Stud. Univ. Babeş-Bolyai Math. 56(2011), No. 2, 299–303

SLAD method for cancer registration

Ioana Chiorean, Liana Lupşa and Luciana Neamţiu

Abstract. The Logical Analysis of Data (LAD) is a method extensively used in Medicine for data classification. The present paper contains a slightly modified approach of this method, called Successive Logical Analysis of Data (SLAD), more appropriate to the data registration in oncology. The corresponding algorithm is also presented.

Mathematics Subject Classification (2010): 68T30, 68T50, 68W10.

Keywords: Logical Analysis of Data (LAD), patterns, data mining, cancer registration, morphological code.

1. Introduction

Cancer registration is a continuous and systematic process of collecting data concerning the occurrence and characteristics of reportable neoplasm. The tumors may be classified according to the International Classification of Diseases for Oncology (ICD-O) and each of them has a corresponding code made by six digits. The first four digits represent the specific histological term, the fifth is the behavior code and the sixth is the grade of differentiation. The book ICD-O contains also a dictionary of codes, where to every numerical code corresponds a group-of-words-in-natural-language. For instance:

Code	Description
8500/3_	Infiltrating duct carcinoma
	Infiltrating duct adeno carcinoma
	Duct adeno carcinoma
	Duct carcinoma
	Duct cell carcinoma
	Duct carcinoma
8480/3_	Mucinous adeno carcinoma
	Gelatinous adeno carcinoma
	Mucous carcinoma
	Colloidal carcinoma

2. The data and the problem

The data

When a patient has a tumor, several investigations have to be done in order to determine its nature. The information is written in a medical record, in a "free" language, which contains, medical terms and other terms (which we call "noisy" terms).

2.1. The problem

In order to process the medical information, for establishing the corresponding code, firstly the information has to be cleaned of "noisy" words, and so the medical terms will be emphasized. Then, if the described tumor is malign, the patient has to be introduced in the cancer register with the morphological and topographical code for the tumor given according to the rules from **ICD-O**. Due to the fact that the medical terms may be consider "patterns", being recognized in the dictionary of codes, the method that we use in our paper to determine the final code for a tumor is based on Logical Analysis of Data (LAD), which is a new methodology used for detecting structural information about datasets. A specific characteristic of LAD is the detection of logical patterns which determine and predict out of a group, a class satisfying specific requirements (see [1], [2]).

Due to the fact that most of the observations in which we have to detect some code, do not contain exactly the group of words which are coded in the dictionary, but others, with the same meaning, we have to construct our own patterns. For this purpose, as in [3], we use our own method, called Successive LAD Method (SLAD), because we have to decide what code we have to give to an observation which contains groups of words belonging to different codes.

3. Constructing the patterns sets

In what follows, we denote by
PG: the set of all expressions corresponding to all codes (all patterns)
WE: the set of all words which appear in the expressions from second column of the dictionary.

The main idea of SLAD method consists in applying the classical LAD method successively, introducing patterns of different levels. They are the following:

1. *The patterns of level 0-* "does the tumor exist"?

In order to answer to this question, we construct the sets:

$PL + 0 = \{exists,\ has,\ etc.\}$,

$PL - 0 = \{does\ not\ exists,\ has\ not,\ etc.\}$.

2. *The patterns of level 1*, denoted by *PL1*, contains 1 key word from the dictionary, and determine the fifth position in the morphological code (e.g. metastatic, carcinoma, limphoma, etc.).

3. *The patterns of level 2, level 3, etc., using SLAD.*

They define the four digits of the morphological code. For every pattern p from PL1 we consider the set $PL2(p)$ made by those patterns $w\Box PG$ with the property that the concatenated patterns pw or wp are to be found in the dictionary.

Example. $p = carcinoma$ from PL1;

$w = duct$ from PG;

Then $pw = duct\ carcinoma$ and has the code 8500/3 in the dictionary. Therefore,

$$duct\Box PL2(carcinoma).$$

4. The algorithms

In [3], the algorithms for the following situations are given:

a) **All the key words in the record appear exactly in the order given in the pattern.**

Example. Let's consider the registration *"Invasive duct carcinoma with extensive papillary component"*.

Step 1. Transform this observation in patterns (key words):

"Invasive duct carcinoma"

Step 2. Apply algorithm *Pattern*:

- looking in the dictionary for the existing patterns, we get:

8500/3, for *duct carcinoma* and,

8503/3, for *intraductal papillary adeno carcinoma with invasion*

Step 3. Computing the final code, as the maximum:

$Code = max 8500/3, 8503/3 = 8503/3$.

Conclusion: Our registration *"Invasive duct carcinoma with extensive papillary component"* will have the code 8503/3.

b) **The words from the record are the same with those in the patterns, but their order differs**

Example. Let us consider the registration: *myxofibrosarcoma*, which is not in PG, but pattern fibromyxosarcoma is, to which corresponds the code 8811/3. Then, the myxofibrosarcoma record will receive the code 8811/3.

In the present paper we present another approach, when:

c) **The record contains key words which are not in the dictionary**

Example. Let's consider the record *"intrusive duct malignant with pap. comp."*

The following key words are not in the dictionary: *intrusive, malignant, pap., comp.* Also, we have some shortenings: *pap.=papillary; comp.=component.*

In order to solve the problem, we propose the following steps:

1. Generate lexicographic dictionary (SINO), which contains all the synonyms
2. Give weights, $w(i)$ to every key word $r(i)$ in the record, where $w(i) \in \{0, 0.1, \ldots, 0.9, 1\}$, for $i = 1$ to n, according with how close is the word to a pattern from the ICD-O dictionary
3. Compute $WI = (w(1) + \ldots + w(n))/n$

4. If $WI >= 0.80$, then the record enters in the dictionary, as another pattern, and receives the corresponding morphological code; if not (i.e. $WI < 0.80$), it has to go back to the physician. He will give the corresponding code and the record, together with this code, will be memorized in the dictionary

Algorithm NewPattern;

Begin
Generate SINO;
For $i = 1$ *to* p *do* {*take a record*}
For $j = 1$ *to* n *do* {*take a key word*}
Lookfor_in_SINO;
Give_Weight$(w(j))$*;*
Endfor;
$WI = (w(1) + \ldots + w(n))/n$;
If $WI >= 0.80$ *then Memo_in_ICD-O;*
Give_code
else Return_to_ Physician
Endif;
Endfor;
End.

Example. Let's consider the registration *"intrusive duct malignant with pap. comp."*

Key words: *intrusive, duct, malignant, pap., comp. so* $n = 5$
- suppose SINO is created, then we have:
Synonyms: $intrusive = invasive; w(1) = 1$
$malignant = carcinoma; w(3) = 1$
- *comp.* and *pap.* get $w(4) = 0.3$, $w(5) = 0.7$
- Compute $WI = (1 + 1 + 1 + 0.3 + 0.7)/5 = 0.8$
- Write in the ICD-O dictionary the new pattern
- Give the record the code 8503/3.

References

[1] Boros, E., Hammer P. L., Ibaraki T., Kogan Al., *Logical analysis of numerical data*, Mathematical Programming, 1997, 79, 163-90.

[2] Boros E., Hammer P. L., Ibaraki T., Kogan Al., Mayoraz E., Muchnik I., *An Implementation of Logical Analysis of Data*, IEEE Transactions on Knowledge and Data Engineering, 2000, 12, 292-306.

[3] Lupşa, L., Chiorean,I., Neamţiu, L., *Use of LAD in establishing morphologic code*, Proceedings of IEEE International Conference on Automation, Quality and Testing, Robots, AQTR 2010, 219-224.

Ioana Chiorean
Babeş-Bolyai University
Faculty of Mathematics and Computer Science
Str. Kogălniceanu, nr. 1
400084 Cluj-Napoca, Romania
e-mail: ioana@math.ubbcluj.ro

Liana Lupşa
Babeş-Bolyai University
Faculty of Mathematics and Computer Science
Str. Kogălniceanu, nr. 1
400084 Cluj-Napoca, Romania
e-mail: llupsa@math.ubbcluj.ro

Luciana Neamţiu
Cancer Institute "I. Chiricuţă"
Str. Republicii, nr. 34-36
400015 Cluj-Napoca, Romania
e-mail: luciana@iocn.ro