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 & Adaptive Informatics Research Centre, Laboratory of Computer and Information Science, Helsinki University of Technology

[2]University of Helsinki, Department of Computer Science

www.cis.hut.fi/projects/mi/

Introduction

Widely occurring problem: little **training data**, but large amounts of potentially relevant **background data**.

If some of the background data are relevant: background data is training data but from a **partially different distribution** than test data.

Both data density and conditional class distributions differ!

The "proper" training data can be used to **search** for relevant background data.

Modeling problem: use "proper" training data to both 1) learn a classifier and 2) estimate which background data are useful for learning that classifier!

We assume background data comes in **sets**.
---> single **weight** for each set:
how strongly to use that set

Optimize weights to maximize classification accuracy of the "proper" training data!

Related (but not identical) problems: transfer learning, multitask learning, semisupervised learning, ...

Some earlier ideas:
In [1] the background data was not divided in sets; our setting is more constrained than [2] and hence may be less prone to overfitting.

[1] P. Wu, T. G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. ICML 2004.
[2] X. Liao, Y. Xue, and L. Carin. Logistic regression with an auxiliary data source. ICML 2005.

The Method

Nonparametric Parzen-windows based classifier (with Gaussian kernel):

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

contribution from "proper" training data

weighted contribution from supplementary sets of background data

normalization

Training data contributes with full weight; supplementary sets contribute with weights w_z. If the weights were all zero the supplementary sets would be ignored; if the weights were all one the supplementary sets would contribute just as much as "proper" training data. We want to optimize the weights and the covariance matrix $\boldsymbol{A}$. (here diagonal).

Objective function:

Soft classification accuracy of the "proper" training data.

$$\sum_{(\boldsymbol{x},c)\in T} p(c|\boldsymbol{x};\boldsymbol{A},\boldsymbol{w})$$

Supplementary sets S_z are used only to build a better classifier for the proper training set T. (We also use a leave-one-out procedure for T to curtail overfitting.)

Here we used conjugate gradient optimization to maximize the objective function with respect to the parameters weights w_z and $\boldsymbol{A}$. (We fixed the trace of $\boldsymbol{A}$ by raparameterization; the trace was chosen with a heuristic.)

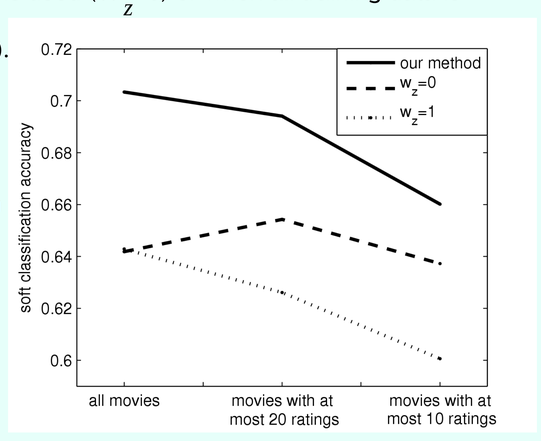
Experiments

Combining collaborative and content-based filtering is an application of our task. Data from a particular user is the "proper" training data for learning to rate new items for that user. We can use background data from other users to help.

We test on movie data: collaborative rankings from the EachMovie database and synopses from the Allmovie database. (134 users; 10% best-rated movies of each user form class 1 and 10% worst-rated movies form class 2. Word histograms of synopses were projected to 10 linear features.)

We compare our method to the case where only "proper" training data is used (w_z=0) or when all training data is used (w_z=1).

We get higher soft classification accuracy.



Summary

Situation: small "proper" training data, large background data but only a part of it comes from the same distribution as test data.

Method: simple nonparametric model that:
1) extracts useful supplementary data sets from background data
2) uses "proper" training data and supplementary data sets together to better classify test data