International Conference Information Systems 2017 Special Interest Group on Big Data Proceedings

Special Interest Group on Big Data Proceedings

2017

The Demand for the Healthcare Services: the Opportunities of Big Data in Predicting Patient Flow

Svetlana Maltseva
National Research University Higher School of Economics, smaltseva@hse.ru

Elizaveta Prokofyeva
National Research University Higher School of Economics, prokofyeva.liza@gmail.com

Roman Zaitsev
Moscow Institute of Physics and Technology, roman.zaitsev@phystech.edu

Recommended Citation

Maltseva, Svetlana; Prokofyeva, Elizaveta; and Zaitsev, Roman, "The Demand for the Healthcare Services: the Opportunities of Big Data in Predicting Patient Flow" (2017). *International Conference Information Systems 2017 Special Interest Group on Big Data Proceedings*. 5.
http://aisel.aisnet.org/icis2017b/5

THE DEMAND FOR THE HEALTHCARE SERVICES: THE OPPORTUNITIES OF BIG DATA IN PREDICTING PATIENT FLOW

Research paper

Maltseva, Svetlana, National Research University Higher School of Economics, Moscow, Russian Federation, smaltseva@hse.ru

Prokofyeva, Elizaveta, National Research University Higher School of Economics, Moscow, Russian Federation, prokofyeva.liza@gmail.com

Zaitsev, Roman, Moscow Institute of Physics and Technology, Moscow, Russian Federation, roman.zaitsev@phystech.edu

Abstract

Nowadays, healthcare field is tightly connected with information technologies, in particular, big data technologies. The simulation of the patients flow in the health care system can significantly enhance the effectiveness of its work. The main purpose of the model improvement is to take into account the features of patient care to enable end users to obtain forecast values at the output. Thus, these values can subsequently become the starting point for decision-making by the management. In this paper, the analysis of the demand for medical services in European hospitals was carried out to determine the required amount of resources within the use of the modern analytical methods. The study contains a description of big data technologies in healthcare; executes the beneficial side of its implementation. The research showed that the science community made a great effort in the development of the industry, however, privacy and security issues, standards establishments still require enormous attention and new efforts to be made in the study area. The study is also focused on the predicting the resources, which are needed for a medical institution.

Keywords: Big data, Healthcare, Technology, Hospitals.

1 Big Data opportunities in healthcare

Rapidly evolving big data technologies play an important role in healthcare. One of the leading trends in this area is a personalized model of services for patients. In a patient-centred model, data analysis technologies allow the user to create an individual portrait, taking into account the medical history, age characteristics, emotional experience and other relevant parameters. Another trend that stimulates the development of personalized services is the Internet of Things in medicine: devices, which are connected to the Internet, exchange the data by means of built-in sensors to monitor the patient's health in real time. There are several sources that are important for applying the technology of big data analysis in healthcare: genomics, electronic medical records (EHR), and devices for monitoring the patient's health, wearable video devices and applications for mobile phones. Therefore, social media updates, patients' chats, digital prescriptions, pharmacy reviews, news feed, online consultations and others are becoming an outstanding source of new valuable knowledge.

1.1 Challenges and barriers

New technology implementation is always associated with a number of challenges. The case with big data technologies in healthcare is not the exception. First, as it was mentioned before, these solution are complex and not user friendly. Besides, they involve advanced programming skills for those, who want to gain real benefits from its implementation. In healthcare field one of the most significant challenges is the variety of data formats, which come from different sources in the unstructured format [1].

Another huge barrier for big data technologies implementation is the psychological issue. The responsible managers still are tending to underestimate the power of advanced analytics solutions. Taking into account that these solutions are rather expensive, the managers attempt to escape these expenditures of the unknown efficiency for decision making. Therefore, the difficulty of funding trustful investors is also another challenge for that specific case.

Besides, the relevancy of gathered data is becoming a challenge, when implementing advanced analytics solution. Since big data tools allow collecting vast amount of data, which is not always valuable, it is up to the responsible managers to decide, what specifically to collect.

Another issue that appears in challenges of managing big data in healthcare is the ownership [1]. Logically, patients would think that they own their personal healthcare data; this may not always be the right conclusion. Healthcare cards, disease stories might also belong to the provider of services.

Finally, the human resources factor presents a barrier, because the technology is rather new on the market, therefore, it is a challenging task to find the data scientists in healthcare field for the appropriate amount of money.

2 Methodology

Modern techniques of data analysis are extremely important for solving the problem of optimizing the workload of medical institutions. The purpose of this study is to establish the dependency between the indicators of the effectiveness of the medical institution and its resource base.

Hospital discharges by diagnosis; hospital days of in-patients and in-patient average length of stay were selected as the performance indicators and the demand of the medical facility. The hospital beds by type of care, medical technology (magnetic resonance tomography, gamma cameras, angiographic complexes and lithotripters) and physicians characterized the resource provision of medical institutions for the developed models.

The data source for the research was an open database of the statistical service Eurostat. The choice of the source is because the databases contain complete and open information necessary for research tasks in the field of public health. In addition, the statistical database has a user-friendly interface that allows one to build analytical reports. The study provides information on 28 European for the period from 2007 to 2016.

3 Data analysis and models development

At the initial stage of the study, a basic regression model was constructed for the three variables: patient discharges from hospital, hospital days of patients, and average length of stay in the hospital. Regression analysis shows that the highest coefficient of determination R2 = 0.97 for the hospital discharges model, which indicates a strong relationship between the variables, close to the functional dependence. The model of the average length of stay in the hospital has the lowest coefficient of determination; therefore, this model does not reflect the real dependency between the analyzed variables. In order to confirm the absence of strongly correlated repressors, the VIF (variance inflation factor) values were calculated to detect multicollinearity. The results in models did not exceed the acceptable value: VIF1 (rs_bds) = 4.15, VIF2 (rs_equip) = 7.71, VIF3 (rs_phys) = 7.35.

The following models after regression analysis were obtained:

1. The model of "Hospital Discharges by Diagnosis":

Y_1=21,24 ·rs_bds + 543 ·rs_equip + 10,9 ·rs_phys ;

2. The model of "Hospital Days of In-patients":
Y_2=1920124+167,87 ·rs_bds + 30299,97 ·rs_equip - 72 ·rs_phys ;

3. The model of "In-patient Average Length of Stay":
Y_3=8,15+0002 ·rs_equip - 0,000012 ·rs_phys

The following notation is used in the above formulas:

Y_1- Hospital discharges by diagnoses, when the patient officially leaves the medical institution after the treatment. The causes of discharge include the completion of the official treatment schedule, the completion of treatment of patient's own decision, transfer to another institution or death. The indicator is measured in number of discharges;

Y_2- Hospital days of in-patients: the total number of days, when the registered patients occupied beds in the medical institution from the time they were admitted to discharge date. The indicator is measured in the number of days;

Y_3- In-patient average length of stay, ALOS: one of the main statistical indicators in a medical institution [2]. This indicator is calculated by dividing the number of days (hospital days of patients) in the facility by the number of discharges (including deaths) during the year;

rs_bds – hospital beds by type of care;

rs_equip – medical technology: the number of medical equipment units, including magnetic resonance tomography, gamma cameras, angiographic complexes and lithotripters;

rs_phys (physicians) – number of physicians in a medical facility.

The influence of the number of hospital doctors on the total number of discharges is almost half-smaller than the influence of the number of beds. In addition, the development of technology and the acquisition of new medical equipment should positively influence the dynamics of hospital discharge, which confirms the obtained model. For example, the availability of expensive modern equipment in the hospital allows specialists to carry out procedures that are more complex and to accept patients who require long-term treatment. The influence of this repressor on the total number of discharges is 25 times greater than the first explanatory factor. Thus, according to the results of this model, the key factor in increasing the number of discharges from the medical institution is the equipment and development of the technological component.

3.1 Cluster analysis application

An attempt to improve the quality and the interpretation of the models was made by cluster analysis of the investigated set of countries.

The main idea was to assess the similarity of the joint behavior of the variables throughout the time under consideration to identify groups of similar countries and to construct the separate regression models for them. Therefore, the original time series were used as the objects of clustering. For such a time period it is more possible to estimate only the similarity of the shape of the time series curves, a nonparametric measure of similarity was chosen on the basis of the time correlation coefficient [3]. Time correlation allows one to estimate the joint monotonicity of time series for coincident lags, without requiring the stationarity of the series.

This measure and its modifications show effectiveness when solving such a class of problems even in comparison with traditionally used the Euclidean distance and the dynamic time warping (DTW) algorithm [4].

In addition, the choice of nonparametric difference measures was reasonable due to additional causes. First, the use of a parametric model to compare time series could lead to additional errors caused by an inaccurately chosen type of parameterization. Also, since the length of the time series is sufficiently small in comparison with the step of the series, it seems practically impossible to single out any reliable periodic components. Finally, the nonparametric measure of the difference between time series described above has a low (linear) computational complexity, which simplifies the spread of the approach applied in the article to big data.

The hierarchical agglomerate algorithm k-medoids was used, in which the sampling objects themselves were used as the centers of the obtained clusters, since determining the centroid when working with time series involves additional difficulties. To estimate the optimal number of clusters, the silhouette coefficient was used [5].

4 Discussion

The programming language R was chosen for the necessary data analysis tasks. This language is well suited for research purposes, since it contains a rich library of packages for various scenarios. The use of the R language helps to visualize the data to understand the general picture of the studied subject area [6].

For all countries included in the study, with the most accurate and complete data for the period under review, predictive models were developed based on historical panel data.

Forecast for the first model "Hospital Discharges by Diagnosis" allowed to get the number of the discharged patients. The discharges from hospital stored in the Eurostat database for the period from 2007 to 2014 were used as the input values. The forecast was created for 2015 and 2016 in order to assess the error and reliability of the developed model. To obtain the primary results of the study, the time series prediction function was used. The value of MAPE (average absolute percentage error) was calculated to estimate the prediction error, and for the first model the value was about 8%, which allows considering this forecast to be useful for resource planning of the medical institution in the short run.

After the cluster analysis it was possible to significantly improve the predictive power of the models: for example, in the one of the clusters, MAPE error was only 0,82%, which makes it possible to conclude that this forecast is highly reliable in the short term. For the first cluster, the MAPE index has become 4.48%, which is almost twice better than the original model. As a result, it is important to note that if the model is built at once over the entire sample, then the error will be quite large, and some abnormal objects (for example, Malta) might increase the error from 8% to 32%. Thus, the value of using cluster analysis is to reduce the forecast error in the short term perspective taking into account the features of the sampling objects.

Conclusion

Currently, data analysis has a huge potential, which allows to significantly improving healthcare services. Medical institutions that are the first to introduce these technologies will certainly have a competitive advantage. Managers of the facilities will be able to access full information, which will allow them to make more informed decisions.

This study showed that an application of modern data analysis methods in healthcare is an important direction to improve performance of the medical facility. As a result of the research, the models for estimating the demand for medical services in European countries were obtained: patient discharge from hospitals and the length of their stay at the medical facility were analyzed. A cluster analysis of the objects was also carried out to improve the quality and interpretation of the obtained models. As clustering methods, a hierarchical agglomerate algorithm was used, as well as k-medoids, a modification of the k-means algorithm. The obtained predicted values of the developed models have a relatively low error and can be used to make decisions on the resource provision of the hospital by medical personnel.

References

1. Burghard C. (2012) Big Data and Analytics Key to Accountable Care Success. IDC Health Insights.
2. Carter, E.M., Potts, H.W.W. (2014) Predicting length of stay from an electronic patient record system: a primary total knee replacement example. BMC Medical Informatics & Decision Making. 14: 26. doi:10.1186/1472-6947-14-26.
3. Chouakria A. D., Nagabhushan P. N. (2007) Adaptive dissimilarity index for measuring time series proximity.Advances in Data Analysis and Classification, vol. 1, no 1, pp. 5-21.
4. Zaitsev R., Britkov V. (2015) Primenenie yazyka R dlya mnogomernoj klasterizacii vremennyh ryadov s cel'yu analiza dinamiki nauchno-tekhnicheskogo razvitiya [The use of R for multidimensional clustering of time series to analyze the dynamics of scientific and technological development]. Proceedings of the Second Youth Scientific Conference "Problems of Modern Informatics", pp. 92-98.
5. Rousseeuw P. J. Silhouettes (1987) A graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, vol. 20, pp. 53-65.
6. Balan, S., Otto, J. (2016) Big Data Analysis of Home Healthcare Services. Information Technology and Management Science, vol. 19, pp. 53-56. doi: 10.1515/itms-2016-0011.