Investigating the Role of Simpson's Paradox in the Analysis of Top-Ranked Features in High-Dimensional Bioinformatics Datasets [1]

Alex A. Freitas

School of Computing, University of Kent, UK

A.A.Freitas@kent.ac.uk

Abstract

An important problem in bioinformatics consists of identifying the most important features (or predictors), among a large number of features in a given classification dataset. This problem is often addressed by using a machine learning-based feature ranking method to identify a small set of top-ranked predictors (i.e. the most relevant features for classification). The large number of studies in this area have, however, an important limitation: they ignore the possibility that the top-ranked predictors occur in an instance of Simpson's paradox, where the positive or negative association between a predictor and a class variable reverses sign upon conditional on each of the values of a third (confounder) variable. In this work, we review and investigate the role of Simpson's paradox in the analysis of top-ranked predictors in high-dimensional bioinformatics datasets, in order to avoid the potential danger of misinterpreting an association between a predictor and the class variable. We perform computational experiments using four well-known feature ranking methods from the machine learning field and five high-dimensional datasets of ageing-related genes, where the predictors are Gene Ontology terms. The results show that occurrences of Simpson's paradox involving top-ranked predictors are much more common for one of the feature ranking methods.

Keywords:

Gene Ontology, machine learning, classification, feature ranking, ageing-related genes

Introduction

Bioinformatics data analysis problems are often cast as a classification problem from a machine learning perspective [1], [2]. In this scenario, the dataset consists of a set of instances (e.g. genes) to be classified into a set of pre-defined categorical classes. For example, each instance can represent a gene, which is described by a set of features (predictors, or predictive variables) and a special class variable indicating whether or not a gene is involved in some disease. In this context, an important bioinformatics problem consists of ranking the features in terms of how relevant they are for predicting the value of the class variable.

Many feature ranking methods have been proposed for this task, and typically these methods are also used as feature selection methods [3], [4] – i.e., after features are ranked, the top K features (where K is typically a user-defined parameter) can be selected, and then the selected features can be used in two major ways: (a) those features could be used as input by a classification algorithm, which would learn a classification model based on the selected features; or (b) the selected features can be directly analysed by users as a form of discovered knowledge by itself, to try to get more insight about the data and the underlying application domain [5],[6]. In this work we focus on the latter approach (b), i.e., the use of classification algorithms is out of the scope of this paper.

Hence, we focus on feature ranking methods that are independent from any classification algorithm that could be applied later to the data, which is called the "filter approach" in the field of feature selection in machine learning [3], [4]. More precisely, the four feature ranking methods used in this work are Information Gain, Gain Ratio, ReliefF and Correlation-based Feature Selection [3], [4], [7], as discussed in the Datasets and Methods section.

Despite extensive previous research on feature ranking and selection [3], [4], [6], [7], the conventional approach for feature ranking in machine learning has an important limitation: it ignores the possible occurrences of Simpson's paradox, which can lead to potentially wrong conclusions about the data. Simpson's paradox occurs when the direction (or sign) of an association between two variables X and Y at the population level is reversed in all the sub-groups produced by partitioning that population according to

[1] *Briefings in Bioinformatics*, Published online (ahead of print) on 9 Jan. 2019, DOI: 10.1093/bib/bby126
https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby126/5280899

the values of a third variable (a confounder) Z [8], [9]. In other words, the (positive or negative) association between X and Y reverses sign upon conditioning on each of the values of Z.

An example of such paradox is shown in Table 1, adapted from [8]. The top of the table divides patients into two groups: who took and did not took a drug. For each patient group, the third row shows the number of patients in that group who recovered (*Rec*) and the number who did not recover (*not Rec*), and the recovery rate in that group (*Rec* / (*Rec* + *not Rec*)). Note that, at the aggregated level, patients who took the drug have a *higher* recovery rate than patients who did not. By contrast, the bottom part of the table shows the figures when each group of patients is further divided based on Gender. Note that, for both males and females separately, patients who took the drug have a *lower* recovery rate than patients who did not, which is a counter-intuitive reversal of the association between drug and recovery suggested by the aggregated data.

Table 1: Simpson's paradox example

	Drug = yes			**Drug = no**		
	Rec.	not Rec.	Rec. Rate	Rec.	not Rec.	Rec. Rate
Aggregated data	25	25	0.50	20	30	0.40
	Drug = yes			**Drug = no**		
	Rec.	not Rec.	Rec. Rate	Rec.	not Rec.	Rec. Rate
Group: Gender = male	22	13	0.63	9	4	0.69
Group: Gender = female	3	12	0.20	11	26	0.30

Mathematically, this phenomenon of reversal of the sign of an association is not really a paradox, since it can be explained by the fact that the distribution of the third variable Z differs across the two groups associated with the two values of the target variable Y [8], [10], [11]. In the case of Table 1, the Gender distribution differs between the two groups of people who took or did not take the drug. That is, 70% (35/50) of the patients who took the drug are males, whilst only 26% (13/50) of the patients who did not take the drug are males. Since males are more like to recover than females in general (both when taking and when not taking the drug), this gives the misleading impression that the drug is effective, whilst in reality both males and females have a lower recovery rate when taking the drug.

However, this phenomenon is usually called a paradox since it is surprising in general (except for people with statistical training) [9], [12]. In particular, Pearl [9] distinguishes between the mathematically simple reversal of association associated with the paradox and the paradox itself – described as (quoting verbatim): "*a psychological phenomenon that evokes surprise and disbelief*". The main explanation for this surprise among non-statistically trained users seems to be that, in the absence of data involving confounder variables, people tend to implicitly interpret the statistical association between two variables at the aggregated level as a causal association [9], [12], which is in conflict with the finer-grained data at the group level, considering the confounder variable. A misinterpretation of an association at the aggregated level can lead to wrong decisions and policies [12]; and in the field of bioinformatics, detecting occurrences of the paradox can lead to better, more informed biological conclusions [13], [14], [15]. In addition, it should be noted that Simpson's paradox seems to be more common than normally thought [12], [16], [17].

In this context, the goal of this paper is not to propose a new machine learning or bioinformatics method, but rather to review and investigate the effect of Simpson's paradox occurrences in the analysis of top-ranked features in high-dimensional classification datasets in the field of bioinformatics, in order to avoid the potential danger of misinterpreting an association between a predictor and the class variable. To the best of our knowledge the effect of Simpson's paradox in machine learning-based feature ranking, in order to identify the most relevant features for classification, has not been studied yet.

To support our investigation, we perform computational experiments to investigate the occurrence of Simpson's paradox involving the top-ranked features identified by four well-known feature ranking methods from the field of machine learning, namely ReliefF [18], [19], Information Gain, Gain Ratio (a popular variation of Information Gain) [20], and Correlation-based Feature Selection [3], [7]. The experiments involve five bioinformatics classification datasets, four of them obtained from [21], and the other from [22]. In general, in these datasets the instances to be classified are genes, the class variable indicates whether a gene has some ageing-related effect on an organism, and the predictive features represent Gene Ontology (GO) terms. We focus on this type of features because GO terms are very popular in bioinformatics.

In addition, detecting Simpson's paradox occurrences involving GO features is particularly important for two reasons. First, GO term annotations are a type of observational data, where it is much harder to control

for confounders than data resulting from (randomized) controlled experiments [23] (see the Background section). Second, datasets using GO terms as features are typically high-dimensional datasets (with hundreds or thousands of features), which increases the opportunity for occurrences of Simpson's paradox, due to the large number of potential confounder variables.

The remainder of this paper is organized as follows. The Background section presents a definition of Simpson's paradox and discusses relevant related concepts. The Datasets and Methods section describes the five datasets of ageing-related genes or proteins (containing GO terms as features) used in our experiments, presents an overview of the four feature ranking methods used in our experiments, and explains how the occurrences of Simpson's paradox were detected in the used datasets. The Computational Results section discusses how the number of occurrences of the paradox varies across different combinations of feature ranking methods and datasets. Finally, the Conclusions section summarizes the results of the experiments, with an associated recommendation for checking whether the paradox occurs when analyzing the most relevant features (predictors) identified by a feature ranking method; and also suggests some future work, to extend the analysis presented in this paper.

Background on Simpson's Paradox

We use the following notation. Let a population of objects (in this work, genes) be partitioned into two mutually exclusive and exhaustive groups, denoted G_1 and G_2, according to the value of a given binary variable X, taking values x_1 and x_2. That is, all objects in G_1 (G_2) take the value x_1 (x_2) for X. Let Y be a binary target variable ($Y \neq X$), taking the value y_{int} when an event of interest has occurred, and $y_{non\text{-}int}$ otherwise. Let $p(y_{int} \mid x_1)$ and $p(y_{int} \mid x_2)$ be the conditional probability of observing the event of interest y_{int} in an object given that the object belongs to G_1 or G_2, respectively. That is, $p(y_{int} \mid G_1)$ is the ratio of the number of objects in G_1 where the target variable Y takes the value y_{int} over the total number of objects in G_1; and $p(y_{int} \mid G_2)$ is analogously defined.

Let us now consider the scenario where each of the two groups, G_1 and G_2, is further partitioned, according to the values of another categorical (discrete) variable Z ($Z \neq X, Y$). Here we assume Z is a binary variable (more general cases are mentioned below). This produces four groups of objects, where $G_{i,j}$ – for *i* in {1,2}, *j* in {1,2} – denotes the group where all objects take values x_i and z_j for X and Z respectively. Let $p(y_{int} \mid x_i,z_j)$ be the conditional probability of observing the event of interest y_{int} for an object given that it belongs to $G_{i,j}$. Finally, Simpson's paradox occurs when one of the following dual conditions is met:

If $Pr(y_{int} \mid x_1) > Pr(y_{int} \mid x_2)$ and $Pr(y_{int} \mid x_1,z_j) < Pr(y_{int} \mid x_2,z_j)$ – for *j* in {1,2}; or

If $Pr(y_{int} \mid x_1) < Pr(y_{int} \mid x_2)$ and $Pr(y_{int} \mid x_1,z_j) > Pr(y_{int} \mid x_2,z_j)$ – for *j* in {1,2}

That is, the paradox occurs [8], [9] if the probability of the event of interest y_{int} *increases* (*decreases*) from G_1 (where $X = x_1$) to G_2 (where $X = x_2$) but, surprisingly, the probability of y_{int} *decreases* (*increases*) from G_1 to G_2 both for objects with $Z = z_1$ and for objects with $Z = z_2$. Note that the above equations use strict inequalities ("<" or ">") in both sides of an "and" statement, whereas other studies may use a little less strict definition of the paradox – e.g. in [16], in the second term of an "and" statement the operators used are instead "≤" or "≥".

The previous mathematical inequalities defining an occurrence of Simpson's paradox refer to the scenario where the Z variable is binary, which is by far the most common scenario addressed in the literature, and it is also the case in all datasets used in this work (see the Datasets and Methods section). However, for a more general definition of Simpson's paradox, the above inequalities can be easily generalized to the case where Z takes more than two categorical values, and it is also possible to detect analogous occurrences of the paradox for continuous variables [12], [14].

In most of the literature on Simpson's paradox, particularly the literature on causality and statistics, the variables X, Y and Z are usually called the treatment, outcome and confounder variables [9], [10]. However, the "treatment" term implies some kind of controlled experiment, instead of observational data (the focus of this paper), and the terms "treatment" and "outcome" ignore the nature of the classification task being addressed in this paper. Hence, from hereafter we will refer to variables X, Y and Z as the predictor, class and confounder variables, to be more consistent with our context of the classification task of machine learning – the term "confounder" has also been used in the context of classification [17].

Simpson's paradox has been extensively studied from a statistical perspective in the field of causal data analysis [8], where the goal is to identify causal relationships between variables. Actually, the paradox is currently considered a "resolved" problem in causality theory [9], but the application of such theory (based on the "Do calculus" [8]) is normally suitable for the analysis of experimental data, resulting from

randomized controlled experiments – where we can observe the effect of varying a variable whilst controlling for confounders. In practice, many (maybe most) datasets in bioinformatics and biomedicine are instead observational data, where it is much harder to control for confounders – particularly in high-dimensional datasets, where there is a large number of potential confounders. For instance, to study the effect of a drug on patients, it would not be feasible to control for all potential confounders in randomized clinical trials, since there are too many potential confounders, and it would be unethical to try to control for many of them – e.g., it would be unethical to randomly assign people to a group of smokers or non-smokers.

Turning to machine learning, Simpson's paradox has been, by comparison, much less studied in the context of the classification task, a kind of supervised machine learning, where the goal is to discover predictive relationships between features (predictors) and the class variable. The main exceptions are the two studies discussed next.

[17] has shown that Simpson's paradox can be very frequent in a text mining dataset and that, by using a technique from causal data analysis to control for a confounding variable, predictive accuracy in a text mining classification task can be improved. In addition, in a classification task involving the prediction of regulatory gene interactions, [15] reported a kind of Simpson's paradox in the results obtained by several different types of classification algorithms – including Support Vector Machines, logistic regression and decision tree algorithms.

Datasets and Methods

The experiments used five high-dimensional bioinformatics datasets previously used in research on the biology of ageing. The first four datasets were used in [21], and each of them has data about a different model organism: *S. cerevisae* (yeast), *C. elegans* (worm), *D. melanogaster* (fly) and *M. musculus* (mouse). Each of these datasets involves a classification problem where each instance (object to be classified) is a gene, the binary class variable (to be predicted) indicates whether a gene has a pro-longevity or anti-longevity effect in an organism, and the features (predictors) represent Gene Ontology (GO) terms describing the biological process(es) in which the gene is involved. All features are binary variables, and each feature indicates whether or not a gene is annotated with a certain GO term. The special event of interest (Y_{int}) was defined as the less frequent class label out of the two class labels, which is pro-longevity for the yeast and worm datasets, and anti-longevity for the fly and mouse datasets. The rationale for this choice is that the least frequent class label in each dataset is in general harder to predict (since it has fewer training instances), so its prediction tends to be more interesting than the prediction of the most frequent class label.

The fifth dataset was used in [22]. This dataset also involves a classification problem where the instances to be classified are genes or proteins, and the binary class variable indicates whether or not a gene is ageing-related. The special event of interest (Y_{int}) was defined as the least frequent class label, namely 'ageing-related'. The vast majority of features (296) are GO terms, whilst the remaining 14 features are numerical (real-valued or integer-valued), representing protein-protein-interaction network properties. In this work we used only the 296 GO term features, since the definition of Simpson's paradox used here refers to categorical attributes. In addition, the focus on GO terms makes the experiments more coherent, since the other four datasets also use GO terms as features, and it is the large number of GO terms that is responsible for the large dimensionality of this dataset (which is a typical problem in datasets using GO terms as features).

Table 2 shows the main characteristics of the used datasets – see [21] for a more detailed description of the first four datasets, and [22] for more details about the fifth dataset.

Table 2: Main characteristics of the datasets used in the experiments

Dataset (organism)	No. of instances (genes)	No. of features (GO terms)	Class label of interest & relative frequency
Yeast	248	698	pro-longev (16.1%)
Worm	478	764	pro-longev (39.7%)
Fly	119	585	anti-longev (32.8%)
Mouse	89	886	anti-longev (29.2%)
Human	20,183	296	ageing (1.5%)

To rank the features in each dataset, we performed experiments with four feature ranking methods (briefly described below). These methods follow the filter (rather than wrapper) approach for ranking features in a data preprocessing phase of the classification task of machine learning [3], [4]. That is, these methods evaluate the quality of each feature independently from the classification algorithm to be applied to

the data. This gives them two advantages, as follows. First, they are in general much faster and more scalable to large datasets than typical wrapper methods, which need to run the target classification algorithm many times. Second, since filters evaluate the intrinsic predictive power of features in general, regardless of any classification algorithm, the top-ranked features can be interpreted as the most relevant features in a more generic way, by comparison with the features selected by wrappers (which are features customized to a particular classification algorithm). As mentioned earlier, in this work we are not using any classification algorithm; so the filter approach is more appropriate for our experiments. Results of classification algorithms applied to the first four datasets used in this work can be found in [21], [24], [25], whilst results of classification algorithms applied to the fifth dataset can be found in [22] (Table 5 in that article).

The four feature ranking methods used in our experiments are well-known methods in machine learning and bioinformatics, namely ranking based on the Information Gain, Gain Ratio, ReliefF and Correlation-based Feature Selection methods [4], [7]. The Information Gain measures the predictive power of a feature by computing the reduction in the entropy (a measure of disorder) of the class distribution that is obtained when the dataset is partitioned based on the values of the feature being evaluated, by comparison with the entropy of the class distribution in the dataset as a whole (ignoring the values of that feature) [20].

Gain Ratio is a popular variation of information gain, where essentially a feature's information gain (its reduction of the class entropy) is divided by the feature's own entropy (regardless of the class distribution) [20]. Hence, the Gain Ratio measures the proportion of the entropy reduction associated with a feature that is useful for classification.

ReliefF essentially estimates the predictive power of a feature based on how well that feature's values distinguish between instances that are near to each other but belong to different classes [18], [19]. The key ideas underlying ReliefF are that a feature's weight is *increased* to the extent that instances of *different* classes which are close to each other in the data space (nearest misses) have different values for that feature; and conversely, a feature's weight is *decreased* to the extent that instances of *the same class* which are close to each other in the data space (nearest hits) have different values for that feature.

Correlation-based Feature Selection (CFS) performs a search in the space of candidate feature subsets, to try to select the best possible feature subset according to an evaluation function. This function rewards the selection of features that have a high correlation with the class variable and have a low correlation with other selected features. Note that, unlike the previously discussed feature selection methods, CFS is a multi-variate method that not only tries to maximize the relevance (correlation with class) of selected features, but also tries to minimize the redundancy among the selected features.

For each of these feature ranking methods, and for each dataset, we select a predefined number of that method's top-ranked features for our experiments with Simpson's paradox detection. We used the implementation of these feature ranking methods in the well-known WEKA data mining tool.

More precisely, our experimental methodology is as follows. For each of the top 15 features in the rankings produced by each feature ranking method, we run an algorithm that finds the list of all occurrences of Simpson's paradox (if any) where that feature is the predictor X and the event of interest Y_{int} is one of the class labels in the original dataset – see the Background section. That paradox-detection algorithm simply checks if one of the dual conditions described in the Background section are met, for each pair of a feature (predictor) X and a confounder Z such that $X \neq Z$, where X can be any of the top-15 features (GO terms) as identified by a feature ranking method, and Z can be any of the GO terms in the dataset. Details of such algorithm (including a pseudocode) can be found in [16], and the algorithm used in this work is a small variation of the one described in our previous work in [16], since in this work we modified the algorithm to use only strict inequalities in the definition of the paradox, as explained in the Background section. The code of the algorithm for detecting Simpson's paradox occurrences will be freely available on the web after the article is published.

Recall that in our experiments the event of interest (Y_{int}) is the class label 'pro-longevity' for the yeast and worm datasets; whilst it is the label 'anti-longevity' for the fly and mouse datasets – as shown in Table 2. In addition, in the fifth dataset, Y_{int} is the class label "ageing-related".

In addition, we use a Chi-squared (χ^2) test of independence between two categorical variables [26] to investigate whether the confounder (Z) variables in detected occurrences of the paradox are statistically significant, in terms of their association with the class variable (Y). The null hypothesis for the test is that Z and Y are independent (no association). We reject the null hypothesis, and conclude that Z is significantly associated with Y, if the value of the χ^2 statistic is greater than the critical value, which is determined based on the specified significance level (α) and the calculated degrees of freedom. We use the conventional significance level of 5% ($\alpha = 0.05$). The number of degrees of freedom is calculated as $(r - 1)(c - 1)$, where r and c are the numbers of rows and columns in the corresponding contingency table. In our case, $r = c = 2$, since Z and Y are binary variables, so we have one degree of freedom. The test is used for each unique combination of GO term acting as the confounder Z and dataset. Note that a GO term can act as a confounder

Z in more than one paradox occurrences for the same dataset, but those occurrences are associated with the same χ^2 value.

Computational Results

The results of searching for all occurrences of Simpson's paradox in the aforementioned five datasets of ageing-related genes are shown in Table 3. The columns in this table show: (a) the method used to rank the features (GO terms): Information Gain (IG) or ReliefF (RelF) – it turned out that no occurrence of Simpson's paradox was found for Gain Ratio and Correlation-based Feature Selection when using the methodology defined in the Datasets and Methods section; (b) the organism whose genes are being studied (defining the dataset); (c) the rank of the feature (GO term) acting as predictor X in a paradox instance (the lower the rank, the more relevant the feature); (d) the Id and name of that GO term; (e) the number of paradox instances where that GO term has the role of predictor X.

In addition, Supplementary Files 1 through 5 contain tables showing the 15 top-ranked features (GO terms) identified by each feature ranking method, with each file referring to a different dataset.

Table 3: GO terms occurring in the role of predictor X in at least one instance of Simpson's paradox, among the top-15 GO terms according to the rankings by IG and ReliefF, for each organism (dataset).

Rank meth.	Organism (dataset)	Rank No.	GO term Id and name	No. of Parad.
IG	Mouse	8	GO:0071375 Cell. response to peptide hormone stimulus	1
		9	GO:1901653 Cellular response to peptide	1
RelF	Yeast	11	GO:0043170 Macromolecule metabolic process	8
		12	GO:0044260 Cellular macromolecule metabolic process	4
		14	GO:0044238 Primary metabolic process	5
RelF	Worm	4	GO:0009058 Biosynthetic process	1
		6	GO:0009059 Macromolecule biosynthetic process	1
		8	GO:0044260 Cellular macromolecule metabolic process	5
		11	GO:0044237 Cellular metabolic process	52
		13	GO:0008152 Metabolic process	3
		14	GO:0071704 Organic substance metabolic process	3
RelF	Fly	14	GO:1901700 Response to oxygen-containing compound	7
RelF	Mouse	4	GO:1901576 Organic substance biosynthetic process	6
		5	GO:0031326 Regulation of cellular biosynthetic process	31
		6	GO:0009889 Regulation of biosynthetic process	31
		7	GO:0044249 Cellular biosynthetic process	3
		8	GO:0009058 Biosynthetic process	3
		9	GO:0006807 Nitrogen compound metabolic process	6
		10	GO:1901362 Organic cyclic compound biosynt. process	11
		13	GO:0018130 Heterocycle biosynthetic process	5
		14	GO:0034654 Nucleobase-contain. comp. biosynt. proc.	3
		15	GO:0010468 Regulation of gene expression	3
RelF	Human	5	GO:0043167 Ion binding	1
		15	GO:0043169 Cation binding	2

When using the IG method for feature ranking, as shown in Table 3, only two instances of Simpson's paradox were found with the property that the predictor X is one of those 15 most important features – both these paradox instances occurred in the Mouse dataset. By contrast, Simpson's paradox instances were found much more often when using the ReliefF method for feature ranking. More precisely, as shown in Table 3, when using ReliefF, the number of selected features in the role of predictor X was 3, 6, 1, 10 and 2, for the yeast, worm, fly, mouse and human datasets, respectively. Note that the vast majority of GO terms in Table 3 take the role of predictor X in multiple instances of Simpson's paradox (with different GO terms taking the role of the confounder Z), and a few predictors occur in many (around 30-50) instances of the paradox .

The fact that the top-15 features in the ranking by ReliefF lead to many more instances of Simpson's

paradox than the top-15 features in the ranking by IG and Gain Ratio is interesting and somewhat unexpected, since ReliefF is a more sophisticated method. Actually, ReliefF determines the relevance of a feature in a partly multi-variate way, because, in order to evaluate the relevance of a feature, ReliefF has to find many nearest neighbor instances, and this process involves measuring the distance between instances using all features. In this sense, when evaluating the relevance of a feature, ReliefF takes into account local contextual information (nearest hit and nearest miss instances), which in principle allows ReliefF to evaluate well the relevance of features in datasets with strong dependencies between features [19]. However, the notion of nearest instances is not so effective in high-dimensional data.

By contrast, IG and Gain Ratio, like most feature ranking methods, determine the relevance of a feature in a completely univariate way, because they simply measure the degree of association between each feature and the class variable, ignoring dependencies between features. That is, they do not consider any local context, they simply provide a "global" perspective of each feature's relevance.

ReliefF combines the above kind of local contextual information with the global perspective of measuring association between each feature and the class [19], which in theory should increase its effectiveness for selecting a subset of features to be used as input by classification algorithms. However, for the purposes of simply measuring the degree of relevance of a feature as a predictor by itself for classification, which is the problem being investigated in this paper, the local contextual information seems to produce a feature ranking where several of the top-ranked features occur as the predictor X in an instance of Simpson's paradox – which can potentially lead to misinterpretation of the association between the feature and the class variable, as discussed earlier.

Table 4: Simpson's Paradox occurrence in the Fly dataset. The class label of interest (y_{int}) is 'anti-longevity' (anti_long). The predictor X is GO:1901700 ('Response to oxygen-containing compound'), ranked 14th by ReliefF. The confounder variable Z is GO:0040007 ('Growth')

	GO:1901700 = yes			**GO:1901700 = no**		
	anti-long	Total	anti-long %	anti-long	Total	anti-long %
Aggregated data	10	28	35.7	29	91	31.9
	GO:1901700 = yes			**GO:1901700 = no**		
	anti-long	Total	anti-long %	anti-long	Total	anti-long %
GO:0040007 = yes	4	7	57.1	4	6	66.7
GO:0040007 = no	6	21	28.6	25	85	29.4

An example of an instance of Simpson's paradox is shown in detail in Table 4. In this paradox instance, the predictor X is GO:1901700 ('Response oxygen-containing compound'), which was ranked 14th by the ReliefF feature selection method in the Fly dataset; whereas the confounder Z is GO:0040007 ('Growth'). The class label of interest (y_{int}) is 'anti-longevity'. In this table, the conditions defining the occurrence of the paradox were satisfied as follows:

(a) The probability of a gene having class label anti-longevity is *greater* for genes annotated with the term GO:1901700 (35.7%) than for genes not annotated with that term (31.9%). That is: $p(y_{int} \mid x_1) > p(y_{int} \mid x_2)$, where x_1 denotes "GO:1901700 = yes" and x_2 denotes "GO:1901700 = no"; and

(b) In each of the two groups of genes defined by the two values of the confounder Z, where z_1 denotes "GO:0040007 = yes" and z_2 denotes "GO:0040007 = no", the probability of a gene having class label anti-longevity is *smaller* for genes annotated with the term GO:1901700 than for genes not annotated with that term. I.e.: $p(y_{int} \mid x_1,z_1) = 57.1\% < p(y_{int} \mid x_2,z_1) = 66.7\%$; and $p(y_{int} \mid x_1,z_2) = 28.6\% < p(y_{int} \mid x_2,z_2) = 29.4\%$.

Hence, if we simply relied on the result of the feature ranking and considered only the association between the predictor GO:1901700 and the class variable, ignoring the confounder GO:0040007, we would conclude that the GO term GO:1901700 has a *positive* association with the class label "anti-longevity" – i.e., a gene annotated with this term is more likely to have an anti-longevity effect on flies than a gene not annotated with this term. In reality, however, that association is a *negative* one for both genes with and genes without the GO:0040007 term annotation. This illustrates the potentially misleading conclusions that can be drawn by analyzing the association between a top-ranked feature and a class label.

This paradox is explained by two facts: (a) 25% (7/28) of the genes annotated with the term GO:1901700 are also annotated with GO:0040007, whilst only 6.6% (6/91) of the genes *not* annotated with GO:1901700 are annotated with GO:0040007; and (b) the GO:0040007 term annotation has a strong positive association

with "anti-longevity" (y_{int}) in general, i.e., both for genes annotated with and genes not annotated with the term GO:1901700 – the relevant probabilities are: $Pr(y_{int}$ | 'GO:0040007 = yes', 'GO1901700 = yes') = 0.571 and $Pr(y_{int}$ | 'GO:0040007 = yes', 'GO1901700 = no') = 0.667. Both these probability values are substantially larger than the marginal probability of y_{int}, $Pr(y_{int}) = 0.328$, confirming GO:0040007's relevance.

Actually, although the term GO:0040007 ('Growth') was not among the top-ranked terms in the ranking produced by ReliefF, there are several studies based on machine learning methods that showed the relevance of GO terms very related to 'Growth' for ageing, in both fly and other organisms, as follows.

In [22], three GO terms related to growth were among the top 15 GO terms most relevant for human ageing, as estimated by a machine learning algorithm (gradient boosted trees), namely GO terms: GO:0043567 ('Regulation of insulin-like growth factor receptor signaling pathway'), GO:0019838 ('Growth factor binding'), GO:0040008 ('Regulation of growth'). Also, in [27], in a list of top 30 genes predicted to be involved in *C. elegans* (worm) longevity, 9 genes have a functional description involving growth, namely 5 genes with description 'Positive regulation of growth rate', and 4 genes with description 'Positive regulation of multicellular organismal growth'. Furthermore, in [21], the terms GO:0030307 ('Positive regulation of cell growth'), GO:0040018 ('Positive regulation of multicellular organism growth'), and GO:0040007 ('Growth') were ranked 20th, 21th and 23rd, respectively, in terms of relevance for ageing in flies.

As another example of Simpson's paradox occurrences with top-ranked GO terms in Table 3, we briefly consider now, in the results for the Worm dataset, the GO term GO:0044260 ('Cellular macromolecule metabolic process'), which was ranked 8th by ReliefF. This GO term occurred as the predictor X in 5 occurrences of the paradox, in which the GO terms occurring as the confounder Z were all terms related to transport, namely: GO:0055085 ('Transmembrane transport'), GO:0015672 ('Monovalent inorganic cation transport'), GO:0034220 ('Ion transmembrane transport'), GO:0006812 ('Cation transport'), and GO:0006811 ('Ion transport'). It is interesting to note that, out of these 5 transport-related GO terms in the role of the confounder Z, two are also highly ranked by ReliefF, namely GO:0055085 and GO:0034220, with ranks 15 and 19, respectively.

Finally, we have also investigated to what extent the confounder (Z) GO terms in the detected occurrences of the paradox are statistically significantly associated with the class variable (Y). For this analysis, we used the Chi-squared (χ^2) test of independence between two categorical variables, with the null hypothesis that Z and Y are independent (no association). As described in more detail at the end of the Datasets and Methods section, we reject the null hypothesis, and conclude that Z is significantly associated with Y, if the value of the χ^2 statistic is greater than the critical value at the significance level of 5% ($\alpha = 0.05$) and one degree of freedom. The test is used for each unique combination of GO term acting as the confounder Z and dataset, which led to 126 uses of the Chi-squared test of independence across all datasets; and 53 (about 42%) of those occurrences have a statistically significant confounder GO term, for the aforementioned settings of the test.

Conclusions

A large body of literature on using feature ranking methods in bioinformatics and classification research evaluates the relevance of a feature in a way that completely ignores the possibility of that feature having the role of the predictor variable in an instance of Simpson's paradox. To investigate this issue, we performed experiments with four well-known feature ranking methods and five datasets of ageing-related genes, where the features are Gene Ontology (GO) terms. Our results have shown that, for several combinations of a feature ranking method and a dataset being analyzed, even some of the top-ranked (most important) features can occur in the role of the predictor variable in an instance of the paradox. This phenomenon was observed much more often in the experiments with ReliefF, which is a popular feature ranking method in bioinformatics and machine learning. The occurrence of such instances of the paradox is important information in the analysis of the top-ranked predictors, since in such cases the direction of the association between the predictor and the class variable is reversed for every value of a confounder variable – which would probably go unnoticed if the paradox occurrence was not detected.

It should be noted that, although the detection of all occurrences of Simpson's paradox in a high-dimensional dataset would be very computationally expensive if any feature could act as either the predictor or the confounder in a paradox instance, this is not a problem in general for the feature ranking scenario considered in this paper. In this scenario, only a relatively small set of top-ranked features can act as the predictor variable, which drastically reduces the number of predictor-confounder pairs to be considered when searching for paradox occurrences. Hence, given the algorithmic simplicity of Simpson's paradox detection, in this scenario it is recommended to check, for each of the top-ranked features, whether that feature acts in the role of the predictor variable in an instance of Simpson's paradox, in order to avoid potential misinterpretations of the association between that predictor and the class variable.

Future work could consist of extending the analysis presented in this paper to other types of feature ranking methods (e.g., using the wrapper approach) and other types of bioinformatics datasets (using features different from the GO terms used here).

Key Points:

In general, feature ranking methods for the classification task (a type of supervised machine learning) ignore the possibility of occurrence of Simpson's paradox involving top-ranked features.

If a top-ranked feature occurs in an instance of Simpson's paradox, the direction of its association (positive or negative association) with the class label of interest may be misleading.

Simpson's paradox occurs more often than intuitively thought among top-ranked features.

One can better interpret the association between a top-ranked feature (as a predictor) and the class variable by automatically checking for occurrences of Simpson's paradox involving that predictor.

Short Biography

Alex A. Freitas is a Professor of Computational Intelligence at the University of Kent, UK. He has a PhD in Computer Science (1997) and an MPhil (master's degree) in Biological Sciences (2011). His main research interests are machine learning, bioinformatics and ageing biology.

References

1. Libbrecht MW and Noble WS. Machine learning applications in genetics and genomics. *Nature Reviews – Genetics 2015*; 16: 321-332.
2. Camacho DM, Collins KM, Powers RK et al. Next-generation machine learning for biological networks. Cell 2018; 173: 12 pages.
3. Li J, Cheng K, Wang S et al. Feature Selection: a data perspective. ACM Computing Surveys 2017; 50(6): Article 94, 45 pages.
4. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods 2016; 111: 21-31.
5. Guyon I and Elisseeff A. An introduction to feature extraction. In: Guyon, I Gunn S, Nikravesh M et al. (Eds.) Feature Extraction: foundations and applications. Berlin: Springer, 2006, 1-14.
6. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23(19): 2507-2517.
7. Hira ZM and Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics, 2015; Article Id 198363, 13 pages.
8. Pearl J. Causality: models, reasoning and inference. Cambridge, UK: Cambridge University Press, 2000.
9. Pearl J. Comment: understanding Simpson's paradox. The American Statistician 2014; 68(1): 8-13.
10. Norton HJ and Divine G. Simpson's paradox … and how to avoid it. Significance 2015; 40-43.
11. Salimi B, Gehrke J, Suciu JD. Bias in OLAP queries: detection, explanation and removal. In: Proceedings of the 2018 International Conference on Management of Data (SIGMOD'18), p. 1021-1035. ACM Press.
12. Kievit RA, Frankenhuis WE, Waldorp LJ, et al. Simpson's paradox in psychological science: a practical guide. Frontiers in Psychology 2013; 4: Article 513, 14 pages.
13. Bansal S and Mittal A. A statistical anomaly indicates symbiotic origins of eukaryotic membranes. Molecular Biology of the Cell 2015; 26: 1238-1248.
14. Brimacombe M. Genomic aggregation effects and Simpson's paradox. Open Access Medical Statistics 2014; 4: 1-6.
15. Petri T, Altmann S, Geistlinger L, et al. Addressing false discoveries in network inference. Bioinformatics 2015; 31(17): 2836-2843.

16. Fabris CC and Freitas AA. Discovering surprising patterns by detecting occurrences of Simpson's paradox. In: Research and Development in Intelligent Systems XVI (Proceedings of ES99, The 19th SGES Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence), p. 148-160. Springer.
17. Landeiro V and Culotta A. Robust text classification in the presence of confounding bias. In: Proc. Thirtieth AAAI Conf. on Artificial Intelligence (AAAI-16), 186-193.
18. Kononenko I, Simec E, Robnik-Sikonja M. Overcoming the myopia of inductive learning algorithms with ReliefF. Applied Intelligence 1997; 7: 39-55.
19. Robnik-Sikonja M and Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 2003; 53: 23-69.
20. Quinlan JR. C4.5: Programs for Machine Learning. Palo Alto, CA. Morgan Kaufmann, 1993.
21. Wan C, Freitas AA, de Magalhaes AA. Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2015; 12(2): 262-275.
22. Kerepesi C, Daroczy B, Sturm A, et al. Prediction and characterization of human ageing-related proteins by using machine learning. Scientific Reports 2018; 8: 4094. 13 pages.
23. Gaudet P and Dessimoz C. Gene Ontology: pitfalls, biases and remedies. In: Dessimoz C and Skunca N (Eds.) The Gene Ontology Handbook. Springer, 2017, 189-205.
24. Wan C and Freitas AA. An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features. Artificial Intelligence Review 2017; 40 pages. DOI: 10.1007/s10462-017-9541-y.
25. da Silva PN, Plastino A, Freitas AA. A novel genetic algorithm for feature selection in hierarchical feature spaces. In: Proceedings of the SIAM International Conference on Data Mining (SDM18), 201 8; 738-746. SIAM.
26. DeGroot MH and Schervish MJ. Probability and Statistics, 3rd Ed. Addison-Wesley, 2002.
27. Li YH, Dong MQ, Guo Z. Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*. Mechanisms of Ageing and Development 2010; 131: 700-709.