International Conference on Applied Internet and Information Technologies, 2016
DOI:10.20544/AIIT2016.30

Improving Readability of Medical Data by Using Decision Rules

Vladimir Brtka[1], Ivana Berkovic[1], Visnja Ognjenovic[1], Eleonora Brtka[1]
Dobrivoje Martinov[2], Tatjana Stojkovic–Jovanovic[2]

[1] Technical faculty "Mihajlo Pupin", Djure Djakovica bb,
23000 Zrenjanin, Serbia
{vbrtka, berkovic, visnjao, norab}@tfzr.uns.ac.rs
[2] General hospital "Djordje Joanovic", dr. Vase Savica 5,
23000 Zrenjanin, Serbia
martinovd@yahoo.com, tatjanastojkovicjovanovic@gmail.com

Abstract. As medical journal abstracts have become more and more difficult to read, there is a burning issue for doctors to get relevant medical information in order to solve a problem in a fast and efficient way. The paper deals with the synthesis of sentences of spoken language from tabular historical data that relate to a specific medical sub–domain. In this case Systematic Syntax Classification of Objects or SSCO algorithm was used in order to generate decision rules which were consequently transformed to natural language and delivered to the user by a machine text reader. The system is "hands–free", reliable, and enables communication by natural language. The experiments were conducted on data sample consisting of patient's conditions after hip surgery procedure and originating from General hospital "Djordje Joanovic", Zrenjanin, Serbia.

Keywords: decision rules, classification, voice communication, rehabilitation.

1. Introduction

Handling data is a difficult task, regardless of the domain from which the data is originating. The task of handling data readability is emphasized as database grows, and becomes even more tough as data structure complicates. More complicated data that comprises of multiple elements which differs in a structure, type and size is a complex task. The task of handling data readability is even more important in the case when there are time and space constraints or data itself is of sensitive nature. This, sensitive nature of data can be related to data secrecy, security and importance, so that data has a great impact to some important study, investigation or conclusion.

This paper deals with the data from medical domain. As medical domain is very important and especially sensitive for humans because somebody's life chances rest on correct and timely diagnosis, the task of correct handling with this kind of data gains in importance. In everyday medical practice there are situations of emergency, time constraints, possible shortage of medical stuff and/or materials which contributes to the importance of a good organization while handling the data. Timely possession of relevant information is of the key importance in the domain of medicine. There are multiple ways to gain some information in the domain of medicine: one possible way is to browse a bunch of paper abstracts in order to find a solution for a problem or to minimize the damage that can be done by the lack of appropriate data or information on subject or situation that has to be resolved. In [1] a study on whether the readability of medical research journal abstracts changed from 1960 to 2010 in ten-year batches had been investigated. Abstracts from medical journals were downloaded from PubMed.org, and their readability was estimated by the Coleman-Liau Index (CLI) readability score [2]. The results here indicate an increase in difficulty grade levels during this time period. So, medical journal abstracts have become more and more difficult to read. This is certainly a problem in domain of medicine; this problem is even intensified by a fact that a solution of a current problem was indeed tackled before by some investigation which results were already published in some journal or conference paper, so that the results are known, and available to professional public. There are also different ways to tackle the problem of medical data readability: if the data is available in the form of some data-table or database then there is no need to search for a problem solution in journals abstract database. It is possible to analyze the database and to infer some conclusions that will hopefully lead to problem solution. This kind of approach requires the existence of some previous knowledge in the form of data-table – the larger the database is, the better. This database comprises of historical data that were collected during some period and usually corresponds to a certain medical situation. For example: database that contains the data about hart rate at rest, blood pressure levels, some hart function anomalies, etc. is usually used in cardiology or cardiovascular sub-domains. If analyzed, this kind of data can yield some important conclusions which will hopefully lead to a problem solution. The question of interest is: How to represent the data after analysis? Fortunately, there are some solutions offered in a literature. As in [3]: "In general, human-readable rule refers to data shown in a format easily read by most humans – normally in the form of IF...THEN rules. This kind of knowledge representation is the most convenient way for doctors to express their knowledge in medical diagnosis.". So, the research presented in this paper concentrates on analysis of table-organized data and representing the results of the analysis in the readable IF…THEN form. The next step is a transformation of multiple IF…THEN statements to a single sentence that is read by a machine text reader.

The paper is organized as follows: Section 2 deals with basic idea and techniques for data analysis which yields IF…THEN rules. Section 3 represents the data sample used for experimental research and results, while Section 4 gives conclusions.

2. Background

The main goal of this investigation is to analyze data-table or database in the domain of medicine, then to generate conclusions of the analysis in the readable IF…THEN form and to deliver these conclusions to the user, typically a doctor. The conclusions are delivered to the user by some machine text reader so that user is not forced to use paper, computer monitor, smartphone or any kind of similar device. Instead, the user can use some “hands–free” device such as headphones or speakers in order to get the results of the analysis “on the way”. Whole process is show on Fig. 1.

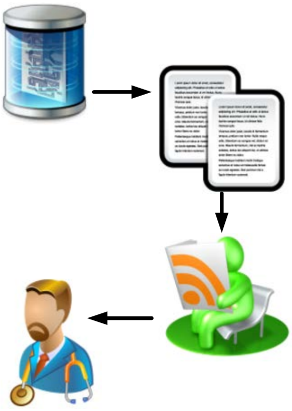

Fig. 1. The database (1) is transformed to a decision rules in the IF…THEN form and consequently to the readable text which contains linguistic terms and sentences (2), then the text is read to the doctor (4) by a machine text reader (3)

There are many IF…THEN rule generators or decision rules generators. Some most common systems are C4.5, C5 (See 5), Weka, etc. which are based on the application of decision trees. However, in this research is used the approach that is based on Systematic Syntax Classification of Objects (SSCO) [4, 5]. This approach is developed at Technical faculty “Mihajlo Pupin” in Zrenjanin, Serbia. The SSCO is comparable to systems that applies the Rough Sets Theory (RST) [6, 7, 8, 9] in order to generate decision rules. The systems which are capable to generate decision rules were used in the domain of medicine [10, 11, 12] and they showed to be suitable for the analysis of medical data.

2.1. Systematic Syntax Classification of Objects

Every object of the universe is described by certain amount of information expressed by means of some attributes used for object description. Let U be a finite set of objects (patients in this case), $Q=\{q_1,\dots ,q_m\}$ is a finite set of attributes, V_q is the domain of attribute q and set V is the union of all domain attribute sets V_q. There is a function f called information function, that defines the value for object’s attribute: $f=U\times Q\rightarrow V$. To every non–empty subset of attributes P is associated an indiscernibility relation on U, denoted by I_P:

$$I_P = \{(x,y) \in U \times U : f(x,q) = f(y,q), \forall q \in P\}. \tag{1}$$

The (1) is mathematical basis of RST, as well as SSCO based algorithms. This relation is an equivalence relation (reflexive, symmetric and transitive). If some attribute q has discrete values then number of classes generated by (1) performed on this attribute is $card(V_q)$, where $card$ stands for set cardinality. The j-th class of i-th attribute $C_{i,j}$ consists of those elements of universe for which some relation R is satisfied:

$$C_{i,j} = \{x : x \in U, R(f(x,q_i), v_{i,j}) = true\}. \qquad (2)$$

In (2) $v_{i,j}$ is j-th value of i-th attribute and f is an information function. If R is relation of equivalence, then indiscernibility (1) is preserved. It is possible to carry the task of state–space search in the manner of depth first search (DFS) algorithm. The root of the state–space graph is set $P \subseteq U$, while nodes are sets $C_{i,j}$. The arcs (operators) of state–space graph are defined by (2). Under the assumption that decision attribute (the attribute in the THEN part of the rule) is the last one, every path of state–space graph which ends with non–empty leaf will produce one decision rule. The attributes which are superfluous (not needed) are omitted from rules. The "roughness" in sense of RST is achieved in the case when consequent (THEN) part of the rule is formed by two or more leafs. Maximal number of nodes in state space is calculated by:

$$\sum_{n=1}^{|C \cup D|} \prod_{i=1}^{n} |V_{qi}| . \qquad (3)$$

In (3) C is a set of attributes that make antecedent part of the rule, while D is a set of attributes that make consequent part of the rule. There is usually only one binary attribute in set D. As (3) produces very large numbers, even for a fairly small data set, the memory consumption is rather high, but this depends on the algorithm that is used for state–space search. The algorithm used here is based on LIFO structure so that memory consumption is optimal.

3. The Experiment

The experiment was conducted on the data set consisting of 180 objects (patients). The data was sampled at General hospital "Djordje Joanovic" in Zrenjanin, Serbia. Attributes and their respective descriptions are given in Table 1. The data relates to patients who were hospitalized because of hip surgery.

Table 1. Attributes and their descriptions

No.	Attribute	Description
1.	Age	Age in years.
2.	LoH	Length of hospitalization in days.
3.	SP	Surgical procedure.
4.	DS	Delay to surgery in hours.
5.	PRBCs	Blood transfusion - The number of units of packed red blood cells which patient received during rehabilitation treatment.
6.	CD	Cardiovascular disease.
7.	DM	Diabetes Mellitus.
8.	CVI	Medical data about cerebrovascular insult.

9.	PD	Psychiatric diagnosis. Mental and behavioral disorders.
10.	D	Dementia was ascertained based on relevant medical record.
11.	LReh	Length of rehabilitation in days.
12.	MC	Medical data about the mental confusion.
13.	Coop	Patient cooperation.
14.	SW	Salvati Wilson’s score on discharge.

The aim of this experiment was to assess the relations among patient characteristics (attributes 1 – 13), and Salvati-Wilson’s (SW) score on discharge (attribute 14). In other words, the aim was to assess how attributes relates to high or low SW score. The SW score is a measure of results of early functional recovery after hip surgery. After applying SSCO based algorithm 175 rules were generated. Some of them are given below:

```
1. IF (Age is [84,*)) and (LoH is 36) THEN (SW is low)
2. IF (Age is [84,*)) and (LoH is 31) THEN (SW is low)
3. IF (Age is [84,*)) and (LoH is 30) THEN (SW is low)

4. IF (Age is [84,*)) and (LoH is 24) and (DS is more48)
       and (PRBCs is 6)
   THEN (SW is high)
5. IF (Age is [84,*)) and (LoH is 24) and (DS is more48)
       and (PRBCs is 3)
   THEN (SW is low)
```

The results are in the form of IF…THEN rules which are fairly readable by humans. The rest of generated decision rules are similar except that they include different attributes in order to estimate the value of SW score. However, the goal is to achieve even more readable form, e.g. first three rules can be condensed so that they form one rule or one sentence expressed by means of linguistic terms:

If a patient is older than 84 years and length of hospitalization is more than 30 days, then SW score is still low.

This means that doctor must undertake some different treatment. Rules 4 and 5 can be also condensed so that they form single sentence:

For a very old patient who stayed in hospital for 24 days and delay to surgery was more than 48 hours it is needed that a number of units of packed red blood cells is bigger or equal to six in order to have high SW score.

Now, there is a possibility to deliver the conclusions of analysis expressed in form of readable sentences to user (usually doctor), by means of automated text reader. There are many text readers that are available online or even incorporated to MS Office package. In this research, the “Natural Reader” system which is available at http://www.naturalreaders.com/index.html was used for initial experiments.

4. Conclusions

As the readability of medical research journal abstracts changed drastically from 1960 to this day in a way that medical journal abstracts have become more difficult to read, there is the urgent need for information source that is reliable, responsive and, above all easy to understand.

This research deals with synthesis of the readable sentences from medical data presented in a form of data table or database. This data is usually historical data that consists of records from some particular medical sub–domain. In this case data sample includes data about patient's conditions after hip surgery procedure and was collected at General hospital "Djordje Joanovic" in Zrenjanin, Serbia. So called SSCO based algorithm was applied to available data and decision rules in the IF…THEN form were generated, few of them are shown in previous section. Decision rules were consequently condensed (joined) to form even more readable sentences which include linguistic terms. These sentences are delivered to the user by the machine text reader.

The importance of solutions presented in this paper are:

- The user (usually doctor) is able to acquire data in a form of spoken words.
- The system enables "hands–free" communication so that the user is able to continue the tasks without interruptions.
- System is fast and reliable.
- Communication by using natural language is enabled.

In a future, the solution will be improved in order to generate even more complex sentences.

Acknowledgment

Ministry of Science and Technological Development, Republic of Serbia financially supported this research, under the project number TR32044: "The development of software tools for business process analysis and improvement".

5. References

1. Severance, S., Bretonnel, Cohen, K.: Measuring the readability of medical research journal abstracts, Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 127–133, Beijing, China, Association for Computational Linguistics, (2015)
2. Coleman, M., Liau, T. L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60 (2), 283, (1975)
3. Daud, N., R., Corne, D., W., in Mastorakis et al. (ed.): Proceeding of European Computing Conference, Lecture Notes in Electrical Engineering 27, DOI: 10.1007/978-0-387-84814-3_79, Springer Science + Business Media LLC, (2009)
4. Brtka V.: Automated Synthesis of Rule Base in Inference Systems, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia, (2008)
5. Ognjenovic, V., Brtka, V., Berkovic, I., Brtka E.: "Comparison of the classification rules generated by See 5.0 and SSCO Systems", Proceedings of the 23rd Central European

Conference on Information and Intelligent Systems - CECIIS, Varaždin, Croatia, September 19-21, ISSN 1847-2001, pp. 71-76, (2012)

6. Pawlak, Z., Grzymala-Busse, J., Slowinski, R., and Ziarko, W.: Rough sets, Association for Computing Machinery. Communications of the ACM 38, (1995)
7. Pawlak, Z., Skowron, A.: Rudiments of rough sets, An International Journal of Information Sciences 177, pp. 3–27, (2007)
8. Pawlak, Z.: Rough set approach to knowledge-based decision support, European Journal of OR 99, pp. 48-57, (1997)
9. Greco, S., Benedetto, M., Slowinski, R.: New Developments in the Rough Set Approach to Multi – Attribute Decision Analysis, Bulletin of International Rough Set Society, Volume 2, Number 2/3, pp. 57–87, (1998)
10. Brtka, V., Stokic, E., Srdic, B.: Automated extraction of decision rules for leptin dynamics - A rough sets approach, Journal of Biomedical Informatics, Vol. 41, No. 4, pp. 667-674, (2008)
11. Stokic, E., Brtka, V., Srdic, B.: The synthesis of the rough set model for the better applicability of sagittal abdominal diameter in identifying high risk patients, Computers in Biology and Medicine, Vol. 40, No. 9, pp. 786-790, (2010)
12. Øhrn, A.: Discernibility and Rough Sets in Medicine: Tools and Applications, Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491 Trondheim, Norway, ISBN 82-7984-014-1, ISSN 0802-6394, (1999)