My book

> **Mirkin Boris: "Clustering for Data Mining: A Data Recovery Approach"**
> Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL; 2005, 266 p.
> Hardcover, 7 Chapters, ISBN 1-58488-534-3

has received a number of reviews that I know of by the time of writing this, December 2007. These reviews are reviewed here.

I list the reviews in the alphabet order of the publication in which they appeared:

1. **BioMedical Engineering OnLine** 2006, 5:34
by Ritaban Dutta

2. **Journal of the American Statistical Association**, Volume 101, Number 474, June 2006, pp. 854-855
by Samuel E. Buttrey

3. **Psychometrika**, vol. 72(1), 109-110, March 2007
by Leslie Rutkowski

4. **Short Book Reviews of the ISI 2007**
by David Hand

5. **Zentralblatt Math** 2007
by Zhizhang Shen

In general, all reviews are rather positive. Further on I quote three types of items from each of the reviews:
(G) what is good in the book,
(B) what is bad in the book, and
(C) general conclusion.

1. BioMedical Engineering OnLine 2006, 5:34 by Ritaban Dutta

Good:
G.1.1. Full of examples, as an example is like a picture which can carry thousand words.
G.1.2. It is very difficult for an engineer to understand the difficult mathematics behind the algorithms, whereas mathematical developments sometimes do not follow real life application scenarios. As an engineer, and not a mathematician, I personally feel that I had a comfortable time during reading this book.
G.1.3. The 'Overall assessment' section at the end of each chapter is a helpful summary to link with the next chapter.

Bad: None

Conclusion:
C.1. I should recommend all engineering students and research engineers to read this book, especially those who are doing research in the field of data clustering, classification algorithms development, intelligent signal processing, biomedical and sensory classification problems.

2. Journal of the American Statistical Association, Volume 101, Number 474, June 2006 , pp. 854-855
by Samuel E. Buttrey

Good:

G.2.1. Mirkin proves a number of interesting points, some on data standardization, many regarding the ANOVA-like decomposition of the "data scatter" into explained (by the clustering) and unexplained parts. The decompositions in particular provide a solid theoretical underpinning for clustering that should take root in the community at large.

Bad:
B.2.1. There is a near-total lack of description regarding how algorithms should actually be implemented.
B.2.2. A flaw regards the limited discussion of cluster validation. How does the user know that a clustering has "succeeded" in the sense of discovering meaningful clusters, and how does he or she interpret the resulting groups? These are difficult and context-dependent questions, to be sure, but Mirkin's final subsection is insufficient to handle this important problem.
B.2.3. Mirkin's sentences are understandable, but they are too often littered with little annoyances like missing articles and obsolete or awkward turns of phrase.
B.2.4. Active data miners who rely on commercial software to do their clustering will find little here to help them.

Conclusion:
C.2. The book presents a number of very interesting ideas and fills a needed niche in the clustering literature. However, the presentation is linguistically uneven and neglects some important areas that the title suggests will be more deeply covered.

3. Psychometrika, vol. 72(1), 109-110, March 2007 by Leslie Rutkowski

Good:
G.3.1.Particularly useful is the section Mirkin devotes to validity and reliability where several common measures and techniques such as bootstrapping are included to test the reliability and validity of identified cluster structures.
G.3.2. In each "method" section, Mirkin includes interpretational aids for making sense of the resulting clusters. Some of these aids are practical for social scientists desiring to perform cluster analysis using commercially available software.
G.3.3. Attention is given to computational demands and computational feasibility of each, K-means and Ward's hierarchy, clustering methods. Social science researchers will find that many of these techniques and issues are of practical interest, even beyond what the title suggests.
G.3.4.The author does an excellent job of providing numerical examples to demonstrate nearly every technique from data preprocessing to performing a cluster analysis to validating the identified structure.

Bad:
B.3.1. A thorough explanation of the data recovery approach and the relevant methods and models are not directly addressed until chapter 5 (i.e., over halfway through the book). Had the data recovery approach been introduced earlier, the author could have used it as a "lens" through which the whole text is viewed.

Conclusion:
C.3.1.Interested in an introduction to the technical aspects of cluster analysis and its use in data

mining? Try Mirkin's book. As an introduction to the ins and outs of cluster analysis, this book is a good place to start. More savvy methodologists who are interested in a wider range of clustering techniques might be better served by consulting Hubert, Arabie, and Meulman (2006) or Brusco and Stahl (2005).

4. Short Book Reviews of the ISI 2007 by David Hand

Good:
G.4.1. Recent years have seen various attempts to formulate a sounder theoretical base, often in the form of model-based approaches. This book also attempts to establish a stronger foundation, though from a rather different perspective.
G.4.2. The particular decomposition studied in this book is the decomposition of the total sum of squares matrix into between and within cluster components, though the book develops this decomposition, and its associated diagnostics, further than I have seen them developed for cluster analysis before.

Bad:
B.4.1. Rather idiosyncratic.

Conclusion:
C.4. The book presents an unusual, perhaps even rather idiosyncratic approach to cluster analysis, from the perspective of someone who is clearly an enthusiast for the insights these tools can bring to understanding data.

5. Zentralblatt Math 2007 by Zhizhang Shen

Good:
G.5.1. After going through this book, there is no doubt in my mind that the author knows what he writes.
G.5.2. A large number (58) of examples taken from many different walks of life are used to demonstrate most of the techniques. I particularly like that at the end of all these examples, a cross reference is given as to where this example will be used later and analyzed, sometimes in as many as 15 different places.
G.5.3. This book also contains an extensive bibliography with 142 items, and a useful, two-level, index.

Bad:
B5.1. There are no exercises assigned anywhere in the book. This might make this book less likely to be adopted as a textbook.

Conclusion:
C5.1. This book is definitely a wonderful resource for those who are interested in the topics as to what clustering is and how it should be applied.

I am thankful to all the reviewers for their reading the text through and their comments that seem mostly fair.
I would comment on this issue, though: **"How does the user know that a clustering has "succeeded" in the sense of discovering meaningful clusters, and how does he or she interpret the resulting groups?"**
I consider this a core issue in clustering.
The thrust of my writing has been in showing that the user should look into the **explained part of the data scatter** if they want addressing this. If the explained part contributes strongly, the clusters are meaningful. If there are strong specific variable-to-cluster contributions, they are to be looked at for the meanings. Otherwise, indeed, the current state of the art is "insufficient to handle this important problem," though I do think that I described all existing ideas in this area.

Comments on two clashes of opinion:

A. Usage by practitioners.
- "Active data miners who rely on commercial software to do their clustering will find little here to help them."
- "Some of these aids are practical for social scientists desiring to perform cluster analysis using commercially available software."

Yes indeed my writing in this regard is rather for scientists or software developers.

I did once lecture marketing researchers that were using package SPSS and the like in their work. They said afterwards that my advice gave them better understanding of clustering in general as well as pointed to some previously unused options in the software, regarding all stages - pre-processing, clustering, and post-processing.

B. Structure of the text.
- "Unfortunately, a thorough explanation of this approach and the relevant methods and models are not directly addressed until chapter 5 (i.e., over halfway through the book)."
- "... has done a great job by writing this book in style. Book chapters explain the algorithms for clustering very well, but most interestingly on the basis of pictures and examples."

The book structure reflects my daily experiences in teaching Computer Science students. An overwhelming majority of them do not like formulas – they hate formulas, I suspect - and prefer examples to general statements. Overall, the book is oriented towards the reader who is not a mathematician and does not like formulas. The data recovery approach features in non-mathematical chapters prominently – just take a look at ScaD and QscaD tables!

I would also point out a dear to me feature missed by the reviewers: in the book, for the first time in the literature, the issue of **simultaneously analyzing both quantitative and categorical features** has been carried through, due to the data recovery approach.