A Review of Use of Data Mining during COVID-19 Pandemic

Ankit Mehrotra[a], Reeti Agarwal[b]

[a]Jaipuria Institute of Management, Lucknow
[b] Jaipuria Institute of Management, Lucknow
[a] ankit.mehrotra@jaipuria.ac.in, [b] reeti.agarwal@jaipuria.ac.in

Article History: *Do not touch during review process(xxxx)*

Abstract: Data mining is one of the promising and continuously evolving fields in the arena of data analytics. Data mining has led to solutions of various unfathomable jobs, events, diseases and evaluations. The rich consortium of techniques that falls under the data mining domain makes it a formidable force for data scientists. The current paper reviews the various papers published on COVID-19 using data mining techniques to address the pandemic in terms of its explanation, assessment and solution. The current paper reviews the work done by various authors using data mining techniques. The paper contributes uniquely to the literature by filling up the gap of review on COVID-19 related work.

Keywords: COVID-19, data mining, review, pandemic, disease

1. Introduction

Data mining (DM) is mining of hidden patterns out of heaps of data spread around us (Dave Smith & Marlow, 2007; J. Han et al., 2011; Mishra et al., 2010; Witten et al., 2005; Zhu & Davidson, 2007). Data mining has a plethora of techniques with the capability to serve the organizations and various domains in a variety of ways. These techniques are what makes data mining applications a tool to study trends and assist in prediction ranging from human behaviours to emergence of a disease and subsequently aiding in finding a solution for these predictions. The various techniques which are of specific interest include clustering, classification, association, regression, summarization and text mining (J. Han et al., 2011). The data mining ability to extract meaningful information from complex raw data provides multiple benefits in the healthcare sector inclusive but not limited detection of drug abuse, diagnosis of patients, suggestions of treatments, early detection of diseases, survivability percentage and approach and the likes (Ogundele et al., 2018).

Data mining has been applied to various health related issues for a long time now and the same has resulted in favourable outcomes for health and medicine. Data mining has been traditionally used for classifying diseases and assisting in treatment and management of diseases (Ogundele et al., 2018). Voluminous data gets generated in the healthcare industry that needs proper storage (Varghese & Tintu, 2015). These data are subjected to various analytical techniques to make sense out of them and this is where data mining has been seen as the promising and result-oriented field. The promise inbuilt in data mining techniques to fight against the odds in any field led to calls for action by White House to various data mining research institutes and technology companies to devise a data mining strategy to fight against the novel Coronavirus breakout (Alimadadi et al., 2020).

The current paper reviews papers published on data mining approaches applied to study COVID-19 pandemic during 2020. The paper highlights the use of data mining and its application by the world to fight against unknown nemesis of SARS family.

2.A Brief on Data Mining and Healthcare

The techniques of data mining have been favored in medicine and have its wide application which has been studied by various authors (Gayathri et al., 2014; Shukla et al., 2014; Sultana et al., 2016). Kunwar et al. (2016) made use of DM techniques such as ANN and Naïve Bayes to study kidney diseases. Chaurasia and Pal (2017) made use of various classification techniques to investigate precision of breast cancer examination. Shakil et al. (2015) in his research pointed out that Naïve Bayes is a better prediction method for dengue disease survival. Shim and Xu (2003) proposed Bayesian Ying Yang (BYY), a classification method, to categorize liver diseases through programmed discovery of medical trends. Islam et al. (Islam et al., 2004) studied lung cancer using decision tree method by grouping of x-rays. Wang et al. (2005) made use of cluster and decision tree methods for classifying mammography into two classes. Cheng et al. (2006) made use of data mining techniques for classification cardiovascular diseases. Bethel et al. (2006) devised a rule based model through an association algorithm on the data of historical breast cancer patients. Bayesian Network was proposed for diagnosing Coronary Heart Diseases. Coronary Heart Disease was also studied by (Abraham et al., 2006; Su et al., 2001;

Xue et al., 2006). Cardiovascular diseases were studied by classification algorithms (Cheng et al., 2006; Karegowda & Jayaram, 2009; Tang & Tseng, 2009). Few authors studied diabetic diseases by applying genetic algorithms (Balakrishnan et al., 2008; Brameier & Banzhaf, 2001; Tang & Tseng, 2009; Xing et al., 2007).

3.Objectives Of The Study

- To review papers published in Scopus database for the year 2020 on COVID-19 which has applied data mining techniques.

4.Methodology

Scopus database was used to extract papers authored on COVID-19 using data mining techniques for the year 2020. A total of 178 papers were listed by Scopus database on searching the keywords data mining and COVID-19. From 178 papers listed by the database, the current study extracted the papers by applying two levels of filters for review in terms of purpose and application of data mining techniques. The first criteria was that the paper has been published in 2020 with specific reference to data mining techquies and second, it has been cited more than 5 times. These criteria led to filtering of 14 papers.

5. Review of Papers on Data mining and COVID-19

Abd-Alrazaq et al. (2020) in their study identified the issues shared by tweeps connected to COVID-19. The authors made use of a text mining approach of data mining to analyze the tweets downloaded over a period of February 2, 2020 March 15, 2020. The authors applied Twitter API, Tweepy Python library and PostgresSQL database as their method to perform sentiment analysis and topic modeling.

Alimadadi et al. (2020) suggested that AI and ML is the major tool to fight against COVID-19. In their article they specified that White House through technology and research companies approached the global AI community to develop and work on various techniques related to data mining skills to support COVID-19 based study for finding a solution to the pandemic.

Tasnim et al. (2020) in their study addressed the issue of rumors and misinformation surging during the COVID-19 pandemic period leading to furtherance of unfounded practices enhancing virus spread and masking the healthy behavior. Tasnim further advocated use of advanced application of data mining approach like natural language processing for detection and removal of non-scientific based online content.

Ayyoubzadeh et al. (2020) stressed that data mining algorithms can be used for studying and predicting spread and trends of outbreak of COVID-19 virus across the world. The author used LSTM (Linear regression and long short-term memory) model of data mining approach to study data downloaded from Google Trends website. Through their work, they showed that the search frequency included words like washing and sanitizing of hands and topics related to antiseptic use besides previous day incidence as being the most looked for incidence.

Franch-Pardo et al. (2020) advocated that an interdisciplinary perspective and approaches like data mining, web-based mapping and spatiotemporal analysis is needed to face the challenges posed by COVID-19 pandemic. Their study supports bibliographic queries and understanding of the evolution of tools used in managing the global pandemic.

Li et al. (2020) made use of Chinese microblogging platform Weibo to perform both quantitative and qualitative analysis on collected data. The authors made use of linear regression and content analysis methodologies of data mining to identify classification of news and user generated topics leading to insight on COVID-19 outbreak. The authors stressed on how social media analysis may lead to better understanding of spread of COVID-19.

Qin et al. (2020) suggested that to avert and arrest outbreak of COVID-19, it is imperative to estimate the number of new cases and confirmed cases and resultantly worked upon data collected from social media search indexes (SMSI) for dry cough, fever, chest distress, coronavirus and pneumonia during a period of 40 days. The authors made use of lagged series SMSI to predict new suspects of COVID-19 cases.

Kumar (2020) discussed the use of Artificial Intelligence, as part of modern technology apart from ML and NLP, in fighting with COVID-19 crisis at various levels based on medical data. The authors strongly advocated use of AI to identify, track and forecast outbreaks.

Han et al. (2020) explored public opinion by analyzing Sina-Weibo text by applying Latent Dirichlet Allocation (LDA) model - a topic extraction technique and Random Forest algorithm - a classification model. The microblogging site texts were analyzed in terms of space, time and content.

Marhl et al. (2020) made use of publication mining to extract common physiological contexts of investigating diabetes and COVID-19 simultaneously.

Ren et al. (2020) studied traditional Chinese medicine by making use of data mining and association network models to suggest potential treatment of COVID-19.

Huang et al. (2020) conducted data mining activities on 485 patients extracted through Sina Weibo who were suspects of confirmed cases of COVID-19. The study aimed at analyzing the suspected or confirmed cases of COVID-19 who sought help through Sina Weibo. The authors extracted 9878 posts during a period of February 3 to February 20 of 2020. The authors suggested that social media through data mining analysis could be a tool to provide them early help.

Amin et al. (2020) suggested use of classification model to aid in the process of COVID-19 drug discovery.

Sarker et al. (2020) through semi-automatic filtering curated reports positive test patients based on tweets extracted from twitter on COVID-19 related keywords. The authors mapped the extracted symptoms through UML (Unified Medical Language) and evaluated the results to the ones reported in previous studies.

The table below lists down the summary of techniques used by top cited papers studied in this study.

Table.1. Techniques - Frequency

Techniques	Frequency
Classification and Regression	(Amin et al., 2020; Ayyoubzadeh et al., 2020; X. Han et al., 2020; Li et al., 2020; Qin et al., 2020)
AI & ML	(Alimadadi et al., 2020; A. Kumar et al., 2020)
Text analysis and NLP	(Abd-Alrazaq et al., 2020; Franch-Pardo et al., 2020; X. Han et al., 2020; A. Kumar et al., 2020; Li et al., 2020; Marhl et al., 2020; Sarker et al., 2020; Tasnim et al., 2020)
Clustering	(S. Kumar, 2020)
Association	(Ren et al., 2020)

6.Conclusion

The paper focused on reviewing data mining related techniques used to study COVID-19 pandemic. The data mining techniques have played a vital role in the healthcare industry ranging from diagnosing diseases to suggesting cures. The world looked up to data scientists for exploring data mining techniques to study various patterns associated with novel virus as well as behavioral patterns of masses across the world and suggest a way forward to counter the disease. The current paper studied a few most frequently cited papers as on date of writing this paper and brings out the fact that various data mining techniques were used for studying different types of issues associated with COVID-19 ranging from prevention mechanisms to solution finding, to studying sentiments. The current paper also specifies that data mining has been a preferred area for disease prediction or cure as per previous applications.

References

1. Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hai, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: A surveillance study. *Journal of Medical Internet Research*, *22*(4). Scopus. https://doi.org/10.2196/19016
2. Abraham, R., Simha, J. B., & Iyengar, S. S. (2006). A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier. *9th International Conference on Information Technology (ICIT'06)*, 235–236.
3. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., & Cheng, X. (2020). Artificial intelligence and machine learning to fight covid-19. *Physiological Genomics*, *52*(4), 200–202. Scopus. https://doi.org/10.1152/physiolgenomics.00029.2020
4. Amin, S. A., Ghosh, K., Gayen, S., & Jha, T. (2020). Chemical-informatics approach to COVID-19 drug discovery: Monte Carlo based QSAR, virtual screening and molecular docking study of some in-house molecules as papain-like protease (PLpro) inhibitors. *Journal of Biomolecular Structure and Dynamics*. Scopus. https://doi.org/10.1080/07391102.2020.1780946

5. Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Niakan Kalhori, S. R. (2020). Predicting COVID-19 incidence through analysis of Google trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health and Surveillance*, *6*(2). Scopus. https://doi.org/10.2196/18828
6. Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., & Samikannu, R. (2008). SVM ranking with backward search for feature selection in type II diabetes databases. *2008 IEEE International Conference on Systems, Man and Cybernetics*, 2628–2633.
7. Bethel, C. L., Hall, L. O., & Goldgof, D. (2006). Mining for implications in medical data. *18th International Conference on Pattern Recognition (ICPR'06)*, *1*, 1212–1215.
8. Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, *5*(1), 17–26.
9. Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol*, *2*.
10. Cheng, T.-H., Wei, C.-P., & Tseng, V. S. (2006). Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 165–170.
11. Dave Smith, S. A. S., & Marlow, U. K. (2007). *Data Mining in the Clinical Research Environment*. PhUSE.
12. Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment*, *739*, N.PAG-N.PAG. https://doi.org/10.1016/j.scitotenv.2020.140033
13. Gayathri, V., Mona, M. C., Chitra, S. B., & Chitra, S. B. (2014). A survey of data mining techniques on medical diagnosis and research. *International Journal of Data Engineering*, *6*(6), 301–310.
14. Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83–124.
15. Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *International Journal of Environmental Research and Public Health*, *17*(8). Scopus. https://doi.org/10.3390/ijerph17082788
16. Huang, C., Xu, X., Cai, Y., Ge, Q., Zeng, G., Li, X., Zhang, W., Ji, C., & Yang, L. (2020). Mining the Characteristics of COVID-19 Patients in China: Analysis of Social Media Posts. *Journal of Medical Internet Research*, *22*(5), N.PAG-N.PAG. https://doi.org/10.2196/19087
17. Islam, M. R., Chowdhury, M., & Khan, S. (2004). Medical image classification using an efficient data mining technique. *Complex 2004: Proceedings of the 7th Asia-Pacific Complex Systems Conference*, 34–42.
18. Karegowda, A. G., & Jayaram, M. A. (2009). Cascading GA & CFS for feature subset selection in medical data mining. *2009 IEEE International Advance Computing Conference*, 1428–1431.
19. Kumar, A., Gupta, P. K., & Srivastava, A. (2020). A review of modern technologies for tackling COVID-19 pandemic. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(4), 569–573. Scopus. https://doi.org/10.1016/j.dsx.2020.05.008
20. Kumar, S. (2020). Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis. *Annals of Data Science*, *7*(3), 417–425.

21. Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 300–305.
22. Li, J., Xu, Q., Cuomo, R., Purushothaman, V., & Mackey, T. (2020). Data mining and content analysis of the Chinese social media platform weibo during the early COVID-19 outbreak: Retrospective observational infoveillance study. *JMIR Public Health and Surveillance*, *6*(2). Scopus. https://doi.org/10.2196/18700
23. Marhl, M., Grubelnik, V., Magdič, M., & Markovič, R. (2020). Diabetes and metabolic syndrome as risk factors for COVID-19. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(4), 671–677. Scopus. https://doi.org/10.1016/j.dsx.2020.05.013
24. Mishra, D., Das, A. K., Mishra, M., & Mishra, S. (2010). Predictive data mining: Promising future and applications. *Int. J. of Computer and Communication Technology*, *2*(1), 20–28.
25. Ogundele, I. O., Popoola, O. L., Oyesola, O. O., & Orija, K. T. (2018). A Review on Data Mining in Healthcare. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*.
26. Qin, L., Sun, Q., Wang, Y., Wu, K.-F., Chen, M., Shia, B.-C., & Wu, S.-Y. (2020). Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *International Journal of Environmental Research and Public Health*, *17*(7). Scopus. https://doi.org/10.3390/ijerph17072365
27. Ren, X., Shao, X.-X., Li, X.-X., Jia, X.-H., Song, T., Zhou, W.-Y., Wang, P., Li, Y., Wang, X.-L., Cui, Q.-H., Qiu, P.-J., Zhao, Y.-G., Li, X.-B., Zhang, F.-C., Li, Z.-Y., Zhong, Y., Wang, Z.-G., & Fu, X.-J. (2020). Identifying potential treatments of COVID-19 from Traditional Chinese Medicine (TCM) by using a data-driven approach. *Journal of Ethnopharmacology*, *258*, N.PAG-N.PAG. https://doi.org/10.1016/j.jep.2020.112932
28. Sarker, A., Lakamana, S., Hogg-Bremer, W., Xie, A., Ali Al-Garadi, M., & Yang, Y.-. (2020). Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. *Journal of the American Medical Informatics Association*, *27*(8), 1310–1315.
29. Shakil, K. A., Anis, S., & Alam, M. (2015). Dengue disease prediction using weka data mining tool. *ArXiv Preprint ArXiv:1502.05167*.
30. Shim, J.-Y., & Xu, L. (2003). Medical data mining model for oriental medicine via BYY binary independent factor analysis. *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03.*, *5*, V–V.
31. Shukla, D. P., Patel, S. B., & Sen, A. K. (2014). A literature review in health informatics using data mining techniques. *International Journal of Software and Hardware Research in Engineering*, *2*(2), 123–129.
32. Su, J.-L., Wu, G.-Z., & Chao, I.-P. (2001). The approach of data mining methods for medical database. *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *4*, 3824–3826.
33. Sultana, M., Haider, A., & Uddin, M. S. (2016). Analysis of data mining techniques for heart disease prediction. *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1–5.
34. Tang, P.-H., & Tseng, M.-H. (2009). Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification. *2009 International Conference on Machine Learning and Cybernetics*, *5*, 3070–3075.

35. Tasnim, S., Hossain, M., & Mazumder, H. (2020). Impact of rumors and misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, *53*(3), 171–174. Scopus. https://doi.org/10.3961/JPMPH.20.094
36. Varghese, D. P., & Tintu, P. B. (2015). A survey on health data using data mining techniques. *International Research Journal of Engineering and Technology (IRJET)*, *2*(07), 2395–0056.
37. Wang, S., Zhou, M., & Geng, G. (2005). Application of fuzzy cluster analysis for medical image data mining. *IEEE International Conference Mechatronics and Automation, 2005*, *2*, 631–636.
38. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). Practical machine learning tools and techniques. *Morgan Kaufmann*, 578.
39. Xing, Y., Wang, J., & Zhao, Z. (2007). Combination data mining methods with new medical data to predicting outcome of coronary heart disease. *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 868–872.
40. Xue, W., Sun, Y., & Lu, Y. (2006). Research and application of data mining in traditional Chinese medical clinic diagnosis. *2006 8th International Conference on Signal Processing*, *4*.
41. Zhu, X., & Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.