RESEARCH

Machine learning-based analysis of the relationship between bone density and drinking patterns of primates

Pablo Rivas[1*], Urszula Iwaniec[2], Kathleen A. Grant[3] and Erich Baker[4]

*Correspondence:
Pablo.Rivas@Marist.edu
[1]School of Computer Science and Mathematics, Marist College, 3399 North Road, 12601 Poughkeepsie, NY, USA
Full list of author information is available at the end of the article

Abstract

Background: Alcohol use disorders (AUDs) represent one of the largest causes of death in the world. More research is needed to understand of the underlying relationships between diverse factors contributing to AUDs, which may lead to models for prevention and treatment of alcohol abuse-related disease, such as bone damage.

Methods: We use a machine learning techniques to investigate the relevance of features from a dataset of monkeys that were subject to an experiment where they self-administered alcohol. We use support vector machines for regression to predict bone mineral density from variables related to drinking and we rank all the features and sets of features to determine which features and which sets of features are more predictive of bone damage.

Results: Empirical evidence suggests that the age at which the monkeys were intoxicated for the first time is one of the best predictors along with information about the maximum bout volume and the percentage of ethanol consumed during a period of two hours in which the monkeys had unlimited access to alcohol. We proved that such features are the best predictors using Friedman's test and Nemenji's test with confidence of $\alpha = 0.05$ and $\alpha = 0.01$ respectively.

Conclusion: Machine learning strategies can successfully inform about the relationship between bone mineral density and alcohol consumption data from primates. Specifically, the investigation of which features are more relevant, both by themselves and as sets, providing the means for better understanding the contribution of some variables over others in the modeling of AUDs.

Keywords: alcohol use disorders; machine learning; support vector regression; forward selection; backward selection; relevance feature ranking

Background

The mining and modeling of biological data is not trivial either in cases where the data is abundant or scarce. This creates problems of large-scale learning and analysis of Big Data as well as feature selection in smaller-scale problems where there is risk of finding under-determined solutions or ill-posed models. Biological data varies greatly depending of the field of study since measurements and meaning of the attributes are extremely different. Our research deals with the specific problem of data analysis for alcohol drinking pattern in primates as we aim to provide a better understanding of the different aspects that model alcohol use disorders (AUDs).

Currently, AUDs is a worldwide issue that has been officially presented as one of the most deadly disorders in the United States [1]. Recent studies indicate that there is a clear increase in the number of AUD cases [2], and we need a better understanding of all the influencing factors in AUDs, such as sex, income, and other disorders that are concurrently pathological [3]. To alleviate the need of understanding all the variables that play an important role in AUDs we developed a non-human primate (NHP) macaque model of oral alcohol self-administration that has identified a range of daily ethanol intakes that encompass categorical levels of drinking severity, *i.e.*, Low Drinkers (LD), Binge Drinkers (BD), Heavy Drinkers (HD), and Very Heaver Drinkers (VHD) [4]. These categories are a reflection of AUDs severity, especially with respect to HD and VHD categories since they are associated with problems of dependence and other brain pathologies [5, 6, 7]. However, such categories do not prescribe by themselves all that is necessary for modeling AUDs. Therefore, the analysis of other factors using machine learning techniques is ultimately important.

In this research we report on the study of drinking patterns of monkeys that were subject to a process of alcohol self-administration. The protocol followed has been established previously [8], and it is known as "schedule-induced polydipsia". From this induction protocol we recorded data for further analysis. In particular, we explore different feature ranking strategies that allow us to understand non-trivial associations using machine learning algorithms. We make use of correlation, entropy, density, and forward-backward selection of features as indicators of the relationship between features and their relevance in the problem of modeling bone damage in primates [9].

Using support vector machines for regression as a control model for prediction [10], we rank the feature selection methods to further identify which features are more predictive of bone mineral density (BMD). The study of the relationship between alcohol consumption patterns and BMD is important for the understanding of the underlying system that causes permanent bone damage. In this paper we will show that the amount of alcohol consumed in a period of two hours of open access to ethanol is one of the most determining factors of BMD in monkeys that were intoxicated early in life.

Methods

Primate Subjects

We studied the post-mortem BMD of primates from the Oregon National Primate Research Center (ONPRC) to determine which parameters of alcohol consumption are more predictive of BMD values. We considered a total of 68 monkeys out of which 14 are females and 54 are males. They are of two species known as Cynomolgus and Rhesus. Table 1 summarizes the information about the monkeys of each cohort. Our study consists of nine cohorts in total. The table also indicates the drinking category count for each cohort. Most monkeys fall into the category of low drinkers (LDs). Note, however, that not all cohorts had a category assigned to them, as is the case of cohort "INIA Cyno 8", which, at the moment of this research, has partial drinking information. Nonetheless, there are other alcohol consumption parameters available that are considered further.

With partially available drinking categorical data the task of directly associating BMD with a specific drinking category is cumbersome. Figure 1 depicts the average

Table 1 Summary of monkey cohorts. This study includes two species of primates: Cynomolgus and Rhesus. Most monkeys are males and low drinkers. Drinking category was not available for some Cynomolgus monkeys at the date of this research.

Cohort Name	Total	Females	Males	BD	LD	HD	VHD
INIA Cyno 2	12	0	12	0	10	0	1
INIA Rhesus 7a	8	0	8	1	3	2	2
INIA Cyno 9	11	0	11	0	6	2	0
INIA Cyno 8	3	3	0	0	0	0	0
INIA Rhesus 7b	5	0	5	1	3	1	0
INIA Rhesus 4	10	0	10	4	5	1	0
INIA Rhesus 5	8	0	8	1	0	3	4
INIA Rhesus 6a	6	6	0	0	0	0	6
INIA Rhesus 6b	5	5	0	0	3	1	1
Total	68	14	54	7	30	10	14

BMD of monkeys broken down by gender and drinking category. The figure suggests no apparent BMD difference between drinking categories. Although it may seem that low drinking females have more BMD when compared to high drinking females, the standard deviation, indicated in the figure, suggests that such conclusion could be misleading.

The next step is to consider other factors that could be associated with BMD. This research considers not only sex as a possible determining factor of BMD, but it also considers investigating ethanol (EtOh) consumption patterns in sets that are potentially predictive of BMD. The EtOh consumption data comes from experiments carried at the ONPRC, which are described next.

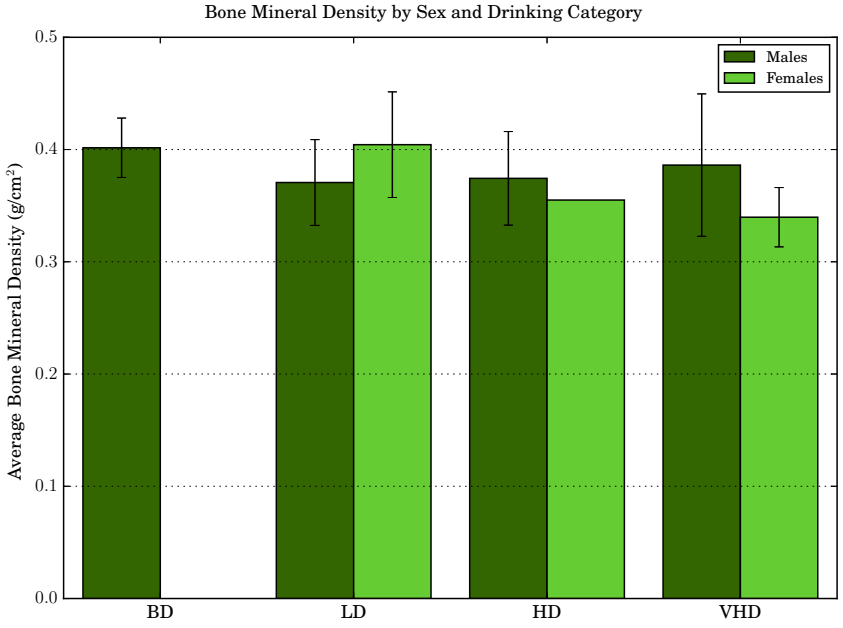
Housing and Development

Pedigreed monkeys were born into a population remaining with their mothers until they were about two years old and completely independent from their mothers for survival. After that, the monkeys were part of an experiment designed to study alcohol consumption and their physiological and biological effects. Monkeys were continually housed at the ONPRC and entered into an internal housing laboratory with individual caging at least three months prior to the onset of ethanol self-administration according to established protocols [11]. The age of onset is defined as the number of days a monkey has lived until their blood ethanol concentration (BEC) is greater than 80 mg/dl [12, 13].

Some monkeys are members of the Cynomolgus cohort that, in our previous research [8], have been identified as chronic heavy drinkers. The age of the monkeys in this experiment varies starting from late adolescence to early middle age. Our recent findings indicate that age was also a contributing factor for chronic levels of alcohol intake of the Cynomolgus cohort [11].

The preparation for the ethanol consumption study involved the setting of four individual cages belonging to a single rack, two above and two below in a horizontal arrangement. Each individual cage had a panel providing the means for drinking water or alcohol. For the comfort of heavier monkeys, over 10kg, they were allowed to occupy two horizontal cages but only one drinking panel was available. A single cohort occupied a single space using racks as necessary, providing visual and auditory contact with other monkeys. In the case of females, tactile access was available by permitting entry to a common area two hours a day, which was made by removing the horizontal divisory barrier between two horizontally adjacent cages. In the

Figure 1 Bone mineral density by drinking category and sex. There is no significant difference between drinking categories or sex. It may seem that high drinking females have lower bone density compared to low drinking ones; however, the standard deviation indicator makes such assumption unreliable. The horizontal axis shows the following drinking categories: low drinker (LD), binge drinker (BD), heavy drinker (HD), and very heavy drinker (VHD). There are no females reported as BDs.



case of males, tactile interchange was available since adjacent males were accessible through groom panel inserts in the common wall of the cage. Details of the panels for drinking are described in full in our previous research [14, 8]. The function of such panels is described next.

Ethanol Self-Administration

In this study we had panels permanently affixed to individual cages allowing access to food and fluid. We collected data every time food or fluid was self-administered by monkeys. During our research we followed a previously established procedure that we defined as "schedule-induced polydipsia" [8], which we will refer to as "induction", for short. In such process we induced monkeys to self-administer 4 w/v (%) EtOh in water. The defining feature of the induction procedure is a monotonic increase in the maximum volume of EtOh allowed for consumption. This increase occurs every 30 days in the following order: 0 g/kg/d, which is a volume of water equivalent to 1.5 g/kg of EtOH, then to 0.5 g/kg/d, next to 1.0 g/kg/d, and finally to 1.5 g/kg/d. In this manner, all monkeys drank to levels that saturated metabolic capacity elevating their BEC over 50 mg/dl.

After the 30th session of 1.5 g/kg EtOH under induction the monkeys had concurrent access to 4 w/v (%) EtOh and a 22 h/d "open-access" to water and food in the form of 1g of banana flavored pellets (Noyes), provided at least three times a day during meals, and at least two hours apart with the first meal provided at the beginning of the session. The open-access phase comprises 12 months. Blood ethanol

concentration levels were measured every fifth day 30, 60, and 90 minutes after the beginning of the induction sessions and approximately every fifth day seven hours after the start of the open-access sessions, which corresponds to the onset of the dark phase of the 11 hour day and 13 hour night diurnal cycle.

Computational Resources for Analysis

All computations were executed on MATRR's web and database servers. These are twin dedicated computers each running four Intel Xeon E5620 processors at 2.4 GHz, with four cores per processor, *i.e.*, a total of 16 cores per server. Each server has 47 GiB of random access memory (RAM) and 1.7 TiB of disk space in a redundant array of independent disks (RAID) configuration. Web services use Django's object-relational mapping (ORM) for database access. We carried out all statistical analysis and data processing using Pandas, NumPy, and SciPy, which are packages for the Python programming language. For all machine learning algorithms we used Python's Scikit-Learn package. Graphical results, such as the figures in this paper, were produced with the Python package called Matplotlib.

Drinking Features for Analysis

In our research we studied five factors contributing to the correct prediction of BMD. From such factors, also known as attributes, we produced a total of 14 attributes. In machine learning research, we call this attributes "features". All 14 features are explained by groups in the next paragraphs.

Drinking Category

Each monkey can be classified to belong to one of the following four categories, if enough data is available: low drinker (LD), binge drinker (BD), heavy drinker (HD), and very heavy drinker (VHD). We provide specific details of the defining criteria for each category in [4]. The data collected has monkeys categorized as follows: BDs: 11.5%, LDs: 49.2%, HDs: 16.4%, and VHDs: 22.94%.

Sex

This feature represents the sex of each monkey. The distribution of the monkeys in this feature is as follows: Females: 20.6% and Males: 79.4%.

Age at First Intoxication (in Days)

The number of days a monkey has lived until the first time its BEC level is greater than 80 mg/dl describes the age at onset or "age at first intoxication". In this research the average age at first intoxication is $\mu = 2292$ days with a standard deviation of $\sigma = 510$ days.i The age at first intoxication is often associated with the drinking category of monkeys [15].

Maximum Bout Volume (in mL)

The volume in mL of each bout of EtOh a monkey drinks is recorded and organized by session. One session is equivalent to one day. The volume of the maximum bout per each session is available per monkey. We particularly focus in the maximum bout volume that occurs during the first 120 minutes after monkeys had open-access to

EtOh. The overall average maximum bout volume is $\mu = 0.728$ mL with a standard deviation of $\sigma = 0.365$ mL.

From this data we extracted the following five features per monkey for the first 120 minutes of all sessions: a) average maximum bout volume (μ), b) standard deviation of the maximum bout volume (σ), c) median of the maximum bout volume, d) maximum of the maximum bout volume (max), and e) minimum of the maximum bout volume (min).

EtOh During Induction (in %)
During the period of induction, explained earlier, monkeys are exposed to limited amounts of EtOh. Such limits decrease gradually until the monkeys pass the period of induction to full unrestricted access. However, during the period of induction we recorded, in percentage, the amount of EtOh the monkeys ingest out of what they are allowed to drink.

Percentages of EtOh during induction are available per monkey and per session. From all this data we extract the following six features: a) average % of EtOh during induction (μ), b) median % of EtOh during induction, c) total sum of the % of EtOh during induction (Σ), d) standard deviation of the % of EtOh during induction (σ), e) minimum % of EtOh during induction (min), and f) maximum % of EtOh during induction (max).

Bone Mineral Density (in g/cm^2)
This feature is the object of our study. In machine learning this feature is called the target. Since the target variable is not a categorical one, regression algorithms are preferred for analysis that go beyond basic statistical examination.

The average BMD is $\mu = 0.376$ g/cm^2 and the standard deviation is $\sigma = 0.049$g/cm^2. In the following paragraphs we explain the methodologies we use to analyze the importance or contribution of the above features in predicting BMD or in finding the parameters of a machine learning model predictive of BMD.

Estimation of Importance and Feature Ranking

In data mining and machine learning in general, researchers aim to have models that are computationally inexpensive and robust. One factor that reduces the risk of poor performance and high complexity is the selection of a set of features that together provide the best way of modeling the desired output. The most popular way of selecting "good" features is known as "feature ranking" [9].

Feature ranking can be done in an individual manner where each feature is evaluated by its relevance in predicting the target data. Usually, individual relevance ranking yields good results, especially when the features are independent and identically distributed (IID). When the data is not IID, there is a risk of discarding features that, if combined with others, may have greater relevance.

In this research we explored methods that evaluate features independently and in sets of features to determine their relevance. The determination of their relevance will give us insight as to which features are associated with BMD, providing a better understanding of how alcohol consumption relates to BMD.

The next paragraphs briefly describe two methods for individual relevance ranking, Pearson correlation, entropy, and density; and also three methods for combined

relevance ranking, which are based on the well-known support vector machines for regression (SVRs) [10].

Pearson Correlation

Let $\mathbf{x}$ be a feature vector of dimension n, $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, where $n = 14$ is the total number of features in the original dataset. A single feature vector $\mathbf{x}$ is also known as an observation. Let y be a the target, or desired output, of the learning model and that is paired in a tuple with a single feature vector $\mathbf{x}$. Then, we can define the dataset as $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$, where m is the total number of samples. It is important to remark that feature selection becomes crucial for cases where $m < n$, or when the dataset is small in general.

The Pearson correlation coefficient, C is a classical measure of relevance for individual feature ranking. If we define $\mathbf{x}_j$ as a vector that contains all the observations corresponding to the individual feature j, then the Pearson correlation coefficient of feature j with respect to the target vector $\mathbf{y} = [y_1, y2, \ldots, y_m]$ is defined by the following equation:

$$C(j) = \frac{|\sum_{i=1}^{m}(x_{i,j} - \bar{x}_j)(y_i - \bar{y})|}{\sqrt{\sum_{i=1}^{m}(x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x}_j$ denotes the average of $\mathbf{x}_j$ and $\bar{y}$ denotes the average of $\mathbf{y}$, both over the the index i.

Coefficient C is directly related to absolute value of the cosine between vectors $\mathbf{x}_j$ and $\mathbf{y}$ after they have been centered with respect to their means, and is also associated with the Fisher coefficient, and is furthermore related to the T-test statistic and Naïve Bayes [9].

In this research features will be ranked according to their C, where the less relevant are the ones with the smallest C and the more relevant are the ones with the largest value of C.

The overall complexity of the algorithm for calculating the Pearson correlation coefficient is for all features is $\mathcal{O}(mn)$.

Differential Entropy

The concept of entropy has been widely used in the area of information theory shortly after Shannon's definition of entropy was accepted. A measure of entropy is a measure of the amount of information that is contained in a random variable [16].

In the case of our research, the entropy of each feature informs about the amount of information each feature contains. If a feature is categorical, *e.g.*, drinking category or sex, the entropy coefficient will determine how much information each categorical feature has; in the case of the categorical feature "sex", since there are more males than females, we expect a low entropy, which indicates that the values in the categorical feature "sex" are very predictable; on the other hand, the entropy of the categorical feature "drinking category" should be higher than "sex", meaning that is less predictable than "sex".

If we assign a different real value to each category and define X_j as the random variable corresponding to the vector $\mathbf{x}_j$, where $X \in \mathbb{R}$, then the entropy of the $j-$th feature, according to Shannon's definition, is the following:

$$H(X_j) = -\sum_{\mathbf{x}_j} p(\mathbf{x}_j) \log_2(p(\mathbf{x}_j)), \tag{2}$$

where $p(\mathbf{x}_j)$ is the probability mass function (PMF) of the categorical feature $\mathbf{x}_j$. PMFs are easy to compute from categorical features because they are simple discrete variables; however, most of the features used in this research are defined as continuous random variables. For such cases, we must estimate the "differential entropy" for continuous variables.

The differential entropy can be defined as follows:

$$H(X_j) = -\int_{\mathbf{x}_j} p(\mathbf{x}_j) \log_2(p(\mathbf{x}_j)) d\mathbf{x}_j, \tag{3}$$

where $p(\mathbf{x}_j)$ is the probability density function (PDF) of the $j-$th continuous real-valued feature. In the case of entropy for discrete random variables, PMFs can be easily obtained; however, PDFs for continuous random variables need to be approximated, especially, when their distributions are completely unknown.

In this research, we have enough information about the features to know that they approximate uniform distributions, and the estimation of the PDFs is performed accordingly. In a general sense, features with a very predictable value are less likely to be relevant; thus, low entropy features are lowly ranked, whereas features with more information, *i.e.*, high entropy, will rank higher.

Note that calculating the entropy requires no knowledge of the target variable Y, which makes this process faster to compute, in comparison with Pearson correlation coefficient. However, the overall complexity of estimating the differential entropy is the same as for Pearson's, that is, $\mathcal{O}(mn)$.

Density

A feature that is highly correlated with many other variables is said to be in a high density region. The fact that one single feature is correlated with another, or with a group of features, does not imply that the feature is redundant [9]. On the contrary, such feature may complement another, or a group, providing a holistic improvement.

One way of calculating the density of a feature is to average the Pearson correlation coefficient of one feature with respect to the others. Formally, if we let $\mathcal{F}$ be the set of indices of the features, $\mathcal{F} = \{1, 2, \ldots, n\}$, then we can define the density of feature j as follows:

$$\bar{D}(j) = \frac{1}{n-1} \sum_{k \in \mathcal{F} \setminus j} \left(\frac{|\sum_{i=1}^{m} (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)|}{\sqrt{\sum_{i=1}^{m} (x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^{m} (x_{i,k} - \bar{x}_k)^2}} \right). \tag{4}$$

Because of the nature of our research, we are not interested in investigating features that are by themselves predictive of BMD, since we know that the patterns

of alcohol consumption are more meaningful when combined with other factors [8, 4, 15]. Therefore, features with a high density are ranked higher than the rest.

Due to the nature of the density coefficient, the complexity is higher, being $\mathcal{O}(m^2 n)$; however, it can be optimized by storing already computed averages and centered vectors leading to a complexity of $\mathcal{O}(mn)$ in the average case.

Linear SVR

Support vector machines (SVMs) are optimal maximum-margin classifiers whose learning algorithms are based on the solution to optimization problems. These algorithms have been extended to solve regression problems where the target output is a real-valued variable. Such are called support vector machines for regression (SVRs).

SVRs have been extensively developed to reduce the high complexity of the learning algorithms, especially when kernel mappings of the data are used [10]. In problems related to classification, SVMs have been successfully used [17, 18, 19]. Similarly, they can be also used for regression tasks with minor adjustments, as we describe next.

From a dataset $\mathcal{D}$, an SVR model attempts to find the solution to the problem of finding a function $f(\mathbf{x}_i)$ that takes the $i-$th feature vector $\mathbf{x}_i$ as input and produces a response, ideally, equal to y_i. It tries to do so by finding the parameters $\mathbf{w}$ and b in the regression equation: $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b = y_i$. Since the solution to the problem may not exist exactly, errors are permitted in the model up to the point determined by a new parameter defined as ϵ. Then, an SVR in its primal form is defined as follows:

$$\begin{array}{lll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} & \frac{1}{2}||\mathbf{w}||_2^2 + C\left(\nu\epsilon\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right)\right) & \\ \text{s.t.} & \left\{\begin{array}{rcl} y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b & \leq & \epsilon + \xi_i \\ \mathbf{w}^T\phi(\mathbf{x})_i + b - y_i & \leq & \epsilon + \xi_i^* \\ \boldsymbol{\xi}, \boldsymbol{\xi}^* & \geq & \mathbf{0} \\ \epsilon & \geq & 0 \end{array}\right. & \\ \text{for} & i = 1, 2, \ldots, m & \end{array} \tag{5}$$

where $0 \leq \nu \leq 1$ is a parameter that controls the number of support vectors and training errors. The summation in the cost function accounts for the $\epsilon-$insensitive training error. The constant $C > 0$ describes the trade off between the training error and the penalizing term $||\mathbf{w}||_2^2$. The term $||\mathbf{w}||_2^2$ is penalized to enforce a sparse solution on $\mathbf{w}$. The variables ξ_i and ξ_i^* are two sets of non-negative slack variables that describe the $\epsilon-$insensitive loss function. Usually, $\mathbf{x}$ is transformed using a mapping function $\phi(\cdot)$ known as a kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. However, using a kernel mapping, not only increases the complexity of the algorithm by a factor of m^2 but will also force us to redefine $\mathbf{w}$ as a parameter that is no longer linearly associated to $\mathbf{x}$.

For the case of feature selection the most important parameter is $\mathbf{w}$, because it directly points to the relevance of each feature [17, 18, 19]. Therefore, we will use a linear mapping, *i.e.*, $\phi(\mathbf{x}) = \mathbf{x}$. Then, we use grid search to determine the best set of

parameters ν and C that yield the best score overall using 10-fold cross validation. The best score is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2}, \tag{6}$$

where $\hat{y}$ is the predicted target. Here, a value of 1 is the best score, and anything less is proportionally worse.

After the SVR is trained with the best set of parameters, we observe the values in the vector $|\mathbf{w}|$. The smallest values indicate that the feature associated with such low value is being inhibited, thus, irrelevant. While the opposite is also true, the highest values indicate higher relevance in the process of prediction. We use this information to rank the features accordingly.

The complexity of finding a solution to the SVR optimization problem, using the SMO algorithm with a linear kernel, is $\mathcal{O}(m^2 n)$. This is without considering the cost of the grid search.

SVR Backward

In feature selection research there are two major strategies for ranking sets of features, forward selection and backward selection [9]. Backward selection starts with a model that considers the set of all the features first and starts removing the feature that produces the smallest contribution to the overall performance. It continues to remove features until only one is left. The feature that is removed first is considered the less relevant and the last feature standing is considered the most relevant.

In this research, we perform backward selection using the SVR model explained previously; however, notice that in this case we no longer look at $\mathbf{w}$, allowing us to use a kernel function, which after several experiments we found that a radial-basis function (RBF) was the best choice. This time, instead of looking at $\mathbf{w}$ we look at the R^2 score defined in (6) and how it varies as we remove features progressively.

The procedure is the following. First, start with all n features and test the hypothesis that there is one feature, j, that is the less relevant of the entire set. We test the hypothesis by removing one feature from the set and record the score, then put it back and remove another one doing the same until all features have been removed once. Second, eliminate from the set the feature that, when it was first removed, produced the worst score and rank it last. Third, repeat the previous steps until there is only one feature left, which should be ranked as number one. At the end, all features are ranked.

The advantage of backward selection is that it looks at sets of features at the same time it indicates the contribution of individual features to the current set. This is particularly important when we look at the set of features that produced the overall best performance, as it gives insight about which features as a group predict better the BMD.

The disadvantage of this methodology is that it is more computationally expensive. The complexity of the SVR itself is now $\mathcal{O}(m^3 n)$ due to the kernel function. And since the process of having sets of $n-1, n-2, \ldots, 1$ features selected corresponds to a series that converges to $n(n-1)/2 = \mathcal{O}(n^2)$, then the total overall complexity becomes $\mathcal{O}(m^3 n^2)$. This complexity does not consider the cost of grid search and leave-one out cross validation.

Table 2 Feature ranking results. Each row is represents a feature considered for predicting BMD. Columns two to seven show the relevance feature ranking methodology considered. The last column indicates the average ranking of each feature. Age at first intoxication, median of maximum bout volume, and average of maximum bout volume are the top three features. The critical difference, $\delta_{CD} = 1.82$ with $\alpha = 0.01$, suggests that the top five ranked features are significantly better than the rest.

$\mathcal{F}$	Feature	C	H	D	Linear SVR	SVR Backward	SVR Forward	Avg. Rank
1	Drinking Category	14	11	4	8	6	12	9.16
2	Sex	2	13	5	13	13	1	7.83
3	Age at First Intoxication	1	6	2	7	7	2	4.16
4	μ of Maximum Bout Vol.	5	3	14	1	5	4	5.33
5	σ of Maximum Bout Vol.	4	2	12	3	3	11	5.83
6	Median of Maximum Bout Vol.	3	4	10	2	4	3	4.33
7	$\max$ of Maximum Bout Vol.	11	8	7	10	8	9	8.83
8	$\min$ of Maximum Bout Vol.	7	5	3	12	14	14	9.16
9	μ % of EtOh During Ind.	13	10	13	6	1	13	9.33
10	Median % of EtOh During Ind.	9	12	8	5	9	10	8.83
11	Σ % of EtOh During Ind.	6	7	9	4	2	8	6.00
12	σ % of EtOh During Ind.	10	1	11	9	10	6	7.83
13	$\min$ % of EtOh During Ind.	8	9	6	11	11	5	8.33
14	$\max$ % of EtOh During Ind.	12	14	1	14	12	7	10.00

SVR Forward

The forward selection of features, as opposed to backward selection, begins with an empty set of features and starts adding individual features one by one. The feature that, when added, produced the best score is added permanently to the current set of features. This process is repeated until all features have been added to the current set.

The advantage of forward selection is that individual features are considered first and then their contribution to an existing set after the first iteration. This process is faster at the beginning since it starts with an empty set, and could potentially be stopped early if the performance worsens instead of becoming better. However, potential disadvantages include that it could miss holistic relationships among features, and that it is also expensive in the computational sense.

Since we also use the SVR approach as in SVR backward selection, the complexity remains the same, $mathcalO(\mathcal{O}(m^3n)$.

Results and Discussion

For each feature ranking methodology we ranked all 14 features in our data set. The results of the ranking are shown in Table 2. The left most column in the table shows the index assigned to each feature forming the set of indices $\mathcal{F}$. The second column indicates the feature name and the right most column indicates the average ranking of such features. The rest of the columns indicate the specific raking each feature obtained for a particular relevance ranking methodology. From the average column we see that the top five features are: 1) age at first intoxication, 2) median of maximum bout volume, 3) μ of maximum bout volume, 4) σ of maximum bout volume, and 5) Σ % of EtOh during induction.

Based on the ranking of the features, we performed the Friedman test [20], and determined the Friedman statistic to be $\chi_F^2 = 17.3143$ with $p = 0.1853$, for a level of significance of $\alpha = 0.05$. Furthermore, the critical difference δ_{CD} using the Nemenji's test was calculated to determine if two features are significantly different if the corresponding average ranks differ by such amount [21]. The value of the critical difference for a confidence level of $\alpha = 0.01$ corresponds to $\delta_{CD} = 1.82$.

Algorithm 1 Training $\nu-$SVRs with Ranked Sets of Features

Input: The dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$. A vector, $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, $r_i \in \mathbb{Z}^+$, whose elements are indices of the ranked features.
Output: A vector $\mathbf{e} = [e_1, e_2, \ldots, e_n]$, $e_i \in \mathbb{R}^+$ whose elements are leave-one-out crosvalidation mean absolute errors for each set of features.

$\mathcal{F} \leftarrow \varnothing$
for $i \leftarrow 1 \ldots n$ **do**
  $\mathcal{F} \leftarrow \mathcal{F} \cup r_i$
  $C, \gamma, \nu \leftarrow \textsc{GridSearch}_{\mathrm{SVR}}(\mathcal{D}, \mathcal{F})$
  $e_i \leftarrow \textsc{TrainSVR}_{\mathrm{LOO}}(\mathcal{D}, \mathcal{F}, C, \gamma, \nu)$
end for

Using the δ_{CD} value we observe that the top four features are not significantly different form each other. However, the top five features are significantly better than the rest features. This suggest that the top five features are most contributing features.

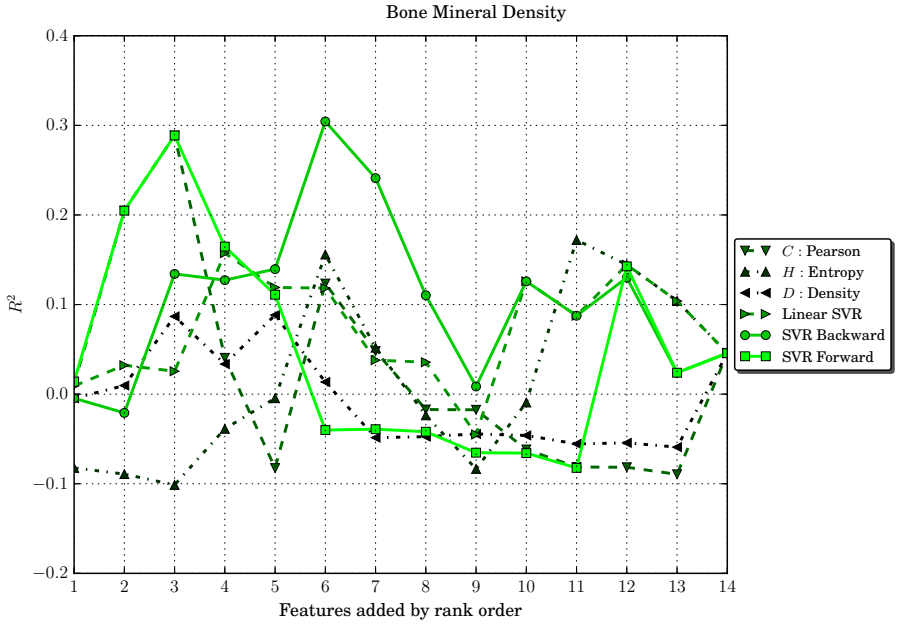
Then, after knowing which features are seemingly relevant now the next step is to fully test sets of features with robust algorithms. The sets of features are created as size-increasing sets based on their ranking. Algorithm 1 shows the steps followed to test the performance of the ranked features. The input to the algorithm is the dataset $\mathcal{D}$ and a vector $\mathbf{r} = [r_1, r_2, \ldots, r_n]$, where r_1 is the index of the feature ranked as the best and r_n is the worst ranked feature. Note, however, that $\mathbf{r}$ is different for every different relevance ranking methodology. *E.g.*, consider the column entitled "C" in Table 2, the corresponding $\mathbf{r}$ would be: $\mathbf{r} = [3, 2, 6, 5, 4, \ldots, 1]$

As shown in Algorithm 1, we perform a traditional grid search to find the best set of hyper-parameters $\{C, \gamma, \nu\}$ that produce the best 10-fold cross validation R^2 score. The search is conducted in the logarithmic spaces $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-14}, \ldots, 2^{3}\}$; and the linear space $\nu \in \{0.05, 0.10, \ldots, 0.95\}$. Then, with the best set of hyper-parameters, we trained $\nu-$SVRs with the leave-one-out (LOO) cross validation strategy, which has proven to provide more accurate estimates of error in smaller datasets [22].

Results from performing Algorithm 1 for all the relevance feature ranking methods are depicted in Figure 2. The figure plots all $\mathbf{e}$ vectors starting from one feature to all 14. Note that since the first feature added, *i.e.*, the best, is distinct for every method, the initial score varies. However, all converge to the same score once all 14 features are considered together. Ideally, we want the set of features that achieves the highest scores with the smallest number of features. From the figure we observe that the best relevance feature ranking strategies are Pearson correlation coefficient, C, SVR forward, and SVR backward. Particularly, SVR forward and Pearson correlation score high with only three features, while SVR backward performs high with six features.

The next logical step is to analyze which features are in the set that produced the highest score for all methods in order to determine which are more frequent. Table 3 presents a summary of which features are frequently part of the best set of features for a given ranking methodology. The table also presents a weighted frequency count using the rank of each method. The rank was determined as the best overall score of a given method regardless of the number of features. From Figure 2 it can be determined that the best score corresponds to SVM backward and the worst to Density.

Figure 2 Performance of the sets of features using the R^2 score. Pearson correlation coefficient, SVR backward, and SVR forward score the highest among all. With SVR forward and Pearson correlation coefficient the scores are achieved with only three features, while SVR backward with six features.



Careful inspection of Table 3 reveals that, considering the frequency count of the features, the two features that are frequently part of the best sets of features are, (with a count of 5) median of maximum bout volume and (with a count of 4) age at first intoxication. On the other hand, if we consider the weighted frequency count which ponders the count using the rank, we observe that the top two features are, again, median of maximum bout volume with $w_{\text{Freq.}} = 2.25$, and three ties with $w_{\text{Freq.}} = 1.45$ for μ of maximum bout volume, σ of maximum bout volume, and Σ % of EtOh during induction.

In a general sense, Table 3 suggests that features related to drinking patterns are associated with the information of age at onset of intoxication in predicting BMD. This is consistent with the findings shown in Table 2.

Figure 3 depicts how two features, namely "median maximum bout volume" and "age at first intoxication", interact with each other in predicting BMD. The figure suggests that lower BMD is associated with a median maximum bout volume between 10 and 150 mL for young monkeys whose age is 1400 and 1700 days of age. The risk of having low BMD is less for older monkeys of ages greater thatn 2400 days. This two features do not appear to have a linear relationship in a two-dimensional plane; however, using the kernel method in SVRs, the possibility of having a linear relationship in a higher-dimensional space is often assumed.

A similar analyis is shown in Figure 4. In this case the two features analized are "median maximum bout volume" and "sum of % EtOh during induction". The relationship between these two features appears to be quasi-linear, suggesting a more evident dependence or correlation. From the figure, we can observe that if the median maximum bout volume is less than 100 mL there is a higher risk of low

Table 3 Analysis of sets of features producing the best score. Each column corresponding to a ranking method has a format b/r, where b is either a zero or a one, indicating that the feature was part of the set that produced the best score for that method; r is the rank of that method compared to the others. The method with the best overall score is ranked one and the worst is ranked as six; in case of a tie, the ranks are averaged. The last column shows the frequency count of how many times that feature was part of a best set; in parenthesis, $w_{\text{Freq.}}$ indicates the sum of r/b for each feature as a weighted frequency count based on the rank. Quantities resulting in zero have been omitted.

$\mathcal{F}$	Feature	C	H	D	Lin. SVR	SVR Bwd.	SVR Fwd.	Freq. ($w_{\text{Freq.}}$)
1	Drinking Category		1/4	1/6		1/1		3 (1.42)
2	Sex	1/2.5		1/6			1/2.5	3 (0.97)
3	Age at First Intoxication	1/2.5	1/4	1/6			1/2.5	4 (1.22)
4	μ of Maximum Bout Vol.		1/4		1/5	1/1		3 (1.45)
5	σ of Maximum Bout Vol.		1/4		1/5	1/1		3 (1.45)
6	Median of Maximum Bout Vol.	1/2.5	1/4		1/5	1/1	1/2.5	5 (2.25)
7	$\max$ of Maximum Bout Vol.		1/4					1 (0.25)
8	$\min$ of Maximum Bout Vol.		1/4	1/6				2 (0.42)
9	μ % of EtOh During Ind.		1/4			1/1		2 (1.25)
10	Median % of EtOh During Ind.							
11	Σ % of EtOh During Ind.		1/4		1/5	1/1		3 (1.45)
12	σ % of EtOh During Ind.		1/4					1 (0.25)
13	$\min$ % of EtOh During Ind.		1/4					1 (0.25)
14	$\max$ % of EtOh During Ind.			1/6				1 (0.17)

BMD for almost any sum of % of EtOh, while the risk decreases in the opposite direction.

Conclusion

Using machine learning techniques known as relevance feature ranking we explored a dataset containing several attributes concerning drinking patterns in primates. Such attributes were categorical and quantitative. We determined that the "median maximum bout volume", "sum of % of EtOh during induction", "μ of maximum bout volume", "σ of maximum bout volume", and "age at first intoxication" are the features most predictive of bone mineral density. Further analysis indicated that there are features significantly better than others in a statistical sense; this was later demonstrated, also, through a full training of SVRs over the data, adding features as they were ranked by each method.

We also analyzed the complexity of the feature ranking algorithms and observed that the Pearson correlation coefficient is significantly less expensive and performs well for very up to three features, after that, Pearson does not consider the inter-dependence of the data. On the other hand, the SVR Backward selection strategy was the best of all, considering all the features at once and their relationship as a group; however, it is more expensive than Pearson's coefficient.

Competing interests

The authors declare that they have no competing interests.

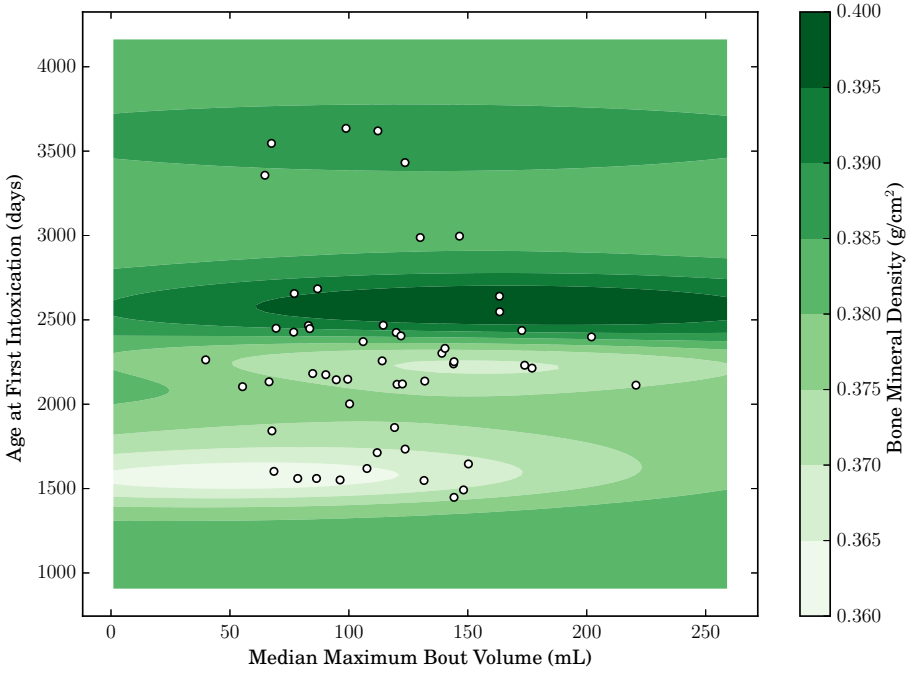
Author's contributions

PRP designed and performed the relevance feature ranking study. All authors were actively involved in the project. EB provided data and experimental design. UI and KG provided the data. PRP conducted the analysis. EB, UI and KG revised early drafts of the manuscript. All authors commented on and approved the final version of the manuscript.

Acknowledgements

This work was supported by NIAAA grant AA019431. This work was supported in part by the National Council for Science and Technology (CONACyT), Mexico, under grant 193324/303732 provided to PRP.

Figure 3 BMD values using "median maximum bout volume" and "age at first intoxication" as predictors. Predicted BMD suggests that young monkeys are at higher risk of low BMDs if their median maximum bout volume is in the range of 10 and 150 mL.



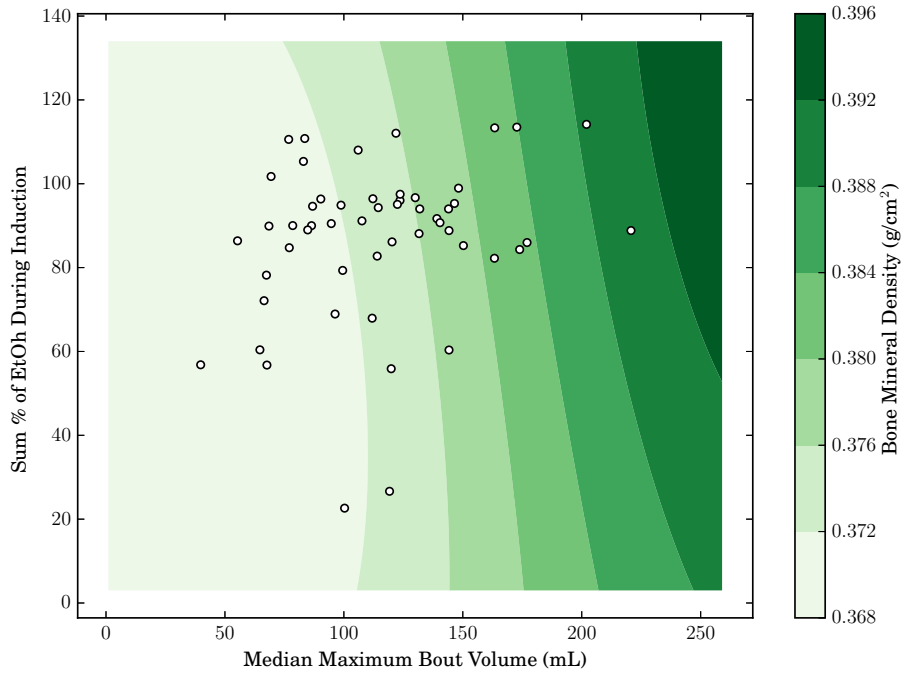
Author details

[1]School of Computer Science and Mathematics, Marist College, 3399 North Road, 12601 Poughkeepsie, NY, USA. [2]College of Public Health and Human Sciences, Oregon State University, 108 Milam Hall, 97331 Corvallis, OR, USA. [3]Department of Behavioral Neuroscience, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, 97239 Portland, OR, USA. [4]Department of Computer Science, Baylor University, One Bear Place #97356, 76798 Waco, TX, USA.

References

1. Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L.: Actual causes of death in the united states, 2000. Jama **291**(10), 1238–1245 (2004)
2. Grant, B.F., Stinson, F.S., Dawson, D.A., Chou, S.P., Dufour, M.C., Compton, W., Pickering, R.P., Kaplan, K.: Prevalence and co-occurrence of substance use disorders and independentmood and anxiety disorders: Results from the national epidemiologic survey on alcohol and relatedconditions. Archives of general psychiatry **61**(8), 807–816 (2004)
3. Grant, B.F., Goldstein, R.B., Saha, T.D., Chou, S.P., Jung, J., Zhang, H., Pickering, R.P., Ruan, W.J., Smith, S.M., Huang, B., *et al.*: Epidemiology of dsm-5 alcohol use disorder: Results from the national epidemiologic survey on alcohol and related conditions iii. JAMA psychiatry **72**(8), 757–766 (2015)
4. Baker, E.J., Farro, J., Gonzales, S., Helms, C., Grant, K.A.: Chronic alcohol self-administration in monkeys shows long-term quantity/frequency categorical stability. Alcoholism: Clinical and Experimental Research **38**(11), 2835–2843 (2014)
5. Kroenke, C.D., Rohlfing, T., Park, B., Sullivan, E.V., Pfefferbaum, A., Grant, K.A.: Monkeys that voluntarily and chronically drink alcohol damage their brains: a longitudinal mri study. Neuropsychopharmacology **39**(4), 823–830 (2014)
6. Chen, G., Cuzon Carlson, V.C., Wang, J., Beck, A., Heinz, A., Ron, D., Lovinger, D.M., Buck, K.J.: Striatal involvement in human alcoholism and alcohol consumption, and withdrawal in animal models. Alcoholism: Clinical and Experimental Research **35**(10), 1739–1748 (2011)
7. Siciliano, C.A., Calipari, E.S., Carlson, V.C.C., Helms, C.M., Lovinger, D.M., Grant, K.A., Jones, S.R.: Voluntary ethanol intake predicts κ-opioid receptor supersensitivity and regionally distinct dopaminergic adaptations in macaques. The Journal of Neuroscience **35**(15), 5959–5968 (2015)
8. Grant, K.A., Leng, X., Green, H.L., Szeliga, K.T., Rogers, L.S., Gonzales, S.W.: Drinking typography established by scheduled induction predicts chronic heavy drinking in a monkey model of ethanol self-administration. Alcoholism: Clinical and Experimental Research **32**(10), 1824–1838 (2008)
9. Guyon, I., Ekusseeff, A.: An introduction to feature extraction. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.) Feature Extraction: Foundations and Applications vol. 207, pp. 1–25. Springer, New York (2008)

Figure 4 Performance of the sets of features using the R^2 score. Pearson correlation coefficient, SVR backward, and SVR forward score the highest among all. With SVR forward and Pearson correlation coefficient the scores are achieved with only three features, while SVR backward with six features.



10. Rivas-Perea, P., Cota-Ruiz, J., Rosiles, J.G.: An algorithm for training a large scale support vector machine for regression based on linear programming and decomposition methods. Pattern Recognition Letters **2013**(34), 439–451 (2013)
11. Daunais, J.B., Davenport, A.T., Helms, C.M., Gonzales, S.W., Hemby, S.E., Friedman, D.P., Farro, J.P., Baker, E.J., Grant, K.A.: Monkey alcohol tissue research resource: banking tissues for alcohol research. Alcoholism: Clinical and Experimental Research **38**(7), 1973–1981 (2014)
12. Li, T.-K., Hewitt, B.G., Grant, B.F.: The alcohol dependence syndrome, 30 years later: a commentary. Addiction **102**(10), 1522–1530 (2007)
13. Dawson, D.A., Goldstein, R.B., Patricia Chou, S., June Ruan, W., Grant, B.F.: Age at first drink and the first incidence of adult-onset dsm-iv alcohol use disorders. Alcoholism: Clinical and Experimental Research **32**(12), 2149–2160 (2008)
14. Vivian, J., Green, H., Young, J., Majerksy, L., Thomas, B., Shively, C., Tobin, J., Nader, M., Grant, K.: Induction and maintenance of ethanol self-administration in cynomolgus monkeys (macaca fascicularis): Long-term characterization of sex and individual differences. Alcoholism: Clinical and Experimental Research **25**(8), 1087–1097 (2001)
15. Salo, A., Baker, E., Grant, K., Rivas-Perea, P.: Identifying future drinkers: Behavioral analysis of monkeys initiating drinking to intoxication is predictive of future drinking classification. Nat Genet **13**, 266–267 (2015)
16. Torkkola, K.: Information-theoretic methods. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.) Feature Extraction: Foundations and Applications vol. 207, pp. 167–185. Springer, New York (2008)
17. Chen, Y.-W., Lin, C.-J.: Combininb svms with various feature selection strategies. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.) Feature Extraction: Foundations and Applications vol. 207, pp. 315–323. Springer, New York (2008)
18. Rakotomamonjy, A.: Variable selection using svm-based criteria. Journal of Machine Learning Research **3**, 1357–1370 (2003)
19. Chang, Y.-W., Lin, C.-J.: Feature ranking using linear svm. Causation and Prediction Challenge Challenges in Machine Learning **2**, 47 (2008)
20. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)
21. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences **180**(10), 2044–2064 (2010)
22. Wong, T.-T.: Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition **48**(9), 2839–2846 (2015)