Fall 2011 Comments and Suggestions---Textbook, Slides, and Class Presentation

You are encouraged to give comments on the draft of each chapter of the "Data Mining: Concepts and techniques" 3rd edition.

If you have new ideas and revisions on the textbook draft, class slides and class presentations, please put them in and mark them in different colors to raise our attention.

Thanks!
-Jiawei Han

Chapter 1. Introduction

Suman Nayak reported 4 errors or suggestions (corrected/adopted) 8/26-27/10
Nikita Spirin [spirin2@illinois.edu] 8/30/10: 4 adopted suggestions

Page 4: Tera- or peta-bytes of data, pour daily  Tera- or peta-bytes of data pour daily (remove comma)
JH: updated, thanks!

Page 4: medicine, and almost every aspect of our daily life  medicine, and almost every other aspect of daily life (add 'other', remove 'our')
JH: updated, thanks! but kept "our" which sounds a little intimate.

Page 4: tools are badly needed to automatically uncover  tools are critical in automatically uncovering (more professional)
JH: I prefer to keep the original since we would like to be closer to readers (less professional :-).

Page 4: (last sentence before 1.1, suggested rewrite) "Data mining will continue to assist in the journey from the data age towards the coming information age."
JH: I prefer the original one.

Page 4: It found a close relationship  It finds a close relationship (all other sentences are in present tense)
JH: I prefer the original one. It emphasizes the past action: "found" and the others simply state the truth.

Page 5: After the establishment of  After the creation of (more natural)
JH: I prefer the original one. It emphasizes the field/technology of database system is established.

Page 5: The steady and dazzling progress of computer hardware  The great innovation in computer hardware (more professional)
JH: I prefer the original one. It reads more lively.

Page 7: Internet-based global information base such as  Internet-based global information bases such as (should be plural)
JH: updated, thanks!

Page 8: major components in the picture  major components of the words (logical error)
JH: I prefer the original one. It reads more lively. It is not an error.

Page 8: In analog, data mining Analogously, data mining (sounds more natural)
JH: I prefer the original one. It reads more natural to me.

Page 10: will certainly continue to embrace new data types  will incorporate new data types (more natural)
JH: I prefer the original one. It reads more lively.

Page 11: (second sentence under 1.3.1: change the two semicolons to commas)
JH: I prefer the original one. It makes distinction between ; and , in such a long sentence.

Page 12: Through the use of relational queries, you can ask things like  You can use relational queries for statements like
JH: I prefer the original one. It reads more lively.
Page 15: For example, given the knowledge that printers

JH: I prefer the original one. It reads more lively.

Page 16: and how to mine patterns that carrying rich and how to mine patterns that carry rich

JH: I prefer the original one. It reads more lively. MK+JH: Take student's suggestion. Corrected on 10/17/10!

Page 16, upper part of the page - application => applications (add S to the word application)

JH: Done. Thanks.

Page 17: of software products whose sales was increased of software products whose sales were increased

JH: Done. Thanks.

Page 19: such as milk and bread which are frequently bought such as milk and bread, which are bought together

JH: I prefer the original one. I do not see anything wrong there. MK+JH: Take student's suggestion. Corrected on 10/17/10!

Page 19: and have purchased a laptop (computer) at and have purchased a laptop at

JH: I prefer the original one.

Page 22: (I recommend removing the opening "What is cluster analysis" under 1.4.4)

JH: Done. Thanks.

Page 23: the rare events can be more the rare events are more

JH: I prefer the original one.

Page 25 - make the names of research areas look similar => Data Mining, but Machine learning, Database technology (because in consequent subsection capitalized notation is used)

JH: I redraw the whole picture (we plan to redraw it anyway). MK: Pls check to make sure this is what we want to show.

Page 29: multimedia data have been accumulated and made available online due multimedia data have accumulated online due to (redundancy)

JH: I prefer the original one.

Page 29: intelligence, bench marking, and intelligence, benchmarking, and (spelling)

JH: Done. Thanks.

Page 29 - maybe substitute Presentations => applications?!

Suggestion: Applications of data mining in knowledge-intensive domains,...

JH: This really should be Presentations since it discussed User Interaction. "Applications" does not fit here and we have discussed in the applications subsection.

Page 31: Such severe skewed data Such severely skewed data

JH: Done. Thanks.

Page 31: This involves the investigation of new kinds This involves investigating new kinds (consistency with rest of sentence)

JH: I prefer the original one.

Page 32 - whether it be the Web => substitute be with is

JH: I think "be" here could be better since it implies "could be".

Page 33 - Cloudy computing => Cloud Computing (wrong usage of a term)

JH: You are right but it is fixed by Suman and we have updated it on Sat.

Page 33 - in parallel => parallel to (mistaken preposition, not sure, non-native speaker ;) )

JH: "in parellel" is better English.

Page 34 - from complex data objects to temporal data, biological sequences, sensor data...
cause I think that temporal data, bio-sequences are exactly complex!

JH: You are right, changed to "simple data objects".

Page 34: for diverse applications for diverse applications (spelling)

JH: You are right but it is fixed by Suman and we have updated it on Sat.

page 35 - mousing clicking => mouse clicking

JH: Done. Thanks.

page 36 -

"Data warehouse systems provide multidimensional data analysis capabilities, collectively referred to as OLAP (online analytical processing)." to

"Data warehouse systems provide multidimensional data analysis capabilities, collectively referred to as OLAP (online analytical processing).

JH: Done. Thanks.

**Chapter 2. Getting to Know Your Data**

Suman Nayak reported one error on p. 36 (corrected) 8/29/10

Efe C. Karakus reported one error on p. 29 (corrected) 8/29/10

Page 4: more sophisticated methods such as tree-map more sophisticated methods such as treemapping (treemapping is the method that produces a treemap)

JH: Thanks. The original one seems ok since visualization method also use nouns to represent "method".

Page 4: you will know the different attribute types, and basic you will know the different attribute types and basic (only two items, so no comma)

JH: Thanks. The original text seems ok since the other part is really long and could be confusing if not using ".".

Page 14: in the order of Minimum, Q1, Median, Q3, Maximum in the order of minimum, Q1, median, Q3, maximum (I see no reason to capitalize these three words.. I could be wrong)

JH: Thanks. The original text seems ok since we use bold to imply "definition" and here they are not definition but emphasis.

Rick Barber (barber5)- I can’t seem to find Figure 2.2 as referenced on page 12

JH: We invite interested students to work out a figure!

On page 8, it is mostly clear from context, but it might be useful to reiterate that Ratio-scaled attributes can have a mean, median, mode calculated as well

JH: Done! Thanks!

Figure 2.10 is incorrectly captioned with the same caption as figure 2.11

JH: Thanks. Changed to the new caption: "Figure 2.10: Pixel-oriented visualization of four attributes by sorting all customers in incoming ascending order."

Peter Wubbels (wubbels1) on 8/30/10 - Ch2, pg. 8 "An attribute is countably infinite if the set of possible values in infinite..." should be "An attribute is countably infinite if the set of possible values is infinite..."

JH: Done! Thanks!

Han-Wen Yeh (yeh7) - Table of contents entry (page 1) and section heading (page 31) for 2.4.1 - "Data Matrix vs. Disimilarity Matrix" should be "Data Matrix vs. Dissimilarity Matrix"

JH: Done! Thanks!

- Page 24 (caption for Figure 2.13) - Suggestion: "A scatter plots visualization of a 2-dimensional data set." to "A visualization of a 2-dimensional data set using scatter plots." The former may seem to have a grammatical error to people at first glance. Same as the caption for Figure 2.14

JH: Done! Thanks!

- Nikita Spirin

page 3 - ...following*: What* are... => following*: what* are (grammatically incorrect)
JH: Sorry, I could not see what is wrong.

page 14 - orderMinimum, Q1, Median, Q3, Maximum (fix the LaTeX sources, split formula)

JH: Done! Thanks!

page 16 - one should indicate that N is a populations size, not sample size

JH: It seems there is no need since it works as well if it is a sample.

page 18 - Example 2.14 Quantile-plot => Quantile-Quantile-plot (misplaced name)

JH: Done! Thanks!

page 43 - fix the LaTeX source, x'y = ... (outside the page frame)

JH: Done! Thanks!

Nikita Spirin

page 37 - definition of metric is redundant, one can exclude non-negativity since it can be derived from triangle inequality and symmetry

JH: I think we may still retain the non-negativity since this has been popularly presented (see Wikipedia).

Hans Hermans (?):

page 38 - It represents the Manhattan distance when h = 1 (i.e. L1 norm) and Euclidean distance when h = 2 (i.e. L2 norm).

=> I think distance and norm are related but essentially different concepts. Distance is a function defined on two data points, while norm is a function defined on a single data point. Mathematically speaking, metric (distance) is induced by norm. For example, Euclidean distance is induced by L2 norm and Manhattan distance is induced by L1 norm. Please see http://en.wikipedia.org/wiki/Metric_(mathematics) for more details. So I suggest replace "(i.e. L1 norm)" with "(induced by L1 norm)"; or just delete it.

Similarly, I suggest modify "The supremum distance (also referred to as L_max, L_req norm, and the chebyshev distance) is ..." to "The supremum distance (also referred to as the chebyshev distance) is..." simply.

page 42 Chapter 2.4.7 Cosine Similarity the third sentence from the bottom of the paragraph-should it read "this does not make them dissimilar." vs "similar."

JH: I think the original one is right.

Jihoon Choo: page 46 on class presentation: smaller -> smaller

JH: Done. Thanks.

Rami Reddy

Page 37: Consider a be measured in either meters, say, or inches height attribute, for example, which could --- I think 'say' can be removed

JH: Done. Thanks.

Seokje Seo

Page 9: In example 2.6, the data given has typo. 9th data has to be 63 instead of 62.

JH: Done. Thanks.

Nikita

page 15 - 379 replace to 361

JH: The original one is the correct number. We use approx to represent it now, it should be \sqrt{379.17} = 19.47.

Chapter 3. Data Preprocessing

Han-Wen Yeh (yeh7)

Page 3: The sentence "Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance" may flow better as "Data reduction can reduce the data size through methods such as aggregation, clustering, and the elimination of redundant features."

JH: I prefer the original one.

- Near the bottom of page 4 - "Other data may not be included simply because it was not considered important at the time of entry." to "...because
they were not...” I am assuming here that the term “data” is being treated as a plural, it being the trend throughout the book so far.

JH: Done! Thanks!

- Page 9 (last sentence of point number 6) - "Decision trees and Bayesian inference are described in detail in Chapters 8 and 9, respectively, while regression is introduced in Section 3.4.5." to "...while regression is introduced..."

JH: Done! Thanks!

- Page 10 (first sentence of point number 2) - "Data can be smoothed by fitting the data to a function, such as with regression." - awkward sentence. A possible alternative is, "Data smoothing can also occur by conforming data values to a function, a technique known as regression."

JH: Sounds good to change. I changed it to "Data smoothing can also occur by conforming data values to a function, a technique known as regression." Micheline, please have a look.

- Page 12 - "Errors in instrumentation devices that record data, and system errors, are another source of discrepancies." to "...are other sources of..." or "Other sources of discrepancies include errors in data-recording instrumentation devices and system errors."

JH: I prefer the original one. MK, pls have a look.

- Following sentence - "Errors can also occur when the data are (inadequately) used for purposes other than originally intended." to "...for purposes that were not originally intended."

JH: I prefer the original one. MK, pls have a look.

- Page 17 - "Note that correlation does not imply causality." to "Note that correlation does not imply causation." A more common phrase, more likely to "click" with the reader's mind.

JH: I prefer the original one. MK, pls have a look.

Han-Wen Yeh (yeh7)

Page 20 (3.4, second paragraph) - "we first present an overview of data reductions strategies" to "...data reduction strategies".

JH: Done! Thanks!

Page 21 (near the top) - "There are several lossless algorithms for string compression, however, they typically allow..." Period instead of comma before "however". The comma makes the reader unsure of which part of the sentence "however" belongs to.

JH: I prefer the original one. I think it is clear.

Page 23 (bottom) - "Because the components are sorted according to decreasing order of "significance," " to "...sorted in decreasing order of..."

JH: I prefer the original one. I added a "the" in front of decreasing ...

Page 24 (near bottom) - "In addition, the added volume..." - "On the other hand" instead of "In addition"? The sentence was not really an addition to the previous statement, but more of a "on the contrary" to it.

JH: I prefer the original one. I think it is clear.

Page 25 (Fig. 3.6) - There are weird boxes after every line of text in the table.

JH: Agree. This figure has to be redrawn!

Suman Nayak:

page 35 the very first line of the page. "may lead to a very different results" => may lead to very different results.

JH: Done! Thanks!

p.35 "In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such attributes greater effect or "weight"." =>
"In general, expressing an attribute in smaller units will lead to a larger range for that attribute, which tends to give such an attribute greater effect or "weight"."

"In general, expressing an attribute in smaller units will lead to a larger range for that attribute. This tends to give such attributes greater effect or weight."

JH: maybe it is better to change it to: "In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or "weight"."?

p.37. "A histogram partition the values for an attribute, A, into disjoint ranges called buckets" => A histogram partition’s the values of an attribute, A, into disjoint ranges called buckets"

JH: Done! Thanks!

p39. "A user or expert can easily............." => "A user or an expert can easily............."
JH: Done! Thanks!

Nikita

page 3 - distance measurements => distance measures (what was meant? two possible senses)
JH: I prefer the original one.

page 4 - Section 3.1.2 outlines the major tasks in data preprocessing (add dot at the end of the sentence)
JH: Done! Thanks!

page 19 - company, If => company. If (punctuation)
JH: Done! Thanks!

page 24 - For example, if the task is to classify customers as to whether or not they are likely to purchase a popular new CD at AllElectronics when notified of a sale,...
JH: Done! Thanks! changed to "If the task is to classify customers based on a probability of positive reaction on a discount offer,..."

page 44 - Take the IRIS data set, obtained from http://www.ics.uci.edu/ml/mlearn/MLRepository.html... (fix LaTeX source, exerted formula)
JH: Will be fixed by the publisher. Thanks.

Rami Reddy

Page 6: Is there any way I can reduce the size of my data set, without jeopardizing the data mining results?" This sentence can be rephrased to "Is there a way to reduce the size of the data set without jeopardizing the data mining results?"
JH: Done! Thanks!

Page13: This process, however, is error-prone and time-consuming. Need a space between iserror
JH: Done! It is a formatting error, not a text error. Thanks!

Nikita

page 35 – I would extend z-score normalization description. Although formula is correct it is true only if we know the true mean and std for the dataset, in this case it is Standard Z-Score Transform.

But in real problems we face with unknown mean and std and hence the appropriate z-score formula is Student Z-Score normalization, where we use sample mean and std/sqrt(N). (It is the same as central statistics in interval estimation (confidence interval) theory)
JH: Thanks. It is a nice extension and it is somewhat beyond the scope of the book. We may later put a similar discussion in a footnote.

Chapter 4. Data Warehouse and Online Analytical Processing

Han-Wen Yeh (yeh7)

Chapter in general - Due to the long chapter title, even page numbers from page 4 onwards look like part of the title. Perhaps change it to "Data Warehousing and OLAP"? Or make it two lines?
JH: No worry. Publisher will handle formatting problems.

Page 4 - “This short*, but comprehensive definition” - the comma between "short" and "but" is unnecessary.
JH: I prefer the original one.

Page 4 - "The four keywords, ..." - The "the" is not necessary. After all, we are introducing the keywords for the first time, and the "the" makes it sound as though we have mentioned them in passing before.
JH: I prefer the original one. We did mention them in the previous sentence in the same paragraph.

Page 5 - "In sum" to "In summary". "In sum" is not very commonly used and sticks out like an outlier.
JH: I prefer the original one.

Pages 5&6 - "(4) managing the customer relationships" - the "the" is unnecessary, especially when the previous three points didn't have it either.

JH: Done! Thanks!

Page 6 - "Data warehousing is also very useful from the point of view of heterogeneous database integration." to ".useful from a heterogenous database integration point of view." Not fully elegant, but I think flows better.

JH: I prefer the original one.

Page 6 - "...wrappers and integrators (or mediators), on top of..." - Comma is unnecessary.

JH: I prefer the original one.

Page 6 - "competes for resources with processing at local sources." - "competes for resources with processes at local sources"? "competes for processing resources with local sources" works, too.

JH: Thanks. I changed it to "competes with local sites for processing resources"

Han-Wen Yeh (yeh7)

Page 11 - "This high-level model, although it will need to be refined ... data marts, will greatly reduce future integration problems." to "This high-level model, which will need to be refined..."

JH: The original text seems to be OK.

Page 13 - "Metadata play a very different role than other data warehouse data..." to "Metadata play a very different role to data in other data warehouses..." - "than" to "to", and "data warehouse data" sounds slightly awkward.

JH: The original text is correct.

Rick Barber (barber5)

Example 4.9 says year=20010, might be unintentional

JH: Right. Done.

Page 44 first paragraph of chapter 4.5 begins "Conceptually, data cube" should say a data cube

JH: The original text seems to be OK since "data cube" plays more role like a concept here.

Jose Luis Fernandez (frndzg2)

Page 11 - "... that are UNIX/LINUX- or Windows-based. The implementation cycle of ..." dash at the end of LINUX intentional? I don't know if there is a standard, but Linux is usually seen written with capital L and the rest of the letters lowercase.

JH: OK. I changed it to lower cases.

Nikita Spirin

all pages - page header: 4CHAPTER 4. DATA... => CHAPTER 4. DATA... (make layout like in previous chapters, otherwise page numbers concatenate with subtopic names)

JH: No worry, the publisher will take care of this.

page 4 - where we store data about our data => may be change to:

where we store additional information about our data (otherwise the word data appears twice in one sentence)

JH: The original sentence is right, it explains what is meta data, which is data about data.

page 32 - to automate as much of the process as possible => to automate the process as much as possible (text suggestion)

JH: The original sentence is slightly better, emphasizing some part of the process will be automated..

page 32 - data summary and comparison => data summarization and comparison (to make things parallel from lexical point of view, two times on a page)

JH: A good point. Done..

page 33 - facilities, and reporting and = > facilities, reporting and (excessive conjunction)

JH: The original sentence is slightly better, since it is a necessary group of multiple kinds of functionalities.

page 42 - selection constant "year = 20010" = may be 2010 ?!
JH: Right. Done.

page 47 - Now that the data => Now when the data (misused interrogative pronoun)

JH: The original sentence is right, "how" is asked.

page 53 - needs to perform => (possible variants: needs performing OR is needed to perform)

JH: The original sentence seems to be better.

page 54 - they are synchronously all generalized => they are both synchronously generalized

JH: "all" is better here since it means all the data in these classes. I modified the surrounding sentences slightly.

page 54, pointer 1 - a target class and one or a set of contrasting class(es) =>

a target class and one or a set of contrasting classes.

JH: I think your suggestion is better than the original and I updated it.

page 58 - OLAP servers may use relational OLAP (ROLAP), or multidimensional OLAP (MOLAP), or hybrid OLAP (HOLAP) - noun is needed (suggestion: architecture)

JH: Thanks. I added: "implementation".

Chapter 5. Advanced Data Cube Technology

Nikita Spirin

page 36 have be => have to be

JH: Right. Done.

Rick Barber (barber5)

Page 3 Very first sentence starts Data warehouse integrates...should this be Data warehousing?

JH: Thanks. I changed it to: "Data warehouse systems".

Page 13 Third paragraph near the end the example. I believe the 10 x 4000 should actually be 40 x 1000 [Note: The same question is also raised by Jay: smittkmpp@illinois.edu]

JH: Thanks. This is an obvious stupid mistake but I could not recall when it happened. I corrected it.

Page 36 Second full paragraph first sentence contains "semantically similar cells maybe" should be "semantically similar cells may be"

JH: Cannot find it. Maybe we have a different version of the text?

Page 36 Same paragraph, last sentence ends with "may bring about an answer that does not semantically meaningful." should be "is not semantically meaningful."

JH: Cannot find it. Maybe we have a different version of the text?

Page 45 First paragraph after start of 5.3.2 near the middle a sentence begins "Let's examine such a new way to defined data cubes" should be "define data cubes"

JH: Cannot find it. Maybe we have a different version of the text?

Page 46 second paragraph after start of 5.3.3 the first sentence "Recently, researchers start paying attentions on" should probably be "Recently, researchers have started paying attention to"

JH: Cannot find it. Maybe we have a different version of the text?

mcclosk1

Page 3, 2nd paragraph: "Although data cube" should be "Although the data cube" or "Although the data cube approach".

JH: Thanks. I changed it to "Although the data cube concept" ...

Page 6, first paragraph: broken reference: A method of full cube computation is given in Section ??.

JH: Thanks. It should be Section 4.4.
Slide 25, should be lossless not loseless.

JH: Thanks. Done.

Carlos P. (pavon)

Page 45. "In this way, the PBC method reduces", I didn't find a reference for PBC, should it be PBE?

JH: Thanks. Done.

JH: Done.

Figure 5.11 from the algorithm, when I recursively call the function and go to the a1 as root stage, the function should insert all its children to the corresponding position at C's star tree, that means that, we need to add both b* and b1 to the BCD tree, right? But the figure just show left half the the tree BCD, if it is in the stage of d* node and start to go back, then at this time, the BCD tree should at least have a node which is b1 on the right hand.

Chapter 6. Mining Frequent Patterns & Association: Basic Concepts

Nikita

page 7 - huge a number => huge number

JH: Thanks. The sentence seems to be OK.

page 8 - in data set => in a data set (two lines below is OK, the same rule - just to make it looks similar)

JH: Thanks. Done.

page 9 - properties, as we shall see following. = properties, as we shall see.

JH: Thanks. I changed it to "as we shall see below."

page 26 - confidence, correlation => confidence, correlation

JH: Thanks. Done.

page 18 - a I3-conditional FP-tree=> an I5-conditional FP-tree

JH: Thanks. It is great you catch it. Done.

Chapter 7. Mining Frequent Patterns & Association: Advanced Methods

Nikita Spirin

page 15 - indicated => indicate (change the tense from past to present)

JH: Thanks. Done.

page 22 - which at explores => which explores (remove at)

JH: Thanks. Done.

page 10 - difficulties => may be "drawbacks" OR "shortcomings" is better?

JH: Thanks. I changed it to "drawbacks".

page 29 - From Example 7.13 => From Example 7.11 (typo in the number, wrong reference)

JH: Thanks. Done.

page 30 - Pattern-Fusion assumes available an initial pool of small frequent patterns is available. (remove first occurrence of a word available)

JH: Thanks. Done.

page 31 - maybe you should replace the word "sheer" to "huge", because sheer has different meaning

JH: Thanks. Done.

page 31 - Too low a value can => Too low value can (remove a)

JH: Thanks. I think the original one is OK.
page 36 - would be redundant To => (put a dot at the end of the sentence)
JH: Thanks. Done.

page 40 - p; and = p; (remove and)
JH: Thanks. But I cannot find where is the problem.

page 42 - Patterns mining => Pattern mining (remove s)
JH: Thanks. Done.

page 23 - count(I) ? 10 monotonic - is it true? I think it is not a monotonic constraint because by adding some item we can decrease count.
for

Unknown macro: \{count <= and >=\}

it is necessary to reverse yes and no in a table on page 24
JH: Thanks. But I cannot find where is the problem.

Nate Dykens:

Page 17, shouldn't one of the following be 100/200 instead of 99/200, since there is only one transaction with both A and B. Same question a few sentences later.
\[ \text{sup}(A \; B) \times \text{sup}(A \; B) = 99/200 \times 99/200 = 0.245 \]
JH: Thanks. The book is correct. Since there are 99 trans containing A only and 99 only B only, and one containing both A and B.

Page 22, change "Based on how a constrain" to "Based on how a constrain"
JH: Thanks. Done.

Chapter 8. Classification: Basic Concepts

Slide #41 - "Rule Extraction from a Decision Tree" - At the bottom of the slide, the very last extracted rule should read "IF age = old..." and NOT "IF age = young..." It looks like the book (see Example 8.7) doesn't need to be changed. NOTE: This fix is different than your earlier mentioned fix where the decision tree leaves should have ">40,excellent" as "yes" and ">40,fair" as "no." (Brian Leege, leege1)
JH: Thanks. Done.

Nikita Spirin

page 3 - main class if interest => main class of interest (preposition)
JH: Thanks. Done.

page 4 - the first paragraph in 8.1.1. section looks as if all space between words were removed, text is very dense (check it please)
JH: Thanks. Done.

page 15 - For example, suppose that instead of the discredited version of age above, we instead have... (remove repetitive instead)
JH: Thanks. Done.

page 16 - formulae 8.6 (fix the bracket)
JH: Thanks. Done.

page 25 - the CART, C4, and SPRINT algorithms (may be C4.5, because C4 is a clustering algorithm)
JH: Thanks. Done.

page 27 - need be maximized => need to be maximized
JH: Thanks. Done.

page 30 - resulting in P(student= yes|buys computer = no) => fix the LaTex source (dangling formulae)
JH: Thanks. Done.

page 35 - the resulting rule should cover more of the "accept" tuples.
May be you should add "relatively more" because absolute number of tuples shrinks with an additional constraint.

But recall increases.

JH: Thanks. Done.

page 39 - Section ?? describes (fix the reference)

JH: Thanks. Done.

page 49 - Section ?? we (fix the reference)

JH: Thanks. Done.

page 51 - false positive For comparison (add a period after positive)

JH: Thanks. Done.

page 59 - Section ??, for example... (fix the reference)

JH: Thanks. Done.

page 61 - F beta (fix the LaTeX source, add backslash)

JH: Thanks. Done.

page 63 - Outline methods for addressing the class imbalance problem.

[from MK: Can someone propose some programming assignment for this?]

For example, one can ask to write a program that generates 2 random classes with the overlap, mutual penetration.

Then sequentially increase the size of one class and draw the decision boundary, observe how it will vary with

the relative ratio of class sizes.

JH: Thanks. We will handle this when we get time.

Ji Hoon Choo (choo5)

Slide page 24 - AVC-set on Age

| Age | Yes | No | -> | Age | Yes | No |
| <=30 | 3 | 2 | -> | <=30 | 2 | 3 |

JH: Thanks. Done.

Efe Karakus (karakus1)

Page 18:

at the very bottom of the page the '-' sign is missing it should be 0.459-0.357 = 0.102, currently it only says 0.459 0.357 = 0.102

Chapter 9. Classification: Advanced Methods

Nate Dykens:

Page 38: Remove the space between "medium_income"

JH: Thanks. Done.

Page 38: Fix the missing spacing here: Given a tuple to classify, more than one fuzzy rule may apply. Each applicable rule

JH: Thanks. Done.

Page 47: Change from "yet the class-labels are scare" to "scarce"

JH: Thanks. Done.

Rick Barber:

Page 20, right before page break: It says 'From the above, we can obtain a formulae' should be a formula (formulae is plural)

JH: Thanks. Done.
Chapter 10. Cluster Analysis: Basic Concepts

David McCloskey:

Pg18, Right before 10.3: “local optimal” should be “local optimum” (optimal is an adjective).

JH: Thanks. Done.

Pg19: “chess and checker” should be “checkers”.

JH: Thanks. Done.

Pg19: “We introduce two such methods, namely BIRCH, Chameleon” should be “BIRCH and Chameleon”

JH: Thanks. Done.

Pg20, first sentence: Should be no comma before “and compute clusters”.

JH: Thanks. Done.

Pg20, next sentence: “Probabilistic methods uses” should be “use”. Should be no comma before “and measure”.

JH: Thanks. Done.

Pg20, 10.3.1 2nd paragraph: “form its own cluster, and iteratively merge” should have no comma. The subject at the beginning of the paragraph is plural (“methods”), but in the 4th sentence it changes to singular (“it”). In next paragraph, the same thing is done, beginning with “divisive methods” then going into “this method”.

JH: Thanks. Done.

Pg27, 10.3.4: “In Chameleon, cluster similarity is assessed based on how well-connected objects are within a cluster and on the proximity of clusters.” Should be “well connected”.

JH: Thanks. Done.

Nikita Spirin

formula 10.1 change $m_i$ to $c_i$ or replace $c_i$ to $m_i$ in the paragraph below to make the notation “uniform”

JH: Thanks. Done.

page 15 - An third => the third

JH: Thanks. Done.

page 23 - Farthest neighbor algorithms tend to minimize the increase in diameter of the clusters at each iteration as little as possible. (double minimization in a sentence, second modifier is excessive)
Wooil Kim

page 16: concern with Figure 10.4

The explanation above figure 10.4 says the case that o_i is replaced by o_random. But figure shows the case that o_j is replaced by o_random.

Each explanation is correct, but they do not match.

JH: Thanks. Added a note to caption and will update the graph,

Suman Nayak

p. 15. given that the mean of cluster

is 3 <- the mean of the cluster is 2

JH: Thanks. Done.

Nikita Spirin

page 19 - give => given

page 20 - all objects one => all objects IN one

JH: Thanks. Done.

page 52 - spatial static => spatial statistics

JH: Thanks. I changed it to spatial statistic (singular)

Chapter 11. Cluster Analysis: Advanced Methods

Chapter 12. Outlier Analysis

Chapter 13. Trends and Research Frontiers in Data Mining