Methodology for Preprocessing Semi-Structured Data for Making Managerial Decisions in the Healthcare*

Elena Makarova[0000-0002-5410-5890] and Dmitriy Lagerev[0000-0002-2702-6492]

Bryansk State Technical University, Bryansk, Russia
lennymbear@gmail.com, LagerevDG@mail.ru

Abstract. This paper describes the process of supporting management decision-making in healthcare based on data mining. The authors described various problems and specifics of data in medical information systems, leading to the complexity of their analysis and integration, such as: the presence of a large number of specific abbreviations, errors in the data and their poor structure. The paper demonstrates an approach to the search and further disclosure of abbreviations in texts, built on a combination of machine and human processing. A method for extracting features from semi-structured fields using an expert in the subject area and using various visualizations is proposed. The proposed abbreviation search and disclosure methods, based on a hybrid approach combining the strengths of processing with the help of a machine and an expert, can increase the number of abbreviations found automatically and significantly reduce the time spent by experts on processing the remaining reductions. In addition, the method for automated feature extraction during integration can significantly increase the amount of useful input data, while reducing the time of the expert.

Keywords: Natural Languages Processing, Data Integration, Healthcare.

1 Introduction

The digitalization of Russian medicine poses new challenges for managers and engineers - implementation, security and support issues, large data storage and processing systems. But the collection of this data in digital format, in turn, opens up new opportunities for researchers and healthcare managers through the use of data analysis technologies.

Over the years of informatization of various cities and regions of the Russian Federation, in medical information systems (hereinafter referred to as MIS), more and more data has been accumulating on various aspects of the work of medical organizations - from medical histories and prescriptions of specific patients to various aspects related to providing medical institutions with necessary medicines and supplies materials.

* The reported study was funded by RFBR, project № 20-04-60185.

Thus, the improvement of analysis technologies and the filling of databases with medical data makes it relevant to use data mining for tasks related to healthcare management [1], such as: planning material and human resources in healthcare, forecasting statistical indicators, optimal and timely provision of resources, tracing outbreaks and spreading of diseases.

These processes explain the relevance of solving the problem of creating an automated system to support managerial decision-making in healthcare and solving side problems associated with the implementation of this system.

The emphasis is on the need for both strategic (development of the healthcare sector in the region) and operational (decisions at the level of a medical organization, response to outbreaks of diseases, etc.)

2 Management decision-making in healthcare

The forecasting task is very relevant for healthcare: it is necessary to predict the incidence rate, assess the required resources to maintain the effective operation of the system, etc. Researchers devote much attention to the problems of disease prognosis in specific regions [2].

Forecasting is a task that can be solved in many ways - from classical statistical methods to models based on machine learning technology. Recently, the neural network approach has become widespread. [3]. For example, recent research show that the accuracy of predicting many diseases using CNN is greater than, for example, the KNN and NB algorithms. [4].

In addition, a research that used deep learning to process textual medical data (various models of embedded representations of words were used) showed an increase in forecasting accuracy [5]. To predict a number of diseases, classic time series are best used, using only numerical values (for example, injuries, SARS, etc.). However, in which groups of diseases it is necessary to apply more complex approaches to the analysis. For example, when predicting malignant neoplasms, it is necessary to understand not only the general characteristics, but also the number of patients at different stages of cancer. In a research on the prediction of breast cancer using machine learning, patients were divided into cohorts depending on the stage of the disease and other parameters, which allowed better identification of factors contributing to patient survival [6].

To make a decision on the distribution of resources between various medical institutions, the regional health department needs to make a forecast about the development of certain diseases and act according to long-term planning. The process of creating such a forecast can simplify the development of an automated system to support management decisions. (Fig 1)

The integration and setting up of this process requires the investment of certain human resources, however, given the need to regularly make such management decisions and constant updating of data in the regional information system (RIS), in the long term, these labor costs will be justified. The general scheme of this process is shown in Figure 2.

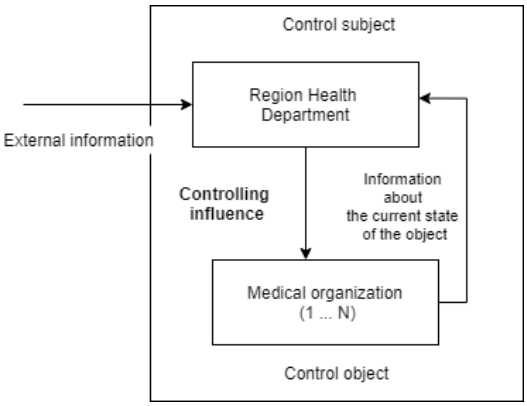

Fig. 1. The management decision-making in healthcare

Health resource management conceptual model:

$$S = < R, M, D, Z; I>, \quad (1)$$

where: R - available resources (budget for all medical institutions);

M - budget for a specific medical organization (MO) for various articles (equipment, maintenance of an inpatient hospital, procurement of medicines, rates of health workers, etc.);

D - effectiveness of the development of this budget;

Z - requests for resources (the current need for various MO in them);

I - information available for analysis and forecasting resource requirements.

The most time-consuming step in data mining is still the process of collecting, cleaning and pre-processing data before analysis. According to various researchers, this process takes from 60% to 80% of the time [7].

In previous works of the authors [8] concerning the process of collecting and processing semi-structured data, much attention was paid to the “hybrid” approach in the field of developing data analysis systems. In this approach, human expertise is used in conjunction with automatic analysis methods, which allows, on the one hand, to improve the quality of the system on tasks that cannot be solved without human intervention, and on the other, to relieve the expert from solving typical, routine tasks. This was achieved using various methodologies for pre-processing and data visualization, which helped the expert make faster decisions about the inclusion / exclusion of a particular data source [8].

Data collection is only the first step in the data preparation process for use in ensembles of data mining models. In addition to the general problems for all subject areas arising at the stage of data preprocessing, when constructing analytical models for the analysis of biomedical data, researchers and developers encounter a number of problems specific to the described data, which will be discussed in more detail in the next section.

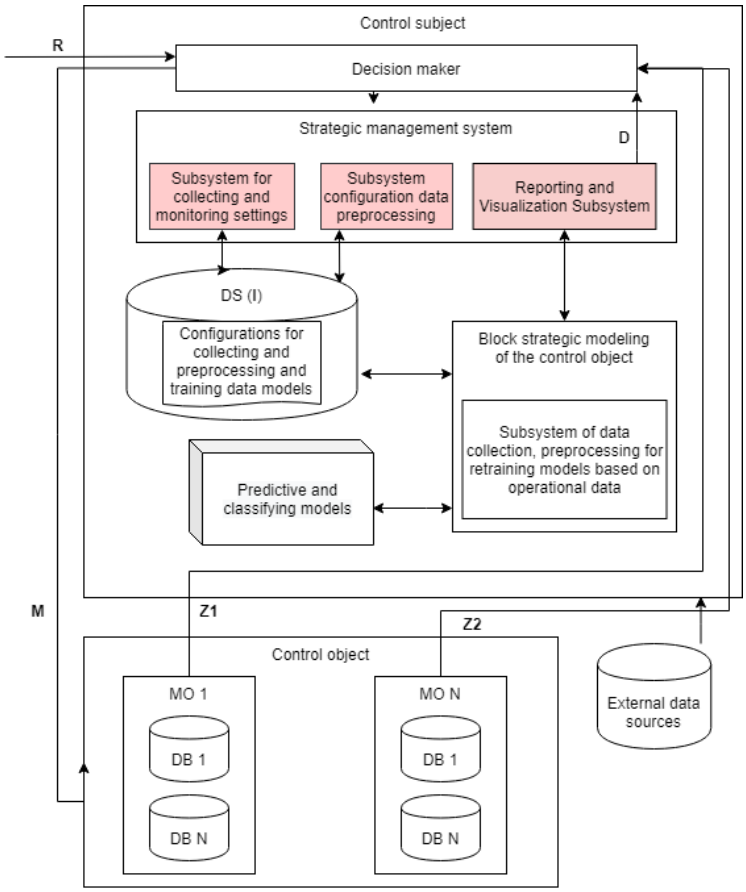

Fig. 2. The general scheme of management decision-making in healthcare

3 Pre-processing of medical data for analysis

Sample Heading The most common methods for implementing data integration are:

1) file-based sharing;
2) data replication;
3) Web services technology;
4) Service Oriented Architecture (SOA);
5) integration servers [9].

In the USISH project (Unified State Information System in Healthcare), an document-based approach is proposed as the main method of integration between different RIS [10]. Because Since various medical documents are poorly structured, we are faced with the problem of correlating various data. For example, processing the parameters “general patient indicators”, which may be different depending on the medical institution - weight, height, blood pressure, blood glucose, etc. Resolving such ambiguities is

also an important part of solving the integration problem. For example, the same indicator in one database is called “growth”, and in another - “body length”.

The problem of poorly structured data is that the ICD-10 classification does not contain details on the diagnosis, doctors must fill out this information on their own according to pre-created forms in the system or in free form.

Also in the field of free entry is usually indicated: degree (stage) of the disease, form, prescribed medications and other. For example, regarding the classification of malignant neoplasms according to ICD-10, the classifier does not reflect the stage of the disease. Usually it is indicated in natural language in another field. However, the availability of these data could constitute a better prognostic model for the stages of the disease. So, for example, for patients with a malignant formation of the first stage, there is a significant risk under certain circumstances, an article by patients with a second stage of cancer, second - third are at the second - go to the third, etc.

Here are a few examples of such an ICD-10 diagnosis uncertainty. For example, the ICD-10 code “S82.6” (fracture of lateral malleolus) should have at least an explanation of the right leg or left leg (which can be expressed in free form as «левая», «слева», «левый», «лев», «л.», etc.), but also an indication of whether the fracture is closed or not, complete or incomplete.

For example, Table 1 presents some examples of how certain important terms are indicated. In addition, in some cases it is not clear that this is an abbreviation, a specific term or word spelled out with errors.

The database also contains specific grammatical constructions that make it difficult to extract features. For example, when identifying symptoms by standard methods, the phrase from the patient’s history “did not have hepatitis, tuberculosis”, information about hepatitis and tuberculosis without mentioning “not sick” could fall into the patient’s model, which would worsen the quality of the prediction models.

Table 1. Examples of data

Abbreviation	Whole word	Close tokens
“отр”	отрицательный	“отриц”
“хр"	хронический	“хрон”, “хроничекий”
“бер”	беременность	“бер-ть”, “бер-сть”, “берем’
“нед”	неделя	“ндл”, “неделль”
“отр”	отрицательный	“отриц”

4 A hybrid approach to finding and revealing abbreviations, incorrect spelling of words

Based on the available data, various approaches to finding abbreviations have been tried, from a standard approach based on regular expressions and a dictionary of commonly used abbreviations. Since many abbreviations in the sample are specific, the

combined methods based on a combination of heuristic, vocabulary, and statistical approaches gave the greatest increase in accuracy. A detailed description of this approach and the results of its use are described in a previous work of the authors [12].

However, it is not yet possible to reveal specific abbreviations absolutely precisely in a fully automated mode. One way or another, when solving this problem, you will have to turn to knowledge by a competent person. By analyzing the context of abbreviations, it is possible to significantly reduce the degree of expert intervention, if we train the word embedding model on the available data and predict the overall meaning of abbreviations depending on the context. There is enough data when training the model to establish syntagmatic and paradigmatic relationships.

The semantic similarity between linguistic units is calculated as the distance between vectors. In studies on distributive semantics, the most often used cosine measure, which is calculated by the formula

$$\text{sim} = \frac{\sum_{i=1}^{n}(\boldsymbol{A}_i \times \boldsymbol{B}_i)}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{A}_i)^2} \times \sum_{i=1}^{n}(\boldsymbol{B}_i)^2}, \quad (2)$$

where A and B are the vectors of words, the distance between which is calculated.

In this study, vectorization was implemented using the Bag of Words method [13], but other methods are also possible. For this sample, a sufficiently high threshold is a value of 0.7

$$\text{sim} \geq 0{,}7 \quad (3)$$

In addition, in order to automatically recognize the word appropriate context abbreviation, three conditions are necessary:

the abbreviation coincides with the beginning of a semantically close word;

the found word is not an abbreviation and is present in the dictionary of used words of the Russian language;

the found word is the only one satisfying the first and second conditions in the range of semantic proximity from 0.7 to 0.99.

For words that do not coincide with the abbreviation but are close in terms of the cosine measure, the Tanimoto coefficient is used with a match value greater than or equal to 0.5 [14]. For example, we calculate the syntactic proximity of the abbreviation "стд" and the word "стадия":

$$k = \frac{c}{a + b - c}, \quad (4)$$

where a, b are the number of elements in the “стд” token and the “стадия” token, respectively;

c is the number of common elements in the "std" and "stage" tokens.

In the standard setting, k is taken to be large 0.5 for words with a non-matching beginning and ending (“стд” and “стадия”) and 0.35 for words with a matching beginning and ending and containing a hyphen (“бр-ть” and “беременность”). In previous

work of the authors was presented word embedding visualization technique for these tasks analytics, which was also implemented in this case.

Of the available sample of depersonalization records from *integrated electronic medical records* (IEMR) of approximately 1.4 million residents of the Bryansk region, a sample of 60,000 records was created, balanced by diagnosis and length of text description, of which 3,000 records were similarly selected. Each of abbreviations was manually specified to verify the developed methodology. Next, a comparison was made of the results of a fully manual approach, a fully automated and the hybrid approach described above. Results are presented in table 2.

Table 2. Search and disclosure of abbreviations

Approach	Processing time 60,000 records	Number of found and disclosed abbreviations
Fully manual filling	About 410 hours	Close to 100%
Fully automated approach	5 to 10 minutes*	Up to 53%
Hybrid approach	20 to 67 minutes *	83-90%

The interface of a specialist in the subject area for marking up data when implementing manual (expert) control is presented in Figure 3. The proposed approach checks not only all cases that do not go beyond the boundaries of automatic marking, but also 5-10% of instances automatically classified by the system in order to verify the correct operation of the algorithms and their settings, if necessary. The large spread in labor costs for the expert when marking up the data is explained by the number of checked examples and the severity of the thresholds for automatic marking depends enough. In the described experiment, the choice of the percentage of data considered by experts and trusted by the system depends on the accuracy requirements and is limited by available resources.

Fig. 3. The interface of a specialist in the subject area for marking up data

5 Visual interface for data extraction

The solution to the problem of bringing various abbreviations with a single value is only one way to reduce the number and improve the quality of features that are input to analytical models. Using a large number of features (some of which will be duplicate, some will be useless) is irrational. There are problems such as overfitting, an increase in processing time and the presence of "noise" and "garbage data".

In addition to selecting features that will go into the model, they often need to be additionally extracted from poorly structured data. In this case, we also use a hybrid approach (Fig. 4).

In this approach, word2vec models are used (to select contextually close tokens) and a visual editor (Fig. 5). The main visualization metric here is coverage of features. It is calculated on a weighted limited sample. In this case, 3000 records to quickly recount the results and provide interactive visualizations.

As the main extraction method used rules based on the principles of regular expressions. A user who is an expert in this field, but does not understand regular expressions and word processing, is provided with a visual editor of these expressions and instructions for use. After re-calculation, several random entries are also presented for manual evaluation of the created rules. The results of these evaluations are saved for automatic validation when the rules change.

Table 3 presents the results of an experiment conducted on a sample of oncological diagnoses, where important metadata, such as the "stage", were described in the free entry fields.

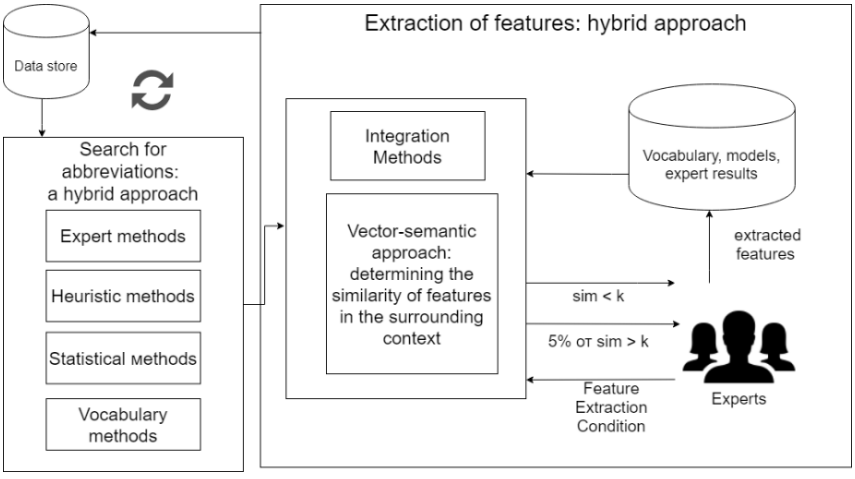

Fig. 4. Extraction of features: hybrid approach

Table 3. Feature extraction in database integration.

Approach	Processing time 9,000 records	Number of extracted features
Fully manual filling	About 18 hours	Close to 100%
Fully automated approach	5 to 10 minutes *	About 40%
Hybrid approach	30 to 120 minutes *	Up to 90%

6 Conclusions

To effectively manage healthcare resources, it is necessary to collect, save and analyze data received from all regions of the Russian Federation. At the moment, one of the main methods of data integration in the USISH project is integration through documents. Since documents are a poorly structured source of information, with such integration there are problems associated with the interpretation of various text data, as well as problems of their quality: the presence of specific abbreviations, errors, difficulties in extracting various features, etc. The presence of a large number of noise, duplicates, and incorrect features degrades the quality of data analysis models.

The proposed abbreviation search and disclosure methods, based on a hybrid approach combining the strengths of processing with the help of a machine and an expert, can increase the number of abbreviations found automatically by 21%, as well as detect in automated mode up to 55% of cases (with a probability of correctness higher 70%) and significantly reduce the time spent by experts on processing the remaining reductions.

In addition, the method for automated feature extraction during integration can significantly increase the amount of useful input data, while reducing the time of the expert.

Using a hybrid approach to preprocessing poorly structured data increases the efficiency of managerial decisions in the field of healthcare by increasing the reliability of data mining models and reducing the time spent by experts on their creation and support. A further line of work in this area will be directed to the development of methods for the semi-automatic selection of features for analytical models.

Fig. 5. The interface for features extraction

References

1. Zakharova, A.A., Lagerev, D. G., Podvesovskii, A. G.: Multi-level Model for Structuring Heterogeneous Biomedical Data in the Tasks of Socially Significant Diseases Risk Evaluation. In: 3rd Conference on Creativity in Intelligent Technologies and Data Science, CIT and DS 2019, pp. 461-473, Volgograd (2019)
2. Choporov, O.N., Zolotuhin, O.V., Bolgov, S.V.: Algoritmizaciya intellektual'nogo analiza dannyh o rasprostranennosti zabolevanij na regional'nom i municipal'nom urovnyah. In: Modelirovanie, optimizaciya i informacionnye tekhnologii № 2 (9), (2015)
3. Lazarenko, V.A., Antonov, A.E.: Diagnostika i prognozirovanie veroyatnosti vozniknoveniya holecistita na osnove nejrosetevogo analiza faktorov riska. In: Issledovaniya i praktika v medicine. №4(4), pp. 67-72. (2017) https://doi.org/10.17709/2409-2231-2017-4-4-7
4. Dahiwade, D., Patle, G., Meshram, E.: Designing Disease Prediction Model Using Machine Learning Approach. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215, Erode, India (2019) https://doi.org/10.1109/ICCMC.2019.8819782
5. Christensen, A., Frandsen, A., Glazier, S., Humpherys, J.: Machine Learning Methods for Disease Prediction with Claims Data. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 467-474, New York, NY (2018). https://doi.org/10.1109/ICHI.2018.00108

6. Shukla, N, Hagenbuchner, M., Win, T. K.: Breast cancer data analysis for survivability studies and prediction. In: Computer Methods and Programs in Biomedicine (2017) https://doi.org/10.1016/j.cmpb.2017.12.011
7. Lohr, S.: For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights, http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0 Last accessed 14 July 2020
8. Makarova, E., Lagerev, D., Lozbinev, F.: Approaches to visualizing big text data at the stage of collection and pre-processing. In: Scientific Visualization N. 11.4, pp. 13–26, (2019). https://doi.org/10.26583/sv.11.4.02
9. Karpov, O.E., Gavrishev, M.YU., SHishkanov, D.V.: Integraciya medicinskoj informacionnoj sistemy i sistemy administrativno-hozyajstvennoj deyatel'nosti kak instrument optimizacii processov medicinskoj organizacii. Otdel'nye problemy i puti ih resheniya. In: Sovremennye naukoemkie tekhnologii. № 9-1. pp. 46-50. (2016)
10. Portal of operational interaction of USISH participants http://portal.egisz.rosminzdrav.ru/materials Last accessed 14 July 2020
11. Kreuzthaler, M., Oleynik, M., Avian, A., Schulz, S.: Unsupervised Abbreviation Detection in Clinical Narratives. In: Studies in Health Technology and Informatics. v. 245, pp. 539–543 (2016)
12. Lagerev, D., Makarova, E., Features of preliminary processing of semi-structured medical data in Russian for use in ensembles of data mining models. 2020. T. 17, № 7. pp. 43–53. https://doi.org/ 10.14489/vkit.2020.07.pp.043-053
13. Zellig, S. H.: Distributional Structure. v.10. pp. 146-162, Word (1954), https://doi.org/10.1080/00437956.1954.11659520
14. Tanimoto, T.T.: IBM Internal Report 17th Nov. IBM. Corp, New York (1957).