Roller Bearing Outlier Detection with Random Forests

Based on the *semisupervised* Condition Monitoring scenario introduced in 1.3, only features representing normal conditions of the roller bearing were used to construct a *Random Forest*. For evaluation of the *Random Forest* classifier, both normal and fault condition data was used.

In unsupervised training mode of Random Forests, all available training data is considered to belong to a single class. Based on the training data, a synthetic dataset representing a second class is created, according to the algorithm described in 5.2.5. This balanced dataset is then used to construct a two-class Random Forest including additional features, such as the Proximity Matrix or the Attribute Importance measure.

The *Outlier Measure* of *Random Forest* is only defined for the training data set and cannot be applied to unseen test data (5.2.6). It can however be used to identify critical or implausible data in the training set. Such data can then be removed or modified before a *Random Forest* retraining run.

Fig. 12 shows the *Outlier Measures* for a *Random Forest* trained with 427 roller bearing normal feature vectors.

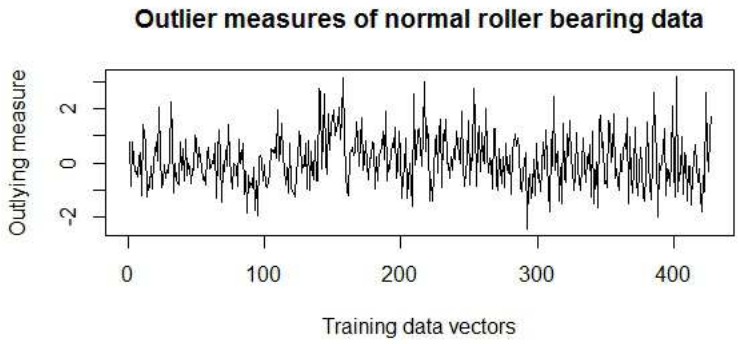

Fig. 1 Outlier Measures of Roller Bearing normal training data

According to a rule of thumb given in (Breiman & Cutler, Random Forests), data points with an *Outlier Measure* beyond a threshold of about $|10|$ require closer inspection. Figure 12 shows clearly, that the *Outlier Measures* for all training data points are smaller than $|4|$. Consequently, no further processing of the training data set was required.

Another useful feature which provides some insight into certain aspects of training data, is the *Attribute Importance* measure introduced in 5.2.2. Figure 13 shows the *Attribute Importance Measures* calculated during a *Random Forest* training with 427 roller bearing normal samples.

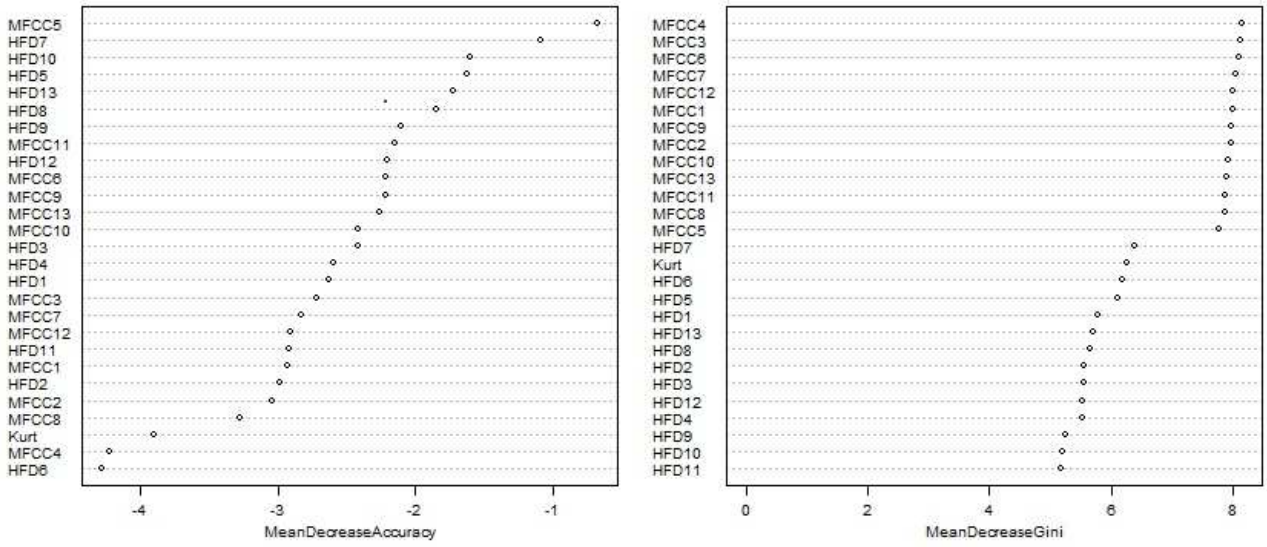

Fig. 2 Attribute Importance of Roller Bearing normal features training set

The *Attribute Importance* can be defined as mean decrease in *accuracy* or as mean *gini* decrease, with both measures resulting in a different importance order of the attributes, as can be seen in Fig, 13. The *Attribute Importance* results were used later, to retrain the *Random Forest* using only the n most important attributes.

Since the built-in *Outlook Measure* is exclusively defined on the training data set, which in this semi-supervised scenario consisted of only normal condition data, a different approach had to be found for the classification problem. In this thesis, a clustering approach based on the *class prototype* (5.2.4) for the normal class was chosen, involving the following steps:

1. Separation of the normal feature set into a training set and a test set
2. Construction of an unsupervised *Random Forest* with the normal feature training set
3. Calculation of a normal class *prototype*
4. Calculation of the *Euclidean Distances* between the normal class *prototype* and the vectors of the normal feature training set
5. Definition of a classification threshold as the largest *Euclidean Distance* value of all the distances calculated in 4
6. Classification of test samples with a Euclidean distance to the prototype smaller than or equal to the classification threshold as normal, or as outlier otherwise.

Figure 14 shows the Euclidean distances between the normal class prototype and a test set consisting of 120 samples from the normal, the ball fault, the Inner Raceway fault and the Outer Raceway fault feature sets.

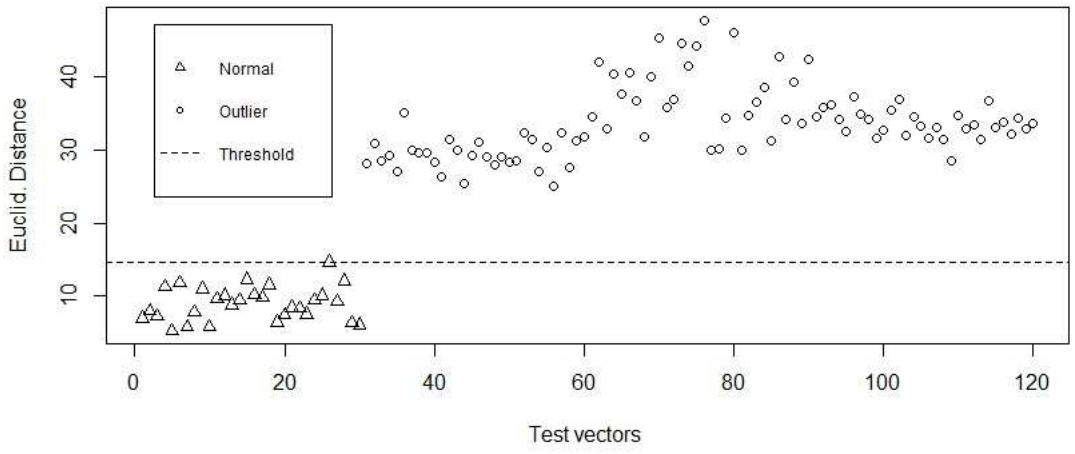

Fig. 3 Euclidean Distances between test data and normal prototype

The plot of Euclidean distances shows, that a constant classification threshold simply defined as the highest Euclidean distance among all normal data is enough to clearly separate normal data from outlier data in this scenario.

In a run of the method described above, the Roller Bearing normal condition feature set with 457 data vectors and 27 attributes was separated into a training set of 382 and a normal test set of 75 samples. The normal test set was then combined with 25 samples of each of the fault state feature sets, to form a complete test set comprising 150 samples.

The training set was used to construct an unsupervised *Random Forest*, including the *Variable Importance* measure and the *Proximity Matrix.* Based on the *Proximity Matrix*, the following normal class prototype was calculated:

MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13
93.814	-4.799	0.910	-7.189	-1.213	-4.746	-6.296	6.458	-1.015	1.593	4.271	-1.854	-5.961
HFD1	**HFD2**	**HFD3**	**HFD4**	**HFD5**	**HFD6**	**HFD7**	**HFD8**	**HFD9**	**HFD10**	**HFD11**	**HFD12**	**HFD13**
1.205	1.264	1.323	1.389	1.463	1.551	1.663	1.781	1.865	1.893	1.888	1.871	1.847
Kurt												
2.773												

Next, the Euclidean Distances between the training vectors and the normal class prototype where calculated and the largest of these distances was used as the classification threshold. Samples from the test set with a Euclidean distance to the class prototype smaller than this threshold were classified as normal or as outlier otherwise. The confusion matrix for this classification is shown in the following table:

	True Classes		
Predicted Classes		**Outlier**	**Normal**
	Outlier	75	0
	Normal	0	75

Based on the confusion matrix, the error rates were calculated as the ratios of rejected normal data (normal data classified as outliers) and the ratio of accepted outlier data (outlier data classified as

normal). Both error types were 0, which proves that the simple classification method introduced here worked perfectly for the given feature sets in a semi-supervised setting.

With the *attribute importance* measures calculated after construction of the *random forest*, some of the least important attributes were removed from the data sets and several reruns of the complete classification process were conducted with the reduced data sets. The results of these reruns showed, that almost half of the attributes could be removed without a significant deterioration of the error rates.