Enliven Archive
Intensifying Thoughts

Review Article www.enlivenarchive.org Enliven: Surgery and Transplantation

ISSN: 2379-5719

Discovering the Cancer Disease Subtype using Integrative Sparse K-Means to Enhance Pharmacotherapy Treatment

George Benson

Department of Surgery, International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur, Malaysia

*Corresponding author: George Benson, Department of Surgery, International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur, E-mail: benson_g@tutanota.com

Received Date: 16th September 2018
Accepted Date: 1st November 2018
Published Date: 2nd November 2018

Citation: George B (2018) Discovering the Cancer Disease Subtype using Integrative Sparse K-Means to Enhance Pharmacotherapy Treatment. Enliven: Surg Transplant 5(1): 002.

The ability to provide initialed or individualized medicine to cancer patient largely depend on the ability to discover the different cancer subtypes. As advanced biological knowledge databases and multi-level omics datasets continue to accumulate, there is urgent need to integrate the current extensive biological knowledge with the omics data to decipher a natural mechanism that is responsible for deadly diseases [1]. The researchers sought to recommend an integrative sparse K-Means (IS-Kmeans) method to facilitate discovery of subtypes of diseases and cancer drugs mostly guided by past biological knowledge. Additionally, to achieve quick optimization, they relied on an algorithm that used an alternating direction method of multiplier (ADMM) [2]. The researchers made a comparison of the IS-Kmeans using three actual applications and simulations together with the current methods to illustrate its computational effectiveness, purposeful annotation of identified molecular features, feature selection, and high accuracy while performing clustering [3]. Despite being regarded as a single type of disease, contemporary transcriptomic investigations prove that each cancer category might include numerous sub-categories characterized by diverse response to treatment, rates of survival, and disease mechanism [4]. The cancer subtypes that have been comprehensively investigated include the ovarian, colorectal, and breast cancers as well as glioblastoma, lymphoma, and leukemia [5]. Since they demonstrate diverse outcomes and respond differently to treatment, these cancer subtypes usually have a strong clinical significance [6]. Nonetheless, a single omics or cohort analysis, for instance, the transcriptome usually is characterized by reproducibility and sample size limitation concerns.

Lately, there has been a massive accumulation of huge volume of omics data in public depositories and databases such as the Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) among others [7]. Such sets of data offer extraordinary opportunities to disclose mechanisms of cancer since they combine multiple-level omics data types and numerous cohorts [8]. There have been tremendous successes which have been achieved in several applications which use omics integrative analysis [9]. Alternatively, the public databases have accumulated an incredible volume of biological information [10]. A suitable utilization of such past details for instance miRNA targeting gene database and pathway information can greatly help to provide guidance to ensure the omics integrative analysis is appropriately modeled [11].

In this case, different forms of clustering methods have been applied to achieve high-throughput experimental data such as microarray to establish the new subtypes of the disease [12]. Good examples of such methods include the mixture model-based techniques, nonparametric methods for assessment of single transcriptomic research, K-means clustering, and hierarchical clustering [13]. The reliability of the clustering analysis has been enhanced through the use of resampling and ensembling techniques [14]. The researchers used the iCluster method which integrates omics data to conduct clustering of cancer samples as well as a baseline method to make a comparison of the study [15]. The researchers are focused on identifying cancer sub-categories by using current biological knowledge to improve clarification and precision as well as concurrently perform integration of multi-level omics datasets [16].

The researchers are confident that discovering the cancer subtype is a critical step towards making sure that the cancer patients receive personalized treatment [17]. The current biological information is perfectly integrated into the IS-Kmeans model and subsequent sparse features can additionally be utilized to describe the features of the cancer sub-category in medical application [18]. The proposed model has several benefits, for instance, the integrative evaluation enhances the understandable regulatory flow, statistical power, and clustering precision between the diverse omics datasets [19]. Furthermore, the researchers used the overlapping tosso to take into consideration the current biological information [20]. The full use of external biological and inter-omics regulatory information greatly improved the analysis and precision of the cancer sub-category results [21]. The study findings determined that the adoption of the IS-Kmeans model is computational effective given the ADMM'S closed-form iteration updates as well as K-Means' EM algorithm [22].

The researchers took less not more than fifteen minutes to compute seven hundred subjects and 15,000 omics data over on a normal personal computer which contained a solitary computing thread; alternatively, the iCluster took nearly four hours [23]. Furthermore, they established that the IS-Kmeans led to superior interpretations of the sparse features [24]. Nonetheless, it was determined that it had several limitations, for instance, the current biological information is susceptible to inaccuracies and the user can update them more frequently [25]. In addition, there is a possibility of including false biological details which might weaken the information which is contained in the data as well as produce compromised study results.

Comparison of different methods using metabric breast cancer (K=5). G2 represents feature groups (gene symbols) where all two types of features are selected. The clustering result is compared with PAM50 subtype definition in terms of ARI. Survival p-value obtained from the log rank test is given for the clustering assignment for each method.

Method	ARI	nfeature	G1	G2	p-value	Time
ISKmeans	0.233	1882	1494	194	8.29×10^{-17}	38.4 mins
SparseKmeans	0.22	2004	2004	0	3.04×10^{-13}	34.3 mins
iCluster	0.0572	2471	2471	0	0.143	11.8 hours

Table 1: Comparison of different methods using metabric breast cancer (Source: Tseng GC, Wing HW (2005) Tight Clustering: A Resampling-Based Approach for Identifying Stable and Tight Patterns in Data. Biometrics 61: 10-16.)

Reference

1. Huo Z, Tseng G (2017) Integrative Sparse K-Means with Overlapping Group Lasso in Genomic Applications for Disease Subtype Discovery. Ann Appl Stat 11: 1011.
2. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
3. Fan X, Kurgan L (2014) Comprehensive Overview and Assessment of Computational Prediction of Microrna Targets in Animals. Brief Bioinfor 16: 780-794.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286: 531-537.
5. Laurent Jacob, Guillaume Obozinski, Jean-Philippe Vert (2009) Group Lasso with Overlap and Graph Lasso. Proceedings of The 26th Annual International Conference On Machine Learning: 433-440
6. He BS, Yang H, Wang SL (2000) Alternating Direction Method with Self-Adaptive Penalty Parameters for Monotone Variational Inequalities. Journal of Optimization Theory and Applications 106: 337-356.
7. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
8. Hubert L, Arabie P (1985) Comparing Partitions. Journal of Classification 2: 193-218.
9. Huo Z, Tseng G (2017) Integrative Sparse K-Means with Overlapping Group Lasso in Genomic Applications for Disease Subtype Discovery. Ann Appl Stat 11: 1011.
10. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
11. Laurent Jacob, Guillaume Obozinski, Jean-Philippe Vert (2009) Group Lasso with Overlap and Graph Lasso. Proceedings of The 26th Annual International Conference On Machine Learning 433-440
12. Kaufman L, Peter R (2009) Clustering by Means of Medoids. Hoboken, NJ: John Wiley & Sons.
13. Kaufman L, Rousseeuw PJ (2009) Finding Groups in Data: An Introduction To Cluster Analysis. Hoboken, NJ: John Wiley & Sons.
14. McLachlan GJ, Bean RW, Peel D (2002) A Mixture Model-Based Approach to the Clustering of Microarray Expression Data. Bioinformatics 18: 413-422.
15. Kim EY, Kim SY, Ashlock D, Nam D (2009) MULTI-K: Accurate Classification of Microarray Subtypes Using Ensemble K-Means Clustering. BMC Bioinformatics 10: 260.
16. Maitra R, Ramler IP (2009) Clustering in the Presence of Scatter. Biometrics 65: 341-352.
17. Richard D Baird, Carlos Caldas (2013) Genetic Heterogeneity in Breast Cancer: The Road to Personalized Medicine?. BMC Medicine 11: 151.
18. Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, et al. (2008) An International Standardization Programme Towards the Application of Gene Expression Profiling in Routine Leukaemia Diagnostics: The Microarray Innovations in Leukemia Study Prephase. Br J Haematol 142: 802-807.
19. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, et al. (2011) Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies. J clin invest 121: 2750-2767.
20. Tseng GC, Wong WH (2005) Tight Clustering: A Resampling□Based Approach for Identifying Stable and Tight Patterns in Data. Biometrics 61: 10-16.
21. Tseng GC (2007) Penalized and Weighted K-Means for Clustering with Scattered Objects and Prior Information in High-Throughput Biological Data. Bioinformatics 23: 2247-2255.
22. Tibshirani R, Guenther W (2005) Cluster Validation by Prediction Strength. Journal of Computational and Graphical Statistics 14: 511-528.
23. Robert Tibshirani, Guenther Walther, Trevor Hastie (2001) Estimating the Number of Clusters in A Data Set Via the Gap Statistic. J R Statist Soc Series B 63: 411-423.
24. Richard S, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. Journal of the National Cancer Institute 95: 14-18.

25. Shen K, Tseng GC (2010) Meta-Analysis for Pathway Enrichment Analysis When Combining Multiple Genomic Studies. Bioinformatics 26: 1316-1323.
26. Verhaak RG, Wouters BJ, Erpelinck CA, Abbas S, Beverloo HB, et al. (2009) Prediction of Molecular Subtypes in Acute Myeloid Leukemia Based on Gene Expression Profiling. Haemtologica 94: 131-134.
27. Wang SL, Liao LZ (2001) Decomposition Method with a Variable Parameter for A Class of Monotone Variational Inequality Problems. Journal of Optimization Theory and Applications 109: 415-429.
28. George J, Lim SJ, Jang SJ, Cun Y, Ozretic L, et al. (2015) Comprehensive Genomic Profiles of Small Cell Lung Cancer, Nature 524: 47.