Example Session for Supervised Classification

Andreas Borg

March 23, 2010

This document shows an example session for using supervised classification in the package *RecordLinkage* for deduplication of a single data set. Conducting linkage of two data sets differs only in the step of generating record pairs.

See also the vignette on Fellegi-Sunter deduplication for some general information on using the package.

1 Generating comparison patterns

For supervised classification, a training set is necessary for which the true matching is known. In this session, a training set with 50 matches and 250 non-matches is generated from the included data set `RLData10000`. Record pairs from the set `RLData500` are used to evaluate the trained classifiers.

```
> data(RLdata500)
> data(RLdata10000)
> train_pairs = compare.dedup(RLdata10000,
+     identity = identity.RLdata10000, n_match = 500,
+     n_non_match = 500)
> eval_pairs = compare.dedup(RLdata500,
+     identity = identity.RLdata500)
```

2 Training

`trainSupv` handles training of supervised classificators, the method of classification is set by the argument `method`. In the following, a simple decision tree, a bootstrap aggregation of decision trees and a support vector machine are trained.

```
> model_rpart = trainSupv(train_pairs, method = "rpart")
> model_bagging = trainSupv(train_pairs,
+     method = "bagging")
> model_svm = trainSupv(train_pairs, method = "svm")
```

3 Classification

`classifySupv` handles classification for all supervised classificators, taking as arguments the structure returned by `trainSupv` which contains the classification model and the set of record pairs which to classify.

```
> result_rpart = classifySupv(model_rpart,
+     eval_pairs)
> result_bagging = classifySupv(model_bagging,
+     eval_pairs)
> result_svm = classifySupv(model_svm, eval_pairs)
```

4 Results

4.1 Rpart

alpha error 0.000000

beta error 0.042237

accuracy 0.957780

	N	P	L
FALSE	119433	0	5267
TRUE	0	0	50

4.2 Bagging

alpha error 0.000000

beta error 0.001203

accuracy 0.998798

	N	P	L
FALSE	124550	0	150
TRUE	0	0	50

4.3 SVM

alpha error 0.000000

beta error 0.003913

accuracy 0.996088

	N	P	L
FALSE	124212	0	488
TRUE	0	0	50