Constrained Parameter Estimation for Semi-supervised Learning: The Case of the Nearest Mean Classifier

Marco Loog*

Pattern Recognition Laboratory
Delft University of Technology
Delft, The Netherlands
m.loog@tudelft.nl,
prlab.tudelft.nl**

Abstract. A rather simple semi-supervised version of the equally simple nearest mean classifier is presented. However simple, the proposed approach is of practical interest as the nearest mean classifier remains a relevant tool in biomedical applications or other areas dealing with relatively high-dimensional feature spaces or small sample sizes. More importantly, the performance of our semi-supervised nearest mean classifier is typically expected to improve over that of its standard supervised counterpart and typically does not deteriorate with increasing numbers of unlabeled data. This behavior is achieved by constraining the parameters that are estimated to comply with relevant information in the unlabeled data, which leads, in expectation, to a more rapid convergence to the large-sample solution because the variance of the estimate is reduced. In a sense, our proposal demonstrates that it may be possible to properly train a known classification scheme such that it can benefit from unlabeled data, while avoiding the additional assumptions typically made in semi-supervised learning.

1 Introduction

Many, if not all, research works that discuss semi-supervised learning techniques stress the need for additional assumptions on the available data in order to be able to extract relevant information not only from the labeled, but especially from the unlabeled examples. Known presuppositions include the cluster assumption, the smoothness assumption, the assumption of low density separation, the manifold assumption, and the like [6,23,30].

While it is undeniably true that having more precise knowledge on the distribution of data could, or even should, help in training a better classifier, the question to what extent such data assumptions are at all necessary has not

* Partly supported by the Innovational Research Incentives Scheme of the Netherlands Research Organization [NWO, VENI Grant 639.021.611].

** Secondary affiliation with the Image Group, University of Copenhagen, Denmark.

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part II, LNAI 6322, pp. 291–304, 2010.

been studied to a great extent. Theoretical contributions have both discussed the benefits and the absence of it of the inclusion of unlabeled data in training [4,13,24,25]. With few exceptions, however, these results rely on assumptions being made with respect to the underlying data. Reference [25] aims to make the case that, in fact, it may be so that no extra requirements on the data are needed to obtain improved performance using unlabeled data in addition to labeled data.

A second, related issue is that in many cases, the proposed semi-supervised learning technique has little in common with any of the classical decision rules that many of us know and use; it seems as if semi-supervised learning problems call for a completely different approach to classification. Nonetheless, one still may wonder to what extent substantial gains in classification performance are possible when properly training a known type of classifier, e.g. LDA, QDA, 1NN, in the presence of unlabeled data.

There certainly are exceptions to the above. There even exist methods that are able to extend the use of any known classifier to the semi-supervised setting. In particular, we would like to mention the iterative approaches that rely on expectation maximization or self-learning (or self-training), as can for instance be found in [16,18,19,26,29,27] or the discussion of [10]. The similarity between self-learning and expectation maximization (in some cases equivalence even) has been noted in various papers, e.g. [1,3], and it is to no surprise that such approaches suffer from the same drawback: As soon as the underlying model assumptions do not fit the data, there is the real risk that adding too much unlabeled data leads to a substantial decrease of classification performance [8,9,19]. This is in contrast with the supervised setting, where most classifiers, generative or not, are capable of handling mismatched data assumptions rather well and adding more data generally improves performance.

We aim to convince the reader that, in a way, it may actually also be possible to guarantee a certain improvement with increased numbers of unlabeled data. This possibility is illustrated using the nearest mean classifier (NMC) [11,17], which is adapted to learn from unlabeled data in such a way that some of the parameters become better estimated with increasing amounts of data. The principal idea is to exploit known constraints on the these parameters in the training of the NMC, which results in faster convergence to their real values. The main caveat is that this reduction of variance does not necessarily translate into a reduction of classification error. Section 4 shows, however, that the possible increase in error is limited.

Regarding the NMC, it is needless to say that it is a rather simple classifier, which nonetheless can provide state-of-the-art performance, especially in relatively high-dimensional problems, and which is still, for instance, used in novel application areas [15,14,21,28] (see also Subsection 4.1). Neither the simplicity of the classifier nor the caveat indicated above should distract one from the point we like to illustrate, i.e., it may be feasibility to perform semi-supervised learning without making the assumptions typically made in the current literature.

1.1 Outline

The next section introduces, through a simple, fabricated illustration, the core technical idea that we like to put forward. Subsequently, Section 3 provides a particular implementation of this idea for the nearest mean classifier in a more realistic setting and briefly analyzes convergence properties for some of its key variables. Section 4 shows, by means of some controlled experiments on artificial data, some additional properties of our semi-supervised classifier and compares it to the supervised and the self-learned solutions. Results on six real-world data sets are given as well. Section 5 completes the paper and provides a discussion and conclusions.

2 A Cooked-Up Example of Exponentially Fast Learning

Even though the classification problem considered in this section may be unrealistically simple, it does capture very well the essence of the general proposal to improve semi-supervised learners that we have in mind.

Let us assume that we are dealing with a two-class problem in a one-dimensional feature space where both classes have equal prior probabilities, i.e., $\pi_1 = \pi_2$. Suppose in addition, the NMC is our classifier of choice to tackle this problem with. NMC simply estimates the mean of every class and assigns new feature vectors to the class corresponding to the nearest class mean. Finally, assume that an arbitrarily large set of unlabeled data points is at our disposal. The obvious question to ask is: Can the unlabeled data be exploited to our benefit? The maybe surprising answer is a frank: Yes.

To see this, one should first of all realize that in general, when employing an NMC, the two class means, m_1 and m_2, and the overall mean of the data, μ, fulfill the constraint

$$\mu = \pi_1 m_1 + \pi_2 m_2 \,. \tag{1}$$

In our particular example based on equal priors, this mean that the total mean should be right in between the two class means. Moreover, again in the current case, the total mean is exactly on the decision boundary. In fact, in our one-dimensional setting, the mean equals the actual decision boundary. Now, if there is anything one can estimate rather accurate from an unlimited amount of data for which labels are not necessarily provided, it would be this overall mean. In other words, provided our training set contains a large number of labeled or unlabeled data points, the zero-dimensional decision boundary can be located to arbitrary precision. That is, it is identifiable, cf. [5].

The only thing we do not know yet is which class is located on what side of the decision boundary. In order to decide this, we obviously do need labeled data. As the decision boundary is already fixed, however, the situation compares directly to the one described in Castelli and Cover [5] and, in a similar spirit, training can be done exponentially fast in the number of labeled samples.

The key point in this example is that the actual distribution of the two classes does in fact not matter. The rapid convergence takes place without making any

assumptions on the underlying data, except for the equal class priors. What really leads to the improvement is proper use of the constraint in Equation (1). In the following, we demonstrate how such convergence behavior can generally be obtained for the NMC.

3 Semi-supervised NMC and Its (Co)variance

One of the major lacuna in the example above, is that one rarely has an unlimited amount of samples at ones disposal. We therefore propose a simple adaptation of the NMC in case one has a limited amount of labeled and unlabeled data. Subsequently, a general convergence property of this NMC solution is considered in some detail, together with two special situations.

3.1 Semi-supervised NMC

The semi-supervised version of NMC proposed in this work is rather straightforward and it might only be adequate to a moderate extent in the finite sample setting. The solution suggested simply shifts all K sample class means m_i ($i \in 1, \ldots, K$) by a similar amount such that the overall sample mean $m' = \sum_{i=1}^{K} p_i m_i'$ of the shifted class means m_i' coincides with the total sample mean m_{t}. The latter has been obtained using all data, both labeled and unlabeled. In the foregoing p_i is the estimated posterior corresponding to class i.

More precisely, we take

$$m_i' = m_i - \sum_{i=1}^{K} p_i m_i + m_{\mathrm{t}} \tag{2}$$

for which one can easily check that $\sum_{i=1}^{K} p_i m_i'$ indeed equals m_{t}.

Merely considering the two-class case from now on, there are two vectors that play a role in building the actual NMC [20]. The first one, $\Delta = m_1 - m_2$, determines the direction perpendicular to the linear decision boundary. The second one, $m_1 + m_2$, determines—after taking the inner product with Δ and dividing it by two—the position of the threshold or the bias. Because $\Delta = m_1 - m_2 = m_1' - m_2'$, the orientations of the two hyperplanes correspond and therefore the only estimates we are interested in are $m_1 + m_2$ and $m_1' + m_2'$.

3.2 Covariance of the Estimates

To compare the standard supervised NMC and its semi-supervised version, the squared error that measures the deviation of these estimated to their true values is considered. Or rather, as both estimates are unbiased, we consider their covariance matrices.

The first covariance matrix, for the supervised case, is easy to obtain:

$$\mathrm{cov}(m_1 + m_2) = \frac{C_1}{N_1} + \frac{C_2}{N_2}\,, \tag{3}$$

where C_i is the true covariance matrix of class i and N_i is the number of samples from that class.

To get to the covariance matrix related to the semi-supervised approach, we first express $m_1' + m_2'$ in terms of the variables defined earlier plus m_u, the mean of the unlabeled data, and N_u, the number of unlabeled data points:

$$\begin{aligned} m_1' + m_2' &= m_1 + m_2 - 2\frac{N_1 m_1 + N_2 m_2}{N_1 + N_2} + 2\frac{N_1 m_1 + N_2 m_2 + N_\mathrm{u} m_\mathrm{u}}{N_1 + N_2 + N_\mathrm{u}} \\ &= \left(1 - \tfrac{2N_1}{N_1+N_2} + \tfrac{2N_1}{N_1+N_2+N_\mathrm{u}}\right) m_1 \\ &+ \left(1 - \tfrac{2N_2}{N_1+N_2} + \tfrac{2N_2}{N_1+N_2+N_\mathrm{u}}\right) m_2 + \tfrac{2N_\mathrm{u}}{N_1+N_2+N_\mathrm{u}} m_\mathrm{u}\,. \end{aligned} \tag{4}$$

Realizing that the covariance matrix of the unlabeled samples equals the total covariance T, it now is easy to see that

$$\begin{aligned} \mathrm{cov}(m_1' + m_2') &= \left(1 - \frac{2N_1}{N_1 + N_2} + \frac{2N_1}{N_1 + N_2 + N_\mathrm{u}}\right)^2 \frac{C_1}{N_1} \\ &+ \left(1 - \frac{2N_2}{N_1 + N_2} + \frac{2N_2}{N_1 + N_2 + N_\mathrm{u}}\right)^2 \frac{C_2}{N_2} \\ &+ \left(\frac{2N_\mathrm{u}}{N_1 + N_2 + N_\mathrm{u}}\right)^2 \frac{T}{N_\mathrm{u}}\,. \end{aligned} \tag{5}$$

3.3 Some Further Considerations

Equations (3) and (5) basically allow us to compare the variability in the two NMC solutions. To get a feel for how these indeed compare, let as consider the situation similar to the one from Section 2 in which the amount of unlabeled data is (virtually) unlimited. It holds that

$$\lim_{N_\mathrm{u}\to\infty} \mathrm{cov}(m_1' + m_2') = \left(1 - \frac{2N_1}{N_1 + N_2}\right)^2 \frac{C_1}{N_1} + \left(1 - \frac{2N_2}{N_1 + N_2}\right)^2 \frac{C_2}{N_2}\,. \tag{6}$$

The quantity $(1 - \frac{2N_i}{N_1+N_2})^2$ is smaller or equal to one and we can readily see that $\mathrm{cov}(m_1' + m_2') \preceq \mathrm{cov}(m_1 + m_2)$, i.e., the variance of the semi-supervised estimate is smaller or equal to the supervised variance for every direction in the feature space and, generally, the former will be a better estimate than the latter. Again as an example, when the true class priors are equal, $1 - \frac{2N_i}{N_1+N_2}$ tends to be nearer zero with increasing number of labeled samples, which implies a dramatic decrease of variance in case of semi-supervision.

Another situation that provides some insight in Equations (3) and (5) is the one in which we consider $C = C_1 = C_2$ and $N = N_1 = N_2$ (for the general case the expression becomes somewhat unwieldy). For this situation we can derive that the two covariance matrices of the sum of means become equal when

$$T = \frac{(4N + N_\mathrm{u})C}{2N}\,. \tag{7}$$

What we might be more interested in is, for example, the situation in which $2NT \preceq (4N + N_{\mathrm{u}})C$ as this would mean that the expected deviation from the true NMC solution is smaller for the semi-supervised approach, in which case this would be the preferred solution. Note also that from Equation (7) it can be observed that if the covariance C is very small, the semi-supervised method is not expected to give any improvement over the standard approach unless N_{u} is large.

In a real-world setting, the decisions of which approach to use, necessarily has to rely on the finite number of observations in the training set and sample estimates have to be employed. Moreover, the equations above merely capture the estimates' covariance, which explains only part of the actual variance in the classification error. For the remainder, we leave this issue untouched and turn to the experiments using the suggested approach, which is compared to supervised NMC and a self-learned version.

4 Experimental Results

We carried out several experiments to substantiate some of the earlier findings and claims and to potentially further our understanding of the novel semi-supervised approach. We are interested to what extent NMC can be improved by semi-supervision and a comparison is made to the standard, supervised setting and an NMC trained by means of self-learning [16,18,29].

The latter is a technique in which a classifier of choice is iteratively updated. It starts by the supervised classifier, labels all unlabeled data and retrains the classifier given the newly labeled data. Using this classifier, the initially unlabeled data is reclassified, based on which the next classifier is learned. This is iterated until convergence.

As the focus is on the semi-supervised training of NMC, other semi-supervised learning algorithms are indeed not of interest in the comparisons presented here.

4.1 Initial Experimental Setup and Experiments

As it is not directly of interest to this work, we do not consider learning curves for the number of labeled observations. Obviously, NMC might not need too many labeled examples to perform reasonably and strongly limit the number of labeled examples. We experimented mainly with two, the bare minimum, and ten labeled training objects. In all cases we made sure every class has at least one training sample.

We do, however, consider learning curves as a function of the number of unlabeled instances. This setting easily disclosed both the sensitivity of the self-learning to an abundance of unlabeled data and the improvements that may generally be obtained given various quantities of unlabeled data. The number of unlabeled objects considered in the main experiments are 2, 8, 32, 128, 512, 2048, and 8192.

The tests carried out involve three artificial and eight real-world data set all having two classes. Six of the latter are taken from the UCI Machine Learning

Table 1. Error rates on the two benchmark data sets from [7]

data set	`Text`		`SecStr`		
number of labeled objects	10	100	100	1000	10000
error NMC	0.4498	0.2568	0.4309	0.3481	0.3018
error constrained NMC	0.4423	0.2563	0.4272	0.3487	0.3013

Repository [2]. On these, extensive experimentation has been implemented in which for every combination of number of unlabeled objects and labeled objects 1,000 repetitions were executed. In order to be able to do so on the limited amount of samples in the UCI data sets, we allowed to draw instances with replacement, basically assuming that the empirical distribution of every data set is its true distributions. This approach enabled us to properly study the influence of the constraint estimation on real-world data without having to deal with the extra variation due to cross validation or the like. The artificial sets do not suffer from limited amounts of data.

The two other data sets, `Text` and `SecStr`, are benchmarks from [7], which were chosen for their feature dimensionality and for which we followed the protocol as prescribed in [7]. We consider the results, however, of limited interest as the semi-supervised constrained approach gave results only minimally different from those obtained by regular, supervised NMC (after this we did not try the self-learner). Nevertheless, we do not want to withhold these results from the reader, which can be found in Table 1. In fact, we can make at least two interesting observations from them. To start with, the constrained NMC does not perform worse than the regular NMC, for none of the experiments. Compared to the results in [7] both the supervised and the semi-supervised perform acceptable on the `Text` data set when 100 labeled samples are available and both obtain competitive error rates on `SecStr` for all numbers of labeled training data, again confirming the validity of the NMC.

4.2 The Artificial Data

The first artificial data set, `1D`, consists of a one-dimensional data set with two normally distributed classes with unit variance for which the class means are 2 units apart. This setting reflects the situation considered in Section 2. The two top subfigures in Figures 1 and 2 plot the error rates against different numbers of unlabeled data points for the supervised, semi-supervised, and self-learned classifier. All graphs are based on 1,000 repetitions of every experiment. In every round, the classification error is estimated by a new sample of size 10,000. Figure 1 displays the results with two labeled samples, while Figure 2 gives error rates in case of ten labeled samples. Note that adding more unlabeled data indeed further improves the performance.

As second artificial data set, `2D correlated`, we again consider two normally distributed classes, but now in two dimensions. The covariance matrix has the form $\left(\begin{smallmatrix} 4 & 3 \\ 3 & 4 \end{smallmatrix}\right)$, meaning the features are correlated, which, in some sense, does not

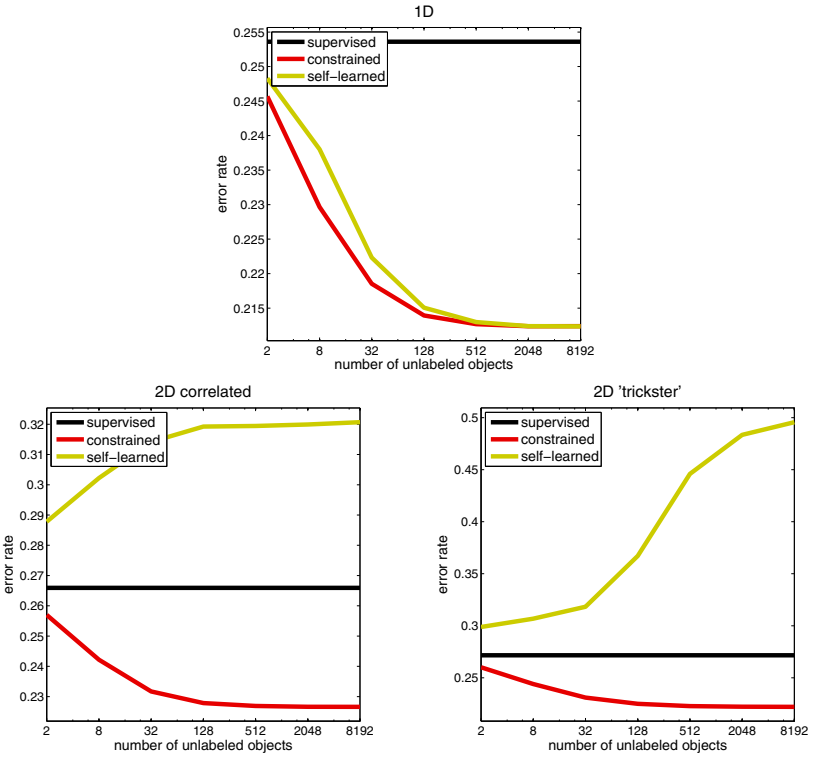

Fig. 1. Error rates on the artificial data sets for various unlabeled sample sizes and a single labeled sample per class. Top subfigure: `1D` data set. Left subfigure: `2D correlated`. Right: `2D 'trickster'`.

fit the underlying assumptions of NMC. Class means in one dimension are 4 apart and the optimal error rate is about 0.159. Further results, like those for the first artificial data set, are again presented in the two figures.

The last artificial data set, `2D 'trickster'`, has been constructed to trick the self-learner. The total data distribution consists of two two-dimensional normal distributions with unit covariance matrices whose means differ in the first feature dimension by 1 unit. The classes, however, are completely determined by the second feature dimension: If this value is larger than zero we assign to class 1, if smaller we assign to class 2. This means that the optimal decision boundary is perpendicular to the boundary that would keep the two normal distributions apart. By construction, the optimal error rate is 0.

Both Figures 1 and 2 illustrate the deteriorating effect adding too much unlabeled data can have on the self-learner, while the constrained semi-supervised approach does not seem to suffer from such behavior and in most cases clearly improves upon the supervised NMC, even though absolute gains can be moderate.

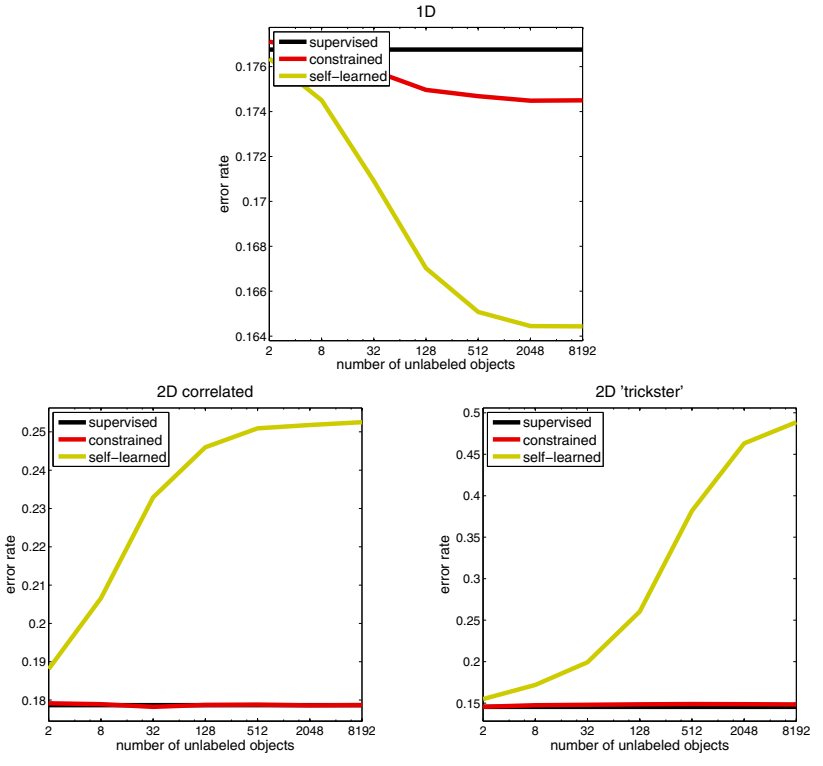

Fig. 2. Error rates on the artificial data sets for various unlabeled sample sizes and a total of ten labeled samples. Top subfigure: `1D` data set. Left subfigure: `2D correlated`. Right: `2D 'trickster'`.

4.3 Six UCI Data Sets

The UCI data sets used are `parkinsons`, `sonar`, `spect`, `spectf`, `transfusion`, and `wdbc'` for which some specifications can be found in Table 2. The classification performance of supervision, semi-supervision, and self-learning are displayed in Figures 3 and 4, for two and ten labeled training objects, respectively.

Table 2. Basic properties of the six real-world data sets

data set	number of objects	dimensionality	smallest class prior
`parkinsons`	195	22	0.25
`sonar`	208	60	0.47
`spect`	267	22	0.21
`spectf`	267	44	0.21
`transfusion`	748	3	0.24
`wdbc`	569	30	0.37

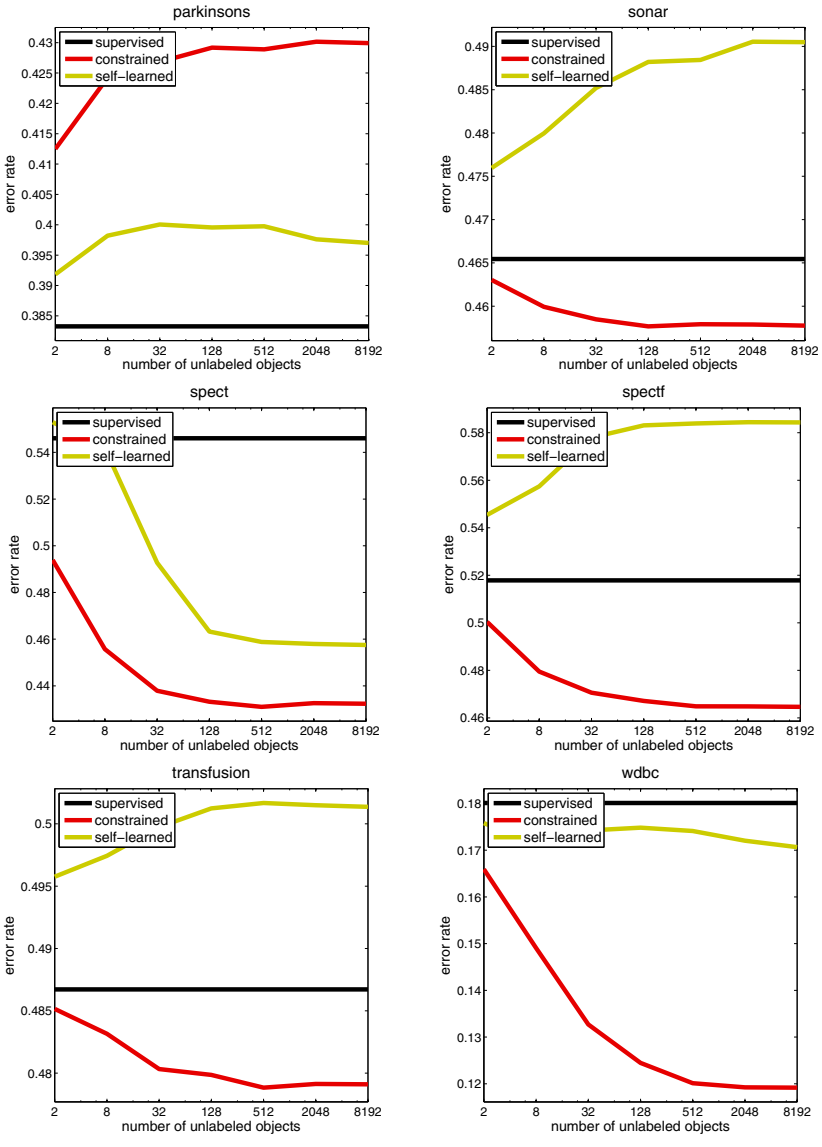

Fig. 3. Error rates for the supervised, semi-supervised, and self-learned classifiers on the six real-world data sets for various unlabeled sample sizes and a single labeled sample per class

In the first place, one should notice that in most of the experiments the constrained NMC performs best of the three schemes employed and that the self-learner in many cases leads to deteriorated performance with increasing unlabeled data sizes. There are various instances in which our semi-supervised approach starts off at an error rate similar to the one obtained by regular supervision, but

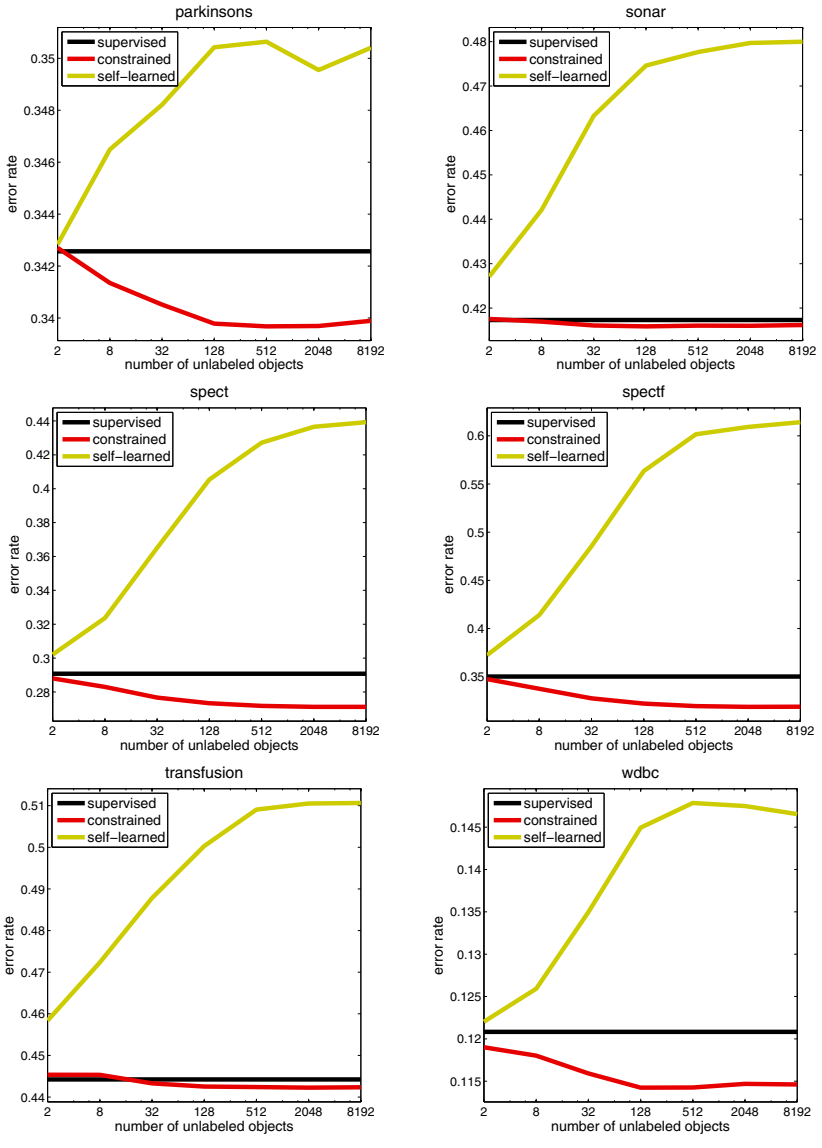

Fig. 4. Error rates for the supervised, semi-supervised, and self-learned classifiers on the six real-world data sets for various unlabeled sample sizes and a total of ten labeled training samples

adding a moderate amount of additional unlabeled objects already ensures that the improvement in performance becomes significant.

The notable outlier is the very first plot in Figure 3 in which constrained NMC performs worse than the other two approaches and even deteriorates with increasing amounts of unlabeled data. How come? We checked the estimates for

the covariance matrices in Equations 3 and 3 and saw that the variability of the sum of the means is indeed less in case of semi-supervision, so this is not the problem.

What comes to the fore here, however, is that a reduction in variance for these parameters does not necessarily directly translate into a gain in classification performance. Not even in expectation. The main problem we identified is basically the following (consider the example from Section 2): The more accurately a classifier manages to approximate the true decision boundary, the more errors it will typically make if the side on which the two classes are located are mixed up in the first place. Such a configuration would indeed lead to worse and worse performance for the semi-supervised NMC with more and more unlabeled data. Obviously, this situation is less likely to occur with increasing numbers of labeled samples and Figure 4 shows that the constrained NMC is expected to attain improved classification results on `parkinsons` for as few as ten labels.

5 Discussion and Conclusion

The nearest mean classifier (NMC) and some of its properties have been studied in the semi-supervised setting. In addition to the known technique of self-learning, we introduced a constrained-based approach that typically does not suffer from the major drawback of the former for which adding more and more unlabeled data might actually result in a deterioration. As pointed out, however, this non-deterioration concerns the parameter estimates and does not necessarily reflect immediately in improved classifier's performance. In the experiments, we identified an instance where a deterioration indeed occurs, but the negative effect seems limited and quickly vanishes with a moderate increase of labeled training data.

Recapitulating our general idea, we suggest that particular constraints, which relate estimates coming from both labeled and unlabeled data, should be met by the parameters that have to be estimated in the training phase of the classifier. For the nearest mean we rely on Equation (1) that connects the two class means to the overall mean of the data. Experiments show that enforcing this constraint in a straightforward way improves the classification performance in the case of moderately to large unlabeled sample sizes. Qualitatively, this partly confirms the theory in Section 3, which shows that adding increasing numbers of unlabeled data, eventually leads to reduced variance in the estimates and, in a way, faster convergence to the true solution.

A shortcoming of the general idea of constrained estimation is that it is not directly clear which constraints to apply to most of the other classical decision rules, if at all applicable. The main question obviously being if there is a more general principle of constructing and applying constraints that is more broadly applicable. On the other hand, one should realize that the NMC may act as a basis for LDA and its penalized and flexible variations, as described in [12] for instance. Moreover, kernelization by means of a Gaussian kernel, reveals similarities to the classical Parzen classifier, cf. [22]. Our findings may be directly applicable in these situations.

In any case, the important point we did convey is that, in a way, it is possible to perform semi-supervised learning without making additional assumptions on the characteristics of the data distribution, but by exploiting some characteristics of the classifier. We consider it also important that it is possible to do this based on a known classifier and in such a way that adding more and more data does not lead to its deterioration. A final advantage is that our semi-supervised NMC is as easy to train as the regular NMC with no need for complex regularization schemes or iterative procedures.

References

1. Abney, S.: Understanding the Yarowsky algorithm. Computational Linguistics 30(3), 365–395 (2004)
2. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
3. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised clustering by seeding. In: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 19–26 (2002)
4. Ben-David, S., Lu, T., Pál, D.: Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In: Proceedings of COLT 2008, pp. 33–44 (2008)
5. Castelli, V., Cover, T.: On the exponential value of labeled samples. Pattern Recognition Letters 16(1), 105–111 (1995)
6. Chapelle, O., Schölkopf, B., Zien, A.: Introduction to semi-supervised learning. In: Semi-Supervised Learning, ch. 1. MIT Press, Cambridge (2006)
7. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press, Cambridge (2006)
8. Cohen, I., Cozman, F., Sebe, N., Cirelo, M., Huang, T.: Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1553–1567 (2004)
9. Cozman, F., Cohen, I.: Risks of semi-supervised learning. In: Semi-Supervised Learning, chap. 4. MIT Press, Cambridge (2006)
10. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
11. Duda, R., Hart, P.: Pattern classification and scene analysis. John Wiley & Sons, Chichester (1973)
12. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. The Annals of Statistics 23(1), 73–102 (1995)
13. Lafferty, J., Wasserman, L.: Statistical analysis of semi-supervised regression. In: Advances in Neural Information Processing Systems, vol. 20, pp. 801–808 (2007)
14. Liu, Q., Sung, A., Chen, Z., Liu, J., Huang, X., Deng, Y.: Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. PLoS ONE 4(12), e8250 (2009)
15. Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C., Eisenberger, M., Carducci, M., Nelson, W., Yegnasubramanian, S., Luo, J., Wang, Y., Xu, J., Isaacs, W., Visakorpi, T., Bova, G.: Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nature Medicine 15(5), 559–565 (2009)

16. McLachlan, G.: Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70(350), 365–369 (1975)
17. McLachlan, G.: Discriminant analysis and statistical pattern recognition. John Wiley & Sons, Chichester (1992)
18. McLachlan, G., Ganesalingam, S.: Updating a discriminant function on the basis of unclassified data. Communications in Statistics - Simulation and Computation 11(6), 753–767 (1982)
19. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Learning to classify text from labeled and unlabeled documents. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 792–799 (1998)
20. Noguchi, S., Nagasawa, K., Oizumi, J.: The evaluation of the statistical classifier. In: Watanabe, S. (ed.) Methodologies of Pattern Recognition, pp. 437–456. Academic Press, London (1969)
21. Roepman, P., Jassem, J., Smit, E., Muley, T., Niklinski, J., van de Velde, T., Witteveen, A., Rzyman, W., Floore, A., Burgers, S., Giaccone, G., Meister, M., Dienemann, H., Skrzypski, M., Kozlowski, M., Mooi, W., van Zandwijk, N.: An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clinical Cancer Research 15(1), 284 (2009)
22. Schölkopf, B.: The kernel trick for distances. In: Advances in Neural Information Processing Systems, vol. 13, p. 301. The MIT Press, Cambridge (2001)
23. Seeger, M.: A taxonomy for semi-supervised learning methods. In: Semi-Supervised Learning, ch. 2. MIT Press, Cambridge (2006)
24. Singh, A., Nowak, R., Zhu, X.: Unlabeled data: Now it helps, now it doesn't. In: Advances in Neural Information Processing Systems, vol. 21 (2008)
25. Sokolovska, N., Cappé, O., Yvon, F.: The asymptotics of semi-supervised learning in discriminative probabilistic models. In: Proceedings of the 25th International Conference on Machine Learning, pp. 984–991 (2008)
26. Titterington, D.: Updating a diagnostic system using unconfirmed cases. Journal of the Royal Statistical Society. Series C (Applied Statistics) 25(3), 238–247 (1976)
27. Vittaut, J., Amini, M., Gallinari, P.: Learning classification with both labeled and unlabeled data. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 69–78. Springer, Heidelberg (2002)
28. Wessels, L., Reinders, M., Hart, A., Veenman, C., Dai, H., He, Y., Veer, L.: A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21(19), 3755 (2005)
29. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pp. 189–196 (1995)
30. Zhu, X., Goldberg, A.: Introduction to Semi-Supervised Learning. Morgan & Claypool Publishers, San Francisco (2009)