MapView:

Graphical Data Representation for Active Learning

Eva Weigl, Alexander Walch, Ulrich Neissl, Pauline Meyer-Heye, Thomas Radauer, Edwin Lughofer, Wolfgang Heidl and Christian Eitzinger

pauline.meyer-heye@profactor.at

i-KNOW, Graz, 18th of October 2016

LEADING
INNOVATIONS

WWW.PROFACTOR.AT

Motivation for the Use of Active and Interactive Learning

- Reducing costs and time in training of supervised classifiers by focusing labeling.
- Selection of interesting samples can be done by the machine learning system or the data labeling expert.
- Human labeling experts hold additional information about the underling real world problem.
- The machine learning system knows the internal decision borders and possibly decision certainties.
- Interaction between learner and expert should benefit the training process.

Objectives of the MapView Development

- A learning scenario of training a classifier from scratch.
- Need for a system that enables the labeling expert to interact with the machine learning system.
- For this we combine active learning elements with user expertise.
- The MapView should help the user by visualizing prediction certainty, misclassifications and sample distribution in the features space.

- The MapView was originally created for usage in an industrial quality control environment.

Method

- Data samples are represented as 2D points in the MapView
- For each sample the information is encoded as follows:

Sample	Map View
Feature-Vector	Position on 2D Map
Label	Center Color
Prediction	Border Color
Prediction Certainty	Transparency
Image & Features	Detail View for single samples

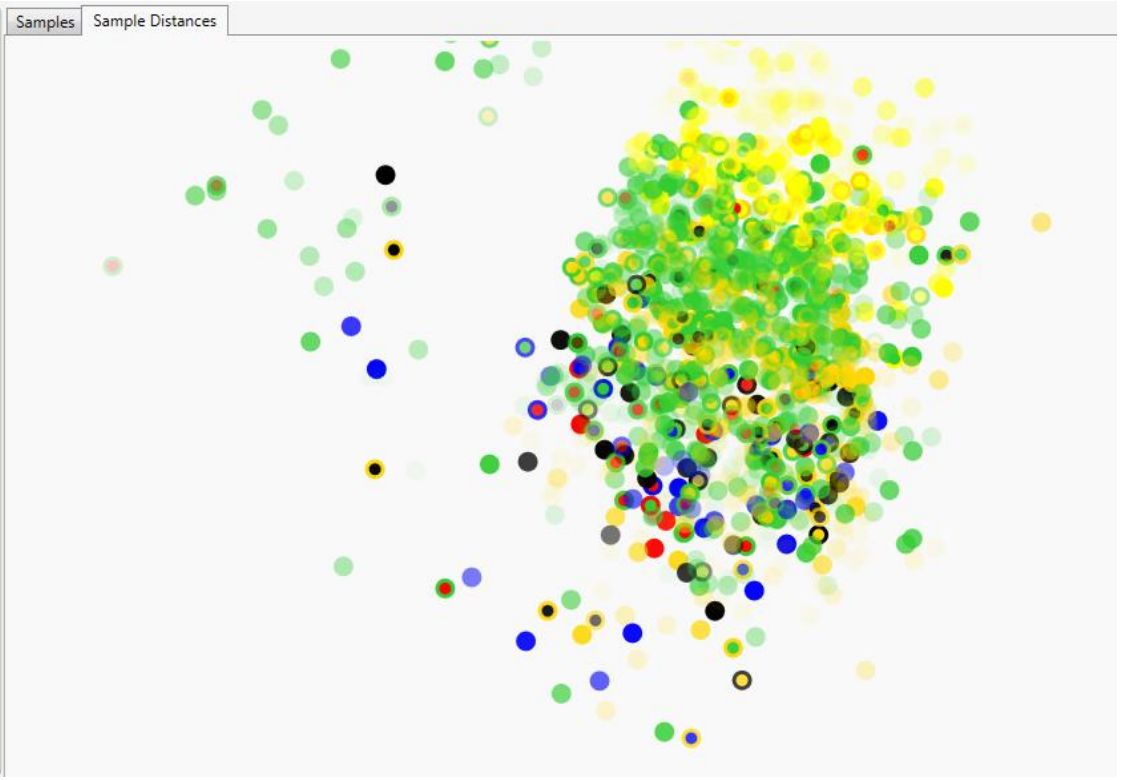
- Main interaction possibilities are: labeling and batch labeling, training, retraining, zooming and filtering.
- Information on the current training state is displayed in form of the classification error.

MapView in the Active Learning Tool

Prediction = Label

Prediction ≠ Label

		01_Dirty	02_Dirty closed	03_Inclusion	04_Surface	05_Bonding Void	08_Fibre	10_Float
		714	8	39	253	39	357	176
unknown	0	0	0	0	0	0	0	0
01_Dirty	620	559	1	4	15	3	37	1
02_Dirty closed	47	34	5	5	2	1	0	0
03_Inclusion	45	8	2	30	0	1	4	0
04_Surface	250	33	0	0	207	0	10	0
05_Bonding Void	69	27	0	0	0	25	17	0
08_Fibre	382	53	0	0	29	9	289	2
10_Float	173	0	0	0	0	0	0	173



Classification Method

- Any classifier that delivers probability measures for each class label can be used.
- A Random Forest classifier was chosen.
- Classification certainty was computed as:

$$Certainty(Sample_N, Label_L) = \frac{\#T_{N,L}}{\#Tree}$$

$$with\ T_{N,L} = \{T \in Tree \mid Prediction(T, Sample_N) = Label_L\}$$

Dimensionality Reduction Method

- Embedding by feature vector.
- Dimensionality reduction method by Sammon‘s Mapping.

$$E = \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}}$$

- Euclidian distances for d^*_{ij} and d_{ij}.
- Computation complexity is quadratic (w.r.t. number of samples)

Dimensionality Reduction Method Pre-Step

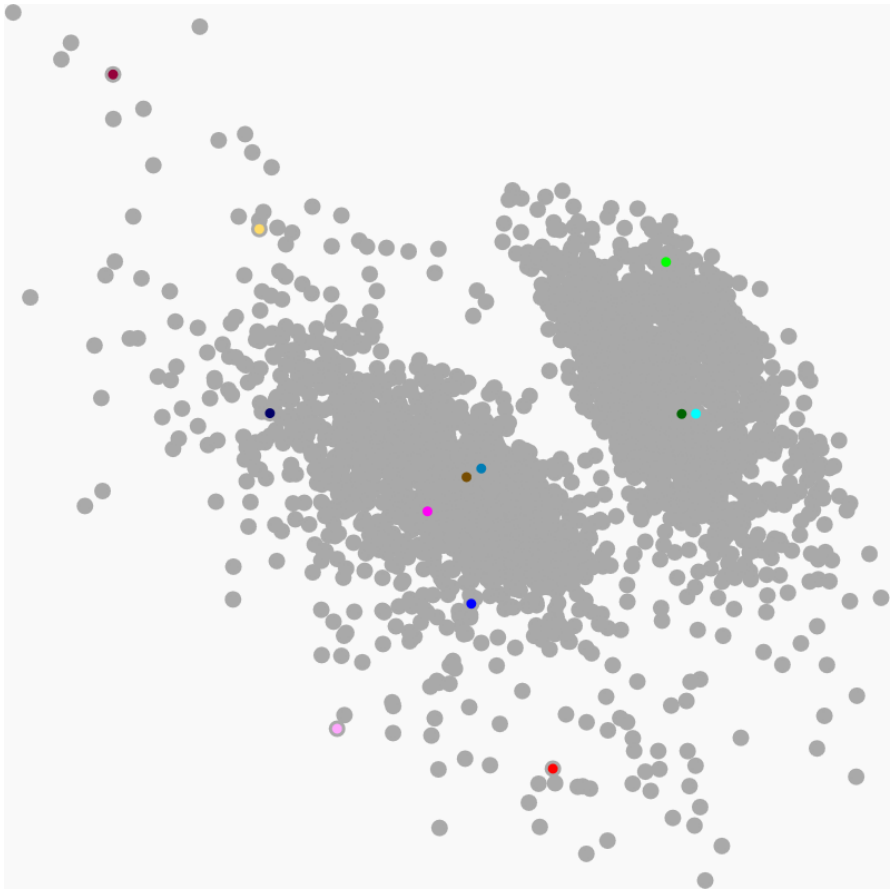
- To reduce computation effort, the dimensionality reduction has been limited to a number of samples that equals the number of classes.
- The samples are selected as cluster centers from a k-Means clustering.
- Only for these samples the location on the map is computed with Sammon's Mapping.
- All other sample coordinates are computed as an affine combinations of the cluster centers.

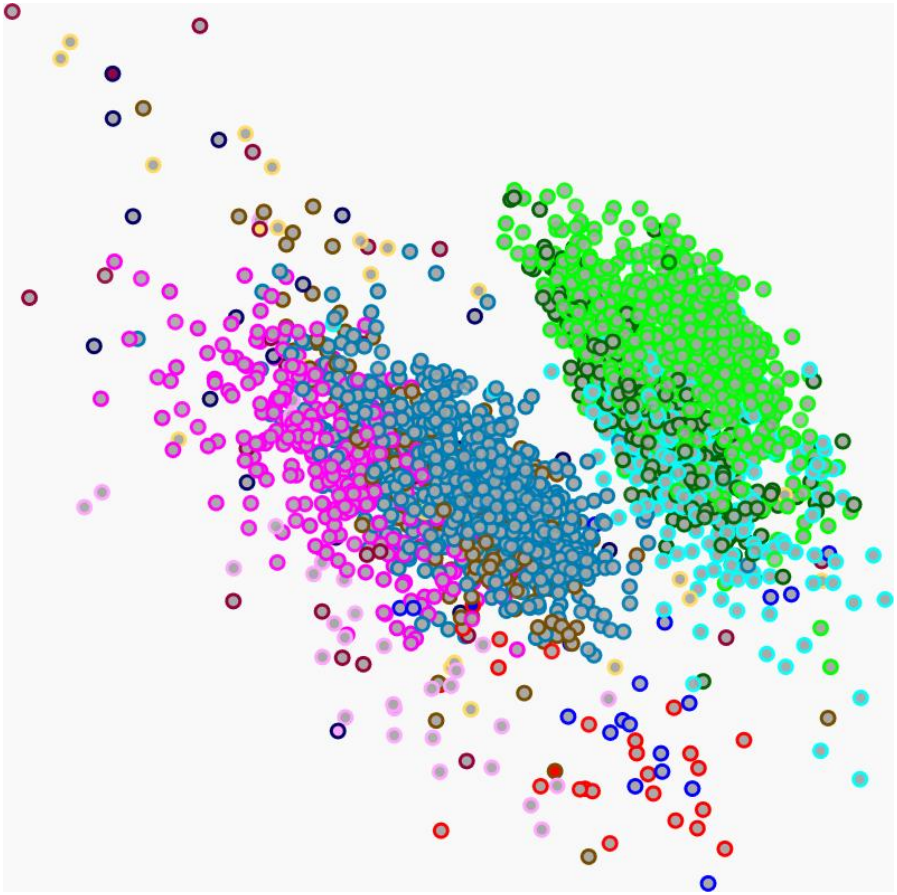
- Computational effort for k-Means is linear (w.r.t. number of samples).
- Computational effort of Sammon's Mapping and matrix inversion is low because of the low number of samples.
- Computational Effort of projecting the feature vectors is linear.

Results – Test on Activity Recognition Data Set

- Test Data Set: “Transition-Aware Human Activity Recognition Using Smartphones” by [Reyes-Ortiz 2016]
- 12 Classes
- ~3000 Samples
- ~500 Features

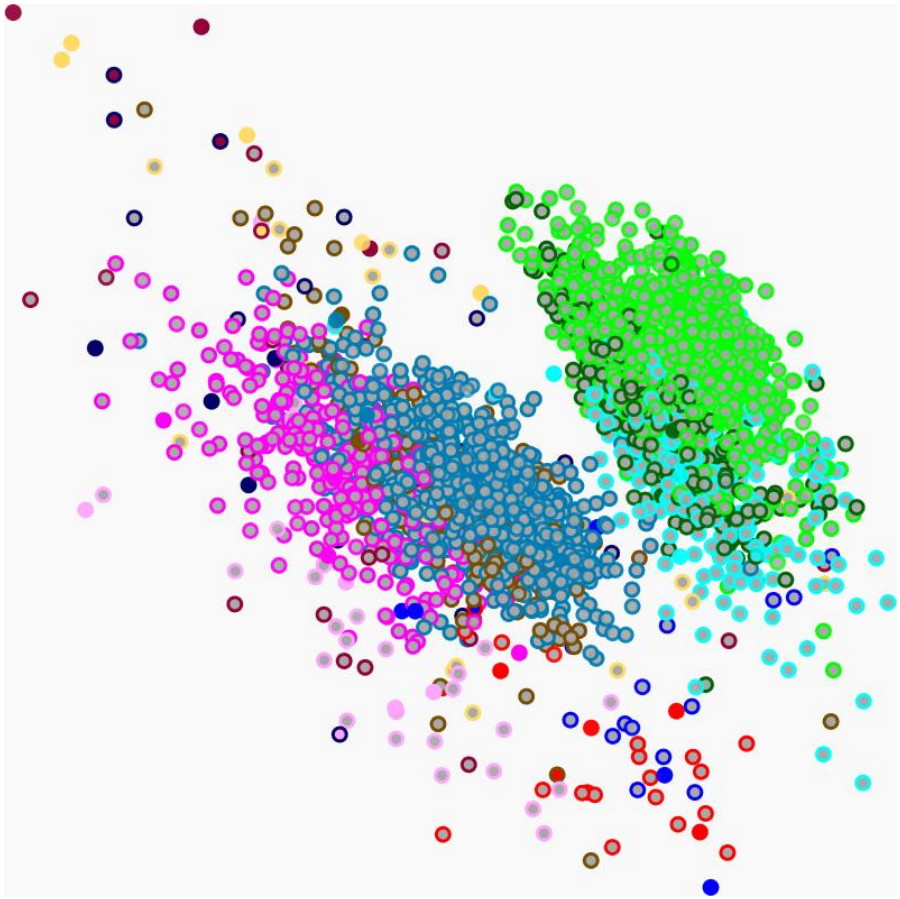
MapView with 1 labeled Sample per Class

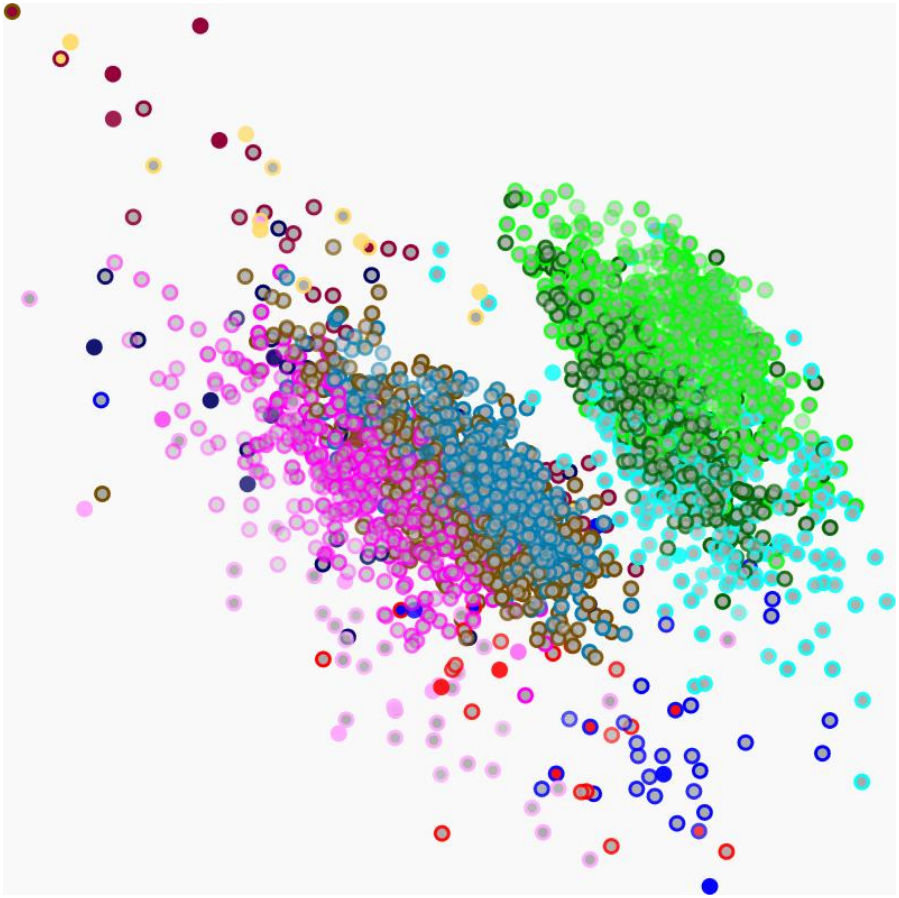




MapView after choosing and labeling 5 more Samples per Class

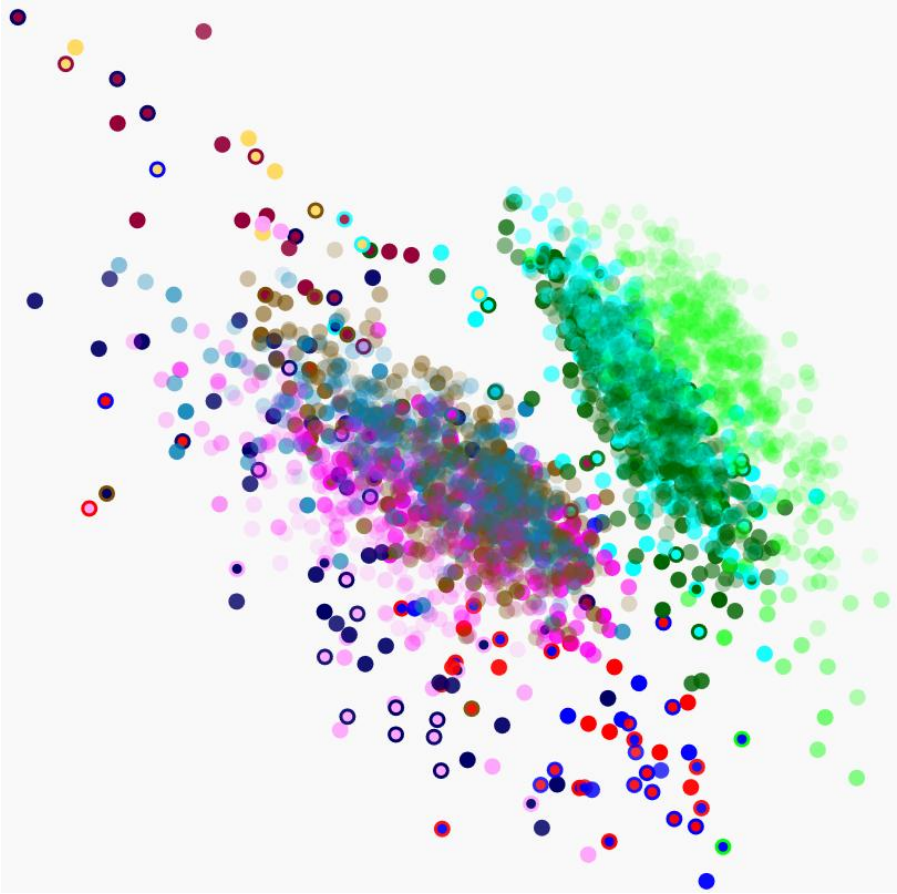
- OOB Error reduced from 100% to 32%





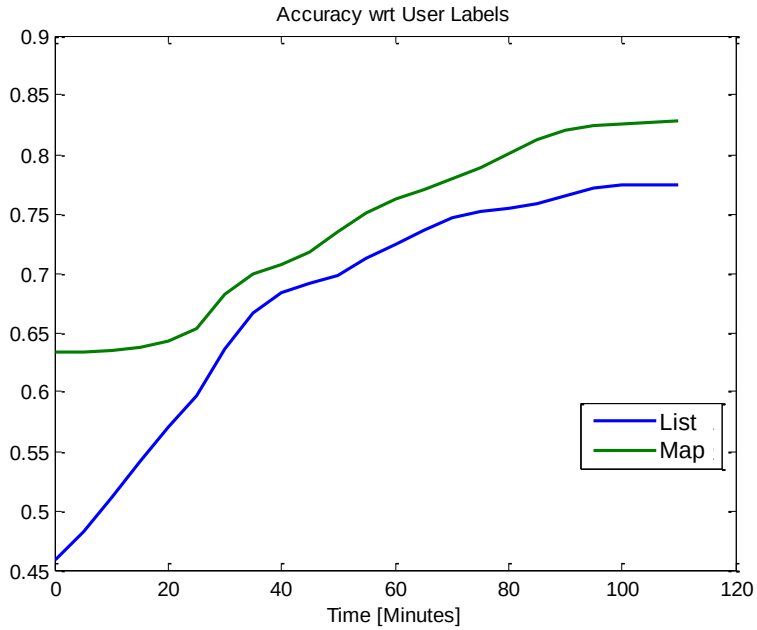
MapView on Ground Truth Data

- OOB Error 5%
- All Samples labeled



Life Labeling Experiment with QC-Experts

- 6 experts trained a classifier from fully unlabeled data for 2h, once with a listed view once with the MapView.
- Comparison showed higher classifier accuracy for the MapView training.



Conclusion

- The MapView helps users to understand the current training state and how to improve classifier accuracy by labeling.
- The transparency of samples labeled with high certainty focuses the annotation to samples that a least certainty approach would have selected.
- The clustering pre-step to dimensionality reduction allows for quick 2D representation (tested on dataset of ~5000 samples).
- Outlook: pre-selection of the most promising features. Further improvement of embedding.

MapView:

Graphical Data Representation for Active Learning

Eva Weigl, Alexander Walch, Ulrich Neissl, Pauline Meyer-Heye, Thomas Radauer, Edwin Lughofer, Wolfgang Heidl and Christian Eitzinger

pauline.meyer-heye@profactor.at

i-KNOW, Graz, 18th of October 2016

LEADING
INNOVATIONS

WWW.PROFACTOR.AT