How Data Mining is Helping in Cancer Research?

Subhasree Majumder
M.Tech Scholar, Department of Computer Science and Engineering, National Institute of Technology, Meghalaya
Correspondence to: Subhasree Majumder (subhasree.majumder24@gmail.com)

MOTIVATION

Cancer is the leading cause of deaths in India and worldwide, having doubled its rate of taking lives since 1990. According to a 2018 report by WHO's IARC, 18.1 million new cases of cancer has been registered with 9.6 million cancer deaths in 2018. [1]

A simple search of "breast cancer" produces more than 2.5 million search results. With this plethora of scientific publications available, and other data like pathological lab reports, clinical data, data mining gives us the tool to assist cancer research, be it in early detection of cancer, mining of latest publications on cancer treatments or cancer prediction without surgical invasion.

In this review, we try to give an overview of the data mining techniques being used in cancer research over the last ten years.

OBJECTIVE

- To show how Text Mining has been used to perform information extraction from medical reports and medical publications.
- To show how Data Mining has been used to gain insights from medical data and in early detection of cancer prediction.

MATERIALS AND METHOD

- We have used Pubmed, IEEE Explorer, ScienceDirect and Google Scholar to look into the latest data mining papers on Cancer research limiting the publishing year to last 10 years.

RESULT

- **EMR Records to Categorical Data using NER**

  There is a high degree of terminology variation in cancer domain such as breast cancer, mammary neoplasm, carcinoma of the breast. Named Entity Recognition, a Text mining method automatically recognizes all variants and map it to a single entity. Metamap is used to map biomedical texts to UMLS, a biomedical thesaurus.

- **Regular Expressions to extract information**

  Information extraction employs pattern matching to extract relevant information from laboratory reports. Cancer is an umbrella term for a number of neoplasmic conditions and hence pattern matching is designed domain specific. For example, prostrate cancer is determined by Gleason Scores (G1 and G2). Reports can have write ups like *"Gleason Grades 4+4, Gleeson score 8"* for which patterns can be coded. [2]

- **Mining Cancer Literature**

  Enormous publication in Cancer research, Text Mining can help us mine the most relevant literature using text classification, NLP and hypothesis generation. [3] IBM Watson Health and Mayo Clinic's cTAKES are the state-of-the-art products using NLP to extract information from patient data. Figure 1.

- **Decrease the number of unnecessary surgeries.**

  Cancer staging determination requires clinical report analysis and pathological report analysis that include high risk invasive biopsies. Data Mining can be used to find a correlation between clinical information and cancer staging. A study [4] was done that used a decision tree to select attributes from clinical dataset from The Cancer Genome Atlas such as gender, age, race, length of smoking, pack-year and reformed smoker and found out association rules to find out TNM (Tumor size – Lymph Nodes-Metastasis) staging.

- **Mining Cancer Biomarkers from Gene expression data**

  Cancer biomarkers are best bet to detect and target therapeutic treatments to patients. Again, the basic biomarker is a gene and its expression level. To validate genes linked to cancer, one has to analyze disease specific gene expression data in silica and reduce the number of molecules to be considered as potential biomarkers. Scientists can validate the same in vitro in lower costs. [5]

- **Mining Nutrigenomics Data to Find functional food for Cancer Treatment**.

  Nutrigenomics is the field of analyzing gene expression data produced from treating cell lines with bioactive compounds such as broccoli nutrients. A study [6] used clustering technique on gene expression data and found a signature of 18 genes that have similar gene expression level on certain food compounds. This genetic expression signature could represent benchmark for food compounds that can provide protection against Cancer.

Figure 1: Products using NLP to draw insights from patient data.

DISCUSSION

- In this poster presentation, we have tried to give a broad overview of the possible applications of Data Mining and Text Mining Techniques in Cancer Domain. Named Entity Recognition has been employed by state-of-the-art health analytics products like IBM Watson and Mayo Clinic's cTAKES for information extraction from EMR.
- Other Data Mining Techniques like decision trees, association rule, classification and clustering can be employed in drawing insights from Cancer datasets and predict Cancer staging as well.

CONCLUSIONS

- With the vast amount of medical data available and increased computational power, data mining can not only reduce cancer diagnosis time by mining through millions of cancer literature but also help in the early prediction of cancer in patients. Data Mining over the years has proved itself to be instrumental in cancer diagnosis, predictions and opening up new paradigms in cancer research

ACKNOWLEDGEMENT

REFERENCES

[1] Bray, Freddie, et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68.6 (2018): 394-424.

[2] Napolitano, G., Fox, C., Middleton, R., & Connolly, D. (2010). Pattern-based information extraction from pathology reports for cancer registration. Cancer Causes & Control, 21(11), 1887-1894.

[3] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., ... & Shen, B. (2013). Biomedical text mining and its applications in cancer research. Journal of biomedical informatics, 46(2), 200-211.

[4] Yang, H., & Chen, Y. P. P. (2015). Data mining in lung cancer pathologic staging diagnosis: correlation between clinical and pathology information. Expert Systems with Applications, 42(15-16), 6168-6176.

[5] Jurca, G., Addam, O., Aksac, A., Gao, S., Özyer, T., Demetrick, D., & Alhajj, R. (2016). Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. BMC research notes, 9(1), 236.

[6] Martín-Hernández, R., Reglero, G., & Dávalos, A. (2018). Data mining of nutrigenomics experiments: Identification of a cancer protective gene signature. Journal of Functional Foods, 42, 380-386.