the classification problem is present:

1. nearby objects are likely to have the same label,
2. objects on the same structure, e.g. a cluster or a manifold are likely to have the same label,
3. the decision boundary should lie in regions of low density [1].

The semi-supervised learning method proposed in this paper is based on the stability of estimated labels for unlabeled objects. In contradiction to the mentioned methods, in particular [6, 8, 7], there is not an implicit clustering step involved in the label propagation process. Therefore, there is no necessity to specify or optimize the number of clusters beforehand.

The layout of this paper is as follows. In section 2, the formal notation and the problem description are introduced, and the proposed algorithm is presented. Section 3 shows advantages and disadvantages of the proposed algorithm based on experiments on artificial and real-world data. Sections 4 presents the discussion and final conclusions.

2 Problem description

Given a partially labeled data set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, \ldots, \mathbf{x}_N\} \subset R^m$, the first l objects are labeled X_l and the remaining objects $\mathbf{x}_i \in X_u$ $(l+1 \leq i \leq N)$ are unlabeled. The goal is to predict the label of the unlabeled objects. The example of such a problem is presented in figure 1.

Our classification model assumes that each data example has a label, for $\mathbf{x} \in X_l$ or a distribution $P(y|\mathbf{x})$ over the class labels for $\mathbf{x} \in X_u$ [2]. These distributions are unknown and represent the parameters to be estimated. Given

[1] The third assumption is related to the second. An example is handwritten digit recognition where one tries to classify e.g. 2 and 5. The probability of having a digit which is between 2 and 5 should be lower than the probability of a distinct 2 or 5.

[2] $P(y|\mathbf{x})$ are also called soft labels

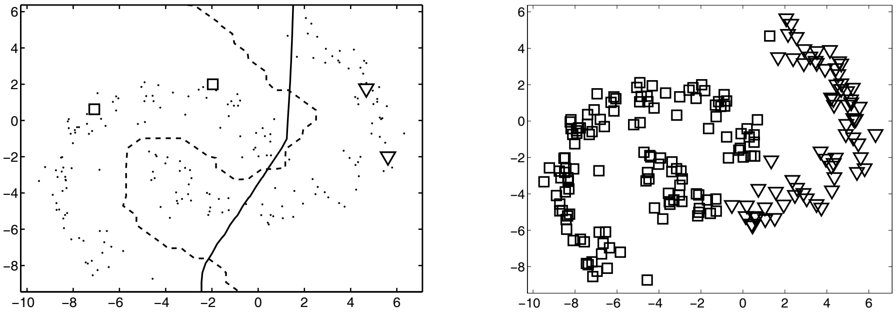


Fig. 1. On the left: A classification problem with four labeled objects denoted by (∇ , $\square$) and many unlabeled objects denoted by ($\cdot$). The continuous line denotes a classifier trained just on the labeled set and the dashed line a classifier trained on labeled and unlabeled objects. On the right: the corresponding classification labels for the classifier trained just on X_l.

an object $\mathbf{x}_k$, which may be labeled or unlabeled, we interpret its label as a weighted sum of crisp and soft labels of its neighbors N_G:

$$P'(y_i|\mathbf{x}_k) = \sum_{\mathbf{x}_i \in N_G(\mathbf{x}_k)} P(y_i|\mathbf{x}_i) p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k) \tag{1}$$

where $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ is the measure of the mutual label relation between the set of examples $\mathbf{x}_i \in N_G(\mathbf{x}_k)$ and the object $\mathbf{x}_k$. In other words $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ is the measure of the contribution of $\mathbf{x}_i$ to the probability that $\mathbf{x}_k$ has the label y_i. $P(y_i|\mathbf{x}_k)$ is computed over ϵ - neighborhood of $\mathbf{x}_k$ defined as follows:

$$N_G(\mathbf{x}_k) = \{\forall \mathbf{x}_i \in \{X_l \cup X_u\} \mid p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k) \geq \epsilon\} \backslash \{\mathbf{x}_k\} \tag{2}$$

In general, $P(y|\mathbf{x}_i)$ are only available for labeled objects X_l and have to be estimated for unlabeled objects X_u. We will now discuss how to estimate $P(y|\mathbf{x}_i)$ for the set X_u and how to choose the measure of the mutual label relation $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$.

2.1 Estimation of soft labels $P(y|\mathrm{x})$

We propose to estimate $P(y|\mathbf{x}_i)$ using the conditional maximum log-likelihood as the criterion. The $P(y|\mathbf{x}_i)$ is estimated for unlabeled objects for the fixed value of $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$:

$$\max_{P(y|\mathbf{x}_i)} \quad \sum_{y}^{C} \log \sum_{\mathbf{x}_i \in N_G(\mathbf{x}_k)} P(y|\mathbf{x}_i) p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k) \tag{3}$$

where for the two-class problem, $C = 2$, $P(y|\mathbf{x}_i) = \{0, 1\}$ for labeled objects and $0 \leq P(y|\mathbf{x}_i) \leq 1$ for unlabeled objects. Since $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ are fixed this objective function is jointly convex in the free parameters and has a unique maximum value. This convexity also guarantees that this optimization is easily performed via the EM algorithm.

2.2 Estimation of the mutual label relation $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$

In the previous subsection we assumed that the mutual label relation $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ was known and fixed. In this section we compute and optimize $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ for the set of label and unlabeled objects in the maximum likelihood sense for the known and fixed sets of soft labels $P(y|\mathbf{x})$.

Consider a set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with a metric $d(\mathbf{x}_k, \mathbf{x}_i) = \|\mathbf{x}_k - \mathbf{x}_i\|$. Since close objects have high value of $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ about their labels and objects far away low value of $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ we can relate the mutual label relation to the distances between objects e.g. as follows $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k) = \exp(-\frac{\|\mathbf{x}_k - \mathbf{x}_i\|}{2\sigma^2})$ [3]. The new estimate of the soft labels $P'(y|\mathbf{x}_k)$ of $\mathbf{x}_k$ can be defined now as:

$$P'(y_i|\mathbf{x}_k) = \sum_{\mathbf{x}_i \in N_G(\mathbf{x}_k)} P(y_i|\mathbf{x}_i) \exp\left(-\frac{\|\mathbf{x}_k - \mathbf{x}_i\|}{2\sigma^2}\right) \tag{4}$$

which is related to the weighted Parzen density estimator. $P'(y_i|\mathbf{x}_k)$ is computed for all labels $y_i \in C$ and normalized to $\sum_{y_i \in C} P'(y_i|\mathbf{x}_i) = 1$.

Now we define how the information about the label of an object $\mathbf{x}_i$ influences the label of an object $\mathbf{x}_k$. In particular, the mutual label relation should decreases when the distance between objects $\mathbf{x}_k$ and $\mathbf{x}_i$ increases. This is related to the choice of σ. For large σ more distant objects have the influence on the soft labels of an object $\mathbf{x}_k$ and for small σ only nearby objects influence the soft labels of an object $\mathbf{x}_k$. We computed σ in equation (4) based on a leave-one out maximum likelihood estimation [9, 10]. The initial estimate σ_l of σ is optimized for just the labeled objects $\mathbf{x} \in X_l$. The final σ_{ul} is optimized for both labeled and unlabeled objects $\mathbf{x} \in \{X_l, X_u\}$. In a series of n EM algorithms σ takes the values:

$$\sigma_l > \sigma_2 > \ldots > \sigma_t > \ldots \sigma_{t-1} > \sigma_{ul}$$

The change in σ from large, σ_l, to small, σ_{ul}, values, during learning of soft labels, changes the stress between the global labels consistency and the local labels consistency.

2.3 Proposed algorithm

The proposed algorithm of the classification with the partially labeled dataset is summarized in algorithm 1.

[3] $p_{MLR}(y_i|\mathbf{x}_i, \mathbf{x}_k)$ can be defined in several ways e.g. as the L_1 norm, a wavelet function or a Gaussian process

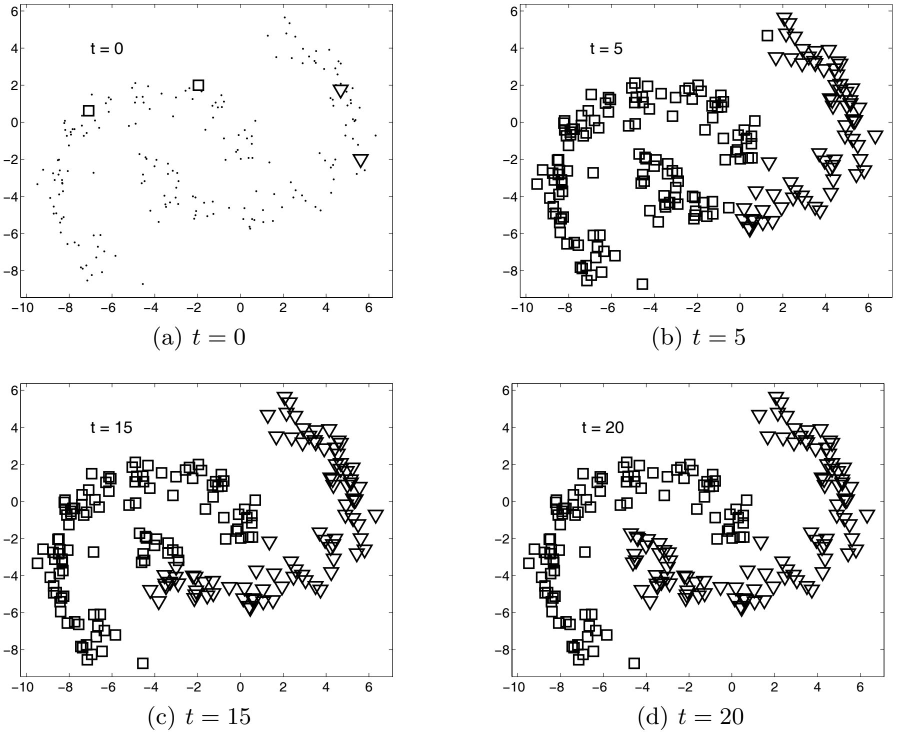
In the initial step of the algorithm the soft labels are computed using only labeled objects, $P(y|\mathbf{x}) = 0 \; \forall \mathbf{x} \in X_u$. In the second step based on the current estimation of σ_t soft labels are optimized $P'(y|\mathbf{x})$ for $\mathbf{x} \in X_u$ using the maximum likelihood criterion. Next, the equation (4) is recomputed using both crisp and soft labels. Step 4 is repeated until the difference between the current estimated labels $P'(y|\mathbf{x}_i)$ and the previous estimated labels $P(y|\mathbf{x}_i)$ is smaller than γ. The procedure is repeated for n different σ-s.

```
 1. set a number of EM algorithms to $n$; compute $\sigma_l$ and $\sigma_{ul}$; set $t = 1$ and a
    stopping criterion $\gamma$;
 2. compute: $\sigma_l > \sigma_2 > \ldots > \sigma_t > \ldots \sigma_{t-1} > \sigma_{ul}$ for each EM algorithm;
while $t \leqslant n$
 3. optimize soft labels $P'(y|\mathbf{x})$ based on equation (4) with a fixed $\sigma_t$; using
    the former $P(y|\mathbf{x})$ as initialization of the labels;
 4. repeat 3 until stopping criterion is reached
    e.g. $\sum_i |P'(y|\mathbf{x}_i) - P(y|\mathbf{x}_i)| < \gamma$; $t = t + 1$;
end
```

Algorithm. 1 *soft*-PARZEN.

3 Experiments

Consider an example (figure 2) of classification with the proposed algorithm. We are given 2 labeled objects per class and 196 unlabeled objects in an intertwining two banana shape patterns. This pattern has a manifold structure where distances are locally but not globally Euclidean, due to the curved arms. Therefore, the pattern is difficult to classify for traditional algorithms using locally defined relations, such as 1-nearest neighbor; figure 1b. We used the proposed algorithm, described in algorithm 1, to incorporate unlabeled data into the Parzen density estimator and scale the Euclidean distance between objects using their soft labels. Figure 2 shows three different timescales. At $t = 5$ the σ is overestimated, therefore there are large, Gaussian clusters and $P(y|\mathbf{x})$ are estimated ruffly. At $t = 15$ because σ becomes smaller local mutual label relations in marginal regions start to change the soft labels. At $t = 20$ almost all objects, apart of one, have correct labels.
Next, we evaluated the performance of the presented algorithm on some of the UCI repository datasets [11]: *waveform*, *satellite*, *letter*, *ecoli*. Datasets were divided into two parts: labeled set X_l and the unlabeled set X_u constituted from remaining objects, the ratio $\frac{X_l}{X_u}$ is indicated by numbers on the abscissa. The label propagation was performed on X_u and the obtained classifier was



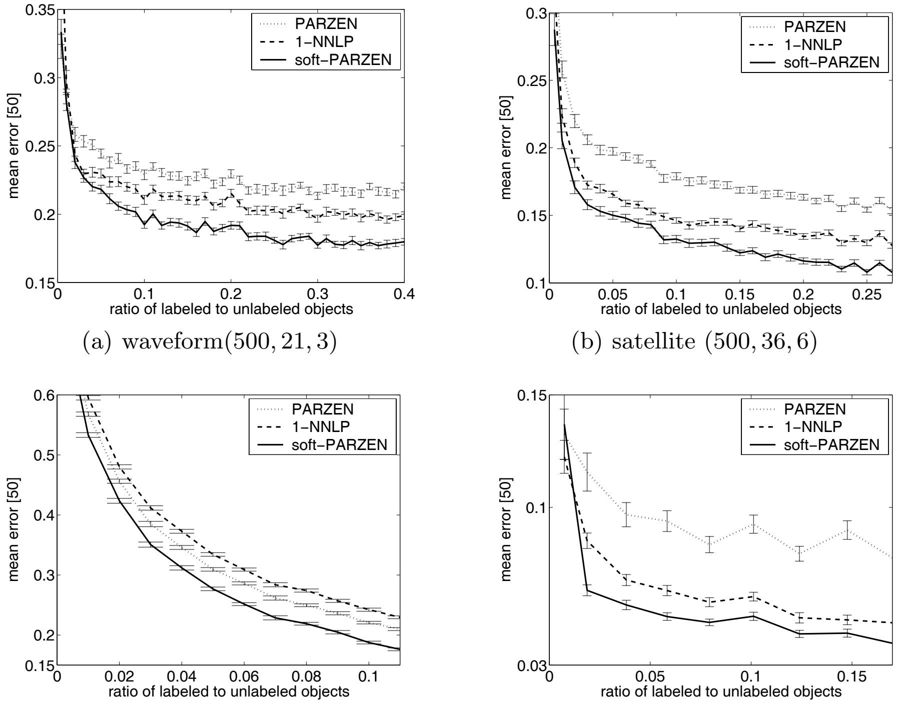
(a) $t = 0$ (b) $t = 5$

(c) $t = 15$ (d) $t = 20$

Fig. 2. Label estimates for the *soft*-PARZEN algorithm for the banana shape dataset. Labeled (soft and crisp labels) objects denoted by (∇ , $\square$) and unlabeled objects denoted by ($\cdot$).

tested on the same set of unlabeled data X_u. The random division was repeated 50 times for each ratio $\frac{X_l}{X_u}$. The performance of the proposed algorithm (*soft*-PARZEN) is compered with the 1-nearest neighbor label propagation (1-NNLP) [12, 8] and the Parzen classifier trained just on labeled objects (PARZEN). The mean error and the standard deviation are shown in figure 3. It can be seen, that the proposed *soft*-PARZEN algorithm outperforms both: 1-NNLP and the Parzen classifier trained on just labeled objects, on considered classification problems. In case of *waveform* and *ecoli* the performance of *soft*-PARZEN is close to 1-NNLP and for *satellite* and *letter* there is significant improvement.

The *soft*-PARZEN and 1-NNLP perform similar if distances between objects in pure clusters and between clusters differ significantly. However, if in the data there is not a clear cluster structure the *soft*-PARZEN might outperform the 1-NNLP significantly.



(a) waveform$(500, 21, 3)$

(b) satellite $(500, 36, 6)$

(c) letter$(500, 16, 26)$

(d) ecoli$(272, 7, 3)$

Fig. 3. Mean square error and standard deviation for *soft*-PARZEN, 1-NNLP compared with a classifier trained just on a labeled dataset PARZEN for UCI repository datasets: *waveform*, *satellite*, *letter*, *ecoli*. Numbers in brackets indicate the size of X_u and the number of features and classes in a dataset.

The performance of the proposed method depends on the quality of the labeled data and their relation to the structure of the unlabeled dataset. If the clusters of the unlabeled data are not related to the class information it is hard to expect that the proposed method performs well. For a broader discussion about merits and disadvantages of the semi-supervised learning we point reader to the paper [13].

4 Conclusions

The proposed algorithm based on expectation maximization of soft labels and the mutual label relation *soft*-PARZEN provides a robust variable resolution approach to classifying data sets with significant cluster structure and very few labels. When the cluster structure in absent or unrelated to the classification task, the proposed method can be expected to derive particular but small improvement over a classifier trained just on the labeled dataset. In such cases the performance is strongly related to the quality of the already labeled set.

In future work we will test the proposed algorithm on large, high-dimensional datasets and explore theoretical connections to network information theory.

Acknowledgments

This work was partly supported by the Dutch Organization for Scientific Research (NWO).

References

1. Parzen, E.: On the estimation of a probability density function and the mode. Annals of Mathematical Statistics **33** (1962) 1065–1076
2. Vapnik, V.N.: Statistical learning theory. Wiley, NY (1998)
3. Jaakkola, T., Meila, M., Jebara, T.: Maximum entropy discrimination. In: NIPS. Volume 12. (1999) 470–477
4. Tishby, N., Slonim, N.: Data clustering by markovian relaxation and the information bottleneck method. In: NIPS. (2000) 640–646
5. Roberts, S., Holmes, C., Denison, D.: Minimum entropy data partitioning using reversible jump markov chain monte carlo. PAMI **23** (2001) 909–914
6. Chapelle, O., Weston, J., Schöelkopf, B.: Cluster kernels for semi-supervised learning. In: NIPS. Volume 15. (2002) 585–592
7. Szummer, M., Jaakkola, T.: Partially labeled classification with markov random walks. In: NIPS. Volume 14. (2001) 945–952
8. Blum, A., Lafferty, J., Rwebangira, M.R., Reddy, R.: Semi-supervised learning using randomized mincuts. In: ICML. (2004)
9. Duin, R.P.W.: On the choice of the smoothing parameters for parzen estimators of probability density functions. IEEE Transactions on Computers **25** (1976) 1175–1179
10. Lissack, T., Fu, K.S.: Error estimation in pattern recognition via $L^{\alpha}-$ distance between posterior density functions. IEEE Transitions on Information Theory **22** (1976) 34–45
11. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998)
12. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: ICML, Morgan Kaufmann, San Francisco, CA (2001) 19–26
13. Cohen, I., Cozman, F., Sebe, N., Cirelo, M., Huang, T.: Semi-supervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. PAMI **26** (2004) 1553–1566