Hsiao-Hui Li[1], Shu-Li, WANG[2,*], Yung-Yen CHIANG[2], Shin-Kuo, FU[1]

National Chung-Cheng University (1), Central Taiwan University of Science and Technology (2)

Using Data Mining to Investigate Dental Cancer Claiming Data

Abstract. *In the view point of preventive medicine, the dangerous factors related to the life style have a close relationship to cancers. Some dangerous factors can be effectively collected by investigating the human health status, the physical examination, family's disease history, etc. The literature reported that the oral cancer is closely related the habits, such as smoking, drinking beer and chewing betel nut. For effectively controlling the treatment course of cancers and improving the quality of the life, this paper adopted the cancer registration data to perform data mining, knowledge being excavated to probe the causes of diseases and looking for the key indicators relatively to the oral cancer.*

Streszczenie. *W artykule przedstawiono analizę danych, dotyczących osób chorych na raka jamy ustnej, w tym m. in. historię choroby, przyzwyczajenia i nawyki pacjenta, sprzyjające rozwojowi choroby nowotworowej. Pozwoli to na określenie wpływu wymienionych czynników na pojawienie się raka. (**Analiza danych dentystycznych na przypadki raka jamy ustnej**).*

Keywords: data mining, neural network, oral cancer.
Słowa kluczowe: analiza danych, sieć neuronowa, rak jamy ustnej.

Introduction

As healthcare service and health technologies progress, the disease type has changed to some extent in decades that the chronic diseases become the majority of diseases rather than the acute infective disease. Cancer is the most serious chronic disease with the characteristics that it is hard to cure, body function is damaged gradually, and various kinds of disabilities or even death are incurred. For effectively controlling the treatment course of cancers and improving the quality of the life, the construction of an effective and simple cancer prediction model show the urgency of the demand than even. In recent years the use of computers in medicine has increased dramatically [1]. The medical community is aware that long-term health data collection should be useful in extracting different knowledge [2]. That is, knowledge discovery in databases (KDD) methods can be developed to identify patterns or trends within the data to be exploited. In these decades, decision tree analysis, neural network, logistic regression and so on have been applied to the medical fields, such as in the prediction of diseases, treatment program, pattern recognition of cares. The results of these methods offer clinical medicine structure and build the medical model for medical issues. In this paper, we construct an oral cancer prediction model to predict and assess the disease status of the people with the oral cancer.

Literature Review

Since the use of computers has increasing a lot and large volumes of medical data are available to the medical community, data collection in medical has become easier than ever. However, it is difficult to retrieve relevant knowledge directly from the medical database, since this raw data is unorganized for extraction for decision support or exploration. Traditionally, the knowledge is retrieved by the professional analyst manually. Clearly, these manual processes easily break down when the size of the data grows and the number of dimensions increases. In dealing with the scale of data manipulation and exploration, the computing technologies for automating the process are desired [3, 4].

In the past, the data mining technology used in medicine is mostly through the three methods, namely, artificial neural networks (ANN), decision tree, and logistic regression analysis: ANN is known as biologically inspired and sophisticated analytical techniques that can model complex non-linear functions. Informally stated, ANN is an analytic technique that models the processes of learning in the cognitive system and the neurological functions of the brain, capable of predicting new observations on specific variables from the previous observations. ANN consists of a lot of artificial neurons, in which the neurons connects with each other as a network to imitate mankind's neural structure. Decision tree theory is very suitable to medical science to predict the result and analyze the outcome. The decision tree is embedded with a powerful classification algorithm that becomes more popular with the growth of data mining in the field of information systems. Decision tree recursively cuts the data records into several sets, each of which has a similar characteristic. The main function of the decision tree is to set up a tree structure by recursively dividing raw data into two or several subparts according the relativeness of the raw data. After the decision tree is established, each leaf node denotes expert's knowledge, forming the rules from root to leaf. Famous decision tree algorithms include Quinlan's ID3, C4.5, and C5 [1]. Logistic regression is a kind of method in the statistical analysis and is primarily used in predicting binary or multi-class dependent variables. Rather than predicting point estimate of the event itself, logistic regression builds the model to predict the odds of occurrence. Even to this day, logistic regression is one of the most important methods to analyze classification data.

Construction of Prediction Model

The incidence of oral cancer varies widely throughout the world. Compared to other areas in the world, the occurrence of oral cancer in the south of Asia, especially in Taiwan, is most serious. The reason is that the people in this area like to eat betel nuts, which have shown to the important stimulus of oral cancer. In accordance with the cancer reports from 1996 to 2007 by Department of Health (DOH), Taiwan, the oral cancer has become one of the top 10 causes of deaths from cancers since 1991, and the ranking has not changed over the past 26 years. This research collected the cancer claim data. The first collected data set is the set of cancer surveillance data of the patients with cancers, each of the data containing 65 attributes needed by the Department of Health, Taiwan. The second collected data set is the set of examination data associated with these oral cancer patients, containing the examination results of patients, such as WBC (white blood cell), CRP (inflammation index), Hb (hemochrome), platelet (blood platelet), etc. In addition to the above two data sets, the final collected data set is the set composed of the admission and discharge time of such patients with oral cancer. In most cases, the data from different data sources will require a lot of manipulation of data before data mining

being performed. That is, the data needs to be filtered (selected), grouped (aggregated) and then analyzied. With the above sets of data, three steps are needed to merge these data sets into a big flat table for later data mining, namely, (1) extracting data from outside sources; (2) transforming the extracted data to fit the needs of data mining; and (3) loading it into the data warehouse. Figure 1 shows the ETL (extracting, transforming and loading) processes performed in the paper.

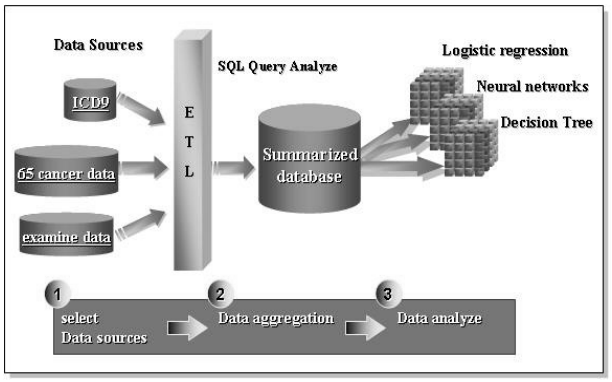

Fig. 1. Acquirement, process and analysis of the raw data.

In detail, the first step of ETL process is to select the needed data from the source systems. The needed data includes the patient demorgraphics (such as name, sex, ICD 9 codes), treatment (such as admission date since of oral cancer, staging, treatment plans), operation and exam (such as surgery, CT and MRI results), and so on. Note that not all the above data are stored in the same system. The selected data may be named differently across different systems; or the data has the same value but is denoted in different metrics (units). Data standardization is the second step to consolidate data from different data sources. For example, all data for later processing must agree on the time format, year/month/day. The step also needs to clean the data if the data obviously violates the needed, such as the disease code of the patients not falling in the desired range. The third step is to handle the missing data, which may be attributed to not being recorded, not being applicable, or being reluctant to provide by patients. For example, not all the cancer claiming data, such as “other staging-in clinical” and “other staging-pathology”, have the values. We filled such a missing data with an impossible value like “9999” to denote the missing data. The fourth step is to process the redundant data. For example, a patient with a cancer is admitted to hospital and discharged again many times. To eliminate the redundancy, we used the first admission date and the last discharged (or expired) date as the admission and discharge dates. The final step is to merge the cleaned data from different data sources into the data warehouse. The joining of several tables is performed to form a big flat table for further data mining.

This research uses the data mining tool through Visual studio analysis service project and Microsoft SQL Server 2005. We adapted three provided data mining methods, i.e., decision tree, neural networks and logistic regression analysis. In order to minimize the bias of the methods, we adopt k-fold cross-validation to evaluate the model accuracy. That is, for validating the three data mining methods, this research used 10-fold random validation. Empirical studies showed that the number 10 seems to be an optimal number of folds, in terms of the time it takes to complete the test while minimizing the bias and variance associated with the validation process. Thus, the derived patient dataset is divided into 10 parts, in which 9 parts are for training and 1 for testing, and each part of the data alternatively takes the role of training and testing.

Results

Traditionally, the performance of the results from the data mining methods is by three criteria (accuracy, sensitivity, and specificity). Before defining the three criteria, four sorts of the prediction results are shown beforhand:

1. True Positive (TP): Result is true and data mining results also predict true;
2. True Negative (TN): Result is true but data mining results predict false;
3. False Positive (FP): Result is false but data mining results predict true;
4. False Negative (FN): Result is false and data mining results also predict false.

Based on the above four sorts of prediction results, we used the following three measures as those in Dursun et al. [6] (2005) to evaluate the performance of the three data mining methods, namely,

1. Accuracy = (TP+TN) / (TP + TN + FP + FN);
2. Sensitivity = TP / (TP + FN);
3. Specificity = TN / (TN + FP).

Table 1. 10 fold cross-validation for all rules

Rul -es	Contents	Accur -acy	Sensit i-vity	Specif i-city
1	External radiation treatment of oral cancer hits 25-30 times	0.78 (88%)	0.90 (90%)	0.76 (76%)
2	Number of regional nodes examined is in the range of 20 to 25	0.82 (82%)	0.90 (90%)	0.84 (84%)
3	External radioactive rays treatments for oral cancer is about 6300 cGy	0.84 (84%)	0.90 (90%)	0.88 (88%)
4	External radiation treatment of oral cancer hits 30-35 times	0.86 (86%)	0.90 (90%)	0.78 (78%)
5	Number of regional nodes examined is in the range of 30 to 35	0.82 (82%)	0.90 (90%)	0.84 (84%)
6	External radiation treatment of oral cancer hits over 35 times	0.78 (78%)	0.85 (85%)	0.76 (76%)
7	Tumor size (pathologically) of oral is about 7cm	0.78 (78%)	0.84 (84%)	0.84 (84%)
8	TNM Path N (pathologically) is N2 (ipsilateral single is in the range of 3cm to 6cm)	0.80 (80%)	0.90 (90%)	0.84 (84%)
9	Tumor size (pathologically) of oral is about 5cm	0.76 (76%)	0.82 (82%)	0.78 (78%)
10	Number of regional nodes examined is in the range of 25 to 30	0.80 (80%)	0.85 (85%)	0.78 (78%)
11	Tumor Size (clinically) of oral is about 7cm	0.80 (80%)	0.90 (90%)	0.90 (90%)
12	Number of regional nodes examined is in the range of 30 to 35	0.72 (72%)	0.69 (69%)	0.66 (66%)
13	External radioactive rays treatments oral cancer is over 6300 cGy	0.84 (84%)	0.88 (88%)	0.83 (83%)
14	Tumor size (pathologically) of oral is about 4cm	0.78 (78%)	0.82 (82%)	0.80 (80%)
15	Number of regional nodes examined is over 35	0.72 (72%)	0.66 (66%)	0.65 (65%)
16	External radiation treatment of oral cancer hits over 40 times	0.84 (84%)	0.90 (90%)	0.90 (90%)
17	Chemical treatment of oral cancer hits about 8 times	0.85 (85%)	0.90 (90%)	0.90 (90%)
18	Oral tumor laterality is in single side	0.88 (88%)	0.90 (90%)	0.90 (90%)

Accuracy = (TP + TN)/(TP + FP + TN + FN); sensitivity = TP/(TP + FN); specificity = TN/(TN + FP)

After defining the three criteria, accuracy, sensitivity and specificity, the following table is the important results after performing 10-fold cross-validation. Since the three data mining methods will produce many rules, this research defines the accuracy as high if its accuracy value is higher than 80%. Thus, in Table 1, the rules with the high accuracy are rules 2, 3, 4, 5, 8, 10, 11, 13, 16, 17, and 18.

Although data mining methods are capable of extracting patterns and relationships hidden in large medical datasets, the results might not be valuable for clinical if without the validation and feedback from the professional. Therefore, in addition to objectively evaluate the results of three data mining methods, we also need to subjectively validate these results. For achieving the subjective evaluation, we design the questionnaire and invite experts majoring in cancer domain to perform the evaluation.

The questionnaires consisted of 18 rules are spread to four specialty doctors (majoring in ENT, oral surgeon, blood tumor, and radioactive tumor) working in the epidemic prevention center of the cancer to evaluate whether our model is correct and valuable. We received 28 replies after sending the 50 questionnaires. We define the results to be accurate if more than a half of the experts (i.e, 14 experts) reply the results being positive, which are shown in Figure 3.

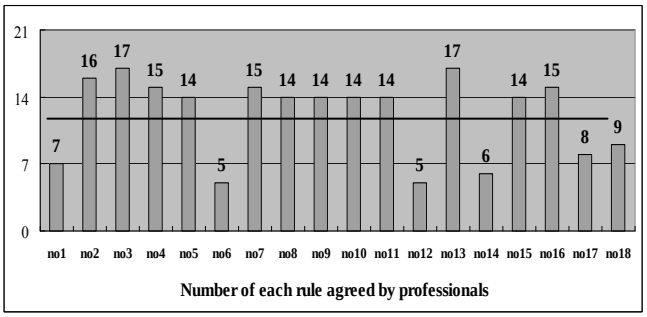

Fig. 2. Evaluation Questionnaire Recycle Result

This research objectively and subjectively evaluates the results, as are summarized as follows:

Part 1: The accuracy of results evaluated by objective evaluation is high, but that by subjective evaluation is low. Such kind of rules are rules 17 and 18. The reason for such kind of rules is that the rules give too little information such that the doctors can't answer the questionnaire with reliability under the limited information. For example, the amount of chemical dosage is not listed in the questionnaire, which is very important in chemistry treats. In addition, the tumor in the left or right side is of no difference, as will not significantly influence the state of oral cancer patient;

Part 2: The accuracy of results evaluated by objective evaluation is low, but that by subjective evaluation is high. Such kind of rules are rules 7, 9, and 15. The reason is that the tumor size of oral cancer and treatment method related to tumor stages is very important. Thus, the approval is relatively high. In contrast to those rules with low accuracy in objective evaluation, the reasons can be attributed that the population in the research is insufficient in quantity and that some important attributes are not recorded. For example, whether the patients eat the betel nut is an indicator but is not recorded;

Part 3: The accuracy of results evaluated by objective evaluation is high, and so is by subjective evaluation. Such kind of rules are rules 2, 3, 4, 5, 8, 10, 11, 13, and 16; Like the reasons in part 2, doctors consider each independent factor and death probability, survival probability are related. According to their experiences and training, the oral cancer survival rate relates to stage, radiological dosage, etc, and the quantity and size of the regional lymph gland should be considered.

From the above evaluations, nine from the eighteen rules are evaluated as high accurate from both objective and subject viewpoints (in part 3); three from the eighteen rules are evaluated as high accurate from subject viewpoints but not from objective viewpoint (in part 2); while two from the eighteen rules are evaluated as high accurate from objective viewpoints but not from subjective viewpoint (in part 1). Or we can say that 66.7% (= (9 + 3) / 18) rules are evaluated as high accurate from the doctors' viewpoint.

Conclusion

The paper developed a model to predict the outcome of an incidence of oral cancer, which is valuable in medical. The paper adopted the cancer registration data to perform data mining. The results of this empirical study are evaluated by the physicians, and 66.7% rules are evaluated as high accurate by the physicians. Our ongoing research efforts are toward incorporating new capabilities into the prediction system to increase the prediction ability. The future direction is to make the system available into the doctor entry system, performing as a clinical decision support system (CDSS) that peeps the diagnosis and treatment from the charts and remind the physician if matching the rules.

REFERENCES

[1] Han J. and Kamber M., "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2011.

[2] Shital C. S., Andrew K. , Michael A., Donnell O., "Patient-recognition data mining model for BCG-plus interferon immunotherapy bladder cancer treatment", Computers in Biology and Medicine 36. pp.634–655, 2006.

[3] Abbass H., "An evolutionary artificial neural networks approach for breast cancer diagnosis", Artificial Intelligence in Medicine 25, pp.265-281, 2002.

[4] Chou S.M., Lee T.S., Shao Y.E., and Chen I.F., "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines", Expert System with Applications, 27, pp.133-142, 2004.

***Authors**: lecturer, Shu-Li, Wang, and Prof. Yung-Yen Chiang, Central Taiwan University of Science and Technology, No. 666, Buzih Road, Beitun District, Taichung City 406, Taiwan, E-mail: mimi@ms83.url.com.tw; lecturer Hsiao-Hui Li, and master student Shin-Kuo Fu, National Chung-Cheng University, No. 168, University Road, Chia-Yi County 62, Taiwany, E-mail:kfwukfwu@gmail.com.*