Diagnosing a disorder in a classification benchmark

James McDermott, corresponding author
Management Information Systems, College of Business, University College Dublin, Ireland
Richard Forsyth,
Independent researcher, Leeds UK

Received 11 September 2015, Accepted 11 January 2016, Available online 22 January 2016

doi:10.1016/j.patrec.2016.01.004

[Please cite as:
McDermott, J. & Forsyth, R.S. (2016). Diagnosing a disorder in a classification benchmark. *Pattern Recognition Letters*, 73, 41-43.
]

Highlights

The UCI BUPA Liver Disorders dataset is a common classification benchmark.
The final variable in the dataset is a train/test indicator, not a classification label.
We have surveyed many papers which use this dataset.
A large majority of surveyed papers misinterpret the final column, with meaningless results.

Abstract

A large majority of the many hundreds of papers which use the UCI BUPA Liver Disorders data set as a benchmark for classification misunderstand the data and use an unsuitable dependent variable.

Keywords

Machine learning; Classification; UCI; BUPA liver disorder; Benchmarks

1. The UCI BUPA Liver Disorders data set

The BUPA Liver Disorders data set was created by BUPA Medical Research and Development Ltd. (hereafter “BMRDL”) during the 1980s as part of a larger health-screening database [10]. At the time the second author was developing machine learning software, including what may be the first tree-structured genetic programming (GP) system [3], and collaborating with the BMRDL researchers who collected the data. He went on to use the data set as a GP benchmark [4]. In 1990 the data set was donated on his behalf to the UCI machine learning repository [5]. It is hosted at https://archive.ics.uci.edu/ml/datasets/Liver+Disorders.

Since then it has been very commonly used as a benchmark for classification algorithms. As an indication, appropriate Google Scholar searches find many hundreds of hits (*1).

It is one of a group of commonly co-occurring classification data sets, which also includes Wisconsin Breast Cancer, Pima Diabetes, Wine, Iris, Vehicle Silhouette, Ionosphere, Sonar, Votes, Hepatitis, Statlog Heart, Australian Credit Approval, and some others. To avoid confusion, we note that it is distinct from the “Indian Liver Patient Data Set”, which is also commonly used as a classification benchmark.

It consists of 345 rows and 7 columns. Each row corresponds to one human male subject. The UCI web page gives the columns as in Table 1 (*2).

The first 5 columns are integer-valued and represent the results of various blood tests which may be of use in diagnosing alcohol-related liver disorders. The next, x6, is real-valued and represents the number of alcoholic drinks (equivalent to half pints of beer) taken per day by the subject, self-reported. The final column x7 is binary. It does not represent presence or absence of a liver disorder: it is a "selector", *intended to be used to split the data into training and test subsets*. It was created, rather than collected, by the BMRDL researchers. There are 145 rows with x7=1 and 200 rows with x7=2.

The binary dependent variable in Forsyth's early experiments with the data set was constructed from x6 by thresholding, with different threshold values in different experiments [4].

2. Misunderstanding

As stated, in the Liver data set, x6 is a dependent variable indicating number of drinks, while x7 is a selector, intended to split the data into train and test subsets for one particular experiment. However, many papers interpret x6 as an independent variable, and x7 as the target for classification. In some cases it is explicitly claimed that x7 represents the presence or absence of a liver disorder. This is incorrect. In fact, the information in this data set which pertains to diagnosis is in the first five variables x1 to x5, which are the results of blood tests a physician might use to inform diagnosis. There is no ground truth in the data set relating to presence or absence of a disorder. Papers which blindly use x7 as a binary target are not doing what the researchers intend.

Table 1.
The 7 columns of the data set.

Var.	Abbreviation	Meaning
x1	mcv	mean corpuscular volume
x2	alkphos	alkaline phosphotase
x3	sgpt	alamine aminotransferase
x4	sgot	aspartate aminotransferase
x5	gammagt	gamma-glutamyl transpeptidase
x6	drinks	number of half-pint equivalents of alcoholic beverages drunk per day
x7	selector	field used to split data into two sets

The misunderstanding is not surprising. It is a common convention to lay data sets out in columns as x1,x2,…,xn,y. The decision to put a train/test selector as the final column was perhaps a poor one. It appears that many researchers have simply downloaded the data set, found the final column to be binary, and jumped to the conclusion, based on the name of the data set and perhaps on an analogy to other medical data sets in UCI and elsewhere, that it described presence or absence of a liver disorder. Other researchers presumably followed their lead.

The information given on the UCI page describing the data set contributes to the confusion. It says "It appears that drinks > 5 is some sort of a selector on this database". It is not clear what "selector" means here. It is also misleading, given the other usage of the term "selector" on the same page: "x7: selector field used to split data into two sets". The latter comment is vague: "two sets" does not clearly distinguish "train/test sets" from "two levels of a variable".

A comprehensive analysis of the many hundreds of papers which have used this data set would not be possible, nor necessary. We have carried out a superficial survey. The methodology was as follows. We took the 23 papers listed at the UCI page as having cited the data set. We added the first 40 hits found by an appropriate Google Scholar search (*3). We eliminated two papers which could not be found and one which did not use the data set in any way. We eliminated one duplicate paper. 59 papers remained.

For each paper, we decided whether the data set is used correctly (i.e. dichotomizing x6) or incorrectly (classifying on x7), or that there was not enough evidence to tell. The criteria we used to decide that it was used incorrectly were: (1) mentioning that there are 6 features or independent variables; (2) mentioning that the dependent variable indicated presence or absence of a liver disorder; (3) mentioning x7, v7 or equivalent as the dependent variable; or (4) mentioning that there are two classes of 145 and 200 items, respectively.

In this survey, we have found just 4 papers which certainly use the data set correctly, as follows. Turney [12] dichotomized using the relation $x6 \geq 3$. Brown et al. [2] discarded x7 and used the other variables for visualization only. Ramana et al. [9] correctly used x6 as the dependent variable but failed to state how they dichotomized it. Tang et al. [11] also failed to state this, but gave the balance of their two classes as 169:176. From this it can be inferred that they also used $x6 \geq 3$ to dichotomize.

We are also aware of one other paper, outside the survey, which uses the data set correctly. Maurelli and Giulio [6] dichotomized using the relation $x6 > 7$.

Of the 59 papers, 7 do not give enough information to decide whether they use the data set correctly or incorrectly. The remaining 48 certainly used the data set incorrectly. We conclude that a large majority of published papers using this data set, but not all, are using it incorrectly.

Since the goal of our paper is not to blame individual researchers, we will not here list the papers which we have found to use the data set incorrectly. Our survey is available as supplementary material. However, we note that many of these papers have been cited more than 50 times. Many appear in top venues, including *Pattern Recognition Letters, the European Journal of Operational Research, Systems, Man, and Cybernetics, Part C, IEEE Transactions on Knowledge and Data Engineering, Neural Processing Letters, ICML, KDD,* and *NIPS*.

In many cases, the data set is just one of many being used, and it is being used as a machine learning benchmark, rather than for clinical purposes. In these cases, the papers' conclusions are not greatly affected. However, it is worrying to see that several of the papers which use the data set incorrectly are published in medical-oriented venues, including *Journal of Medical Systems, Expert Systems with Applications, Computer-Based Medical Systems, Computers in Biology and Medicine*, and *Artificial Intelligence in Medicine*. In some of these papers the data set is the only one in use, or one of only two, and the premise of the paper concerns improving liver disorder diagnosis, which is not possible given the true description of the data. In this case serious doubt must be cast on the conclusions of the paper.

We note that the OpenML project, which archives the results of experimental runs on many data sets, also incorrectly gives x7 as the dependent variable (*4).

3. Discussion

There are several obviously desirable properties of classification algorithms: we want them to fit the training data well, to generalize well, to give interpretable models, and to run fast during training and during classification. Another less obvious but still desirable property is an ability to quantify confidence in the data fit, or in the classification of a particular point. Many researchers failed to detect that they were running classification on an artefactual variable, x7. Should the algorithms have detected this?

One way to think about this is to compare these algorithms' test accuracy to some reasonable baseline, for example the accuracy achieved when we simply predict the majority class. Failure to out-perform such a baseline should draw our attention to a potential problem. Table 2 shows what we can expect in this comparison. By predicting the majority class on x7, we can expect to achieve 200/345 = 58% accuracy. Based on these results, we might conclude that most previous research has succeeded in out-performing this reasonable baseline: reported accuracy values of 55% to 72% are common. Hence there is some signal in the noise, and hence we cannot criticize our algorithms on grounds of "over-confidence".

Table 2.

Contingency table on x6 and x7.	x7 = 1	x7 = 2	∑
x6 <= 5	100	157	257
x6 > 5	45	43	88
∑	145	200	345

However, there is an important caveat. If we carry out 30 randomized train/test splits, and run the "predict majority" method each time, we find there is large variability: 49–65%. Since the vast majority of papers do not specify the procedure and the random seed they use to create the train/test split, it is easy to accidentally or dishonestly boost performance by choosing an "easy" train/test split. Using cross-validation is not always sufficient to deal with this problem: this point has been demonstrated forcefully in recent research on Genetic Programming benchmarks [7]. Taking this into account, it is not so clear that all previous research is beating the low bar set by "predict majority".

Another low bar is set by a simple algorithm like logistic regression. Here we can achieve 60–75% accuracy, across the same 30 train/test splits. This seems to constitute evidence that the x7 variable is not purely random, and indeed, it was created by hand rather than using a random number generator. The fact that the x7 variable does have a partly predictable relationship with the other variables suggests that the choice of training and test cases was biased by the original researchers' judgments of what was interesting to them. Indeed, on this basis it is possible to make a case that forecasting x7 is an interesting exercise. However, it remains clear that forecasting this variable is in no way related to diagnosing liver disorders.

We turn now to the intended use of the data set, and classification on x6. If we dichotomize x6 using the relation x6 > 5, Table 2 shows that we can expect 200/345 = 58% accuracy by predicting the majority class. When running multiple splits, we achieve 68–81% with "predict majority", and when using logistic regression we achieve 67–82%. It is unfortunate that these numbers overlap with those reported above for classification on x7, since it is therefore not possible to glance at published performance numbers and be able to guess which variable is being used.

Code to reproduce these simple results is available for download (*5).

4. Actions

How should researchers respond? There are four options.

- **Continue as before.** As described above, many researchers have reported significant improvements over baseline performance. We can conclude that the x7 variable cannot be entirely random. Moreover, the lift attained by logistic regression and more sophisticated methods, relative to “predict majority”, is much larger for x7 than for the dichotomized (x6 > 5). Therefore, some might argue that there is more to learn from continuing to use x7. According to this argument those who are currently using it as the dependent variable can continue to do so. The disadvantage is that they would continue to report results which are not comparable to those reported by Turney [12] and any others who have used the data set correctly; and that this continues to mislead some readers into thinking we are doing classification on liver disorder diagnoses.

- **Switch to using x6**. This is the “correct” option. It has the significant disadvantage that our new results will not be numerically comparable to the large majority of those reported heretofore. With this option we also have to choose the threshold for dichotomizing. Several values have previously been used. The dataset could then be replaced by a version with the final column removed altogether.

- **Use x6 as a regression target**. Alternatively, we could switch to using x6 as a target for regression rather than classification. This has the advantage that there cannot be any confusion as to which task researchers are attempting, if they give their results. However, it still seems likely to cause confusion.

- **Quit using the data set.** We could quit using the data set entirely. Any paper which uses it will need to spend several lines explaining its position. This tends to defeat the purpose of using well-known data sets as benchmarks.

Our preference is to switch to using x6, dichotomized as x6 > 3.

Regardless of what we choose to do, we should tell our readers (1) that the data set does not contain liver disorder diagnoses, and if we wish to make the argument that forecasting a train/test selector variable is an interesting thing to do, then (2) we have to firstly acknowledge that that is what we are doing. Editors and reviewers should ensure that in future papers which use the data set are not accepted unless they specify exactly what they are attempting to do.

This is not the first time a significant misunderstanding of common machine learning data sets has become widespread. Pearson [8] points out common errors in the Pima data (*6), while Bezdek et al. [1] demonstrate that multiple distinct versions of the Iris data set are in use. In most cases it seems little damage is done, but it seems worthwhile to correct the error. As a first step, we have written to the UCI and OpenML caretakers.

Acknowledgments

Thanks to members of the UCD Natural Computing Research & Applications group for discussion. Thanks also to former members of BMRDL, Sharon Allaway, David Robinson and Roger Smolski, for helping to elucidate the status of the last column in the BUPA data set.

Appendix A. Supplementary material

Supplementary Raw Research Data. This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/.

References

[1]
J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva, N.R. Pal
Will the real Iris data please stand up?
IEEE Trans. Fuzzy Syst., 7 (3) (1999), pp. 368–369
[2]
C.T. Brown, H.W. Bullen, S.P. Kelly, R.K. Xiao, S.G. Satterfield, J.G. Hagedorn, Visualization and Data Mining in an 3D Immersive Environment: Summer Project 2003, *US National Institute of Standards and Technology*, 2003. http://flip1-www.boulder.nist.gov/itl/math/hpcvg/upload/vdm2003.pdf (accessed 29.01.16).
[3]
R. Forsyth
BEAGLE -- A Darwinian approach to pattern recognition
Kybernetes, 10 (3) (1981), pp. 159–166
[4]
R. Forsyth, R. Rada
Machine Learning: Applications in Expert Systems and Information Retrieval, *Halsted Press* (1986)
[5]
M. Lichman, UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml (accessed 29.01.16).
[6]
G. Maurelli, M.O. Giulio
Artificial neural networks for the identification of the differences between “light” and “heavy” alcoholics, starting from five nonlinear biological variables
Subst. Use Misuse, 33 (3) (1998), pp. 693–708
[7]
M. Nicolau, A. Agapitos, M. O’Neill, A. Brabazon
Guidelines for defining benchmark problems in genetic programming
Congress on Evolutionary Computation, IEEE, Sendai, Japan (2015)
[8]
R.K. Pearson
The problem of disguised missing data
ACM SIGKDD Explor. Newsl., 8 (1) (2006), pp. 83–92
[9]
B.V. Ramana, M.S.P. Babu, N.B. Venkateswarlu
A critical study of selected classification algorithms for liver disease diagnosis
Int. J. Database Manag. Syst., 3 (2) (2011), pp. 101–114
[10]
D. Robinson, S.L. Allaway, C.D. Ritchie, O.R. Smolski, A.R. Bailey
The use of artificial intelligence in the prediction of alcohol-induced fatty liver
Proceedings of the Sixth Conference on Medical Informatics (MEDINFO-89), North Holland (1989), pp. 170–174
[11]
Y. Tang, B. Jin, Y. Sun, Y.-Q. Zhang
Granular support vector machines for medical binary classification problems
Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, IEEE (2004), pp. 73–78
[12]
P.D. Turney

Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm
J. Artif. Intell. Res., 2 (1995), pp. 369–409

This paper has been recommended for acceptance by Prof. A. Marcelli.

Corresponding author: Tel.: +353 1 7168031.

Notes:

(*1)

11 August 2015:
https://scholar.google.com/scholar?q=bupa+%22liver+disorders%22+classification
finds “About 735 results”, while
https://scholar.google.com/scholar?q=uci+%22liver+disorders%22+classification+-bupa
finds “About 940 results”.

(*2)

The term “alamine” is a typo in the original for “alanine”.

(*3)

11 August 2015:
https://scholar.google.com/scholar?q=bupa+%22liver+disorders%22+classification.

(*4)

http://openml.org/d/8.

(*5)

https://github.com/jmmcd/ML-snippets/tree/master/Liver.

(*6)

These errors have now been corrected: see
http://exploringdatablog.blogspot.ie/2012/10/characterizing-new-dataset.html.