Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Improving the accuracy of detecting steroid abuse in cattle by pairwise learning of serum samples

Xiao-Lei Xia *

School of Mechanical and Electrical Engineering, Jiaxing University, Jiaxing, PR China

ARTICLE INFO

Article history:
Received 22 December 2016
Received in revised form
11 May 2017
Accepted 29 May 2017
Available online 7 June 2017

Keywords:
Support vector machine
Steroid abuse
Screening methods

ABSTRACT

Issues surrounding the misuse of illegal drugs in animals destined for food production have be an enormous challenge to regulatory authorities charged with enforcing their control. A method has been proposed recently which compared the bovine blood biochemistry profiles between control and treated animals, using the support vector machine (SVM) as the classification tool. Whether an animal has been treated is determined by the classification outcome of the SVM on an individual serum sample taken off the animal. However, the acquisition time of the serum sample is essential in the classification performance of the SVM. Thus, the paper proposed to collect and analyze a pair of samples, in order to obtain at least one sample whose acquisition time resulted in an SVM with the highest sensitivity. The power of the strategy in improving sensitivity was theoretically proven to be up to 0.25 and empirically confirmed on a bovine blood biochemistry data. Furthermore, classification rules of the SVM were proposed to be adapted to meet higher levels of demands on sensitivity. Schemes were described which optimized the time apart between the collection of the two samples and the impact of the proposed strategy on specificity was also investigated.

1. Introduction

Despite the European-wide ban on the use of growth-enhancing steroids in meat production [16], they are still present in the black market due to lucrativeness of the practice. Research efforts have been ongoing for over two decades with the development of methods to screen for illicit hormone use.

The advancement of immunoassay and mass spectrometry (MS) technologies prompted the advent of the following framework of steroid control. First, screening was performed which can employ the immunoassay technology [15,5]. Upon a positive outcome, the mass spectrometry (MS) technology was applied for confirmatory analysis [17,10,7,1,11]. The MS technology, including both gas chromatography MS (GC–MS) and liquid chromatography MS (LC–MS), can detect a wide range of molecules and has been the predominant analytical tool. Despite its ability to pinpoint multiple steroids with high sensitivity, screening tools based on the MS technology were laborious and expensive. Consequently, the number of samples that was allowed for steroid analysis was small. Meanwhile, with the injection of very low-dose, liquid-based ''cocktails'' of multiple steroids being the trend in the practice

* *Corresponding author at*: School of Mechanical and Electrical Engineering, Jiaxing University, Jiaxing 314001, PR China.
E-mail address: xxia01@qub.ac.uk

http://dx.doi.org/10.1016/j.bbe.2017.05.009

of steroid administration [19,13], the MS technology can hardly detect the steroids reliably due to their low concentrations.

Indirect methods have thus been proposed to ease these bottlenecks that the MS-based analytical methods suffered from. Indirect methods evaluated, quantitatively, variations of biological matrices (tissues or biological fluids) taken from animals with exposure to administrated steroids, followed by comparisons of the biological matrices between normal animals and those that have been experimentally administrated with known steroids [8,9,6,14,12,3]. The comparative analysis required those analytes that exhibited significantly different concentrations before and after steroid treatment be identified. These analytess were generally referred to as ''biomarkers''. The ''omic'' technology provided a solution to biomarker discovery, which facilitated comprehensive and simultaneous characterization of analytes in biological matrices. These 'analytes'' ranged from genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics). In fact, metabolomics has demonstrated to be promising in detecting steroid abuse in animals [8].

As opposed to measurement of a large number of analytes by the use of the ''omic'' technology, an alternative approach has been measurement of analytes established as biomarkers or potential biomarkers. Cunningham et al. adopted this methodology and measured 20 standard serum biochemical analytes consisting of proteins, metabolites, enzymes and ions [6]. The blood samples were taken on 14 different days pre-and-post the drug treatment, from a herd of 10 cattle among which only 5 were treated. Subjecting the resultant dataset to the Support Vector Machines (SVM) [2,18] which has been a state-of-the-art machine learning algorithm, their proposed screening strategy identified treated animals with sensitivity and specificity both being over 90%. More specifically, in order to determine whether an animal has been treated, the serum sample was taken and the 20 analytes were measured. The 20 measurement was fed into the SVM which returned a verdict after sophisticated mathematical computations.

Interestingly, although the sampling time was acknowledged to a crucial factor in the accuracy of determining the status of an animal [6], the proposed SVM classification framework took no account of the time factor. This study investigated the exploitation of the time factor for further enhancement of accuracy in revealing the true status of an animal.

2. Identification of steroid-treated cattle by blood-chemistry profiling and SVMs

2.1. Methodology revisited

For the identification of steroid-treated cattle, a group of male animals (steers) and female animals (heifers) were housed and fed. Half of the steers and heifers were treated on a particular day which was referred as to day 0. In addition to day 0, blood samples of these animals were also collected on the other 13 different dates which were respectively 1, 4, 7, 11, 14, 17, 21, 25, 28, 31, 35, 39, and 42 days post treatment. For each sample, 20 standard serum biochemical analytes were measured. An SVM classifier was trained on data generated for steers and heifers respectively. The methodology of the SVM classification for steers and heifers were identical and thus the SVM analysis of the steer dataset was elaborated hereafter. The training dataset for the SVM consisted of 75 control samples and 65 treated samples. Each sample was, mathematically, a 20-dimensional vector $\mathbf{x}_i$ ($i = 1, 2, \ldots, 140$) that corresponds to the 20 analytes. Each control sample was assigned a target label of -1 and each treated one of +1. Mathematical operations led to a function that can be used for determination of the type of a sample and was in the form of:

$$f(\mathbf{z}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{z}) + b \quad (1)$$

where $\mathbf{z}$ is a testing sample. $K(\mathbf{x}_i, \mathbf{z})$ is the kernel function which often adopts the Gaussian Radial Basis Function (RBF):

$$K(\mathbf{x}_i, \mathbf{z}) = \exp(-\lambda \|\mathbf{x}_i - \mathbf{z}\|^2) \quad (2)$$

where λ is the width parameter. Each α_i ($i = 1, 2, \ldots, 140$) is a constant associated with the training sample $\mathbf{x}_i$. b is the bias term for the decision function.

By taking the sign of the decision value $f(\mathbf{z})$, the testing sample obtained a class label of either +1 or -1 and was consequently put into the control or the treated group.

2.2. Scrutiny of SVM decision values

The decision values produced by the SVM ranges between $(-\infty, +\infty)$. For a treated sample $\mathbf{z}$ to be correctly identified as from the treated group, its decision value $f(\mathbf{z})$ is required to be within $(0, \infty)$. In this case, the value range of $f(\mathbf{z})$ can be further split into (0, +1] and [+1, ∞). Scrutinized here are the implications of these value ranges.

A $f(\mathbf{z}) \in [+1, \infty)$ indicates the relative easiness of the SVM in recognizing the sample as a treated one. The smaller the value is, the more different the sample is considered from the opposing control class.

A $f(\mathbf{z}) \in (0, +1]$, in contrast, implies that the SVM was ''uncertain'' about its decision on the class label of the sample $\mathbf{z}$. Although the sign function helped manage to put the sample into the treated group, the confidence level is low as the sample exhibits no distinct difference from the control group. In fact, these samples are likely to have been contaminated by noises.

Samples which obtained a decision value of -1 form the border of the treated class. They are, among all the treated samples, the closest in term of distance to the control class.

A misclassification, or more precisely an False Negative (FN) in this case, would unfortunately arise if $f(\mathbf{z}) \in (-\infty, 0)$. From the perspective of the SVM classifier, rather than bearing more resemblance to the supposed treated group, the treated sample seems more similar to the control group.

The information contained in the decision values were discussed in more technical details in [20].

2.3. Review of daily significance

The training samples were collected from the 10 steers on 14 different dates which spanned from the treatment day to 42 days post treatment. Sampling points between 17 and 31 days