International Journal of Medical Physics, Clinical Engineering and Radiation Oncology, 2014, 3, 235-240
Published Online November 2014 in SciRes. http://www.scirp.org/journal/ijmpcero
http://dx.doi.org/10.4236/ijmpcero.2014.34030

Scientific Research

Extraordinary Potential of High Technologies Applications: A Literature Review and a Model of Assessment of Head and Neck Squamous Cell Carcinoma (HNSCC) Prognosis

Claudio Camuto[1], Nerina Denaro[2,3]

[1]High Technology Department, ASO Santa Croce e Carle, Cuneo, Italy
[2]Oncology Department, ASO Santa Croce e Carle, Cuneo, Italy
[3]Human Pathology Department, Messina University, Messina, Italy
Email: nerinadenaro@hotmail.com

Received 25 August 2014; revised 22 September 2014; accepted 20 October 2014

Abstract

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cause of cancer mortality in the world and the 5th most commonly occurring cancer (Siegel, R. 2014). In the last few decades a growing interest for the emerging data from both tumor biology and multimodality treatment in HNSCC has been developed. A huge number of new markers need to be managed with bio-informatics systems to elaborate and correlate clinical and molecular data. Data mining algorithms are a promising medical application. We used this technology to correlate blood samples with clinical outcome in 120 patients treated with chemoradiation for locally advanced HNSCC. Our results did not find a significant correlation because of the sample exiguity but they show the potential of this tool.

Keywords

Data Mining, Mining Software, Algorithm, Biomarker, Head and Neck Squamous Cell Carcinoma (HNSCC)

How to cite this paper: Camuto, C. and Denaro, N. (2014) Extraordinary Potential of High Technologies Applications: A Literature Review and a Model of Assessment of Head and Neck Squamous Cell Carcinoma (HNSCC) Prognosis. *International Journal of Medical Physics, Clinical Engineering and Radiation Oncology,* **3**, 235-240.
http://dx.doi.org/10.4236/ijmpcero.2014.34030

1. Introduction

1.1. Data Mining

With the term "Data mining", people usually intend a set of algorithms to discover hidden knowledge from a very large amount of heterogeneous data and to group data into categories. Data mining was originally developed for economy field to help managers in their decisions but after few years it was progressively introduced in other fields; in last ten years their application in medicine was largely increased in particular for the elaboration of signals such as EEG, ECG, etc. [1] [2]. The main objective of this paper is to explain the data mining tool and to provide an example of its application in clinical practice. We also provided a brief review of data mining application in clinical practice.

In this study we looked for a correlation among clinical outcome (tumor progression) and blood tests (white blood cell-WBC, C reactive protein-PCR). Therefore we applied data mining algorithms to blood test. Obviously, this tool shouldn't be considered as the absolute method to detect progression but it may play a prognosticator role in providing an elaboration of several variables. Normally blood tests, imaging and biomarkers are used to evaluate patient state of disease. However, recent data suggest that high levels of inflammatory markers indicate a high probability of progression.

1.2. Medical Data

Head and neck carcinoma (HNC) is the sixth most common cancer worldwide [3].

Despite recent advances in the diagnosis and treatment of head and neck squamous cell carcinoma HNSCC, there has been little evidence of improvement in 5-year survival rates over the last few decades [4]. The most important risk factors are heavy exposure to alcohol and smoking and human papilloma virus (HPV) infections. These last two are also prognostic factors [5]. Other common prognostic factors include T and N stage, synchronous multiple primary cancers, patients performance status and age [6].

Correlations with blood sample values have not been reported although the role of PCR and infiammation is now well known to contribute to both pathogenesis and toxic deaths.

The goal of this paper is to provide an example of data mining application in HNSCC treated with chemo-radiation (CRT) or bio-radiation (bio-RT) at the S. Croce General Hospital in the years 2010 and 2011 in daily clinical practice.

2. Methods

We analyzed blood samples results of 120 patients, all patients were treated with chemo-radiation or bio-radiation at the S. Croce General Hospital in the years 2010 and 2011 in daily clinical practice. We analyzed results of white blood cell (WBC), hemoglobin (HB), PCR and lactate of each patients pre during and post treatment.

First steps of this work were loading and cleaning original data; original data format was an excel file with 57 columns, in this file were stored many information about patients, treatments, progressions, exams but not all of these information were helpful for this analysis. We started saving the excel file as csv (Comma separated text format) another format more simple to manage and use in database contexts. The second step was to create a "tablespace" and a user on an Oracle Database Schema (Using Oracle Express Edition 11 g) and load all data in a table called "tmpdata", using PL/SQL Developer Text data import tool, with the same structure of the original data. After this were created two new tables: a table to store patient's information such as Name, Surname, birth date; a second table with exams. In user table was generated a unique code called "id" for each patient, this code is used in the second table to link exams with a specific patient without using his personal data. The exam table contains the following columns:

Patient_code → id of the patient, linked with the patient's table;

Age → the age of the patient when he/she information were recorded for first time;

Exam_age → the age of the patient at the exam's day;

Exam_type → the type of exam (for example S-LDH);

Exam_result →the value of exam;

Target_value → a value used to indicate if the patient at that date and exam was in progression or not (initially empty).

This table was populated from the tmpdata cleaning exams data for example "S-LDH" value was the same of Sldh, SLDH and so on; exam age was calculated from exam data minus patient birth's date.

Mining Software Used and Algorithms

In this analysis, a software called WEKA was used. WEKA is a software freeware and open source developed in Java with a modular structure. It contains several algorithms for mining and it's possible to develop and add new ones. WEKA could work with files in various formats or with databases; we chose the second one because with a database we could easily manipulate data. We analyzed data using two different classification algorithms called "J48-Decision Tree" and "Decision Table"; the first one was choose because it is the best algorithm in many cases and it's commonly used; the second one because the structure of data should be analyzed as table and not as single record so it seem be the best choice in this case.

J48 algorithm is an implementation realized by WEKA's team of C4.5 decision tree algorithm created by Quinlan.

It works as follow:

- It choose in the attribute set the one who best discriminate the target attribute;
- For every value of chose attribute it creates a branch;
- Move data into correct branch;
- For every branch repeat the process until a branch contains only an element or all elements of a branch have the same values (or range of values) and is impossible to determinate a discriminant attribute;

The Decision table algorithm works as follow:

- It choose in the attribute set the one who best discriminate the target attribute;
- It creates a table, in rows the attributes discrimination ranges and in columns the conditions;
- After it creates a second table with conditions and corresponding categories;
- If all record's attributes satisfies all conditions of a category the record is placed in that category;
- For every record it repeats the process adding conditions for a category (if it already exists) or creating a new category.

3. Results

The analyzed population came from a mono Institutional experience with HNSCC patients treated with CRT or bioRT. Among this population Male/female ratio were 91/22 Heavy smokers (more than 10 pack/year were 100/120. Primary site were hypopharynx 28; larynx 24; oral cavity 21; orpharynx 26, rhynopharynx 4, mascellar sinus 2 and 15 unknown primary HNSCC respectively. We considered evaluable those patients with at least 10 records of blood tests during the treatment period.

PCR results were available in 52 patients, we found a correlation among PCR levels and worst outcome in 27/52 patients with PCR higher than 155 (overall survival inferior than 3 months).

LDH levels correlates with tumor progression. On 95 patients evaluable 58 had higher LDH levels who correlates with disease progression (35 local recurrence and 23 distant metastases).

4. Data Mining Analyses

A data mining analysis has typically five steps: Collecting data, preprocessing data, creating a model of data, testing generated model, applying generated model to new/complete data.

4.1. Collecting Data

The main target of first step is to load all heterogeneous data for the analysis and elaborate them to obtain an homogeneous structure for example we can have some data as number in excel, other data as string in a text file, other data as floating number in a database at the end of the process we'll probably have a table in a database where that field is a floating number so the process convert the first thing of data in floating number adding dot zero at the end, the second will be simply transformed in number, the third remain the same; all them will be loaded into a database (or another system such as a csv file).

4.2. Preprocessing Data

Data mining is very powerful to understand the data and their relationships but in the other side these analysis are very hard and slow so usually before send data to real processing they are pre-analyzed to remove not-significant data and to add pre-elaboration info that help the mining process. For example if we are analyzing a set

of patients mining processor don't care of names and surnames so that information could be safely removed without effect negatively the elaboration and increasing the running of algorithm because it has to process less fields. Another important thing is to add pre-calculated data for example if we want to analyze data in witch are relevant the days after an event we should calculate this information before processing so processor has an important field in plus to help it in taking decisions.

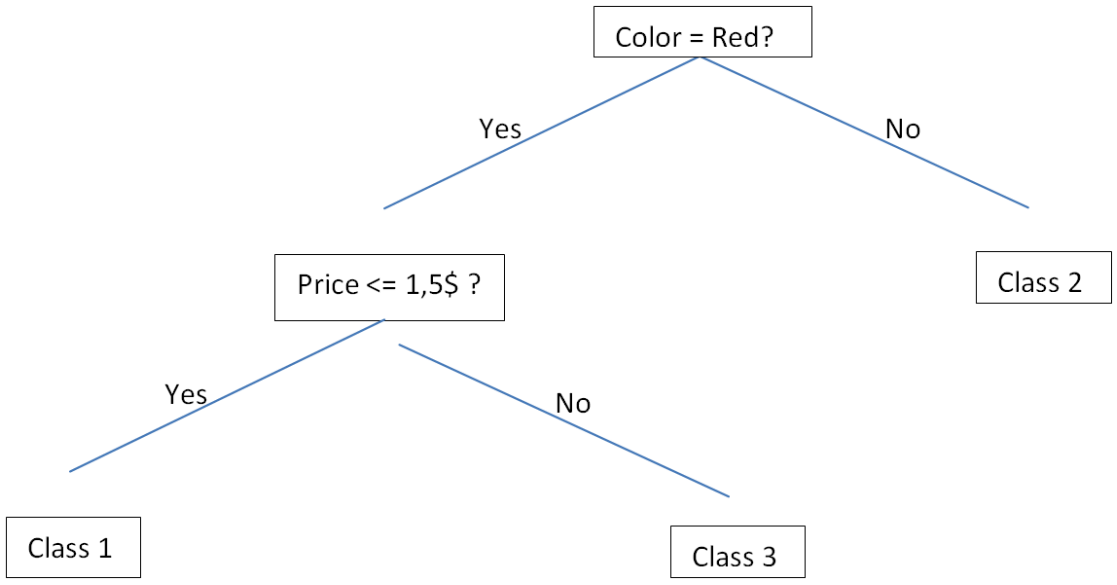
4.3. Creating a Model of Data

Data mining algorithms groups' data into a limited set of groups called "Classes" the basic rules are: an element must stay only in one class; elements in the same class are similar and they are different from element of other classes. [6] To Classification algorithm analyze the attribute of an object (every data element is an object for example a patient with his exams) and decide the class of an element. The core of data mining is the creation of a model of data; it is a decisional model used by mining to choose in which class put new elements. There are several algorithm to generate models one of the most popular is the "Decision Tree" model; it has three elements: "decision node", "leaf", "branch" the model created is similar to a tree where there is an initial node (often called root) with two or more branch, a branch can has a decision node with others branch or a leaf that is the ending point. The most important part is the decision nodes; every decision node has a set of binary rules such as "major than…" "equals to" and so on. To generate this model the algorithm needs some data in witch is know the class (at least one element for each class), this set of data is called "training set". Now an example of data and generated model was given:

Data:

Color	Type	Price	Class
Red	Pencil	1$	1
Blue	Pen	1$	2
Blue	Pencil	2$	2
Red	Pen	1.5$	1
Green	Pen	1$	2
Red	Pen	2$	3

Decision tree:



Based on generated model is irrelevant type of object. This tree is a two level's three.

4.4. Testing Generated Model

To generate a model usually people submit at least 1/3 of total data and use the remaining to test the model. The training data (1/3) contains also the associated class, other data, often called "test set" contains also this information but it isn't submitted to algorithm. The algorithm takes test set and the previous generated model and returns an associated class for each data record; after this the automatic associated data is compared with real association if confidence is better than 90% the model is usable otherwise we retry to generate model using another set of training data (training data are chose selecting random records from full data, remaining data became test set).

4.5. Applying Generated Model to New/Complete Data

When the model is created and tested and it's considered stable (confidence factor equals of better than 90%) the model can be applied to full set of data and to new data to decide the correspondent class for example if we have a model than can distinguish between healthy patient or not we can use it to discover the health status of a submitted patient. A model is never perfect so is a good procedure to update periodically model with new data.

4.6. Clustering

Until now we talk about classification but there is another important group of algorithms for data mining called "cluster algorithms". Main aim of these algorithms is to discover automatically classes and store data into them. They analyze a test set without class and put similar data to same class; a class is generated when a single data is very different from others in other classes; the process is repeated recursively until there is a stable classification. The core of algorithm is the "distance function" a function that takes two data and returns a value that represents the distance between the two data, in other words it represents of much two objects are different. Clustering is used when we don't know classes, for example we could analyze the purchases of users of a credit card to classificate users in some categories.

5. Data Mining Clinical Application

Several data mining approaches are routinely used in research work these include dose-volume metrics, equivalent uniform dose, mechanistic Poisson model, and model building methods using statistical regression and machine learning techniques. Their application in daily clinical practice could quicken the time lost to achieve information from biomarkers or physics or genetic variables [7] [8].

From a brief revision of literature in English language of cancer patients we concluded that software automated analysis will significantly reduce the overall time required to complete daily biological-radiological or physics studies (such as dose volumes studies in radiotherapy, microarray analyses and genetic elaboration). Many tools are available for automated digital acquisition of images of the spots from the microarray slide.

6. Conclusions

This study provides an example of future applications of high technology in oncology. In the era of microarray and personalized medicine these instruments are fundamental. Furthermore as the HNC patients clinical approach is well recognized to necessitate a multidisciplinary team (including ENT surgeons, radiation oncologist, medical oncologist, speech language specialist), the future global approach cannot work without a close cooperation between HT Engineers and biologists.

A correlation among elevated and reduced blood tests was not found. Data are too small to be interpreted but our analyses show the potential of this tool to evaluate correlations among a huge number of records.

References

[1] Banville, D.L. (2009) Mining Chemical and Biological Information from the Drug Literature. *Current Opinion in Drug Discovery & Development*, **12**, 376-387.

[2] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W. and Shen, B. (2013)

Biomedical Text Mining and Its Applications in Cancer Research. *Journal of Biomedical Informatics*, **46**, 200-211. http://dx.doi.org/10.1016/j.jbi.2012.10.007

[3] Siegel, R., Naishadham, D. and Jemal, A. (2013) Cancer statistics, 2013. *CA*: *A Cancer Journal for Clinicians*, **63**, 11-30. http://dx.doi.org/10.3322/caac.21166

[4] Denaro, N., Russi, E.G., Adamo, V. and Merlano, M.C. (2014) State-of-the-Art and Emerging Treatment Options in the Management of Head and Neck Cancer: News from 2013. *Oncology*, **86**, 212-229.

[5] Ang, K.K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D.I., Nguyen-Tân, P.F., Westra, W.H., Chung, C.H., Jordan, R.C., Lu, C., Kim, H., Axelrod, R., Silverman, C.C., Redmond, K.P. and Gillison, M.L. (2010) Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *New England Journal of Medicine*, **363**, 24-35. http://dx.doi.org/10.1056/NEJMoa0912217

[6] Wu, J., Ho, C., Laskin, J., Gavin, D., Mak, P., Duncan, K., French, J., McGahan, C., Reid, S., Chia, S. and Cheung, H. (2013) The Development of a Standardized Software Platform to Support Provincial Population-Based Cancer Outcomes Units for Multiple Tumour Sites: OaSIS—Outcomes and Surveillance Integration System. *Studies in Health Technology and Informatics*, **183**, 98-103.

[7] Naqa, I.E, Deasy, J.O., Mu, Y., Huang, E., Hope, A.J., Lindsay, P.E., Apte, A., Alaly, J. and Bradley, J.D. (2010) Datamining Approaches for Modeling Tumor Control Probability. *Acta Oncologica*, **49**, 1363-73. http://dx.doi.org/10.3109/02841861003649224

[8] Spencer, S.J., Bonnin, D.A., Deasy, J.O., Bradley, J.D. and El Naqa, I. (2009) Bioinformatics Methods for Learning Radiation-Induced Lung Inflammation from Heterogeneous Retrospective and Prospective Data. *Journal of Biomedicine and Biotechnology*, **2009**, 1-14. http://dx.doi.org/10.1155/2009/892863

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.