Homework II Data Mining Fall 2012
Computer Science
Indiana University
Bloomington, IN

Mehmet M. Dalkilic

November 18, 2013

Introduction

- This homework is due Sunday December 1, 2013 11:49 p.m.
- All the work should be your own. You are not allowed to use pre-existing code. You may use C, C++, C#, Python, or Perl, R. No existing packages that do naive Bayes, k-means, ID3. You may reuse your previous k-means code. State explicitly that all the work is yours. In this homework, while I am explicitly asking less, this tacitly means you should be doing more on your on. I am sketching some elements here. The final report must be done in LaTeX. You will turn in *.tex, *.c, *.exe, *.pdf files.

Problems

1. Fraudulent Sales Data.
   (a) This problem examines fraudulent sales data from [1].
   (b) The data is found at: `http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR`
   Click "Data Sets"
   (c) Download the unprocessed data only. The author describes the datas provenance as undisclosed and anonymized. The data captures the behavior of salesman and, in particular, whether sales are fraudulent. The table consists of five attributes, `ID`, `Prod`, `Quant`, `Val`, and `Insp`: the salesmans anonymous ID, the product identifier, the number of units sold, the total value of the sale, whether the human annotated transaction was *deemed* either valid `ok`, fraudulent `fraud`, or unexamined `unkn`. Clearly there are multiple questions that might be posed with this data set. Patently, the questions should center on fraud.

   The final report should be a data mining analysis:

   Description of the problem Π and Δ using k-means, Naïve Bayes, ID3. The description should be in two parts: a laymans description that is easily understood by a non-technical audience and more formal description understood by computer scientists.

- Analysis of original data Δ
- Steps to clean and transform data to $\Delta*$
- Data Mining algorithms employed (and implemented, of course) to solve $\Pi(Delta)$
- Quality of results.
- Analysis of results that includes suggestion of what, if anything, can be done $wrt\,\Pi(Delta)$
- Appendix that includes code, citations (if any), links, *etc.*

2. You've received a strange data set `enigma.txt`. Analyze the data.

3. I am providing you with voting data from the 35^{th} to the 112^{th} US Congressional Sessions. What can you you say about the differences between Republican and Democratic Members? What about Northern Democrat and Southern Republicans? What about Northern Republican and Southern Democrats? The data is in an *.csv file `USCongress`. The description of the file is here:

    1. Congress number
    2. Roll Call number
    3. Month
    4. Day
    5. Year
    6. Number Missing Votes (not voting and not in Congress)
    7. Number Yeas
    8. Number Nays
    9. Number Republican Yeas
    10. Number Republican Nays
    11. Number Democrat Yeas
    12. Number Democrat Nays
    13. Number Northern Republican Yeas
    14. Number Northern Republican Nays
    15. Number Southern Republican Yeas (11 States of Confederacy plus KY and OK)
    16. Number Southern Republican Nays
    17. Number Northern Democrat Yeas
    18. Number Northern Democrat Nays
    19. Number Southern Democrat Yeas (11 States of Confederacy plus KY and OK)
    20. Number Southern Democrat Nays

Suggestions

- Write two reasonably formal statements that use Δ as either clustering or classification problems. Is the data amenable to both clustering and classification or is one method more appropriate?

    1. Is this a large data set?
    2. Discuss the number of attributes with the datas size
    3. For each attribute present a breakdown of the active domain values.
    4. Present some analysis of the individual attributes as well as pairs. Are there correlations of some kind?

- Discuss missing values. Perhaps use one of the three methods to replace missing data.Assess the significance of either keeping or removing the tuples with unknown data. Is the amount of missing data significant?

References

[1] L. Torgo. Data Mining with R Learning with Case Studies. Chapman & Hall, 2010.