ECML PKDD 2014, Nancy

A PEEK INTO THE BLACK BOX

EXPLORING CLASSIFIERS BY RANDOMIZATION

Andreas Henelius[1], Kai Puolamäki,[1] Henrik Boström[2], Lars Asker[2] and Panagiotis Papapetrou[2]

1 Finnish Institute of Occupational Health
{andreas.henelius, kai.puolamaki}@ttl.fi

2 Stockholm University
Department of Computer and Systems Science
{henrik.bostrom, asker, panagiotis}@dsv.su.se

1 BACKGROUND

In predictive data mining, it is important to both find models with high predictive performance, but also to understand what factors that are of importance for the predictions.

Classifiers are often opaque and cannot easily be inspected to gain understanding of which factors that are of importance and how the classifier is utilizing the structure of the data. Many high-performing learning algorithms, such as support vector machines (SVMs) and random forests are complex and can be considered to be black box models.

Randomization testing has been used for detecting if a classifier uses interactions between attributes but this method only detects the existence of attribute interaction, and does not provide information on which attributes interact.

2 A MOTIVATING EXAMPLE

The class variable (+ or −) for this binary toy dataset is given by column *c*. The prediction of a classifier, defined by the binary relation $f(A) = (A_1 \oplus A_2) \vee A_3$, on the original data is given in column *y*. The prediction of the classifier on the randomized data is given in the column $y^* = f(A)$; non-matching predictions that drop fidelity are shown encircled.

c	y	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
+	+	1	0	1	1
+	+	1	0	1	0
+	+	0	1	1	1
+	+	0	1	1	0
+	+	0	0	1	1
+	+	0	0	1	0
+	+	1	1	1	1
+	+	1	1	1	0
+	+	1	0	0	1
+	+	1	0	0	0
+	+	0	1	0	1
+	+	0	1	0	0
−	−	1	1	0	1
−	−	1	1	0	0
−	−	0	0	0	1
−	−	0	0	0	0

y	y^*	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
+	+	0	1	1	1
+	+	0	1	1	0
+	+	0	1	1	1
+	+	1	0	1	0
+	+	0	0	1	1
+	+	1	0	1	0
+	+	1	1	1	1
+	+	1	1	1	0
+	+	1	0	0	1
+	+	1	0	0	0
+	+	0	1	0	1
+	−	0	0	0	0
−	−	0	0	0	1
−	−	1	1	0	0
−	−	0	0	0	1
−	−	1	1	0	0

y	y^*	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
+	+	1	0	1	1
+	+	0	1	1	0
+	+	1	0	1	1
+	+	0	0	1	0
+	+	1	0	1	1
+	+	0	1	1	0
+	+	1	1	1	1
+	+	1	1	1	0
+	−	0	0	0	1
+	−	0	0	0	0
+	+	0	1	0	1
+	+	1	1	0	0
−	+	0	1	0	1
−	−	0	0	0	0
−	−	1	1	0	1
−	+	1	0	0	0

y	y^*	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
+	+	1	0	1	1
+	+	1	0	1	0
+	+	0	1	0	1
+	+	0	1	0	0
+	+	0	0	1	1
+	+	0	0	1	0
+	−	1	1	0	1
+	+	1	1	1	0
+	+	1	0	0	1
+	+	1	0	0	0
+	+	0	1	1	1
+	+	0	1	0	0
−	−	1	1	0	1
−	−	1	1	0	0
−	−	0	0	0	1
−	−	0	0	0	0

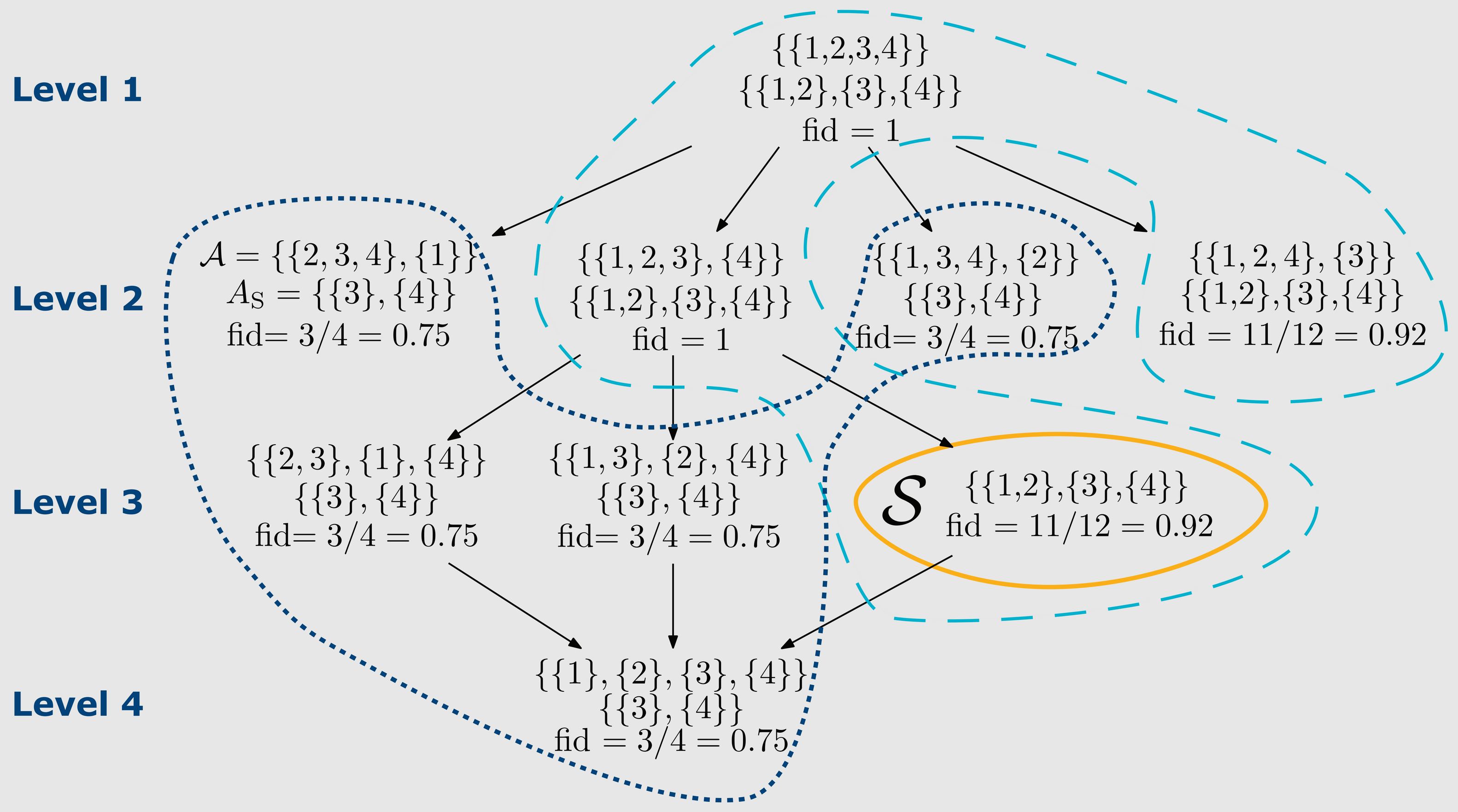
3 THE GOLDENEYE ALGORITHM

We study the novel problem of finding groups of attributes whose interactions affect the predictive performance of a given classifier. The problem is formulated as an optimization problem, using **fidelity** (the fraction of matching predictions between the original dataset and a randomized dataset) as the goodness measure.

The iterative **GoldenEye algorithm** solves the optimization problem and can be used for analyzing what interactions are important for any generic classifier, without any assumptions on the classifier or the underlying distribution of the data.

The goal is to **find groups of interacting attributes**, the breaking of which decreases fidelity. This is realized iteratively, in a top-down and greedy fashion.

The algorithm as available as an R-package
https://bitbucket.org/aheneliu/goldeneye



- Calculate baseline fidelity ("Delta") using an all-singleton grouping
- Remove the attribute from the current grouping that reduces the fidelity the least
  - if fidelity drops below Delta: a group is found
    - store the group and start over using the current grouping
  - else, continue removing attributes
- Using the final grouping, test which singletons can be pruned

4 RESULTS

The groupings reveal interesting patterns in the data. The discovered groupings reflect the asumptions of the classifier and represent the interaction between the classifier and the data.

glass
size: 214
classes: 6
attributes: 9
major class: 0.36

Classifier	original accuracy	final accuracy	Bayes fidelity	final fidelity	mg	al	ri	si	na	fe	ca	k	ba	mg	al	ri	si	na	fe	ca	k	ba
OneR	0.52	0.52	1.00	1.00	·	○	·	·	·	·	·	·	·	4	1	2	5	3	9	7	6	8
JRip	0.55	0.51	0.94	0.91	·	○	·	·	·	·	○	○	·	7	1	5	4	6	9	3	2	8
SMO	0.51	0.50	0.87	0.96	A	A	A	·	·	A	A	○	·	1	2	3	7	8	4	5	6	9
J48	0.58	0.57	0.87	0.97	A	A	A	·	A	A	·	·	○	2	1	5	7	4	6	9	8	3
randomForest	0.73	0.72	0.89	0.99	A	A	A	A	A	A	A	A	·	2	1	3	5	4	8	6	7	9
naiveBayes	0.52	0.51	0.90	0.96	A	A	A	A	A	A	A	A	A	1	4	8	9	7	5	6	3	2
Bagging	0.72	0.69	0.88	0.94	A	A	A	A	A	A	○	·	·	1	2	4	5	3	7	6	8	9
PART	0.63	0.57	0.78	0.90	A	A	A	A	A	○	·	○	○	4	1	7	5	2	8	9	6	3
IBk	0.69	0.55	0.64	0.74	A	A	A	A	○	A	○	·	·	1	2	4	3	5	6	7	8	9
SMO radial	0.66	0.60	0.78	0.89	A	A	A	A	○	A	○	○	·	1	2	4	7	5	3	6	8	9
LMT	0.55	0.52	0.77	0.88	A	A	A	A	○	○	·	A	○	2	1	7	5	4	8	9	3	6
Logistic	0.56	0.43	0.54	0.63	A	○	·	A	A	·	A	○	○	4	6	7	1	3	9	5	2	8
AdaBoostM1	0.47	0.47	1.00	1.00	○	·	·	·	·	·	·	·	·	1	4	2	5	3	9	7	6	8
DecisionStump	0.47	0.47	1.00	1.00	○	·	·	·	·	·	·	·	·	1	4	2	5	3	9	7	6	8
LogitBoost	0.65	0.61	0.78	0.91	○	A	A	A	A	·	A	○	○	4	5	3	6	2	9	1	8	7

5 CONCLUSIONS

The novel algorithm finds groupings of interacting attributes exploited by the different classifiers. These groupings allow for finding similarities among classifiers for a single dataset as well as for determining the extent to which different classifiers exploit such interactions in general.

The method is usable in explorative data mining tasks, as it allows us to peek into black box classifiers and thus aids in the interpration of results.

ACKNOWLEDGEMENTS

AH and KP were partly supported by the Revolution of Knowledge Work project, funded by Tekes. HB and LA were partly supported by the project High- Performance Data Mining for Drug Effect Detection at Stockholm University, funded by Swedish Foundation for Strategic Research under grant IIS11-0053.

FOR DETAILS PLEASE REFER TO

Henelius, A., Puolamäki, K., Boström, H., Asker, L. and Papapetrou, P. A peek into the black box: exploring classifiers by randomization. Data Mining and Knowledge Discovery, 28(5-6): 1503-1529, 2014. **http://dx.doi.org/10.1007/s10618-014-0368-8**.

Stockholm University
Finnish Institute of Occupational Health
BRAIN·WORK RESEARCH CENTRE