Learning when only some of the training data are from the same distribution as test data

Jaakko Peltonen[1,2] **and Samuel Kaski**[1]
[1]Helsinki Institute for Information Technology, Laboratory of Computer and Information Science,
Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland
[2]Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Finland
{jaakko.peltonen, samuel.kaski}@tkk.fi

Abstract

The most difficult learning scenario is when the training and test distributions differ both in the data density and in the conditional class distributions. Learning is still possible assuming that some of the learning samples are known to come from the same distribution as the test samples. We formulate a simple nonparametric learner for this task, and apply it for building a "personalized recommender system" that uses the recommendations of other users as possibly useful parts of the training data.

A widely occurring problem in building classifiers is that we may have only few samples of proper training data, but there are databases full of potentially relevant background data. Assuming that some of the data are relevant, the background data can be considered training data from a (partially) different distribution as the test data. The assumption we make is that the small set of "proper" training data comes from the same distribution as the test data, and hence can be used for searching for relevant background data.

The modeling task is to use the small set of "proper" training data both for learning a classifier, and for estimating which of the backround data are useful to be incorporated in learning the classifier. To be able to solve this problem, we assume that the background data come in sets, and introduce a single weight for each set. The weight tells how strongly the set should be taken into account, and the weights are optimized to maximize classification accuracy on the "proper" training data. This generalizes earlier ideas [1] where the background data was not divided in sets, and is more constrained than [2] and hence may be less prone to overfitting.

This problem is related but not identical to several other learning problems: transfer learning, multitask learning, semisupervised learning. The exact relationships need to be discussed in a longer paper.

The idea is at its clearest in a nonparametric Parzen windows-based classifier, which we will use for a case study. The model could easily be generalized to more general parametric or semiparametric models using kinds of empirical priors. Place normal distributions N (or any other kinds of kernels) over all data points $\mathbf{x}'$; if the class of $\mathbf{x}'$ is c' the kernel contributes to class c'. All "proper" training data T is naturally included. The rest, the "supplementary" sets S_z, of which the background data consists, are included as well but weighted with w_z. The model is

$$p(c|\mathbf{x};\mathbf{A},\mathbf{w}) = Z(\mathbf{x})^{-1}\left(\sum_{(\mathbf{x}',c')\in T}\delta_{c,c'}N(\mathbf{x};\mathbf{x}',\mathbf{A}) + \sum_{z}\sum_{(\mathbf{x}',c')\in S_z}\delta_{c,c'}w_z N(\mathbf{x};\mathbf{x}',\mathbf{A})\right) \quad (1)$$

where Z normalizes the distribution, $\delta_{c,c'}$ is one when $c = c'$ and zero otherwise, and the covariance matrix (here diagonal) $\mathbf{A}$ and the weights w_z are parameters to be optimized. We maximize the sum of $p(c|\mathbf{x};\mathbf{A},\mathbf{w})$ over the "proper" training data T; the classifier thus uses the supplementary sets S_z only to build a better classifier for T. (We also use a leave-out procedure for T to curtail overfitting.)

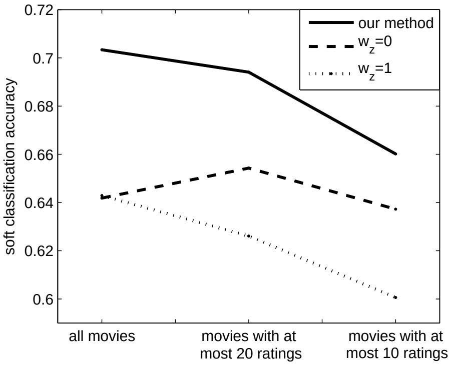

Figure 1: Soft classification accuracies on movie rating data, averaged over users. '$w_z = 0$': use only data from the target user, '$w_z = 1$': use data from all users indiscriminately.

One potential application area is combining collaborative and content-based filtering in predicting interests of a user (see, e.g., [3] for earlier work; we leave comparisons for future work). Collaborative filtering only uses the matrix of grades given by users to items such as movies, which works only if the matrix is reasonably dense. Content-based filtering, on the other hand, requires a large enough learning data set. When building a model for a certain user the data of each of the other users is a supplementary data set, and our goal is to use the supplementary sets to complement the originally small learning set.

Skipping details, compared with using only the "proper" training data ($w_z = 0$ for all z), and indiscriminately using all background data ($w_z = 1$), the new intermediate model (1) gives higher soft classification accuracy for new data (Figure 1). Key to success here is to avoid overfitting; performance will be studied in more detail later. The collaborative filtering (ranking) data is from the EachMovie database; 134 users; the 10% best-rated movies of each user form the first class and the 10% worst-rated movies the second class. Synopses from the Allmovie database form the content for the movies; the texts were treated as word histograms and preprocessed by projecting them onto 10 linear features, each chosen to best differentiate one movie genre from the rest.

In summary, we introduced a simple non-parametric model which extracts useful supplementary data sets, and uses them to better classify test data for which only part of the whole data set comes from the same distribution.

Acknowledgments

S. Kaski and J. Peltonen belong to the Adaptive Informatics Research Centre, a national centre of excellence of the Academy of Finland. They were supported by grant 108515, and by University of Helsinki's Research Funds. This work was also supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. All rights are reserved because of other commitments.

References

[1] Pengcheng Wu and Thomas G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 871–878. Omnipress, Madison, WI, 2004.

[2] Xuejun Liao, Ya Xue, and Lawrence Carin. Logistic regression with an auxiliary data source. In *Proceedings of The Twenty-Second International Conference on Machine Learning (ICML 2005)*, pages 505–512. Omnipress, Madison, WI, 2005.

[3] Justin Basilico and Thomas Hofmann. Unifying collaborative and content-based filtering. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of The Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 65–72. Omnipress, Madison, WI, 2004.