Active Learning for Efficient Labeling and Classification of Imaging Mass Spectrometry Data

Michael Hanselmann[1], Jens Röder[1,2], Ullrich Köthe[1], Bernhard Y. Renard[1], A. Kreshuk[1], Ron M. A. Heeren[3], Fred A. Hamprecht[1]

[1] Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Heidelberg, Germany
[2] Robert Bosch GmbH, CR/AEM5, Robert-Bosch-Straße 200, 31139 Hildesheim, Germany
[3] FOM AMOLF, FOM-Institute for Atomic and Molecular Physics, Amsterdam, The Netherlands

Introduction: What is Active Learning?

Supervised classifiers such as Random Forests or Support Vector Machines (SVMs) have successfully been used for the automated annotation of Imaging Mass Spectrometry (IMS) data. However,

- training of supervised classifiers requires labeled training examples
- labeling is costly and time-consuming, especially if the tissue is heterogeneous

In practice, it may be sufficient to label few highly informative points, but it is unclear which are the informative ones.

Active learning strategies [1, 2] automatically suggest good candidate points.

In comparison to random sampling, our active learning approach

- significantly reduces the number of required labels
- while maintaining classification accuracy

Methods: Active Learning with Random Forests

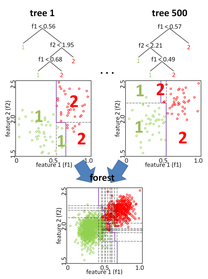

Figure 1: Random Forest.

We combine the Random Forest classifier (cf. Fig. 1), a state-of-the-art ensemble method which has previously been used for the robust, fast and accurate classification of IMS data [3] with a novel, iterative active learning strategy for multi-class scenarios. Our selection criterion is statistically motivated and depends on the uncertainty of the classification and the density of labeled and unlabeled points. This way, the algorithm is geared to select fewer but more informative training samples.
On the data used here (cf. Data section), selecting the next query point required ≈ 3s. The algorithm iterates the following steps:

- select a point that is supposed to contribute most to improving the classifier's performance
- ask the expert for a label
- train a Random Forest using all previously labeled examples

References

[1] Schleif et al. Margin based Active Learning for LVQ Networks. *Neurocomp.*, 70: 1215–1224, 2007.

[2] Zomer et al. Active learning support vector machines for optimal sample selection in classification. *J. Chemometr.*, 18(6):294–305, 2004.

[3] Hanselmann et al. Toward digital staining using imaging mass spectrometry and random forests. *J. Prot. Res.*, 8:3558–3567, 2009.

Results: Comparison to Random Sampling

We compared our active learning algorithm (AL-RF) to random sampling (RS) which means selecting a hitherto unlabeled point at random in each learning step. We repeated both approaches 100 times and averaged the obtained results. Results were compared using sensitivity and positive predictive value (PPV):

$$sensitivity = \frac{true\ pos.}{true\ pos.+false\ neg.}, \quad PPV = \frac{true\ pos.}{true\ pos.+false\ pos.}.$$

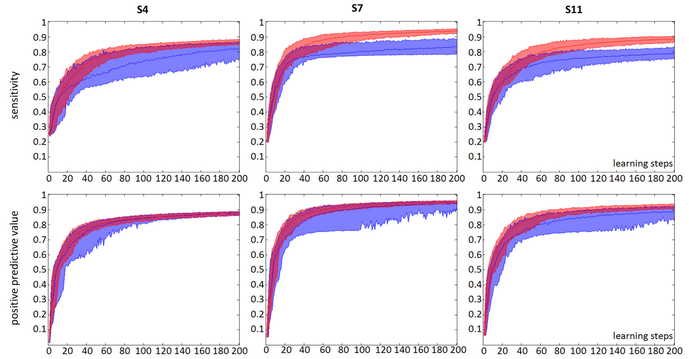

Figure 2: Average results for the first 200 learning steps for AL-RF (red) and RS (blue). We show the median as well as the band between the 95% quantile and the 5% quantile to visualize variance between runs.

After 100 learning steps, our active learning strategy

- features significantly lower variance between repeats
- gains 10% in sensitivity
- gains ≈ 3% in positive predictive value

over random sampling (cf. Fig. 2). This is because our algorithm samples more points that correspond to "difficult" classes, i.e. classes which are highly similar to each other in feature space, and less samples from simpler classes (cf. Fig. 3). Fig. 4 shows how the classification result improves over the iterations.

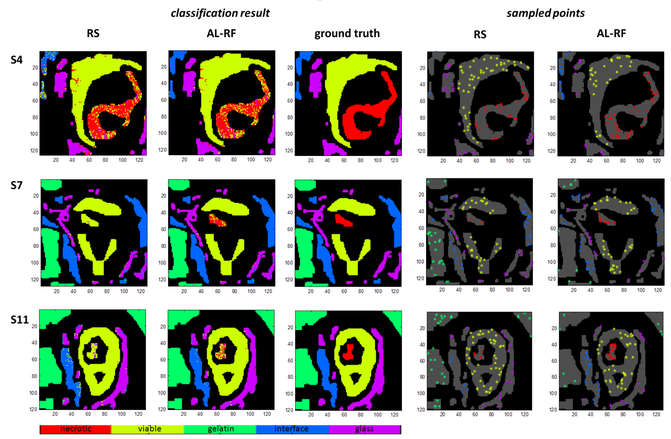

Figure 3: Average results after 100 learning steps (left). AL-RF selects more difficult points (i.e. from the necrotic, viable, interface class which are close in feature space) and less simple points (gelatin, glass class) than RS (right).

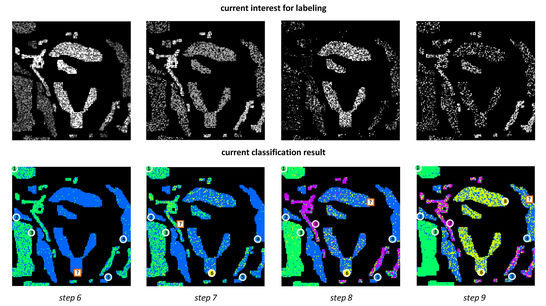

Figure 4: In each learning step, our AL method asks the expert for a label for the most informative point according to our selection criterion (top row, learning steps 6 to 9 from left to right). Over the course of the iterations the classification result gradually improves (bottom row, color-coding as in Fig. 2).

Data and Data Preprocessing

- 3 slices of human breast cancer tissue (MCF-7) grown in mice (S4, S7, S11)
- tissue was embedded in gelatin, flash-frozen, cryo-sectioned and thaw-mounted on a cold indium tin oxide-coated glass slide
- TRIFT II TOF SIMS with an Au+ liquid metal ion gun, mass range: 0-400 Da
- spectra were baseline-corrected and TIC-normalized, features were extracted with a peak-picker based on local maximum detection
- gold standard labels were obtained by Hematoxylin-Eosin staining of parallel slices, five classes were identified: necrotic tissue, viable tissue, interface region, gelatin, glass/hole in tissue

Conclusions

Our active learning approach significantly reduced labeling time without sacrificing classification accuracy. In comparison to random sampling it

- yielded classification results with significantly higher sensitivities and positive predictive values if the same number of learning steps was used
- selected more samples from the difficult classes
- featured less variance between repeats.

It is thus suitable for the efficient annotation of IMS data.

Acknowledgments

We gratefully acknowledge financial support by the DFG (grant no. HA4364/2-1) (MH, BYR, FAH), the HGS Math Comp (MH, BYR, AK), the Robert Bosch GmbH (JR, FAH), and the Netherlands BSIK program 'Virtual Laboratory for e-science' (RMAH). We furthermore like to thank Erika R. Amstalden (FOM-AMOLF, Amsterdam, The Netherlands) and Kristine Glunde (Johns Hopkins, Baltimore, USA) for the IMS data, as well as Bernhard X. Kausler, Xinghua Lou (University of Heidelberg, Germany) and Marc Kirchner (Harvard Medical, Boston, USA) for fruitful discussions.