Short Report
UDC 005.94:[005.511:519.246.8
DOI 10.7251/IJEEC2206036K

Possibility of Applying Data Mining in Health Insurance

Dragoljub Krneta, Sofija Krstev

Dwelt d.o.o, Banjaluka, Bosnia and Herzegovina

E-mail address: dragoljub@dwelt.net, sofija@dwelt.net

Abstract—**In this paper, a great range of possible applications of data mining, data warehousing and knowledge discovery is described. With the increase of costs and complexity of healthcare systems, application of the following techniques and algorithms can help decision makers in health insurance organisations to increase productivity, cut costs and lower the complexity of their daily processes. The application of data mining is described by using CRISP-DM methodology, while example of possible application is given for each type of data mining approach.**

Keywords-component; healthcare systems, data mining, data warehousing, knowledge discovery, CRISP-DM

I. Introduction

Data mining is a set of technologies that combine the application of statistical techniques and mathematical formulas as tools, trying to identify the importance of relationships between historical data that could then be used to predict, perform sensitivity analysis or just identify the importance of relationships between data that we have [1].

In certain literature, the term data mining has been replaced by the term Discovery rules in data. Discovering rules in data is part of business intelligence whose role is to support decision makers in their search for rules or laws in an organization's data. The discovery of regularity in data is similar to classical statistical analysis, expanding the range of algorithms and techniques from multivariate analysis to machine learning [2].

Data mining is a multidisciplinary field because it includes the areas of databases, expert systems, information theory, statistics, mathematics, logic and a whole range of related fields. Data mining is applicable in all those areas where a large amount of data is available and where certain regularities, connections and laws are to be discovered based on it.

Health care is a set of services and activities for the improvement and preservation of human health, prevention of diseases and injuries, early detection of diseases, timely treatment and rehabilitation. The primary level of health care is provided in specialist family medicine clinics, dental clinics, health centers, nursing homes and pharmacies. It includes comprehensive health care, implementation of preventive measures, health education and cooperation with all organizations and institutions that contribute to better health of the population. Family medicine is the basic form of organizing the primary level of health care and the first contact with the patient in the health care system. Secondary level means health care in institutions that provide treatment and specialized care and is focused only on diagnosis and treatment. The tertiary level of health care provides highly specialized health care, which is not provided at the level of secondary health care. The tertiary level of health care is organized in such a way that it complements the secondary health care and provides it with organized and continuous help and support [3].

The main goals of informatization of the modern health care system are, on one hand, medical goals which include the possession of all relevant medical information about persons receiving health care, and on the other economic goals which include collecting and analyzing economic indicators of health care costs and monitoring income and health care costs. In health insurance models based on the principle of non-profit and solidarity among all insured persons, a significant role in the functioning of the health system is played by organizations that provide insured persons with health care rights, which are most often called health insurance funds or institutes.

A. Reasons for the implementation of Data mining in health insurance

Decision-making in the health insurance system requires quality health information that should be provided by the health information system. The fund's transactional database (OLTP - On Line Transaction Processing), burdened with a large amount of data, cannot respond in a timely manner to all requests for information initiated by decision-making needs at the operational, tactical and strategic levels. In addition, there are other data and information in the health system that are not part of the transaction system, and which may be of varying degrees of structure.

The concept of Data Warehouse as a specific database and OLAP (On-Line Analytical Processing) system, as a category of applications and technologies designed to collect, manage, process and present multidimensional data for analysis for management purposes, partially solves this shortcoming transaction systems. Further improvement of the business decision-making process in health insurance and the use of information can be achieved by applying data mining methods to detect patterns in the data.

Data mining is applied in domains where large amount of data can be acquired with the goal of discovering certain regularities, connections and patterns. The health insurance system, with the Health Insurance Fund as part of the system that collects contributions and contracts and pays for health services provided to insured persons, is an area that contains a large amount of data from which it is necessary to extract certain laws in the data.

Decision-making in the health system must be based on a large amount of statistically significant data. Data mining gives results that show relationships and interdependencies in data that are mainly based on different mathematical and statistical relationships. The most important reasons for applying data mining techniques in health care are: reasoning based on past data, predicting results based on the past and making business decisions based on the results of the data mining process.

II. Related Works

In the last few years, the field of health insurance has been a frequent topic of discussion at public, political and scientific gatherings. One of the reasons is certainly the fact that the expenditures of health insurance institutions are increasing every year due increased number of patients with insurance demanding specialized care or expensive treatments

Increasing the elderly population is the biggest challenge that society will have to face in the coming years, which will lead to increased demands for health care [4]. A larger number of provided health services contributes to the increase of health information in health insurance institutions. Due to the large amount of information, decision makers in the health system need the help of a decision support system [5]. A large number of papers at scientific conferences and professional journals are focused on supporting decision-making and discovering legality in data in health care and health insurance.

The review of the latest research and searches of published publications was performed through the Publish or Perish software [6], a tool for searching databases of scientific publications, using the keywords 'data mining' and 'health insurance'. Academic publications in the field can well present a picture of the situation in the field when it comes to state-of-the-art, the application of Data Mining technologies as well as the focus of scientific research for a certain period of time. By analysing the published papers from the past 15 years, one can easily get a picture of where the focus of the academic community is. Using the above keywords for a given period, papers which were cited at least once were identified. In the following, the most cited and, in the opinion of the author, the most significant works for the topic of this paper are highlighted.

Most of the papers deal with the use of data mining techniques to detect fraud in health insurance [7], [8], [9], [10], [11]. The papers present data mining tools and techniques that can be used to detect health insurance fraud by analysing large data sets. Health insurance analysts can, based on data mining results, investigate cases labelled with data mining software in more detail.

In [12], the use of Data mining to predict and prevent claims in the processing of health insurance claims is shown. The authors describe a system that helps reduce these errors using machine learning techniques by predicting claims that will need to be reworked, generating explanations to help the auditors correct these claims, and experiment with feature selection, concept drift, and active learning to collect feedback from the auditors to improve over time.

The paper [13] deals with predicting the cost of the health insurance which has to be paid by the patient. Here various data mining regression algorithms such as decision tree, random forest, polynomial regression and linear regression are implemented to achieve the best prediction analysis. A comparison has been done between the actual and predicted expenses of the prediction premium and eventually, a graph has been plotted on this basis which will enlighten us to choose the best-suited regression algorithm for the insurance policy prediction.

In [14], challenges are presented for building predictive risk models for predicting the cost of health care costs. Empirical studies indicate that this approach is useful and may be useful for further research in limiting health care costs. In [15], the importance of e-health and data mining in health insurance is presented and the possibility of applying clustering and classification techniques in health care is shown. A lot of research deals with the application of data mining in health insurance cost control. In [16], the possibility of applying data warehousing, SOA (Service-Oriented Architecture) and various data mining algorithms in monitoring and auditing real-time health insurance costs was investigated.

The application of data mining in reducing health care costs is presented in [17]. Customer Relationship Management (CRM) provides support to health insurance beneficiaries and the use of patient data throughout the health system [18]. CRM uses information technology and data mining to improve the quality of health services and the relationship between policyholders and health facilities. CRM is a way to better manage complex processes in health insurance.

In some studies, such as [19], the application of different data mining techniques has been tested using proprietary clustering and prediction algorithms over large amounts of data from health system databases.

The potential of future application of data mining to improve health information systems, but also concerns about the protection of patient data privacy are discussed in [20]. Data mining has the potential to improve executive health information systems. Its implementation means better decision-making by management, which will mean better customer service [20].

In the analysed works, no papers were found that use different data mining techniques and algorithms to support health insurance contracting between health insurance funds and health care institutions.

III. Data Knowledge Discovery Process

The process of discovering knowledge from data is extremely important for any business. There is also an international methodology by which this process is conducted, and that is CRISP-DM (Cross Industry Standard for Data

Mining). CRISP-DM methodology, shown visually in figure 1, consists of the following phases [21]:

- Understanding the business problem.
- Data understanding.
- Data preparation.
- Solution modeling.
- Evaluation of the solution.
- Deployment.

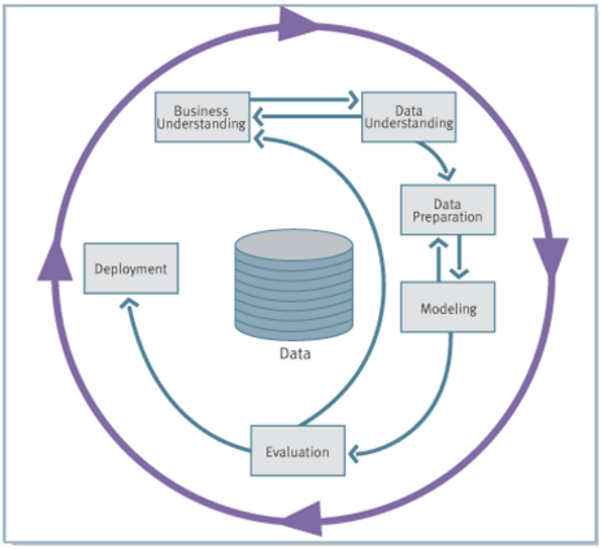

Figure 1. CRISP methodology [21]

Business understanding phase. The business process understanding phase or the research phase in the general case has three parts: a good understanding of the project objectives and project requirements as a whole, the formulation of objectives and constraints, and the preparation of an initial strategy to achieve those objectives.

Data understanding phase. The data understanding phase in the general case has four parts: data collection, acquaintance with data and their relationships based on simple analyses and representations (usually in graphical form), and evaluation of the quality of data accuracy. Optionally, subsets of data can be selected that are considered to possibly contain interesting patterns for a defined problem.

Data preparation phase. The data preparation phase in the general case has four parts and, as a rule, represents the most strenuous and difficult phase. It includes the following activities:

- Preparation of the final data set to be used in the following phases.
- Selection of variables and their attributes that are considered to be beneficial.
- Transformation of certain variables if necessary.
- Data purification and preparation for working with modelling tools.

Modelling phase. When the data is consolidated and adapted to the tools, it moves to the modelling phase. This phase has four steps and requires the least time and activity if the previous phases are well executed, i.e., if the problem is understood and if the data is well prepared. The activities that are carried out are as follows:

- Selection of appropriate modelling techniques.
- Adjustment of model parameters (optional application of other techniques).
- If necessary, return to the data preparation phase for a specific algorithm.

Evaluation phase. When the algorithm is executed, therefore we learn the model and get the results, we move to the evaluation phase, where the accuracy and applicability of the solution is examined before it is included in a business or research problem. It also evaluates the process of detecting regularity in data versus other waiting time measurement techniques. If the results are inadequate or inapplicable, it returns to the previous phase. This phase consists of the following activities:

- Comparison of model results with results of other approaches.
- Evaluation of the results of one or more models in terms of effectiveness or applicability.
- Examining whether the application of the model for the same solves the problems defined in the first phase.
- Examining whether some details of the problem are not covered in sufficient detail and making decisions based on the results.

Deployment phase. The phase refers to the application of solutions in a real system and monitoring of work with additional fine-tuning in production use.

A. The importance of Data warehousing in Data mining process

Although in practice there is often a case of non-existence of data warehouses (after data preparation it directly enters the modeling process without the mediation of data warehouses), in business intelligence systems data mining is most often supported by data warehouses.

Data warehouse (DW) is a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions [22]. A DW can be used to support permanent systems of records and information management (compliance and improved data governance).

The Extract, Transform, Load (ETL) process (figure 2) involves fetching data from transactional systems, cleaning the data, transforming data into appropriate formats and loading the result to a warehouse. In the ETL process, data from data sources is extracted by extraction routines. Data are then propagated to the Data Staging area where they are transformed and cleaned before being loaded to the data warehouse [23].

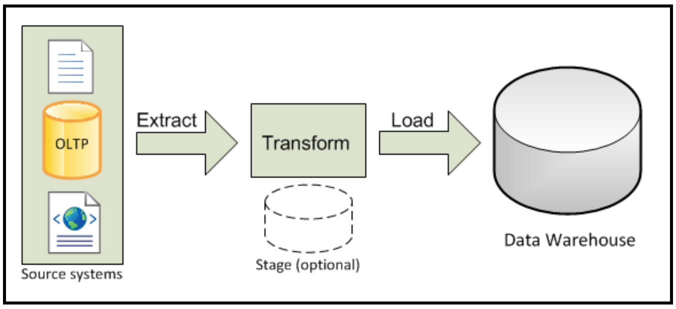

Figure 2. DW with staging

There is a similarity between the data preparation phase and the ETL process in building a data warehouse. The ETL process is a pre-process of the DM project because in the ETL process the data is transformed into a form suitable for DM analysis.

Data mining uses Data warehouse as data source for model training and pattern evaluation (figure 3). The process of model training can happen in multiple iterations, since knowledge from the data warehouse can be extracted using different models or parameters.

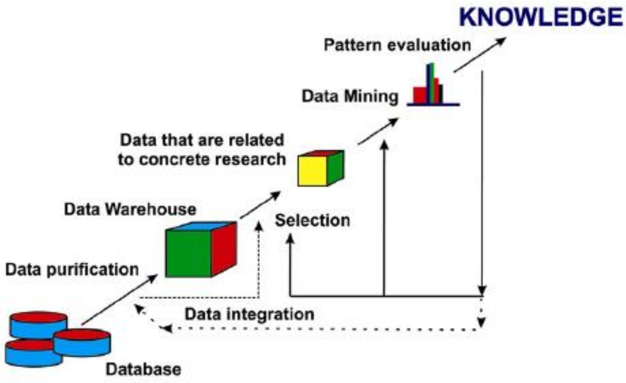

Figure 3. Data warehouse and Data mining [24]

The use of data warehousing in DM processes is reduced to the use of dimensions in cubes that are of interest for analysis and that can be used with the aim of preprocessing data that then enter the analytical methods of data mining. The biggest advantage of using data warehouses during the data mining process is time savings, because data warehouses can prepare data to be used for analysis in a very short time.

Data marts are small data warehouses that contain only a subset of the DW [23]. The data mart provides the platform for Online Analytical Processing (OLAP) analysis. Therefore, OLAP is a natural extension of the data warehouse. The data mart is a part of data storage, but usually contains summarized data [5]. Data mart is usually designed in a star scheme consisting of a single fact table and multiple dimension tables that are all linked to a fact table and not linked to another dimension table. In some situations, it is useful to split the dimension table into multiple dimension tables, when we get a snowflake scheme. In this scheme, the data is normalized, but access to the data is somewhat more difficult.

IV. Tasks that can be Solved by Data Mining in Health Insurance

In order to better understand a business problem, it is important to know that each data mining project aims to fulfill one or more tasks. The data mining methods and algorithms need to accomplish several tasks. All these tasks are performed with the aim of improving the business process [2]. An overview of the tasks that can be solved by data mining in health insurance with examples for each task, as well as an overview of the most well-known algorithms for individual DM tasks are given below.

a) Reduction. Reduction aims to reduce the dimensions of the problem being analysed to the form most appropriate for analysis. Reduction reduces the data to a measure that is suitable for DM analysis. With less data, DM algorithms work faster. Of course, data quality needs to be taken into account. Reduction is done in almost every DM project. In addition to reducing attributes, it is sometimes necessary to reduce rows that are not relevant to the analysis.

For example, in the analysis of health services provided by the health institution, and in order to determine the elements of the contract in the coming period, the age of the insured person is an important element for analysis. As different age groups have different health conditions and spend different amounts for health care, therefore different amounts of capitation will be applied for certain age groups (so-called weight capitation) in accordance with expected health consumption, i.e. frequency of health care requests. On the other hand, the gender of the insured person is not an important element of contracting primary health care and it can be omitted in the reduction process. Capitation represents a certain number of points or a certain amount of money for a certain period of time, for each insured person, who is registered in a certain health institution for the provision of health services.

b) Assessment. The assessment aims to discover the regularity between input and output attributes. In the case of estimation, the input attributes can be of numeric and non-numeric type, and the output attributes are usually of numerical type.

For example, the assessment tasks in health insurance in the part of the process of analysis of the execution of the contract between the fund and the health institution may be as follows:

- How many health services within a certain activity will be provided by family medicine teams in the second quarter of the current year?
- How much of the contracted funds will the health care institution spend for a certain activity in the fourth quarter for insured persons from the age group over 65?

c) Classification. The classification aims to reveal the regularity between input and output attributes. Input attributes can be of numeric and non-numeric type, but unlike regression, output attributes are mostly of non-numeric type.

For example, the task of classification in the analysis of the execution of contracts with health care institutions can be: to find the regularity according to which the health care institution for a certain activity performs fewer services than contracted and for which age groups.

d) Clustering. Clustering aims to discover the regularity by which data are grouped into certain clusters, i.e. classes that are not known in advance. Algorithms that do clustering are called unsupervised learning algorithms because there is no output attribute. On the other hand, algorithms that use output attributes to detect certain patterns are called supervised learning algorithms [2]. For example, the task of clustering in the analysis of the execution of contracts with health care institutions may be as follows:

- In which groups to assign insured persons to whom a certain health care institution provided health care services? What are the similarities and differences between these groups?
- How many health services are provided in institutions within a certain activity?

e) Associative rules. Associative rules aim to detect regularity in the form of IF (cause) THEN (consequence) (IF – THEN). Associative rules make it possible to spot some regularities that are difficult to spot.

For example, when analysing the services performed by a health institution according to individual diagnoses, it is possible to examine which drugs are prescribed for the diagnosis of Helicobacter pylori? Or, are these drugs the same for different age groups?

f) Forecasting. Forecasting aims to find certain regularities in data that contain a time dimension and to extrapolate certain regularities based on past data. Prediction includes all other DM tasks, provided that algorithms and techniques in prediction have the ability to work with the time dimension.

For example, how will the consumption of a particular group of drugs for the next quarter or year in certain municipalities or regions behave? Will some, and which ones, health care institutions implement the provisions of the contract related to preventive health care?

All above mentioned tasks are represented in short in the following table.

TABLE I. DATA MINING TASKS AND ALGORITHMS

Data Mining tasks	***The most famous algorithms***
Reduction	PCA (Principal component analysis), FA (factor analysis)
Assessment	Linear Regression, CART Classification and Regression Tree, Artificial Neural Networks
Classification	Decision trees (ID3, C4.5, CHAID, CART), Logarithmic regression, Discriminant analysis, Associative rules
Clustering	K-means, X-means, MPC K-means, Hierarchical cluster algorithms, DB Scan, Kohonen SOM
Associative rules	A priori
Forecasting	Regression algorithms, Decision trees, Cluster algorithms, Artificial neural networks, Associative rules

V. Conclusion

The introduction gave the foreword of basic terms and domain settlement. After introduction, in the second part of the paper, related works in the application of data mining in healthcare were explored and described. In the means of constant increase of costs in the healthcare systems, a great number of research is found, with the goal of decision support based on data. Short overview of knowledge discovery is shown using CRISP-DM methodology of data mining in the third part of the paper.

Future work of the authors will be focused on application of data mining algorithms in data provided by the health insurance institution from Republic of Srpska, with the goal of extracting knowledge for the broader academic community and decision makers.

In the healthcare insurance domain, it is possible to apply different data mining techniques and algorithms with the goal of problem solving, which is described in forth part of the paper. The application examples are described for each data mining approach. By using data mining techniques and algorithms for knowledge discovery and data laws definition, it is possible to increase productivity and efficacy of decision makers in the healthcare insurance systems.

References

[1] S. Kudyba, and R. Hoptroff, „Data Mining and Business Intelligence: A Guide to Productivity“, Idea Group Publishing, London, 2001.

[2] M. Suknović, and B. Delibašić Boris, „Poslovna inteligencija i sistemi za podršku odlučivanju“, FON Beograd, 2010.

[3] Zakon o zdravstvenoj zaštiti Republike Srpske (Sl. glasnik RS br. 106/2009 i 44/2015).

[4] S. Koch, and M. Hägglund, „Health informatics and the delivery of care to older people“, www.elsevier.com/locate/maturitas, Stockholm, 2009.

[5] D. Krneta, B. Radulović, D. Radosav, „Business Intelligence in health information system“, Infoteh, 2008.

[6] Publish or Perish. http://www.harzing.com/resources/publish-or-perish

[7] M. Kirlido, and C. Asuk, „A Fraud Detection Approach with Data Mining in Health Insurance“, Procedia - Social and Behavioral Sciences, Volume 62, 2012, Pages 989-994, ISSN 1877-0428.

[8] V. Rawte, and G. Anuradha, „Fraud detection in health insurance using data mining techniques“, 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 2015, pp. 1-5.

[9] L. Kuo-Chung, and Y. Ching-Long, „Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance“, International Journal of Engineering andTechnology Innovation, vol. 2, no. 2, 2012, pp. 126-13.7

[10] P. Pawar, „Review on Data Mining Techniques for Fraud Detection in Health Insurance“, International Journal on Emerging Trends in Technology (IJETT), 2016.

[11] P. Pandey, A. Saroliya, and R. Kumar, „Analyses and Detection of Health Insurance Fraud Using Data Mining and Predictive Modeling Techniques“ BSoft Computing: Theories and Applications, 2017.

[12] M. Kumar, R, Ghani, and Z-S. Mei, „Data mining to predict and prevent errors in health insurance claims“, processing KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data miningJuly 2010.

[13] K. Dutta, and S. Chandra, M.K. Gourisaria, GM H, „A Data Mining based Target Regression-Oriented Approach to Modelling of Health Insurance Claims“, 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1168-1175.

[14] S.T. Moturu, G. William, G. Johnson, H. Liu, „Predicting Future High-Cost Patients: A Real-World Risk Modeling Application“, IEEE International Conference on Bioinformatics and Biomedicine, 2007.

[15] S. Rahaman, and S. Biju, „Data Mining Facilitates e-Patients through e-Healthcare: An Empirical Study“, International Conference on New Trends in Information and Service Science, 2009.

[16] D. Qin, „Design of Medical Insurance Supervision System Base on Active DW and SOA“, World Congress on Computer Science and Information Engineering, 2009.

[17] R. Rajkumar, K.J. Shim, J. Srivastava, „Data Mining Based Predictive Models for Overall Health Indices“, Department of Computer Science and Engineering University of Minnesota, Technical report, 2010.

[18] Y. Wan, „Application of CRM in Health Care“, Second International Conference on MultiMedia and Information Technology, 2010.

[19] D. Bertsimas, M. Bjarnadóttir, M. Kane, C. Kryder, R. Pandey, S. Vempala, G. Wang, „Algorithmic Prediction of Health-Care Costs“, Operations Research 56(6), pp. 1382–1392, ©2008 INFORMS.

[20] S. Glover, P.A. Rivers, D.A. Asoh, C.N. Piper, and, K. Murph, „Data mining for health executive decision support: an imperative with a daunting future“, Health Services Management Research, 2010.

[21] C. Shearer, „The CRISP-DM model: the new blueprint for data mining“, J Data Warehousing (2000); 5:13—22.

[22] Inmon H. W.: Building the Data Warehouse. Wiley Computer Publishing. (1992)

[23] D. Krneta, V. Jovanović, Z. Marjanović, A Direct Approach to Physical Data Vault Design. Computer Science and Information Systems, Vol. 11, No. 2, 2014, 569–599.

[24] M. Suknović, M. Čupić, M. Martić, „Data warehousing and data mining – a case study“, Yugoslav Journal of Operations Research, 2005.

[25] B. Delibašić, M. Suknović, M. Jovanović, „Algoritmi mašinskog učenja za otkrivanje zakonitosti u podacima“, FON Beograd, 2009.

Dragoljub Krneta (1966) received PhD degree in Information Systems at Faculty of Organizational Sciences, University of Belgrade in 2016. He has more than 30 years of experience in database design, information system design and software development. His current research interests are in the field of databases, particularly data warehouse and data vault, business intelligence and data mining.

Sofija Krstev (1994) received B.Sc. and M.Sc. degrees at Faculty of Organizational Sciences, University of Belgrade, in 2016 and 2018, respectively. She is currently PhD candidate in multidisciplinary programme of University of Belgrade - Intelligent systems and economy with application in energy industry. She has more than 6 years of experience in software development and machine learning. She has research interests in machine learning, artifical intelligence and application in energy industry.