Interpreting Classifiers through Attribute Interactions in Datasets

Andreas Henelius[1] **Kai Puolamäki**[1] **Antti Ukkonen**[1]

Abstract

In this work we present the novel ASTRID method for investigating which attribute interactions classifiers exploit when making predictions. Attribute interactions in classification tasks mean that two or more attributes together provide stronger evidence for a particular class label. Knowledge of such interactions makes models more interpretable by revealing associations between attributes. This has applications, e.g., in pharmacovigilance to identify interactions between drugs or in bioinformatics to investigate associations between single nucleotide polymorphisms. We also show how the found attribute partitioning is related to a factorisation of the data generating distribution and empirically demonstrate the utility of the proposed method.

1. Introduction

A lot of attention has been on creating high-performing classifiers such as, e.g., support vector machines (SVMs) (Cortes & Vapnik, 1995) and random forest (Breiman, 2001), both of which are among the best-performing classifiers (Fernández-Delgado et al., 2014). However, the complexity of many state-of-the-art classifiers means that they are essentially opaque, black boxes, i.e., it is very difficult to gain insight into how the classifiers work. Gaining insight into machine learning models is a topic that will become more important in the future, e.g., due to possible legislative requirements (Goodman & Flaxman, 2016). Interpretability of machine learning models is a multifaceted problem, one aspect of which is post-hoc interpretability (Lipton, 2016), i.e., gaining insight into how the method reaches the given predictions.

Interpreting black box machine learning models in terms of *attribute interactions* provides one form of post-hoc interpretability and is the focus of this paper. Given a supervised classification dataset $D = (X, C)$, where X is a data matrix with m predictor attributes $x_1, \ldots, x_m$ (e.g., gender, age etc), and C is a vector with a target attributes (class), an interaction between a subset of these m attributes means that the attributes together provide stronger evidence concerning C than if the attributes are considered alone. We say that attributes interact whenever they are *conditionally dependent given the class*. We next motivate attribute interactions from the perspective of interpretability of real-world problems.

Two difficult problems involving interactions concern drug-drug interactions in pharmacovigilance (e.g., Zhang et al., 2017; Cheng & Zhao, 2014) and investigating associations between single nucleotide polymorphisms (SNPs) in bioinformatics (e.g., Lunetta et al., 2004; Moore et al., 2010). Recently, machine learning methods have been applied to investigate drug-drug (Henelius et al., 2015) and gene-gene interactions (Li et al., 2016). The benefit of using powerful classifiers, such as random forest, is that one does not need to specify the exact form of interactions between attributes (Li et al., 2016), which is necessary in many traditional statistical methods (e.g., linear regression models that include interaction terms). To utilise classifiers in this manner for studying associations in the data requires that we have some method for revealing *how the classifier perceives attribute interactions*.

A *grouping* of the attributes in a dataset is a partition where interacting attributes are in the same group, while non-interacting (i.e., independent) attributes are in different groups. In this paper we study two problems. Firstly we want to *determine if a particular grouping of attributes represents the attribute interaction structure in a given dataset*. Secondly, we want to *automatically find a maximum cardinality grouping of the attributes in a given dataset*.

We approach these problems using the following intuition concerning classifiers, which are used as tools to investigate interactions. A classifier tries to model the class probabilities given the data, i.e., the probability $P(C \mid X) \propto P(X \mid C)\, P(C)$. Here $P(X \mid C)$ is the *class-conditional* distribution of the attributes, which we focus on here. Formally, let $\mathcal{S}$ represent a factorisation of $P(X \mid C)$ into in-

[1] Finnish Institute of Occupational Health, Helsinki, Finland. Correspondence to: Andreas Henelius <andreas.henelius@ttl.fi>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

dependent factors, i.e.,

$$P\left(X \mid C; \mathcal{S}\right) = \prod_{S \in \mathcal{S}} P\left(X\left(\cdot, S\right) \mid C\right) \tag{1}$$

where $X\left(\cdot, S\right)$ only contains the attributes in the set S. In other words, interacting attributes are in the same group $S \in \mathcal{S}$ and, hence, in the same factor in $P\left(X \mid C; \mathcal{S}\right)$.

Assume that the dataset D is sampled from a factorised distribution of the form given in Eq. (1) for some $\mathcal{S}$. Further assume that we can generate datasets $D^{\mathcal{S}}$ that are exchangeable with D. Suppose now that we train a classifier f_1 using D and that we train a second classifier f_2 (of the same type as f_1) using $D^{\mathcal{S}}$. Now, if classifiers f_1 and f_2 cannot be distinguished from each other in terms of accuracy on the same test data, it means that the factorisation $\mathcal{S}$ captures the class-dependent structure in the data to the extent needed by the classifier. On the other hand, if f_2 performs worse than f_1, some essential relationships in the data needed by the classifier are no longer present, i.e., D has not been sampled from a distribution of the form given by Eq. (1). To determine whether f_1 and f_2 are indistinguishable, we compute a confidence interval (CI) for the performance of f_2 by generating an ensemble of datasets $D^{\mathcal{S}}$. If the performance of f_1 is above the CI we conclude that the factorisation $\mathcal{S}$ is not valid.

1.1. Related Work

In this paper we combine the probabilistic approach of Ojala & Garriga (2010) studying whether a classifier utilises attribute interactions at all with the method of Henelius et al. (2014) allowing identification of groups of interacting attributes. For a review on attribute interactions in data mining see, e.g., Freitas (2001). Interactions have been considered in feature selection (Zhao & Liu, 2007; 2009). Mampaey & Vreeken (2013) partition attributes by a greedy hierarchical clustering algorithm based on Minimum Description Length (MDL). Their goal is similar to our, but we focus on supervised learning. Tatti (2011) ordered attributes according to their dependencies while Jakulin & Bratko (2003) quantified the degree of attribute interaction and Jakulin & Bratko (2004) factorised the joint data distribution and presented a method for significance testing of attribute interactions.

1.2. Contributions

We present and study the two problems of (i) assessing whether a particular grouping of attributes represents the class-conditional structure of a dataset (Sec. 2.2) and (ii) automatically discovering the attribute grouping of highest granularity (Sec. 2.3). We empirically demonstrate using synthetic and real data how the proposed ASTRID[1] (Automatic STRucture IDentification) method finds attribute interactions in data (Secs. 3–5).

2. Methods

In this section we consider (i) how to determine if a particular attribute grouping is a valid factorisation of the class-conditional joint distribution, and (ii) automatically finding the maximum cardinality attribute grouping.

2.1. Preliminaries

Let X be an $n \times m$ data matrix, where $X(i, \cdot)$ denotes the ith row (item), $X(\cdot, j)$ the jth column (attribute) of X, and $X(\cdot, S)$ the columns of X given by S, where $S \subseteq [m] = \{1, \ldots, m\}$, respectively. Let $\mathcal{C}$ be a finite set of class labels and let C be an n-vector of class labels, such that $C\left(i\right)$ gives the class label for $X(i, \cdot)$. We denote a dataset D by the tuple $D = (X, C)$.

We denote by $\mathcal{P}$ the set of disjoint partitions of $[m] = \{1, \ldots, m\}$, where a partition $\mathcal{S} \in \mathcal{P}$ satisfies $\cup_{S \in \mathcal{S}} S = [m]$ and for all $S, S' \in \mathcal{S}$ either $S = S'$ or $S \cap S' = \emptyset$, respectively.

Here we assume that the dataset has been sampled i.i.d., i.e., the dataset D follows a joint probability distribution given by

$$\begin{aligned} P\left(D\right) &= \textstyle\prod_{i \in [n]} P(X\left(i, \cdot\right), C\left(i\right)) \\ &= \prod_{i \in [n]} \overbrace{P\left(X\left(i, \cdot\right) \mid C(i)\right)}^{P(X \mid C)} P\left(C\left(i\right)\right), \end{aligned} \tag{2}$$

where $P\left(X \mid C\right)$ is the *class-conditional distribution*. We consider a factorisation of $P\left(D\right)$ into class-conditional factors given by the grouping $\mathcal{S} \in \mathcal{P}$ and write

$$P\left(D\right) = \prod_{i \in [n]} \overbrace{\prod_{S \in \mathcal{S}} P\left(X\left(i, S\right) \mid C\left(i\right)\right)}^{\prod_{S \in \mathcal{S}} P(X(\cdot, S) \mid C)} P\left(C\left(i\right)\right). \tag{3}$$

Given an observed dataset D, we want to find the attribute associations in the data and ask: *Has the observed dataset D been sampled from a distribution given by Eq. (3) with the grouping given by $\mathcal{S} \in \mathcal{P}$?*

2.2. Framework for Investigating Factorisations

Our goal is to determine whether the data obeys the factorised distribution of Eq. (3). To do this we compare the accuracy of a classifier trained using the original data with

[1] R-package available: https://github.com/bwrc/astrid-r

the confidence interval (CI) formed from the accuracies of a collection of classifiers trained using permuted data. The permuted datasets are formed such that they are exchangeable with the original dataset if Eq. (3) holds. If the accuracy of the original data is above the CI we can conclude with high confidence that the data does not obey the factorised distribution.

We denote a classifier trained using the dataset D by f_D. Further assume that we have a separate independent test dataset from the same distribution as D, denoted by $D_{\text{test}} = (X_{\text{test}}, C_{\text{test}})$.

Definition 1. Classification Accuracy *Given the above definitions, the accuracy for a classifier trained using D is given by*

$$T\left(D\right) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} I\left[f_D\left(X_{\text{test}}\left(i, \cdot\right)\right) = C_{\text{test}}\left(i\right)\right], \quad (4)$$

where $I\left[\Box\right]$ is the indicator function and n_{test} is the number of items in the test dataset.

Note that T is not the accuracy of f on D, but the accuracy of f on X_{test} when f is trained using D. Because direct sampling from Eq. (3) is not possible as the data generating model is unknown, we generate the permuted data matrices $X^{\mathcal{S}}$ so that they have same probability as X *under the assumption that X is a sample from a factorised distribution as given in Eq.* (3). This means that X and $X^{\mathcal{S}}$ are *exchangeable* under the assumption of a joint distribution that is factorised in terms of $\mathcal{S}$.

We sample datasets using the permutation scheme described in Henelius et al. (2014). A new permuted dataset $D^{\mathcal{S}} = \left(X^{\mathcal{S}}, C\right)$ is created by permuting the data matrix of the dataset $D = (X, C)$ at random. The permutation is defined by m bijective permutation functions $\pi_j : [n] \mapsto [n]$ sampled uniformly at random from the set of allowed permutations functions. The new data matrix is then given by $X^{\mathcal{S}}\left(i, j\right) = X\left(\pi_j\left(i\right), j\right)$. The allowed permutation functions satisfy the following constraints for all $i \in [n]$, $j, j' \in [m]$, and $S \in \mathcal{S}$:

1. permutations are within-a class, i.e., $C\left(i\right) = C\left(\pi_j\left(i\right)\right)$, and
2. items within a group are permuted together, i.e., $j \in S \wedge j' \in S \implies \pi_j\left(i\right) = \pi_{j'}\left(i\right)$.

Let $\mathcal{D}_{\mathcal{S}}$ be the set of datasets that can be generated by the above permutation scheme using the grouping $\mathcal{S}$. We note:

Lemma 1. *Each invocation of the permutation scheme produces each of the datasets in $\mathcal{D}_{\mathcal{S}}$ with uniform probability.*

Lemma 2. *The datasets in $\mathcal{D}_{\mathcal{S}}$ have equal probability under the distribution of Eq.* (3)*, parametrised by $\mathcal{S}$.*

Proof. The proofs follow directly from the definition of the permutation and the probability distribution of Eq. (3). □

Definition 2. Confidence intervals *Given a dataset D, a grouping $\mathcal{S}$, a classifier f and an integer R, let $A = \left\{T\left(D_1^{\mathcal{S}}\right), \ldots, T\left(D_R^{\mathcal{S}}\right)\right\}$ be a vector of accuracies where the datasets $D_i^{\mathcal{S}}$ are obtained by the permutation parametrised by $\mathcal{S}$, and T is as in Eq.* (4)*. The CI is the tuple $C = (c_{lower}, c_{upper})$, where c_{lower} and c_{upper} are values corresponding to the 5% and 95% quantiles in A, respectively.*

We cast the above discussion as a problem:

Problem 1. *Given an observed dataset D, a grouping $\mathcal{S}$ and a classifier f, let a_0 be the accuracy of f (trained using the original data) on the test set. Determine if the upper end of the CI of Def. 2 for the accuracy of a classifier trained using factorised data is at least a_0.*

If the above condition is met, we conclude that the factorisation correctly captures the structure of the data.

2.3. Automatically Finding Groupings (ASTRID)

In the previous section we examined whether a *particular grouping $\mathcal{S}$* describes the structure of the data in terms of the factorisation in Eq. (3). A natural step is now to ask *how to find the grouping best describing the associations in a dataset D*? Here we choose *best* to be the grouping $\mathcal{S}$ of (i) maximum cardinality such that (ii) a classifier trained using data shuffled with $\mathcal{S}$ is indistinguishable in terms of accuracy from a classifier trained using the original, unfactorised data.

Finding the maximum cardinality grouping is motivated by the fact that in this case there are no irrelevant interactions. Also, interpreting attribute interactions in small groups is easier than in large groups. The requirement on accuracy means that no essential information is lost and in practice this means that the upper end of the CI for the accuracy of the classifier f trained using $D^{\mathcal{S}}$ is at least as large as the original accuracy a_0 of f trained using D.

Exhaustive search of all groupings is in general impossible due to the size of the search space. Hence, to make our problem tractable we assume that accuracy decreases approximately monotonically with respect to breaking of groups in the correct solution, i.e., the more the interactions are broken, the more classification performance decreases. Using this property we use a *top-down greedy algorithm* termed ASTRID. For details see the extended description in Henelius et al. (2017). In practice, T in Eq. (4) is susceptible to stochastic variation and for stability we instead use *expected accuracy V* when optimising accuracy in the greedy algorithm:

$$V\left(\mathcal{S}\right) = \frac{1}{N}\sum_{i=1}^{N} T\left(D_i^{\mathcal{S}}\right), \tag{5}$$

where N is the number of samples used to calculate the expectation, $D_i^{\mathcal{S}}$ ($i \in [N]$) is a dataset generated by the permutation parametrised by $\mathcal{S}$ and T is defined as in Eq. (4).

3. Experiments

We use ASTRID to identify attribute interactions. We use a synthetic dataset and 11 datasets from the UCI machine learning repository (Bache & Lichman, 2014)[2]. All experiments were run in R (R Core Team, 2015) and our method is released as the ASTRID R-package, available for download[3]. We use a value of $R = 250$ in Def. 2 and $N = 100$ in Eq. (5). In all experiments the dataset was randomly split as follows: 50% for training (D) and the rest for testing (D_{test}, see Eq. (4)): 25% for computing V (Eq. (5)), and 25% for computing CIs. As classifiers we use support vector machines (SVM) with RBF kernel, random forest (RF) and naïve Bayes (NB).

The datasets are summarised in Table 1. The **UCI datasets** were chosen so that the SVM and random forest classifiers achieve reasonably good accuracy at default settings, since the goal here is to demonstrate the applicability of the method rather than optimise classifier performance. Rows with missing values and constant-value columns were removed from the UCI datasets. The **synthetic dataset** has two classes, each with 500 data points. Attributes 1 and 2 carry meaningful class information only when considered jointly, attribute 3 contains some class information and attribute 4 is random noise. The correct grouping is hence $\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$.

4. Results

The results are presented as tables where each row is a grouping and the columns represent attributes. Attributes belonging to the same group are marked with the same letter, i.e., attributes marked with the same letter on the same row are interacting.

Table 2 shows the results for the synthetic dataset where the highest-cardinality grouping is highlighted and is also shown below the table. Using the SVM and RF classifiers ASTRID identifies the correct attribute interaction structure ($k = 3$). For $k = 4$ the accuracy is clearly lower. For naïve Bayes all groupings (all values of k) are equally valid since the classifier assumes attribute independence. The results mean that the average accuracy of an SVM or RF classifier trained on the synthetic dataset permuted using $\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$ is within CIs. ASTRID reveals the factorised form of the joint distribution of the data, which makes it possible to identify the attribute interaction structure exploited by the classifier in the datasets. This makes the models more interpretable and we, e.g., learn that NB does not exploit interactions (as expected!).

Table 1: The datasets used in the experiments (2–10 from UCI). Columns as follows: Number of items (Ni) after removal of rows with missing values, number of classes (Nc) after removal of constant-value columns, number of attributes (Na). MCP is major class proportion. $\mathbf{T_{SVM}}$ and $\mathbf{T_{RF}}$ give the computation in minutes of the ASTRID method for the SVM and random forest, respectively.

n	Dataset	Ni	Nc	Na	MCP	$\mathbf{T_{SVM}}$	$\mathbf{T_{RF}}$
1	synthetic	1000	2	4	0.50	0.1	0.4
2	balance-scale	625	3	4	0.46	0.1	0.3
3	diabetes	768	2	8	0.65	0.2	1.1
4	vowel	990	11	13	0.09	1.2	56.1
5	credit-a	653	2	15	0.55	0.8	3.5
6	vote	232	2	16	0.53	0.6	0.9
7	segment	2310	7	18	0.14	3.7	14.2
8	vehicle	846	4	18	0.26	1.5	6.8
9	mushroom	5644	2	21	0.62	13.1	19.8
10	soybean	682	19	35	0.13	9.1	29.5
11	kr-vs-kp	3196	2	36	0.52	42.2	41.7

The groupings for the UCI datasets are summarised in Table 3. SVM and RF are in general similar in terms of the cardinality (k), with the exception of `kr-vs-kp` and `soybean`. In many cases it appears that the classifiers utilise few interactions in the UCI datasets. To compare this finding with the results of Ojala & Garriga (2010), we calculated the value of their Test 2, denoted p_{OG} in Table 3. This test investigates whether a classifier utilises attribute interactions. $p_{\text{OG}} \geq 0.05$ indicates that no attribute interactions are used by the classifier, which we find for `diabetes` and `soybean` for SVM and for `diabetes` and `credit-a` for random forest (highlighted in the table). This is in line with the findings from ASTRID, since for these datasets k equals N in Table 3 and no interactions are hence utilised as the dataset can be factorised into singleton groups.

Finally, as an illustrative example of grouping attributes exploited by a classifier we consider the `vote` dataset. This dataset contains yes/no information on 16 issues with the target of classifying if a person is republican or democrat. Using SVM ASTRID finds that the maximum cardinality grouping is of size $k = 8$ (Tab. 3). The grouping consists of **7 singleton attributes** (`water-project-cost-sharing`, `synfuels-corporation-cutback`, `physician-fee-freeze`, `education-spending`, `duty-free-exports`,

[2] Datasets obtained from http://www.cs.waikato.ac.nz/ml/weka/datasets.html

[3] https://github.com/bwrc/astrid-r (R-package and source code for experiments)

Table 2: The `synthetic` dataset. The cardinality of the grouping is k and CI is the confidence interval for accuracy. Original accuracy using unshuffled data (a_0) and the final grouping ($\mathcal{S}$, highlighted row) shown above and below the table, respectively. An asterisk ($*$) denotes that the factorisation is valid.

(a) SVM

$a_0 = 0.908$

k	**CI**	a3	a4	a2	a1
2	[0.900, 0.920] *	(A)	(B	B	B)
3	[0.896, 0.920] *	(A)	(B)	(C	C)
4	[0.696, 0.784]	(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$

(b) Random forest

$a_0 = 0.904$

k	**CI**	a3	a4	a1	a2
2	[0.896, 0.928] *	(A)	(B	B	B)
3	[0.896, 0.928] *	(A)	(B)	(C	C)
4	[0.668, 0.756]	(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$

(c) Naïve Bayes

$a_0 = 0.760$

k	**CI**	a1	a2	a3	a4
2	[0.760, 0.760] *	(A)	(B	B	B)
3	[0.760, 0.760] *	(A)	(B)	(C	C)
4	[0.760, 0.760] *	(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1\},\{2\},\{3\},\{4\}\}$

`export-administration-act-south-africa`, `immigration`) and **one group with 9 interacting attributes** (`crime`, `handicapped-infants`, `religious-groups-in-school`, `superfund-right-to-sue`, `adoption-of-the-budget-resolution`, `mx-missile`, `anti-satellite-test-ban`, `aid-to-nicaraguan-contras`, `el-salvador-aid`). It appears that the 9 attributes in the group roughly represent military and foreign policy issues, and economic and social issues. This means, that the SVM exploits relations between these 9 political issues when classifying persons into republicans or democrats. On the other hand, the singleton attributes seem to mostly represent domestic economic, economic and export issues. The classifier does not use any singleton attribute jointly with any other attribute when making predictions.

Note that ASTRID is a randomised algorithm and the found groupings are hence not necessarily unique. The stability of the results depends on factors such as the used classifier, the size of the data and the strength of the interactions. Also, the results are affected by the number of random samples (R in Def. 2 and N in Eq. (5)) and for practical applications a trade-off between accuracy and speed must be made.

5. Discussion and Conclusion

Interpreting black box machine learning models is an important emerging topic in data mining and in this paper we present the ASTRID method for investigating classifiers. This method provides insight into generic, opaque classifier by revealing how the attributes are interacting. ASTRID automatically finds in polynomial time the maximum cardinality grouping such that the accuracy of a classifier trained using the factorised data cannot be distinguished (in terms of confidence intervals) from a classifier trained using the original data. The method makes no assumptions on the data distribution or the used classifier and hence has high generic applicability to different datasets and problems. This work extends previous research (Henelius et al., 2014; Ojala & Garriga, 2010) on studying attribute interactions in opaque classifiers.

Table 3: Groupings for UCI datasets. Columns as follows: number of attributes in the dataset (N), size of the grouping (k), size of the largest (N_1) and second-largest (N_2) groups, baseline accuracy for the classifier trained with unshuffled data (a_0) and the CI. p_{OG} is the p-value of Test 2 in Ojala & Garriga (2010) ($p \geq 0.05$ highlighted).

Dataset	**N**	**k**	$\mathbf{N_1}$	$\mathbf{N_2}$	$\mathbf{a_0}$	**CI**	$\mathbf{p_{OG}}$
		SVM					
balance-scale	4	3	2	1	0.891	[0.821, 0.897]	0.03
credit-a	15	12	4	1	0.871	[0.847, 0.871]	0.04
diabetes	8	8	1	1	0.714	[0.688, 0.740]	0.59
kr-vs-kp	36	33	4	1	0.917	[0.922, 0.924]	0.00
mushroom	21	15	7	1	0.995	[0.991, 0.995]	0.00
segment	18	3	16	1	0.948	[0.936, 0.948]	0.00
soybean	35	35	1	1	0.844	[0.820, 0.850]	0.26
vehicle	18	3	15	2	0.767	[0.719, 0.781]	0.00
vote	16	8	9	1	0.931	[0.897, 0.931]	0.00
vowel	13	3	11	1	0.806	[0.760, 0.806]	0.00
		random forest					
balance-scale	4	3	2	1	0.821	[0.731, 0.833]	0.02
credit-a	15	15	1	1	0.877	[0.847, 0.883]	0.19
diabetes	8	8	1	1	0.703	[0.698, 0.740]	0.89
kr-vs-kp	36	16	21	1	0.982	[0.972, 0.982]	0.00
mushroom	21	14	8	1	1.000	[0.996, 1.000]	0.00
segment	18	4	15	1	0.986	[0.979, 0.986]	0.00
soybean	35	24	12	1	0.964	[0.946, 0.964]	0.00
vehicle	18	3	13	4	0.752	[0.710, 0.757]	0.00
vote	16	10	7	1	0.948	[0.897, 0.948]	0.00
vowel	13	3	11	1	0.917	[0.901, 0.917]	0.00

Knowledge of attribute interactions exploited by classifiers is important in, e.g., pharmacovigilance and bioinformatics (see Sec. 1) where powerful classifiers are used in data analysis, since they make it possible to simultaneously investigate multiple attributes instead of, e.g., just pairwise interactions. Here ASTRID allows the practitioner to automatically discover attribute groupings, providing insight into the data by making the classifiers more transparent.

Acknowledgements

This work was supported by Academy of Finland (decision 288814) and Tekes (Revolution of Knowledge Work project).

References

Bache, K. and Lichman, M. UCI machine learning repository, 2014. URL http://archive.ics.uci.edu/ml.

Breiman, Leo. Random Forests. *Machine Learning*, 45(1): 5–32, 2001.

Cheng, Feixiong and Zhao, Zhongming. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21(e2):e278–e286, 2014.

Cortes, Corinna and Vapnik, Vladimir. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

Fernández-Delgado, Manuel, Cernadas, Eva, Barro, Senén, and Amorim, Dinani. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

Freitas, Alex A. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, 16(3):177–199, 2001.

Goodman, Bryce and Flaxman, Seth. EU regulations on algorithmic decision-making and a ”right to explanation”. In *ICML Workshop on Human Interpretability in Machine Learning*, 2016. URL http://arxiv.org/abs/1606.08813.

Henelius, Andreas, Puolamäki, Kai, Boström, Henrik, Asker, Lars, and Papapetrou, Panagiotis. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5-6):1503–1529, 2014.

Henelius, Andreas, Puolamäki, Kai, Karlsson, Isak, Zhao, Jing, Asker, Lars, Boström, Henrik, and Papapetrou, Panagiotis. Goldeneye++: A closer look into the black box. In *Statistical Learning and Data Sciences*, pp. 96–105. Springer, 2015.

Henelius, Andreas, Puolamäki, Kai, and Ukkonen, Antti. Finding Statistically Significant Attribute Interactions. *arXiv e-prints, arXiv:1612.07597*, 2017.

Jakulin, Aleks and Bratko, Ivan. Analyzing attribute dependencies. In *PKDD 2003*, pp. 229–240. Springer, 2003.

Jakulin, Aleks and Bratko, Ivan. Testing the significance of attribute interactions. In *ICML 2004*, 2004.

Li, Jing, Malley, James D, Andrew, Angeline S, Karagas, Margaret R, and Moore, Jason H. Detecting gene-gene interactions using a permutation-based random forest method. *BioData mining*, 9(1):14, 2016.

Lipton, Zachary C. The Mythos of Model Interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*, 2016. URL https://arxiv.org/abs/1606.03490.

Lunetta, Kathryn L, Hayward, L Brooke, Segal, Jonathan, and Van Eerdewegh, Paul. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):32, 2004.

Mampaey, Michael and Vreeken, Jilles. Summarizing categorical data by clustering attributes. *Data Mining and Knowledge Discovery*, 26(1):130–173, 2013.

Moore, Jason H, Asselbergs, Folkert W, and Williams, Scott M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

Ojala, Markus and Garriga, Gemma C. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

Tatti, Nikolaj. Are your items in order. In *SDM 2011*, pp. 414–425. SIAM, 2011.

Zhang, Wen, Chen, Yanlin, Liu, Feng, Luo, Fei, Tian, Gang, and Li, Xiaohong. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics*, 18(1):18, 2017.

Zhao, Zheng and Liu, Huan. Searching for interacting features. In *IJCAI 2007*, pp. 1156–1161, 2007.

Zhao, Zheng and Liu, Huan. Searching for interacting features in subset selection. *Intell. Data Anal.*, 13(2):207–228, 2009.