BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(11)

Association rule mining based study for identification of clinical parameters akin to occurrence of brain tumor

Dipankar Sengupta[1]*, Meemansa Sood[1], Poorvika Vijayvargia[1], Sunil Hota[2] & Pradeep K Naik[1]

[1]Dept. of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, H.P., India; [2]DIHAR, Defense Research & Development Organization, Leh, Jammu & Kashmir, India; Dipankar Sengupta – Email: dipankarsengupta.1982@gmail.com; Phone: +91-1792-239313; Fax: +91-1792-245362; *Corresponding author

Received June 13, 2013; Accepted June 16, 2013; Published June 29, 2013

Abstract:
Healthcare sector is generating a large amount of information corresponding to diagnosis, disease identification and treatment of an individual. Mining knowledge and providing scientific decision-making for the diagnosis & treatment of disease from the clinical dataset is therefore increasingly becoming necessary. Aim of this study was to assess the applicability of knowledge discovery in brain tumor data warehouse, applying data mining techniques for investigation of clinical parameters that can be associated with occurrence of brain tumor. In this study, a brain tumor warehouse was developed comprising of clinical data for 550 patients. Apriori association rule algorithm was applied to discover associative rules among the clinical parameters. The rules discovered in the study suggests - high values of Creatinine, Blood Urea Nitrogen (BUN), SGOT & SGPT to be directly associated with tumor occurrence for patients in the primary stage with atleast 85% confidence and more than 50% support. A normalized regression model is proposed based on these parameters along with Haemoglobin content, Alkaline Phosphatase and Serum Bilirubin for prediction of occurrence of STATE (brain tumor) as 0 (absent) or 1 (present). The results indicate that the methodology followed will be of good value for the diagnostic procedure of brain tumor, especially when large data volumes are involved and screening based on discovered parameters would allow clinicians to detect tumors at an early stage of development.

Keywords: Apriori algorithm, Association Mining, Brain Tumor Data warehouse, Data Mining.

Background:

The ability to collect and store data has grown at a dramatic rate in all disciplines over the past decade or so with new techniques being developed for effective storage and analysis. Healthcare has been no exception. The shift toward evidence-based research presents significant opportunities to extract meaningful information and transform into the knowledge from clinical data **[1].** Interpreting data across multiple systems is challenging, and various integration techniques, with varying levels of complexity, have been proposed to solve the problem of data integration and storage **[2]**. However, research reveals that current designs are not efficient for data sets with large numbers of attributes that vary over time **[3]**. The architecture required for clinical data management has been researched in applications such as clinical study data management systems (CDMSs) and clinical patient record systems (CPRSs). They both use an entity-attribute-value (EAV) system as opposed to conventional database design **[4]**. The EAV system has the advantage of remaining stable as the number of parameters increases when knowledge expands, a common situation in the basic sciences and in clinical trials **[3, 5]**. The characteristics of clinical data as it originates during the process of clinical documentation - includes issues of data availability and complex representation models, that can make data mining process challenging. Therefore, data preprocessing and transformation are required before one can apply data mining algorithms on clinical data. The application of classical data warehousing process should be thus able to answer the queries being raised. It should also be able to mitigate issues like appropriate storage structure of clinical data, varied sources of data, reduce the dimensionality constraint, and handle multiple data variables.

The stored data in the warehouse would provide a basis for the analysis of risk factors for the disease. For example, we can compare tumor with non-tumor patients to find patterns associated with occurrence of brain tumor. This method has been common practice in evidence-based medicine, which is an approach where clinician is aware of the evidence in support of clinical practice, and its associated strength **[6]**. In general, medical practitioners and researchers do not care how sophisticated a data mining method is, but they do care how understandable its results are **[6]**. Rules are a type of the most human-understandable knowledge, and therefore it is most suitable for deciphering new rules corresponding to data associated with medical applications. Association rule mining is a general purpose rule observation scheme that has been widely used for observing rules in medical applications **[7]**. The Apriori association algorithm exploits the downward closure property, which states that if an itemset is infrequent, all of its supersets must be infrequent. Each itemset has an associated statistical measure called support. For an itemset $X \subset I$, support(X) = s, if the fraction of transactions in the dataset D containing X = s. The classic framework for association rule mining uses support and confidence as thresholds for constraining the search space. The confidence or accuracy of an association rule X => Y in *D* is the conditional probability of having Y contained in a transaction, given that X is contained in that transaction: confidence(X => Y) = P (Y | X) = support(X Y) /support(X). This method has been used to find disease-disease, disease-finding, and disease-drug co-occurrences in electronic health record data **[8, 9]**. Association rule mining using objective measures and transitive inference for pruning has also been done in the clinical domain to find associations between medications and clinical problems using electronic health record data **[10]**. Studies made by Brossette *et al.* **[7]** & Ordonez *et al.* **[11]**, states about associative rules corresponding to hepatitis & heart.

The objective of this study is to propose for a data mining process, which can be used for storage & assessment of data for patients of brain tumor (primary stage) and observe associative rules based on clinical diagnostic parameters. The data warehouse being designed in this study for storage of clinical data, should be able to render the data in appropriate structures, provide metadata that adequately records semantics of data and reference pertinent medical knowledge. The data in the warehouse is subject to association mining for observing new rules. Based on the associative clinical parameters deciphered, we propose for a predictive model which can be used for an early prediction of brain tumor in suspected patients independent of results from MRI, CT scan, arteriogram or small dime craniotomy. Applying association rule mining to a given clinical data set has the potential to confirm existing knowledge regarding disease co-occurrences as well as to discover new disease relationships that could potentially lead to improved clinical health care.

Methodology:

The path of knowledge discovery process is said to be complete when knowledge has been extracted from pool of data. The said path involves collection, cleaning and storage of data followed by mining of knowledge from this pool. Considering the same, this study focuses on deciphering the clinical parameters which can be associated with the 'STATE' of brain tumor by applying association rule mining algorithm. For a patient not having tumor, 'STATE' is represented as 0 while for diseased as 1. In (**Figure 1),** it depicts approach followed for this study. In this study a data warehouse was designed to store temporal data for patients of brain tumor. The warehouse acts as a data collector, data integrator and data provider for the data mining process which could be used by doctors, physicians and other health professionals. There is varied dimensionality observed in clinical data. Based on consultation with oncologists appropriate data forms were selected. The set of clinical parameters selected for the study focuses on blood analysis result, KFT (Kidney Functionality Test) result, LFT (Liver Functionality Test) result, sugar level, triplets of blood pressure and MRI/CT scan images. The warehouse currently stores information about 550 patients from hospitals across India. Based on the variability of data, the dimensional model was designed considering the static & measurable features. In (**Figure 2),** it depicts the logical data model, based on which the structure was created in the warehouse. The functional schema has date and time dimension which ensures historical data storage for a patient. The model has been implemented and developed as a data warehouse using MySQL 5.019 RDBMS.

Preprocessing of data in the warehouse was done using STATISTICA DATAMINER 9.0, to select the features for mining purpose, considering - 1) Missing value identification; 2) Selection of integrated forms of data; 3) Identification of incorrect values based on prescribed scale (Marshall Clinical Biochemistry, 2008); 4) Feature selection. Data of 200 patients were selected for further analysis which included 124 cases of brain tumor (primary stage) and 76 normal patients. From the feature selection step, the parameters selected for the study are: Haemoglobin_content, Total_Leucocyte_count (TLC), Eosinophils, Neutrophils, Lymphocytes, Monocytes, Platelet count, KFT_Creatinine (Kidney Functionality Test - Creatinine), KFT_BUN (Kidney Functionality Test - Blood Urea Nitrogen), LFT_Sr_Bilirubin (Liver Functionality Test - Serum Bilirubin), LFT_ALP (Liver Functionality Test - Alkaline Phosphatase), LFT_SGOT (Liver Functionality Test - Serum Glutamic Oxaloacetic Transaminase), LFT_SGPT (Liver Functionality Test - Serum Pyruvic Transaminase). Each of the said parameter values was processed into qualitative form & labeled as HIGH, NORMAL or LOW based on prescribed clinical ranges (Marshall 2008). STATISTICA DATAMINER 9.0 was used to calculate the frequency of each itemset with support % criteria of atleast 30 along with head and body iteration rate of 10. All the frequent item set obtained with atleast 30% support criteria were subjected for the observation of association rules. STATE was declared as the response indicator and the remaining parameters were defined as categorical indicators. The confidence to deduce rule was set to atleast 85% and the process was executed with antecedent and precedent iteration rate of value 10.

The parameters found to be associated with occurrence of tumor were selected to build a predictive model using normalized regression approach. Jackknifing was applied for cross-validation on the model obtained and validated by R^2 observed & Y-randomization test to calculate the PRESS (predicted residual sum of squares) value.

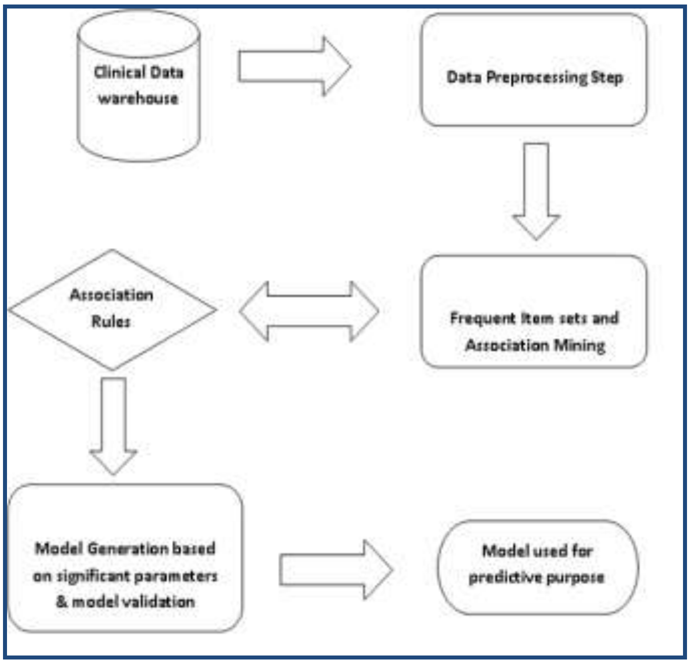
Clinical Data warehouse
Data Preprocessing Step
Association Rules
Frequent Item sets and Association Mining
Model Generation based on significant parameters & model validation
Model used for predictive purpose

Figure 1: Flow chart representation of Knowledge Discovery Process.

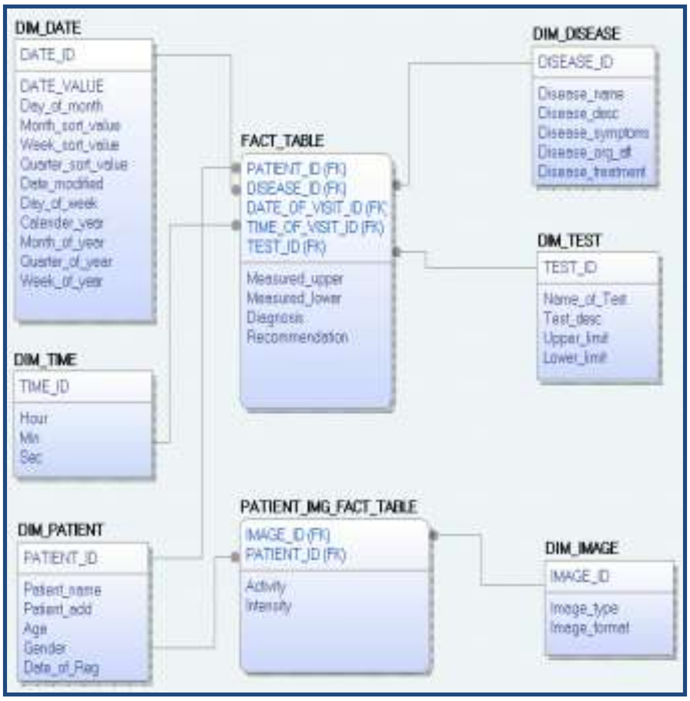

Figure 2: Tumor Data Warehouse Dimensional Model

Results & Discussion:

Haemoglobin_content, TLC, Platelet Count, KFT_Creatinine, KFT_BUN (Blood Urea Nitorgen), LFT_Sr_Bilirubin, LFT_ALP, LFT_SGOT and LFT_SGPT are the parameters that showed support of atleast 30%. Item sets satisfying the support % were subjected to observation of association rules within the specified mining criteria that showcased association of high values of Creatinine, BUN, SGOT & SGPT with presence of tumor in patients. **Table 1 (see supplementary material)** enlists various Association Rules discovered within the defined criteria. Usually high value of creatinine indicates any renal functional impairment (intrinsic renal lesions, decreased perfusion of the kidney, or obstruction of the lower urinary tract), acromegaly and hyperthyroidism, while that of BUN (Blood Urea Nitrogen) indicates acute & chronic intrinsic renal disease or post renal obstruction of urine because of high protein intake. The SGOT (serum glutamic-oxaloacetic transminase) test, also known as an AST test, measures the amount of a protein enzyme called glutamic-oxaloacetic transaminase occurring in blood. The SGOT enzyme can be associated with functioning of skeletal muscles, red blood cells, heart muscles, kidney tissue and with the brain as well. An SGPT blood test is used to measure the amount of the enzyme glutamate pyruvate transaminase (GPT) in blood and usually associated with occurrence of diseases like cirrhosis and hepatitis. However the result of the study suggests that the described factors can also be associated in a combined form with occurrence of the disease - brain tumor. Diagnostic value of Creatinine & Urea nitrogen (BUN) which are usually tested as part of Kidney Functionality test and; SGOT & SGPT which are usually tested as part of Liver Functionality test were found to be unusually high with no abnormalities reported for Kidney or Liver for patients diagnosed by brain tumor in the primary stage. The study suggest Creatinine, Urea Nitrogen, SGOT & SGPT based values can be associated together and used for deterministic analysis for STATE of the disease and its early screening. There are significant associative rules observed corresponding to the discovered parameters with respect to STATE parameter of brain tumor. There is 100% confidence observed corresponding to Creatinine and Blood Urea Nitrogen association with the disease whereas 95% confidence with SGOT and SGPT. Also the study suggests that Haemoglobin content is usually normal along with other blood related parameters in case of patients suffering from brain tumor during the primary stage with 100% confidence.

Based on the parameters identified among the associative rules with 85% (Creatinine, BUN, SGOT, SGPT) & 75% confidence (Hemoglobin Content, Alkaline Phosphatase and Serum Biliuribin), a predictive model is proposed to predict the possible STATE of a individual i.e whether suffering from tumor or not. Most significant model obtained is:

STATE = 0.171 + 0.0491 Haemoglobin_content + 0.0652 KFT_Creatinine +

0.00171 KFT_BUN- 0.0504 LFT_Sr_Bilirubin + 0.000304 LFT_ALP
- 0.00007 LFT_SGPT+ 0.00806 LFT_SGOT

The cross-validation results obtained from Jackknifing - R^2 $_{\text{(calculated)}}$ = 74.66% and PRESS = 1.67, indicates the model to be significant.

Conclusion:

This study primarily focuses on observation of clinical parameters that can be associated with occurrence of brain tumor which is rarely focused upon, by applying association rule mining algorithm. The study highlights four of the clinical factors, usually tested for Kidney & Liver functionality, to be directly associated with occurrence of brain tumor for patients diagnosed in the primary stage. Based on the discoveries made in this study a predictive model is proposed for its early diagnosis. For robustness & higher accuracy, the model proposed in the study needs to be further validated by including data set of patients suffering from other kind of tumors, renal functional impairment, kidney based problems, metastatic brain tumor and other brain related diseases.

Acknowledgement:

The authors would like to thank Prof. (Dr.) M.C.Pant from Ram Manohar Lohia Hospital, Lucknow, India and Dr. Ankur from King's George Medical College, Lucknow, India who have provided the technical guidance corresponding to aspects of brain tumor and other clinical parameters during the course of this study.

Funding Source:

Declared None.

References:

[1] Berger AM & Berger CR, *Comput Inform Nurs.* 2004 **22:** 123 [PMID: 15520581]

[2] Brazhnik O & Jones JF, *J Biomed Inform.* 2007 **40:** 252 [PMID: 17071142]

[3] Dinu V & Nadkarni P. *Int J Med Inform.* 2007 **76:** 769 [PMID: 17098467]

[4] Deshpande AM *et al. J Am Med Inform Assoc.* 2002 **9:** 369[PMID: 12087118]

[5] Anhøj J, *J Med Internet Res.* 2003 **5:** e27 [PMID: 14713655]

[6] Li J *et al. Artif Intell Med.* 2009 **45:** 77 [PMID: 18783927]

[7] Brossette SE *et al. J Am Med Inform Assoc.* 1998 **5:** 373 [PMID: 9670134]

[8] Chen ES *et al. J Am Med Inform Assoc.* 2008 **15:** 87 [PMID: 17947625]

[9] Hanauer *et al. PLoS One.* 2009 **4:** e5203 [PMID: 19365550]

[10] Wright A *et al. J Biomed Inform.* 2010 **43:** 891 [PMID: 20884377]

[11] Ordonez C *et al. Knowledge and Information Systems* 2006 **3:** 1

Edited by P Kangueane

Citation: **Sengupta *et al.*** Bioinformation 9(11): 555-559 (2013)

Supplementary material:

Table 1: Association Rules deciphered for clinical parameters corresponding to occurrence of brain tumor.

Association Rule	**Support %**	**Confidence %**	**Correlation %**
KFT_Creatinine = HIGH ==> KFT_BUN = HIGH	56.75	100	77.45
KFT_Creatinine = HIGH ==> STATE = 1	56.75	100	77.77
KFT_BUN = HIGH ==> STATE = 1	78.37	85.29	90.8
KFT_Creatinine = HIGH, KFT_BUN = HIGH ==> STATE = 1	56.75	100	79.77
LFT_SGOT = HIGH ==> STATE = 1	62.16	98.83	81.72
LFT_SGOT = HIGH ==> LFT_SGPT = HIGH, STATE = 1	62.16	95.83	85.71
LFT_SGPT = HIGH ==> STATE = 1	81.08	88.23	89.56
Haemoglobin_content = NORMAL ==> STATE = 1	59.45	100	81.64