Correspondence

Will the *Real* Iris Data Please Stand Up?

James C. Bezdek, James M. Keller, Raghu Krishnapuram,
Ludmila I. Kuncheva, and Nikhil R. Pal

Abstract—**This correspondence points out several published errors in replicates of the well-known Iris data, which was collected in 1935 by Anderson [1], but first published in 1936 by Fisher [2].**

Index Terms—**Iris data.**

I. Introduction and Conclusions

While preparing Kuncheva and Bezdek [3], these authors discovered that there are (at least) two *distinct* published replicates of the Iris data that have been used as a basis for an unknown number of papers. Subsequently, Bezdek *et al*. [4] discovered two different errors in the version of Iris available through the well-known University of California at Irvine (UCI) machine learning database. Reproduced below, from the preface of Bezdek *et al*. [4] is an account of the problems errors like this cause.

The data: Most of the numerical examples (in [4]) use small data sets that may seem contrived to you—and some of them are. There is much to be said for the pedagogical value of using a few points in the plane when studying and illustrating properties of various models. On the other hand, there are certain risks too. Sometimes conclusions that are legitimate for small specialized data sets become invalid in the face of large numbers of samples, features and classes. And, of course, time and space complexity make their presence felt in very unpredictable ways as problem size grows.

There is another problem with data sets that everyone probably knows about, but that is much harder to detect and document and that problem goes under the heading of, for example, "*will the real Iris data please stand up?*" Anderson's Iris data [1], which we think was first published in Fisher [2], has become a popular set of labeled data for testing—and especially for comparing—clustering algorithms and classifiers. It is, of course, entirely appropriate and in the spirit of scientific inquiry to make and publish comparisons of models and their performance on common data sets and the pattern recognition community has used Iris in perhaps a thousand papers for just this reason—or have we?

During the writing of this book we have discovered (perhaps others have known this for a long time, but we did not) that there are at least two (and, hence, probably half a dozen) different well-publicized versions of Iris. Specifically, vector 90, class 2 (Iris Versicolor) in Iris has the coordinates (5.5, 2.5, 4, 1.3) in Johnson and Wichern [5, p. 566] and has the coordinates (5.5, 2.5, 5, 1.3) in Chien [6, p. 224]. YIKES!! For the record, we are using the Iris data published in Fisher [2] and repeated in Johnson and Wichern [5]. We will use *Iris* (?) when we are not sure what data were used.

TABLE I
The Iris Data: Fisher [2]

Iris sestosa				Iris versicolor				Iris virginica			
Sepal Leng	Sepal Width	Petal Leng	Petal Width	Sepal Leng	Sepal Width	Petal Leng	Petal Width	Sepal Leng	Sepal Width	Petal Leng	Petal Width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

What this means is that many of the papers you have come to know and love that compare the performance of this and that using Iris may in fact be examples of algorithms that were executed using different data sets! What to do? Well, there is not much we can do about this problem. We have checked our own files, and they all contain the data as listed in Fisher [2] and Johnson and Wichern [5]. That is not too reassuring, but it is the best we can do. We have tried to check which Iris data set was used in the examples of other authors that are discussed in this book, but this is nearly impossible. We do not guarantee that all the results we discuss for "the" Iris data really pertain to the same numerical inputs. Indeed, the "Lena" image is the

Manuscript received October 31, 1998. This work was supported by ONR Grant N00014-96-1-0642.

J. C. Bezdek is with the Computer Science Department, University of West Florida, Pensacola, FL 32514 USA.

J. M. Keller is with the Computer Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211 USA.

R. Krishnapuram is with the Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 USA.

L. I. Kuncheva is with the School of Mathematics, University of Wales, Bangor, Gwynedd, LL57 1UT U.K.

N. R. Pal is with the Electronic and Machine Communication Unit, Indian Statistical Institute, Calcutta, 700035 India.

Publisher Item Identifier S 1063-6706(99)04933-4.

Iris data of image processing—after all, the original Lena was a poor quality 6-bit image and more recent copies, including the ones we use in this book, come to us with higher resolution. To be sure, there is only one analog Lena (although *Playboy* ran many), but there are probably many different digital Lena's.

Data get corrupted many ways and in the electronic age it should not surprise us to find (if we can) that this is a fairly common event. Perhaps the best solution to this problem would be to establish a central repository for common data sets. This has been tried several times without much success. Out of curiosity, on September 7, 1998 we fetched Iris from the anonymous FTP site "ftp.ics.uci.edu" under the directory "pub/machine-learning-databases" and discovered not one, but two errors in it! Specifically, two vectors in Iris Sestosa were wrong: vector 35 in Fisher [2] is (4.9, 3.1, 1.5, 0.2), but in the machine learning electronic database it had the coordinates (4.9, 3.1, 1.5, 0.1); and vector 38 in Fisher is (4.9, 3.6, 1.4, 0.1), but in the electronic database it was (4.9, 3.1, 1.5, 0.1). Finally, we are aware of several papers that used a version of Iris obtained by multiplying every value by ten so that the data are integers and the papers involved discuss 10*Iris as if they thought it was Iris. We do not think there is a way to correct all the databases out there which contain similar mistakes (we trust that the machine learning database will be fixed after our alert), but we have included a listing of Iris in Appendix 2 of this book (and, we hope it is right). What all this means for you, the pattern-recognition aficionado, is this: *pattern recognition* **is** *data and not all data are created equally, much less replicated faithfully!*

Table I lists—we hope—the Iris data as published in Fisher [2]. If you think you have the Iris data in your computer or, if you plan to use it in the future, we suggest that you check the version you have or use against these values. Better yet (and we know many of you will check our version this way), return to the source and take the values directly from Fisher's paper.

References

[1] E. Anderson, "The Irises of the Gaspe peninsula," *Bull. Amer. Iris Soc.*, vol. 59, pp. 2–5, 1935.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.

[3] L. I. Kuncheva and J. C. Bezdek, "Nearest prototype classification: clustering, genetic algorithms, or random search?," *IEEE Trans. Syst., Man, Cybern.*, vol. C28, no. 1, pp. 160–164, 1998.

[4] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Norwell, MA: Kluwer, 1999.

[5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[6] Y. T. Chien, *Interactive Pattern Recognition*. Monticello, NY: Marcel-Dekker, 1978.

Comments on "Combinatorial Rule Explosion Eliminated by a Fuzzy Rule Configuration"

Jerry M. Mendel and Qilian Liang

Index Terms—**Combs method, De Morgan's laws, modus ponens, union rule configuration.**

I. Introduction

Combs and Andrews[1] have proven the following logical equivalence (stated here for two antecedents p and q and one consequent r, but easily generalize to an arbitrary number of antecedents and consequents [1])

$$[(p \wedge q) \Rightarrow r] \Leftrightarrow [(p \Rightarrow r) \vee (q \Rightarrow r)]. \tag{1}$$

This is a very significant result because it suggests that we can replace multi-antecedent rules with an interconnection of single antecedent rules, which eliminates the *rule explosion* that is associated with multi-antecedent rules.

Combs and Andrews refer to the left-hand side of this equivalence as an *intersection rule configuration* (IRC) and to its right-hand side as a *union rule configuration* (URC) so that (1) can be expressed as $IRC \Leftrightarrow URC$. In [1], Combs refers to (1) as Combs method; we shall do likewise and shall also use the IRC and URC abbreviations.

From the truth of IRC⇔URC, we now have two distinctly different paths for the design of fuzzy logic systems (FLS's)—the traditional IRC path or the new URC path. The IRC path leads to rule explosion, whereas the URC path does not.

In this correspondence, we discuss four points about the IRC⇔URC. On some points, the interpretation or answer may not be final, so one reason for this correspondence is to generate other responses to it in the spirit of intellectual inquiry. Much of what is in this correspondence is based on many e-mails between the first author and Combs and Andrews.

II. Discussion

Point #1: Generalized Modus Ponens and IRC⇔URC.

The proof of Combs method lies totally within the framework of crisp logic; but an FLS, as described by the sup–star composition, is associated with generalized modus Ponens [8]. Since there is no discussion of this in the above paper or [1], we elaborate on it next in relation to Combs method.

We begin by examining IRC⇔URC for crisp logic's modus Ponens.

1) IRC

Premise: x_1 is A_1 and x_2 is A_2

Implication: IF x_1 is A_1 and x_2 is A_2 THEN y is B

Consequence: y is B.