Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-584

Assessing the Feasibility of Data Mining Techniques for Early Liver Cancer Detection

Mu-Hsing KUO [a,1], Chang-Mao HUNG [b], Jeff BARNETT [c] and Fabiola PINHEIRO[a]
[a] School of Health Information Science, University of Victoria, BC, Canada
[b] Yung-Ta Institute of Technology & Commerce, Taiwan
[c] BC Cancer Agency, Victoria, BC, Canada

Abstract. The objective of this study is to assess the feasibility of a data mining association analysis technique, the FP Growth algorithm, for the detection of associations of liver cancer, geographic location and demographic of patients. For the research, we are planning to use data extracted from electronic health record systems of three healthcare organizations in different geographic locations (Canada, Taiwan and Mongolia). The data are arranged into 'transactions' which contain a set of data items focused around cancer diseases, geographic locations and patient demographics. This analysis produces association rules that indicate what combinations of demographics, geographic locations and patient characteristics lead to liver cancer.

Keywords. Data Mining, FP Growth Algorithm, Live Cancer, Association Rules

Introduction

Liver cancer is one of the least curable cancers and it is probably under diagnosed in many parts of the world [1]. Although it ranks only fourth in cancer incidence in the world, it leads to approximately 1 million deaths each year [2], and its incidence is projected to continue rising, with 12.7 million deaths expected from it in 2030 [3]. It is believed that the earlier the cancer is detected the greater the possibility of cure [1].

Since 2009, the National Taiwan University (NTU) and the Mongolian University of Science and Technology (MUST) have carried out a 3-year "Data Mining (DM) on Healthcare" joint research project. This year, researchers at the School of Health Information Science, University of Victoria (UVic) and the BC Cancer Agency plan to join the Taiwan-Mongolia data mining project as collaborators to form a three-country based research team (we plan to extend the study period to 5 years). Two of the principal investigator's graduate students have been involved in liver cancer data mining research. The main benefit of the collaboration is that researchers can apply DM algorithms to analyze diverse clinical diagnosing records contained in three distinct Electronic Health Record (EHR) systems to discover hidden knowledge related to liver cancer. The joint project expects to achieve the following two goals:

(1) To provide early detection of liver cancer

[1] Corresponding Author: Dr. M. H. Kuo, School of Health Information Science, PO Box 3050 STN CSC, Victoria, BC, V8W 3P4, Canada. E-mail:akuo@uvic.ca.

Researchers will be able to infer relationships from a large number of medical records using DM techniques. This analysis produces association rules that indicate what combinations of demographics, geographic locations and patient characteristics may lead to liver cancer. As a consequence, the resulting system could provide early alerts to patients with high liver cancer risk.

(2) To establish clinical pathways and guidelines

The study will collect medical records from three country's EHR systems, including admissions/discharge diagnoses, chief complaints, physician orders, etc., and proposes to apply sequence clustering methods to discover better clinical pathways and establish standardized clinical guidelines for liver cancer treatment.

The aim of this paper is to assess the feasibility of *FP Growth* algorithm to perform association analysis on the demographics, geographic locations and patient characteristics. This analysis expects to produce association rules that indicate what combinations of the factors lead to liver cancer.

1. Literature Review

Data mining (DM) is a critical step from knowledge discovery in database processes, which refers to a "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data" [4]. In order to perform descriptive and predictive analysis, data mining employs various analysis methods, which include clustering, classification, regression and association analysis, to discover interesting patterns in the given data set that serve as the basis for estimating future trends. In recent years data mining has received considerable attention as a tool that can be applied to cancer detection and treatment [5-20].

On the research of DM application to liver cancer detection and treatment, El-Serag [5] reported that the major liver cancer risk factors are hepatocellular carcinoma (HCC) comprising of chronic hepatitis virus infections, cirrhosis caused by either hepatitis or alcoholisms, and chronic exposure to various cytotoxic substances (e.g. arsenic, polyvinyl chloride (PVC)). He developed DM models for a better understanding the fundamental mechanism leading to HCC development and early HCC detection.

Luk *et al*. [9] used Artificial Neural Network (ANN) and Classification And Regression Tree (CART) algorithms to distinguish HCC from non-tumor liver tissues. They employed 2-dimensional gel electrophoresis to produce protein expression profiles of 66 tumor and 66 non-tumor paired samples. Eventually, they revealed that those classification algorithms were suitable to be applied to the building of classification model based on the hidden pattern in the proteomic dataset. In addition, ANN and CART algorithms generated good predictive abilities in differentiation between tumor and non-tumor tissues for liver cancer.

Similarly, Lin [14] proposed CART and Case-Based Reasoning (CBR) techniques to structure an intelligent diagnosis model aiming to provide a comprehensive analytic framework to raise the accuracy of liver cancer diagnosis. The major steps in applying the model include: (1) adopting CART to diagnose whether a patient suffers from liver disease; (2) for patients diagnosed with liver disease in the first step, employing CBR to diagnose the types of liver diseases. In the first step, the CART rate of accuracy is 92.94%. In the second step, the CBR diagnostic accuracy rate is 90.00%. The experimental results showed that the intelligent diagnosis model was capable of

integrating CART and CBR techniques to support the physician in making decisions regarding liver disease diagnosis and treatment. More recently, Rajeswari and Reena [17] used the liver disease datasets obtained from UCI repository consists of 345 instances with seven different attributes to test three DM algorithms: Naive Bayes algorithm, FT Tree algorithm and KStar algorithm. The study results showed that FT Tree had better classification accuracy compared to other algorithms.

2. The *FP Growth* Association Algorithm

Association analysis has been used extensively in business to analyze customer transactions and to find associations between the products consumers purchase. An association rule is an implication or If-Then rule which is supported by data. A typical rule resulting from such a study could be "90 percent of all customers who buy bread and butter also buy milk". However, its application to liver cancer data analysis is relatively unexplored.

In this study, we propose the *FP Growth* algorithm [21] to execute association analysis on the demographics, geographic locations and characteristics of liver cancer patients. The algorithm employs an iterative approach known as level-wise search, where *k*-itemsets are used to explore *(k+1)*-itemsets. Then, association rules are generated, which is an implication of the form X→Y, where X and Y are disjoint subsets of all the possible data items. The strength of the association rule can be measured by its *support* and *confidence* as follows:

Let $I = \{i_1, i_1, \cdots, i_m\}$ be a set of items and $X, Y \subset I$, then the support of an association pattern is defined as equation (1).

$$support(X \rightarrow Y) = P(X \cap Y) \tag{1}$$

, and the confidence of the association pattern is defined as equation (2).

$$confidence(X \rightarrow Y) = \frac{support(X \cap Y)}{support(X)} \tag{2}$$

Support determines how often the data items in a rule are present together in a transaction in a given data set and is simply the count of transactions that contain X and Y. Support is used to determine if a rule is of interest since high support indicates that the rule occurs often in the data. Confidence is used to determine the reliability of the inference made by the rule and is an estimate of the conditional probability of Y given X. It says “If X is present in a transaction, how likely is it that Y is also present”. The generated rules suggest a strong co-occurrence relationship between the given data item subsets.

3. A Hypothetical Case Study

We are working in partnership with the BC Cancer Agency, National Taiwan University and Mongolian University of Science and Technology, and are using data extracted from their EHR systems. The data are arranged into “transactions” which contain a set of data items focused around a specific event, object, or time period. In this paper, since our human research ethics approval is pending, we use a hypothetical data set with 25 records to test the feasibility of the proposed method. The data are encoded as numbers so that they can be processed by the *FP Growth* algorithm as follows (see Figure 1):

	Country	Gender	Cancer_type	Age_diagnosis	Chemotherapy	Death
2	0	M	1	5	0	1
3	0	M	0	5	0	1
4	0	F	0	3	0	1
5	1	F	2	5	0	1
6	1	F	0	5	0	0
7	1	M	1	5	0	1
8	1	M	0	5	1	1
9	0	M	2	2	1	1
10	2	F	0	4	0	1
11	2	M	1	5	1	1
12	2	M	1	4	1	0
13	2	F	1	5	1	1
14	1	M	2	4	0	0
15	1	F	1	2	0	1
16	2	M	2	3	0	0
17	0	F	2	3	1	1
18	2	M	1	5	1	1
19	2	M	0	3	1	1
20	1	M	1	4	1	1
21	2	M	0	2	0	1
22	1	M	1	3	0	1
23	2	F	0	4	1	0
24	2	F	0	3	1	0
25	1	M	1	4	1	0
26	0	M	1	4	0	0

(1) Country (0="Canada", 1="Taiwan", 2="Mongolia")
(2) Gender (F="female", M="male")
(3) Cancer type (0="breast cancer", 1="liver cancer", 2="lung cancer")
(4) Age at diagnosis (1="20~29", 2="30~39", 3="40~49", 4="50~59", 5="60~69")
(5) Chemotherapy (0="no", 1="yes")
(6) Death (0="no", 1="yes")

Figure 1. A hypothetical data set with 25 records

For this case study, we set a minimum support count of 2 and minimum confidence of 65%. Applying the algorithm to this simple data set has yielded some interesting relationships (rules) as follows:

Rule 1: {(Taiwan, Male, 50~59) → Liver Cancer} *Confidence* = 2/3 =67%
Rule 2: {(Taiwan, Male, 50~59) → (Liver Cancer, Death)} *Confidence* = 2/2 =100%
Rule 3: {(Mongolia, Male, 60~69) → Liver Cancer} *Confidence* = 2/2 =100%
Rule 4: {(Mongolia, Male, 60~69) → (Liver Cancer, Death)} *Confidence* = 2/2 =100%
Rule 5: {(Mongolia, Male, 60~69, Chemotherapy) → (Liver Cancer, Death)} *Confidence* = 2/2 =100%

Rule 1 is interpreted as: Male Taiwanese age between 50 and 59 years old is likely to have liver cancer with confidence 67%. Rule 2 says that the same group of patients die of liver cancer with a confidence of 100%. The interpretations of rule 3 and 4 are similar to that of rule 1 and 2. Rule 5 provides more interesting information for the analysis. It shows that male Mongolian age between 60 and 69 years old is likely to have liver cancer and die of the disease even with chemotherapy treatment. The rule confidence is 100%. However, it is worth to note that these rules obtained from the data set should not be interpreted as indicators of reality.

4. Conclusion and Discussion

As the incidence of liver cancer is associated with several risk factors and varies by geographic region, researchers believe that environmental factors, in addition to patient characteristics, play a significant role in the development of the disease [22]. In this

study, we assess the feasibility of the *FP Growth* algorithm to perform association analysis on the demographics, geographic locations and characteristics of liver cancer patients. The algorithm has high efficiency and generates association rules that have high levels of confidence.

However, our analysis data is not actually extracting from three country's EHR systems. Rules generated from the analysis do not reveal the reality of associations of patient demographics, geographic locations and liver cancer. Future work will use real cancer patient data, and domain experts will be involved in the validation process to interpret the mined patterns. This should increase the value and interesting of the study.

References

[1] Hainaut P. and Boyle P, 2008. Curbing the liver cancer epidemic in Africa. The Lancet 371, 367-368.
[2] Pellegrino A, 2006. Looking at Liver Cancer. Nursing, 36(10), 52-55
[3] Blanchard K. Cancer deaths to double by 2030 without intervention [Cited 2012 January 15], http://www.emaxhealth.com/1020/cancer-deaths-double-2030-without-intervention
[4] Fayyad U, Piatetsky-Shapiro G and Smyth R, 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39(11), 27-34.
[5] El-Serag HB, 2002. Hepatocellular Carcinoma: An Epidemiologic View, J Clin Gastroenterol, 35:S72-S78
[6] Lia L, Tanga H, Wua Z, Gonga J, Gruidlb M, Zoub J, Tockmanb M, and Clarka RA, 2004. Data mining techniques for cancer detection using serum proteomic profiling, Artif Intell Med, 32(2), 71-83
[7] Pospisil P, Kassis AI, Iyer LK and Adelstein SJ, 2006. A combined approach to data mining of textual and structured data to identify cancer-related targets, BMC bioinformatics, 7, 1471-2105
[8] Barker N and Clevers H, 2006. Mining the Wnt pathway for cancer therapeutics, Nature reviews-Drug Discovery, 5(12), 997 -1014
[9] Luk JM,. Lam BY et al., 2007. Artificial neural networks and decision tree model analysis of liver cancer proteomes, Biochemical and biophysical research communications, 361, 68 -73
[10] Yang Y, Iyer LK , Adelstein SJ and Kassis AI, 2008. Integrative Genomic Data Mining for Discovery of Potential Blood-Borne Biomarkers for Early Diagnosis of Cancer, PLoS ONE, 3(11), e3661
[11] Jonsdottir T, Hvannberg ET, Sigurdsson H and Sigurdsson S, 2008. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining, Expert Systems With Applications, 34(1), 108-118
[12] Yeh JY, 2008. Applying Data Mining Techniques for Cancer Classification on Gene Expression Data, Cybernetics and Sstems, 39(6),583-602
[13] Delen D, 2009. Analysis of cancer data: a data mining approach, Expert Systems, 26(1),100-112
[14] Lin R.H., 2009. An intelligent model for liver disease diagnosis, Artif Intell Med, 47(1), 53 -62
[15] Lisboa PJG, Vellido A, Tagliaferri R, Napolitano F, Ceccarelli M, Martin-Guerrero JD and Biganzoli E, 2010. Data Mining in Cancer Research, IEEE Comput Intell Mag, 5(1), 14-18
[16] Xu L, Wang F, Xu XF, Mo WH, Wan R, Guo CY and Wang XP, 2010. Data mining of microarray for differentially expressed genes in liver metastasis from gastric cancer, Frontiers of Medicine in China, 4(2), 247-253
[17] Rajeswari P, Reena GS, 2010. Analysis of Liver Disorder Using Data Mining Algorithm, Global Journal of Computer Science and Technology, 10(14), 57-61
[18] Fabregue M, 2011. Mining microarray data to predict the histological grade of a breast cancer, Journal of biomedical informatics, 44(12), S12
[19] Çakir A and Demirel B, 2011. A Software Tool for Determination of Breast Cancer Treatment Methods Using Data Mining Approach, J Med Syst 35(6), 1503-1511
[20] Chen CM, Hsu CY, Chiu HY and Rau HH, 2011. Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees, International Conference on Natural Computation (ICNC2011), 811-815
[21] Kotsiantis S and Kanellopoulos D, 2006. Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering. 32 (1), 71-82
[22] Barber FD and Nelson JP, 2000. Liver Cancer: Looking to the Future for Better Detection and Treatment. American Journal of Nursing, Supplement: Oncology, 41-46