Dr. Eick

COSC 6335*“Data Mining”*

ProblemSet3 Fall 2020

Reviewing Data Mining Papers and Clustering

Deadlines; Task 7 is due Monday, November 23, 11p

All other tasks are due: Thursday, December 3, 11p

Last updated: November 18, 11a

Task 7: Reviewing a Data Mining Paper

Peer-Reviewed Group Project

In this task you will review an influential data mining paper which was published in 2008 and won the 2017 *IEEE ICDM 10-Year Highest-Impact Paper Award*: Collaborative Filtering for Implicit Feedback Datasets by Yifan Hu, Yehuda Koren and Chris Volinsky, ICDM '08, pp. 263-272.

After you understood what this paper is all about, conduct a web search looking for papers with similar themes as the paper you need to review. Next, write your review, complying with the following review template:

1. Summarize what the research area and the topic of the paper is and what its contributions are 3-4 paragraphs); *write in a neutral or positive tone no matter how bad the paper is*!
2. Evaluate the contributions of the paper and its writing (at 3-6 paragraphs); follow the questions and criteria of the KDD 2012 Reviewing Criteria (you can find at the end of this document as an appendix). In particular, assess the **novelty**, **technical quality**, **significance and impact**, and **clarity of writing of the paper**. If the paper makes contributions that do not fit into these 4 criteria, summarize those in an optional “*other contributions” paragraph*)
3. What are the 3 strongest points of the paper (just one sentence for each point)?
4. What are the 1-3 weakest points of the paper (just one sentence for each point)?
5. Assess the educational value of the paper for graduate students (2-3 paragraphs)! Is the paper a good starting point to do work/research in the area? Does the paper do a good job in introducing the goals and objectives and the methods of the field of research? Does the paper do a good job in getting graduate students excited about working in the research field? What did you learn from reading the paper?
6. Numbered List of Specific Comments and Questions (e.g. *the claim stated in the second paragraph is not clear; I do not agree with the conclusion in the third paragraph…, symbol x was never defined, it is not clear to me what the purpose of Section 4.3.2 is; the author introduced formulas 2.4 that are never used in the remainder of the paper*, I do not understand what the term x means,…). Each review should have 4-7 specific questions/comments!
7. Summarize the findings of the web-search you conducted; also evaluate what you found and how it relates to the paper your reviewed (3-6 paragraphs)
8. Broader Impact (2-3 paragraphs); what real world applications will arise from this work? Assess how the paper will help society to make earth a better place! Does the paper foster new research/new approaches that could be investigated in future research? Does it establish new connections between different, originally disconnected research communities?
9. The paper won the ICDM 10-year Highest Impact Award. Explain why, you believe, this happened! (1-2 paragraphs)
10. Give the paper a numerical score (1-7) using the KDD-2012 Criteria (described in appendix 1); 7 scores (add scores for educational value, broader impact and overall score!)!
11. Assess the usefulness of Task7! (1 paragraph)!

Please comply with the indicated length requirements in your review; violating them will result in penalties. Assume that a paragraph has 4-10 sentences.

Task 8: Clustering with K-Means, EM and DBSCAN

Non-Peer Reviewed Individual Task

Second Draft

**Learning Objectives**:

1. Learn to use popular clustering algorithms, namely K-means, EM and DBSCAN
2. Learn how to summarize and interpret clustering results
3. Learn to write analysis and evaluation functions which operate on the top of clustering algorithms and clustering results
4. Learning how to interpret unsupervised data mining results

**Datasets**: In the project we will use the Complex8 and the z-scored cleaned Pima Indians Diabetes dataset we used for some Task1 subtasks in ProblemSet1. The Complex8 dataset is a 2D dataset which can be found at <http://www2.cs.uh.edu/~ceick/DM/Complex8.csv> and cleaned Pima Indian Diabetes is an 8D dataset; the last attribute of each dataset denotes a class variable which should be ignored when clustering the data sets—however, the class variable will be used in the post analysis of the clusters generated by running K-means, EM and DBSCAN.

**Task 8 Subtasks:**

1. Write an function[[1]](#footnote-1) purity(a,b,outliers=FALSE) that computes the purity of a clustering result based on an apriori given set of class lables, where *a* gives the assignment of objects in O to clusters, and *b* is the “ground truth”. Purity is defined as follows: Let

O be a dataset

X={C1,…,Ck} be a clustering of O with Ci ⊆O (for i=1,…,k), C1∪…∪Ck ⊆O and Ci∩Cj=∅ (for i≠ j)

PUR(X)= (number\_of\_majority\_class\_examples(X)/(total\_number\_examples\_in\_clusters(X))

If the used clustering algorithm supports outliers, outliers should be ignored in purity computations; if you use R-clustering algorithms, you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. If the parameter outliers is set to FALSE, the function just returns a floting point number of the observed purity, if parameter outliers is set to T the function returns a vector: (<purity>,<percentage\_of\_outliers); e.g. if the function returns (0.98, 0.2) this would indicate that the purity is 98%, but 20% of the objects in dataset O have been classified as outliers.\*

1. Run K-means for k=8 and k=12 twice for the Complex8 dataset. Compare the obtained four clusterings! Also compute their purity using the function you developed in Task a. \*\*
2. Run K-means and EM for k=3 for the z-scored cleaned Pima Indian dataset. Compute the purity of the obtained clustering results; also create box plots for all 8 attributes of the obtained 3 clusters for each clustering and report their centroids, means, covariance matrix, and prior of the EM results. Compare the two clusterting results. Finally, summarize based on the obtained boxplots and centroids/cluster means what kind of objects each of the 2x3=6 clusters contains. \*\*\*\*\*
3. Try to obtain a DBSCAN clustering for the z-scored cleaned Pima Indian dataset, having between 2 and 15 clusters with less than 20% outliers. Report its purity! \*\*
4. Develop a search procedure that looks for the “best” clustering by exploring different settings for the (MinPoints, epsilon) parameters of DBSCAN for the Complex8 dataset. The procedure maximizes purity of the obtained clustering, subject to the following constraints:
   * 1. There should be between 2 and 15 clusters
     2. The number of outliers should be 10% or less.

The procedure returns the “best” DBSCAN clustering found and the accomplished purity as its result[[2]](#footnote-2); please limit the number of tested (MinPoints, epsilon)-pairs to 3000 in your implementation! Explain how your automated parameter selection method works and demonstrate your automated procedure using an example!

Apply the procedure you developed to the Complex8 datasets and report the best clustering you found. Are you happy with the obtained solution? \*\*\*\*\*

If you did not succeed in writing the function that seeks of the optimal DBSCAN clustering, you can manually seek for the best clustering for the Complex 8 dataset and report it; also report how you searched manually for it. \*\*

Extra credit: Apply your search procedure also to the z-scored cleaned Pima Indian Diabetes Dataset and report the clusters of the best result and what purity you accomplished. \*\*

Remark: Number of \*’s indicate a tentative assessment of the number of points allocated to each subtask: more \*’s mean more points!

**Deliverables for Task 8:**

1. A Report[[3]](#footnote-3) which contains all deliverables for the 5 subtasks of Task 8.
2. An Appendix which describes how to run the procedure that you developed for Task e, if you developed such a procedure.
3. An Appendix which contains the softwar/code you developed as part of Task 8.

**Appendix 1: KDD 2012 Reviewing Criteria: Research Track**

Below we have provided some guidelines to reviewers on how to write reviews, both the content of reviews and also how the numerical scoring system works. Many of the suggestions below have been liberally borrowed from other conferences - so thanks to the many folks who have contributed to writing these types of "guidance" pages in the past.

**Writing Reviews: Content** (Edited by Ch. Eick)

For each paper you will provide written comments under each of the headings below. Your review should address **both the strengths and weaknesses of the paper** - identify the areas where you believe the paper is particularly strong and particularly weak - this will be very valuable to the PC Chairs and the SPC.

* **Novelty**: This is arguably the single most important criterion for selecting papers for the conference. Reviewers should reward papers that propose genuinely new ideas or novel adaptations/applications of existing methods. It is not the duty of the reviewer to infer what aspects of a paper are novel - the authors should explicitly point out how their work is novel relative to prior work. Assessment of novelty is obviously a subjective process, but as a reviewer you should try to assess whether the ideas are truly new, or area novel combinations or adaptations or extensions of existing ideas, or minor extensions of existing ideas, and so on.
* **Technical Quality**: Are the results sound? Are there obvious flaws in the conceptual approach? Did the authors ignore (or appear unaware of) highly relevant prior work? Are the experiments well thought out and convincing? Are there obvious experiments that were not carried out? Will it be possible for later researchers to replicate these results? Are the data sets and/or code publicly available? Did the authors discuss sensitivity of their algorithm/method/procedure to parameter settings? Did the authors clearly assess both the strengths and weaknesses of their approach?
* **Potential Impact and Significance**: Is this really a significant advance in the state of the art? Is this a paper that people are likely to read and cite in later years? Does the paper address an important problem (e.g., one that people outside machine learning and data mining are aware of) or just a problem that only a few researchers are interested in and that wont have any lasting impact? Is this a paper that researchers and/or practitioners might find useful 5 or 10 years from now? Is this work that can be built on by other researchers?
* **Clarity of Writing**: Please make full use of the range of scores for this category so that we can identify poorly-written papers early in the process. Is the paper clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc? It is the responsibility of the authors of a paper to write clearly, rather than it being the duty of the reviewers to try to extract information from a poorly written paper. Do not assume that the authors will fix problems before a final camera-ready version is published - unlike journal publications, there will not be time to carefully check that accepted papers are properly written. Think of future readers trying to extract information from the paper - it may be better to advise the authors to revise a paper and submit to a later conference, than to accept and publish a poorly-written version.
* **Additional Points (optional)**: this is an optional section on the review form can be used to add additional comments for the authors that don’t naturally fit into any of the areas above.

**Comments that are only for the SPC and PC (optional)**: again this is another optional section. If there are any

**General Advice on Review Writing**: please be as precise as you can in your comments to the authors and avoid vague statements. Your criticism should be constructive where possible - if you are giving a low score to a paper then try to be clear in explaining to the authors the types of actions they could take to improve their paper in the future. For example, if you think that this work is incremental relative to prior work, please cite the specific relevant prior work you are referring to. Or if you think the experiments are not very realistic or useful, let the author(s) know what they could do to improve them (e.g., more realistic data sets, larger data sets, sensitivity analyses, etc).

**Writing Reviews: Numerical Scoring**

For KDD-2012 we are using a 7-point scoring system. We strongly encourage you to use the full range of scores, if appropriate for your papers. Try not to put all of your papers in a narrow range of scores in the middle of the scale, e.g., 3s, 4s, and 5s. Don’t be afraid to assign 1s/2s, or 6s/7s, if papers deserve them. If you are new to the KDD conference (or have not attended for a number of years) you may find it useful to take a look at online proceedings from recent KDD conferences to help calibrate your scores. The scoring system is as follows:

* 7: An excellent paper, a very strong accept. I will fight for acceptance, this is potentially best-paper material.
* 6: A very good paper, should be accepted. I vote and argue for acceptance, clearly belongs in the conference.
* 5: A good paper overall, accept if possible. I vote for acceptance, although would not be upset if it were rejected because of the low acceptance rate.
* 4: Decent paper, but may be below KDD threshold I tend to vote for rejecting it, but could be persuaded otherwise.
* 3: An OK paper, but not good enough. A rejection. I vote for rejecting it, although would not be upset if it were accepted.
* 2: A clear rejection. I vote and argue for rejection. Clearly below the standards for the conference.
* 1: A strong rejection. I'm surprised it was submitted to this conference. I will actively fight for rejection.

1. This function could be an R-function, a Python function or any other function. You might find some implementation of this function online; it is okay to use those implementations, as long as you acknowledge in your report what you use, and not all software you find on the internet is running properly. [↑](#footnote-ref-1)
2. It should report the number of clusters obtained and the percentage of outliers as well. [↑](#footnote-ref-2)
3. Single-spaced; please use an 11-point or 12-point font! [↑](#footnote-ref-3)