Journal of Biomedical Informatics 61 (2016) 267–275

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Constraint based temporal event sequence mining for Glioblastoma survival prediction

Kunal Malhotra [a], Shamkant B. Navathe [a], Duen Horng Chau [a], Costas Hadjipanayis [b,c], Jimeng Sun [a,*]

[a] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[b] Department of Neurosurgery, Emory University School of Medicine, Atlanta, GA, USA
[c] Winship Cancer Institute of Emory University, Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 3 November 2015
Revised 5 March 2016
Accepted 25 March 2016
Available online 5 April 2016

Keywords:
Graph mining
Predictive model
Sequential pattern mining
Classification
Treatment patterns
Glioblastoma

ABSTRACT

Objective: A significant challenge in treating rare forms of cancer such as Glioblastoma (GBM) is to find optimal personalized treatment plans for patients. The goals of our study is to predict which patients survive longer than the median survival time for GBM based on clinical and genomic factors, and to assess the predictive power of treatment patterns.
Method: We developed a predictive model based on the clinical and genomic data from approximately 300 newly diagnosed GBM patients for a period of 2 years. We proposed sequential mining algorithms with novel clinical constraints, namely, 'exact-order' and 'temporal overlap' constraints, to extract treatment patterns as features used in predictive modeling. With diverse features from clinical, genomic information and treatment patterns, we applied both logistic regression model and Cox regression to model patient survival outcome.
Results: The most predictive features influencing the survival period of GBM patients included mRNA expression levels of certain genes, some clinical characteristics such as age, Karnofsky performance score, and therapeutic agents prescribed in treatment patterns. Our models achieved *c*-statistic of 0.85 for logistic regression and 0.84 for Cox regression.
Conclusions: We demonstrated the importance of diverse sources of features in predicting GBM patient survival outcome. The predictive model presented in this study is a preliminary step in a long-term plan of developing personalized treatment plans for GBM patients that can later be extended to other types of cancers.

1. Introduction

Glioblastoma (GBM) is the most lethal and biologically the most aggressive brain cancer with patients having a median survival of 12–15 months [10,29]. Understanding what factors prolong survival and promote treatment responses can be of value to patients and physicians. The Cancer Genome Atlas (TCGA) [17], a project of the National Institutes of Health (NIH), classified Glioblastoma patients into four distinct molecular subtypes affecting biological behaviors, suggesting that no single therapeutic regimen can be equally effective for all subtypes [6]. Patients with certain molecular subtypes may have greater overall survival than other patient subtypes, and analyzing gene expression levels, copy number variation (CNV), and mutations may give us information correlating to survival periods. The current standard of care for new GBM patients involves surgical resection followed by radiation therapy and chemotherapy with the oral alkylating agent Temodar [20]. Krex et al. [12] and Walid [33] have analyzed newly diagnosed GBM patients undergoing therapy and discovered certain clinical and molecular features, which play a significant role in prolonging the survival period. Predictive models have been developed in the past utilizing imaging and clinical features of patients [14] and there also exists ongoing clinical trials on certain drugs to test their effect on survival [34] but to our knowledge there is a lack of comprehensive data-driven work in this space which studies the impact of clinical features, genomic features along with patterns in treatment together on the survival of Glioblastoma patients.

The high mortality rate of GBM patients, where long-term survival is a rare phenomenon, has drawn significant attention to improving treatment of these tumors. After the first line standard of care treatment, there are different treatment combinations

* Corresponding author at: College of Computing, Georgia Institute of Technology, 266 Ferst Dr, Atlanta, GA 30332, USA.
E-mail address: jsun@cc.gatech.edu (J. Sun).

http://dx.doi.org/10.1016/j.jbi.2016.03.020
1532-0464/

chosen by oncologists. The sequence in which the next set of drugs or therapy is prescribed adds to the level of complexity since drugs given in a particular sequence may have a better therapeutic effect than the same drugs given in some other order. Furthermore, other drugs such as steroids and antiepileptics are administered in conjunction while treating GBM, which adds another layer of complexity. We believe analyzing the treatment plans of patients from the TCGA will provide insight into treatment patterns, which may be associated with greater overall patient survival. Based on our knowledge, there is no existing literature that analyzes treatment patterns that may influence survival for new GBM patients. The proposed approach is general and can be used for other clinical settings.

1.1. Contributions

Our study makes the following contributions:

1. We introduce a novel graph approach to extend existing sequential pattern mining algorithms for a clinical predictive modeling application.
2. We extended existing sequential pattern mining algorithms by incorporating two additional constraints called the 'exact-order' and 'overlap', which can generate more clinically meaningful treatment patterns.
3. We followed a data-driven approach to build and evaluate a predictive model for treatment effectiveness of GBM patients by treating temporal treatment patterns as features in addition to the existing clinical and genomic features.

2. Related work

2.1. Influence of genomic factors on GBM

High dimensional gene expression profiling studies in GBM patients have identified gene signatures associated with epidermal growth factor receptor (EGFR) overexpression and survival [5,13,15,16,19,22,25–27]. Genomic abnormalities associated with TP53 and RB1 mutations have been identified in TCGA along with GBM-associated mutations in genes such as PIK3R1, NF1, and ERBB2. CNV and mutation data on TP53, RB, and receptor tyrosine kinase pathways revealed that the majority of GBM tumors have abnormalities in all these pathways suggesting this is a core requirement for GBM pathogenesis [28]. However, no one systematically tests those genomic factors together with clinical and treatment information for predicting GBM survival outcome, which is a focus of this paper.

2.2. Sequential pattern mining

Sequential pattern mining refers to the mining of frequently occurring ordered events or subsequences as patterns [11]. This technique, introduced by Agarwal and Srikant [1] in their 1995 study of customer purchase sequences, led to the development of the Generalized Sequential Pattern mining (GSP) algorithm which is based on the Apriori [35] algorithm to mine frequent itemsets. GSP uses the downward-closure property of sequential patterns and adopts a multiple-pass, candidate generation approach. Initially it finds all the frequent sequences of length one item with minimum support. Subsequently it combines every possible 1-item itemset which has the minimum support for the next pass. Besides GSP, another popular sequential mining algorithm is SPADE (Sequential PAttern Discovery using Equivalent classes) [30] which uses a vertical id-list database format data format and associates each sequence a list of transactions in which it occurs. The frequent sequences can be found by efficiently using intersection on id-lists. Bellazi et al. [31] have worked on generating temporal association rules using an Apriori approach to help improve care delivery for specific pathologies. These rules consist of antecedents and consequents signifying that if the antecedent occurs then the consequent would also occur with a certain probability. Another algorithm, which is based on temporal association rules is KarmaLego [32]. This is a fast time-interval mining method, which exploits the transitivity inherent in temporal relations. The other sequential pattern mining algorithms are based on the 'Pattern Growth' technique of frequent patterns avoiding the need for candidate generation unlike GSP and SPADE which are based on Apriori. This approach involves finding frequent single items, and condensing this information into a frequent pattern tree. PrefixSpan [8,21] is one such algorithm which exploits this approach by building prefix patterns and concatenating them with suffix patterns and concatenating them with suffix patterns to find frequent patterns. SPAM (Sequential PAttern Mining using a bitmap representation) [2] uses a depth-first traversal of the search space with various pruning mechanisms and a vertical bitmap representation of the database enabling efficient support counting. Our approach is very minimally inspired by Apriori and reads the data as a graph of events to mine only those sequences which exist in the graph instead of analyzing all possible combination of events. To properly apply treatment pattern mining, we introduce several important constraints such as 'exact-order' and 'overlap'.

3. Approach

3.1. Data

We constructed a rich dataset of newly diagnosed GBM patients by integrating two different databases called the *TCGA* [17] and the *cBioPortal* [4,7]. TCGA consists of clinical and treatment data pooled together from different research teams, which is publicly accessible. The genomic data for the same patients was obtained from cBioPortal, a web resource of multidimensional cancer genomics data maintained by the Memorial Sloan Kettering Cancer Center.

3.1.1. Features

For our study, we analyzed data from 309 newly diagnosed GBM patients spanning over a period of 2 years from the date of diagnosis. The data was categorized into 'Clinical', 'Genomic' and 'Treatment' domains. The clinical domain includes demographic information about the patient along with basic clinical features such as Karnofsky Performance Score (KPS), histopathology, prior glioma history, and whether the patient is alive or deceased. Under the genomic domain, the mRNA expression levels and CNV data was collected for a specific set of genes which play a role in classifying GBM patients into 4 genomic subtypes, namely, 'Classical', 'Mesenchymal', 'Proneural', and 'Neural' [28]. The log2 copy number values were collected from Affymetric SNP6 for each gene and for mRNA expression, Z-scores were used from Agilent microarray. The methylation status of the promoter region of the MGMT gene was also used for our analysis [9]. The treatment domain consists of treatment plans for each patient, which can be viewed as process data. We use sequential mining algorithms to mine significant patterns in their treatment plans and use them as features in the dataset in addition to clinical and genomic features. Table 1 summarizes the dimensions of the dataset categorized by the domain.

3.1.2. Target variable

The goal of this study is to apply our extended modeling protocol to effectively predict patients used for model validation who survived for greater than 12 months. The pool of patients used