- Published: 09 July 2020

OsamorSoft: clustering index for comparison and quality validation in high throughput dataset

- Ifeoma Patricia Osamor &
- Victor Chukwudi Osamor

Journal of Big Data **volume 7**, Article number: 48 (2020) Cite this article

Abstract

The existence of some differences in the results obtained from varying clustering k-means algorithms necessitated the need for a simplified approach in validation of cluster quality obtained. This is partly because of differences in the way the algorithms select their first seed or centroid either randomly, sequentially or some other principles influences which tend to influence the final result outcome. Popular external cluster quality validation and comparison models require the computation of varying clustering indexes such as Rand, Jaccard, Fowlkes and Mallows, Morey and Agresti Adjusted Rand Index (ARI_{MA}) and Hubert and Arabie Adjusted Rand Index (ARI_{HA}). In literature, Hubert and Arabie Adjusted Rand Index (ARI_{HA}) has been adjudged as a good measure of cluster validity. Based on ARI_{HA} as a popular clustering quality index, we developed *OsamorSoft* which constitutes *DNA_Omatrix* and *OsamorSpreadSheet* as a tool for cluster quality validation in high throughput analysis. The proposed method will help to bridge the yawning gap created by lesser number of friendly tools available to externally evaluate the ever-increasing number of clustering algorithms. Our

implementation was tested alongside with clusters created with four k-means algorithms using malaria microarray data. Furthermore, our results evolved a compact 4-stage *OsamorSpreadSheet* statistics that our easy-to-use GUI java and spreadsheet-based tool of *OsamorSoft* uses for cluster quality comparison. It is recommended that a framework be evolved to facilitate the simplified integration and automation of several other cluster validity indexes for comparative analysis of big data problems.

Introduction

Given dataset points X_n as genes, $x_1, x_2, x_3, \ldots, x_n$, in d dimensional space say R^d, clustering process can be clearly stated as thus:

We are required to find partition subsets X_1, X_2, X_3,…, $X_k \forall\forall x_i$, $i = 1,2,3,\ldots,.n$, such that every gene falls into one of the subsets and no x_i falls into two or more subsets.
Partitions X_1, X_2, X_3,…, X_k satisfy the following: $X_1 \cup\cup X_2 \cup\cup X_3 \ldots \cup\cup X_k = X$ and $X_i \cap\cap X_j = 0 \ \forall\forall\ i \neq\neq j$, where $\cup\cup$ represents union and $\cap\cap$ represents intersection.
In addition, we cluster to form subsets with the goal that data points x_i that are similar as much as possible belongs to same group. This require a similarity measure (or dissimilarity measure) usually given in form of values to represent the degree of resemblance or natural association between one data and another [1,2,3,4,5]. The converse indicates dissimilarity measure ρ which satisfies the following condition:

MathSciNet MATH Google Scholar

1. Batagelj V, Bren M. Comparing resemblance measures. J Classif. 1995;12(1):73–90.

MathSciNet MATH Google Scholar

2. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. A local search approximation algorithm for k-means clustering. Comput Geom. 2004;28(2–3):89–112.

[MathSciNet](#) [MATH](#) [Google Scholar](#)

3. Albatineh AN, Niewiadomska-Bugaj M, Mihalko D. On Similarity indices and correction for chance agreement. J Classif. 2006;23(2):301–13.

[MathSciNet](#) [MATH](#) [Google Scholar](#)

4. Milligan GW, Cooper MC. A Study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behav Res. 1986;21(4):441–58.

[Google Scholar](#)

5. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 1999;9(11):1106–15.

[Google Scholar](#)

6. Tamayo P, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci. 1999;96(6):2907–12.

[Google Scholar](#)

7. Tseng VS, Kao CP. Efficiently mining gene expression data via a novel parameterless clustering method. IEEE/ACM Trans Comput Biol Bioinform. 2005;2(4):355–65.

[Google Scholar](#)

8. Friedler SA, Mount DM. Approximation algorithm for the kinetic robust K-center problem. Comput Geom. 2010;43(6–7):572–86.

MathSciNet MATH Google Scholar

9. Fahim AM, Salem AM, Torkey FA, Ramadan MA. An efficient enhanced k-means clustering algorithm. J Zhejiang Univ Sci A. 2006;7(10):1626–33.

MATH Google Scholar

10. Gerso A, Gray RM. Vector quantization and signal compression. 1992;159.
11. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37.

Google Scholar

12. Scott AJ, Symons MJ. Clustering methods based on likelihood ratio criteria. Biometrics. 1971;27(2):387–97.

Google Scholar

13. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. Pattern Anal Mach Intell IEEE Trans. 1997;19(2):153–8.

Google Scholar

14. Marriott FHC. Practical problems in a method of cluster analysis. Biometrics. 1971;27(3):501–14.

MathSciNet Google Scholar

15. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998;95(25):14863–8.

Google Scholar

16. Cho RJ, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 1998;2(1):65–73.

MathSciNet Google Scholar

17. Chu S, et al. The transcriptional program of sporulation in budding yeast. Science. 1998;282(5389):699–705.

Google Scholar

18. Wen X, et al. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA. 1998;95(1):334–9.

Google Scholar

19. Osamor VC, Adebiyi EF, Oyelade JO, Doumbia S. Reducing the time requirement of k-means algorithm". PLoS ONE. 2012;7:12.

Google Scholar

20. D'Argenio V. The high-throughput analyses era: are we ready for the data struggle? High Throughput. 2018;7:1. https://doi.org/10.3390/ht7010008.

MathSciNet Article Google Scholar

21. Krieger AM, Green PE. A generalized rand-index method for consensus clustering of separate partitions of the same data base. J Classif. 1999;16(1):63–89.

Google Scholar

22. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. Clustering algorithms: a comparative approach. PLoS ONE. 2019;14:1. https://doi.org/10.1371/journal.pone.0210236.

Article Google Scholar

23. Hämäläinen J, Jauhiainen S, Kärkkäinen T. Comparison of internal clustering validation indices for prototype-based clustering. Algorithms. 2017;10:3. https://doi.org/10.3390/a10030105.

MathSciNet Article MATH Google Scholar

24. Pirim H, Ekşioğlu B, Perkins A, Yüceer C. Clustering of high throughput gene expression data. Comput Oper Res. 2012;39(12):3046–61. https://doi.org/10.1016/j.cor.2012.03.008.

MathSciNet Article MATH Google Scholar

25. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846.

Google Scholar

26. Morey LC, Blashfield RK, Skinner HA. A comparison of cluster analysis techniques withing a sequential validation framework. Multivariate Behav Res. 1983;18(3):309–29.

Google Scholar

27. Morey LC, Agresti A. The measurement of classification agreement: an adjustment to the rand statistic for chance agreement. Educ Psychol Meas. 1984;44(1):33–7.

Google Scholar

28. Steinley D. Properties of the hubert-arabie adjusted rand index. Psychol Methods. 2004;9(3):386–96.

Google Scholar

29. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218.

MATH Google Scholar

30. Warrens MJ. On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. J Classif. 2008;25(2):177–83.

MathSciNet MATH Google Scholar

31. Llet R, Ortiz MC, Sarabia LA, Sánchez MS. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Anal Chim Acta. 2004;515(1):87–100.

Google Scholar

32. Milligan GW. A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika. 1981;46(2):187–99.

MATH Google Scholar

33. Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybern. 1974;4(1):95–104.

MathSciNet MATH Google Scholar

34. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.

MATH Google Scholar

35. McClain JO, Rao VR. Clustisz: a program to test for the quality of clustering of a set of objects. J Mark Res. 1975;12(4):456–60.

Google Scholar

36. Saltstone R, Stange K. A computer program to calculate Hubert and Arabie's adjusted rand index. J Classif. 1996;13(1):169–72.

Google Scholar

37. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc. 1983;78(383):553–69.

MATH **Google Scholar**

38. Yeung KY, Ruzzo WL. Details of the adjusted Rand index and clustering algorithms, supplement to the paper 'An empirical study on principal component analysis for clustering gene expression data. Bioinformatics. 2001;17(9):763–74.

Google Scholar

39. Santos JM, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification. Berlin: Springer; 2009.

Google Scholar

40. Alonso-Betanzos A, Bolón-Canedo V, Morán-Fernández L, Sánchez-Maroño N. A review of microarray datasets: where to find them and specific characteristics. Methods Mol Biol. 2019;1986:65–85. https://doi.org/10.1007/978-1-4939-9442-7_4.

Article **Google Scholar**

41. Rogers LRK, de los Campos G, Mias GI. Microarray gene expression dataset re-analysis reveals variability in influenza infection and vaccination. Front Immunol. 2019;10:2616. https://doi.org/10.3389/fimmu.2019.02616.

Article **Google Scholar**

42. Osamor V, Adebiyi E, Doumbia S. Comparative functional classification of *Plasmodium falciparum* genes using k-means clustering, in computer science and information technology-spring conference, 2009. IACSITSC'09. International Association of. 2009; 491–495.

43. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.

Google Scholar

44. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.

Google Scholar

45. Karmakar B, Das S, Bhattacharya S, et al. Tight clustering for large datasets with an application to gene expression data. Sci Rep. 2019;9:3053. https://doi.org/10.1038/s41598-019-39459-w.

Article **Google Scholar**

46. Shirkhorshidi AS, Aghabozorgi S, Wah TY. A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS ONE. 2015;10(12):e0144059. https://doi.org/10.1371/journal.pone.0144059.

Article **Google Scholar**

47. Zhang Z, Fang H. Multiple-vs non-or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. In 2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE). 2016; 219–228.
48. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 2003;1(1):5.

Google Scholar

49. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of

Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol. 2003;4(2):R9.

Google Scholar

50. Roch KG, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 2003;301(5639):1503–8.

Google Scholar

51. Xu Q, Zhang Q, Liu J, Luo B. Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Syst Appl. 2020;151:113367.

Google Scholar

52. Wang H, Mahmud MS, Fang H, Wang C. Wireless Health, SpringerBriefs in Computer Science. 2016; 30

Download references

Acknowledgements

I wish to thank Covenant University for their support towards the publication expenses of this manuscript. Also, I gladly appreciate Akinnusi Opeyemi for his support during the preparation of this manuscript.

Funding

Covenant University is funding the cost of the publication.

Author information

Affiliations

1. **Department of Accounting, Faculty of Management Sciences, Lagos State University, Ojo Campus, Lagos, Nigeria**
   Ifeoma Patricia Osamor
2. **Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria**

Victor Chukwudi Osamor

Contributions

VCO initiated the idea of the work while IPO and VCO did the experiment and wrote the manuscript. All authors read and approved the final manuscript.

Corresponding author

Correspondence to Victor Chukwudi Osamor.

Ethics declarations

Competing interests

The authors do not have any competing interest.

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions

About this article

Cite this article

Osamor, I.P., Osamor, V.C. OsamorSoft: clustering index for comparison and quality validation in high throughput dataset. *J Big Data* **7,** 48 (2020). https://doi.org/10.1186/s40537-020-00325-6

Download citation

- Received02 March 2019
- Accepted02 July 2020
- Published09 July 2020
- DOIhttps://doi.org/10.1186/s40537-020-00325-6

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Provided by the Springer Nature SharedIt content-sharing initiative

Keywords

- ***Clustering index***
- ***Algorithms***
- ***OsamorSoft***
- ***Validation***
- ***Rand***
- ***Automation***

- Sections
- Figures
- References