International Journal of All Research Education and Scientific Methods (IJARESM),
ISSN: 2455-6211, Volume 11, Issue 11, November-2023, Available online at: www.ijaresm.com

Fraud Detection in Health Insurance Claims using ML

Harsha Rajendra, Olin Dilip Dsouza, Sarandha Tiwari, Basireddy Harish Reddy, Lakshmana Sai Kumar Kommireddy, Shubham Khatri, Krishna Mihir Tatavarthi, Mutyala Bhuvan Saieesh , Dhruv Purohit

---*****************--

ABSTRACT

The healthcare industry is a complex system with many moving parts. One issue in this field is the misuse of medical insurance systems, such as Medicare. In this paper, we build a machine learning model to detect when physicians exhibit anomalous behavior in their medical insurance claims. This new research has the potential to give some insight in determining if, and when, physicians are acting outside the norm of their respective specialty, which could indicate misuse, fraud, or lack of knowledge around billing procedures.

Fraud detection in health insurance claims using ML

LITERATURE SURVEY

Process Query-constraint-based mining of association rules for exploratory analysis of clinical datasets in the National Sleep Research Resource.

Authors: Abeysinghe R, Cui L.

Year: 2018 Jul

Background: Association Rule Mining (ARM) has been widely used by biomedical researchers to perform exploratory data analysis and uncover potential relationships among variables in biomedical datasets. However, when biomedical datasets are high-dimensional, performing ARM on such datasets will yield a large number of rules, many of which may be uninteresting. Especially for imbalanced datasets, performing ARM directly would result in uninteresting rules that are dominated by certain variables that capture general characteristics.

Methods: We introduce a query-constraint-based ARM (QARM) approach for exploratory analysis of multiple, diverse clinical datasets in the National Sleep Research Resource (NSRR). QARM enables rule mining on a subset of data items satisfying a query constraint. We first perform a series of data- preprocessing steps including variable selection, merging semantically similar variables, combining multiple-visit data, and data transformation. We use Top-k Nonredundant (TNR) ARM algorithm to generate association rules. Then we remove general and subsumed rules so that unique and non-redundant rules are resulted for a particular query constraint.

Results: Applying QARM on five datasets from NSRR obtained a total of 2517 association rules with a minimum confidence of 60% (using top 100 rules for each query constraint).

The results show that merging similar variables could avoid uninteresting rules. Also, removing general and subsumed rules resulted in a more concise and interesting set of rules.

Conclusions: QARM shows the potential to support exploratory analysis of large biomedical datasets. It is also shown as a useful method to reduce the number of uninteresting association rules generated from imbalanced datasets. A preliminary literature-based analysis showed that some association rules have supporting evidence from biomedical literature, while others without literature-based evidence may serve as the candidates for new hypotheses to explore and investigate. Together with literature-based evidence, the association rules mined over the NSRR clinical datasets may be used to support clinical decisions for sleep-related problems.

Keywords: Exploratory data analysis; National sleep research resource; Query-constraint based association rule mining.

Process mining in healthcare: A literature review

Authors: ARRojas E, Munoz-Gama J, Sepúlveda M, Capurro D.

YEAR: - 2016 June DESCRIPTION:
Process Mining focuses on extracting knowledge from data generated and stored in corporate information systems in order to analyze executed processes. In the healthcare domain, process mining has been used in different case studies, with promising results. Accordingly, we have conducted a literature review of the usage of process mining in healthcare. The scope of this review covers 74 papers with associated case studies, all of which were analyzed according to eleven main aspects, including: process and data types; frequently posed questions; process mining techniques, perspectives and tools; methodologies; implementation and analysis strategies; geographical analysis; and medical fields. The most commonly used categories and emerging topics have been identified, as well as future trends, such as enhancing Hospital Information Systems to become process- aware. This review can: (i) provide a useful overview of the current work being undertaken in this field; (ii) help researchers to choose process mining algorithms, techniques, tools, methodologies and approaches for their own applications; and (iii) highlight the use of process mining to improve healthcare processes.

Text mining for traditional Chinese medical knowledge discovery: a survey

Authors: Xuezhong Zhou 1,YonghongPeng, Baoyan Liu.

Year: 2010 Aug

Description:
Extracting meaningful information and knowledge from free text is the subject of considerable research interest in the machine learning and data mining fields. Text data mining (or text mining) has become one of the most active research sub-fields in data mining. Significant developments in the area of biomedical text mining during the past years have demonstrated its great promise for supporting scientists in developing novel hypotheses and new knowledge from the biomedical literature. Traditional Chinese medicine (TCM) provides a distinct methodology with which to view human life. It is one of the most complete and distinguished traditional medicines with a history of several thousand years of studying and practicing the diagnosis and treatment of human disease. It has been shown that the TCM knowledge obtained from clinical practice has become a significant complementary source of information for modern biomedical sciences. TCM literature obtained from the historical period and from modern clinical studies has recently been transformed into digital data in the form of relational databases or text documents, which provide an effective platform for information sharing and retrieval. This motivates and facilitates research and development into knowledge discovery approaches and to modernize TCM. In order to contribute to this still growing field, this paper presents (1) a comparative introduction to TCM and modern biomedicine, (2) a survey of the related information sources of TCM, (3) a review and discussion of the state of the art and the development of text mining techniques with applications to TCM, (4) a discussion of the research issues around TCM text mining and its future directions.

Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support

Authors: Xuezhong Zhou 1, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, YufengGuo, Hua Zhang, ZhuyeGao, Xiufeng Yan

Year:- 2010

DESCRIPTION

Objective: Traditional Chinese medicine (TCM) is a scientific discipline, which develops the related theories from the long-term clinical practices. The large-scale clinical data are the core empirical knowledge source for TCM research. This paper introduces a clinical data warehouse (CDW) system, which incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and TCM clinical decision support (CDS).

Materials and methods: We have developed the clinical reference information model (RIM) and physical data model to manage the various information entities and their relationships in TCM clinical data. An extraction-transformation-loading (ETL) tool is implemented to integrate and normalize the clinical data from different operational data sources. The CDW includes online analytical processing (OLAP) and complex network analysis (CNA) components to explore the various clinical relationships. Furthermore, the data mining and CNA methods are used to discover the valuable clinical knowledge from the data.

Results: The CDW has integrated 20,000 TCM inpatient data and 20,000 outpatient data, which contains manifestations (e.g. symptoms, physical examinations and laboratory test results), diagnoses and prescriptions as the main information components. We propose a practical solution to accomplish the large-scale clinical data integration and pre processing tasks. Meanwhile, we have developed over 400 OLAP reports to enable the multidimensional analysis of clinical data and the case-based CDS. We have successfully conducted several interesting data mining applications. Particularly, we use various classification methods, namely support vector machine, decision tree and Bayesian network, to discover the knowledge of syndrome differentiation. Furthermore, we have applied association rule and CNA to extract the useful acupuncture point and herb combination patterns from the clinical prescriptions.

Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support

Authors: Xuezhong Zhou 1, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, YufengGuo, Hua Zhang, ZhuyeGao, Xiufeng Yan

Year:- 2010

DESCRIPTION

Objective: Traditional Chinese medicine (TCM) is a scientific discipline, which develops the related theories from the long-term clinical practices. The large-scale clinical data are the core empirical knowledge source for TCM research. This paper introduces a clinical data warehouse (CDW) system, which incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and TCM clinical decision support (CDS).

Materials and methods: We have developed the clinical reference information model (RIM) and physical data model to manage the various information entities and their relationships in TCM clinical data. An extraction-transformation-loading (ETL) tool is implemented to integrate and normalize the clinical data from different operational data sources. The CDW includes online analytical processing (OLAP) and complex network analysis (CNA) components to explore the various clinical relationships. Furthermore, the data mining and CNA methods are used to discover the valuable clinical knowledge from the data.

Results: The CDW has integrated 20,000 TCM inpatient data and 20,000 outpatient data, which contains manifestations (e.g. symptoms, physical examinations and laboratory test results), diagnoses and prescriptions as the main information components. We propose a practical solution to accomplish the large-scale clinical data integration and pre processing tasks. Meanwhile, we have developed over 400 OLAP reports to enable the multidimensional analysis of clinical data and the case-based CDS. We have successfully conducted several interesting data mining applications. Particularly, we use various classification methods, namely support vector machine, decision tree and Bayesian network, to discover the knowledge of syndrome differentiation. Furthermore, we have applied association rule and CNA to extract the useful acupuncture point and herb combination patterns from the clinical prescriptions.

Conclusion: A CDW system consisting of TCM clinical RIM, ETL, OLAP and data mining as the core components has been developed to facilitate the tasks of TCM knowledge discovery and CDS. We have conducted several OLAP and data mining tasks to explore the empirical knowledge from the TCM clinical data. The CDW platform would be a promising infrastructure to make full use of the TCM clinical data for scientific hypothesis generation, and promote the development of TCM from individualized empirical knowledge to large- scale evidence-based medicine.

HARDWARE REQUIREMENTS

- Processor - Pentium –III
- Speed – 2.4 GHz
- RAM - 512 MB (min)
- Hard Disk - 20 GB

- Floppy Drive - 1.44 MB
- Key Board - Standard Keyboard
- Monitor – 15 VGA Colour

SOFTWARE REQUIREMENTS

Functional requirements for a secure cloud storage service are straightforward:

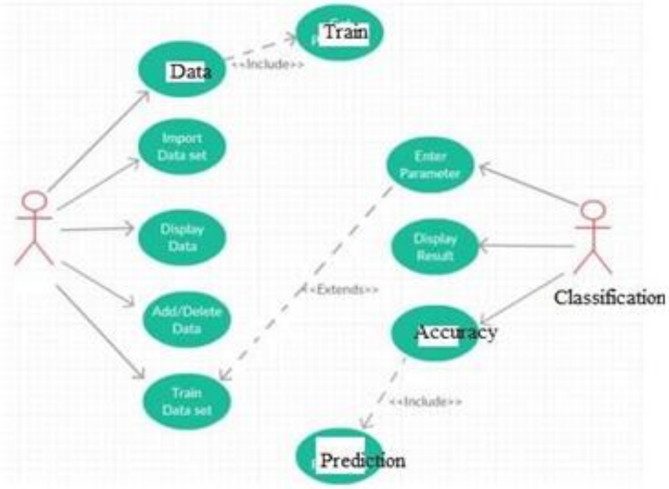
1. The service should be able to store the user's data;
2. The data should be accessible through any devices connected to the Internet;
3. The service should be capable to synchronize the user's data between multiple devices (notebooks, smart phones, etc.);
4. The service should preserve all historical changes (versioning);
5. Data should be shareable with other users;
6. The service should support SSO; and
7. The service should be interoperable with other cloud storage services, enabling data migration from one CSP to another.

- Operating System: Windows
- Coding Language: Python 3.7

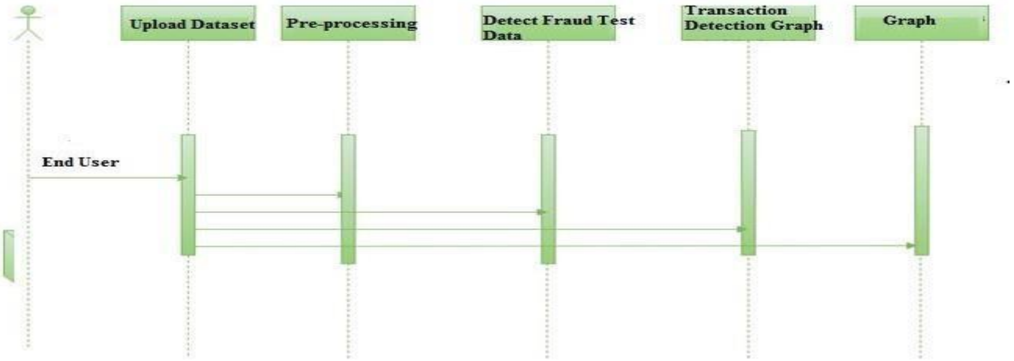
Detailed Design:

UML is an acronym that stands for Unified Modelling Language. Simply put, UML is a modern approach to modelling and documenting software. In fact, it's one of the most popular business process modelling techniques. It is based on diagrammatic representations of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representations, we are able to better understand possible flaws or errors in software or business processes. UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems. The need arose for a more unified way to visually represent those systems and as a result, in 1994-1996, the UML was developed by three software engineers working at Rational Software. It was later adopted as the standard in 1997 and has remained the standard ever since, receiving only a few updates.

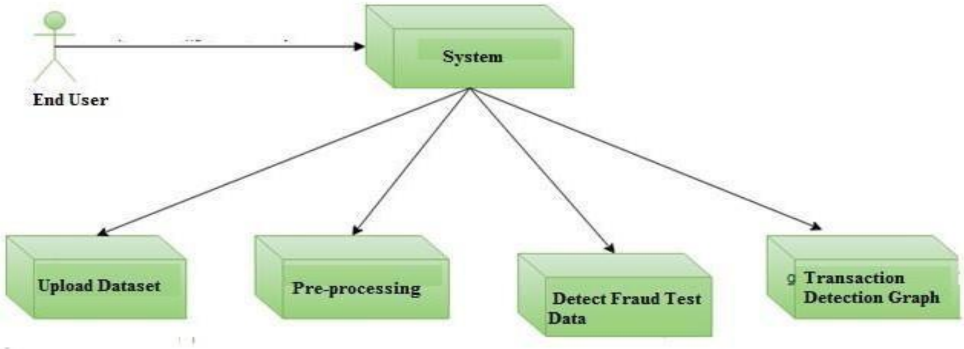
Use Case Diagram:



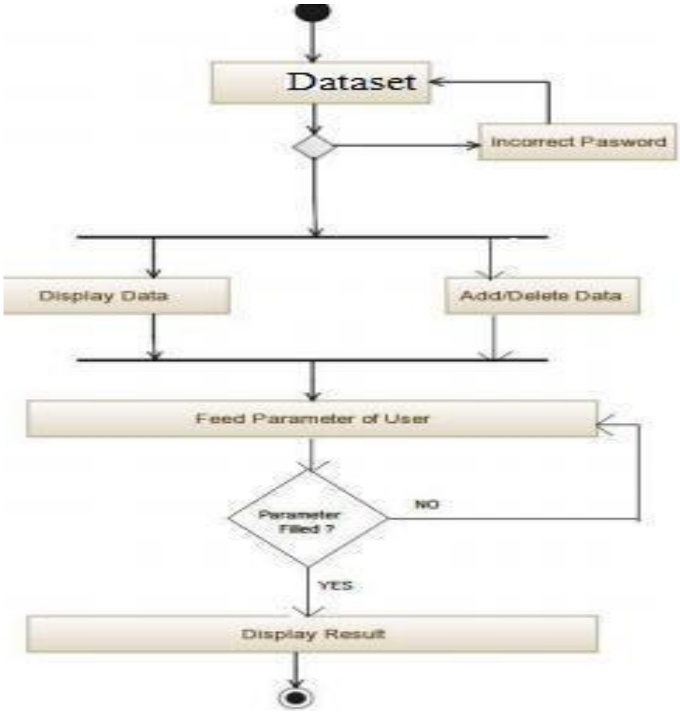
Sequence Diagram:



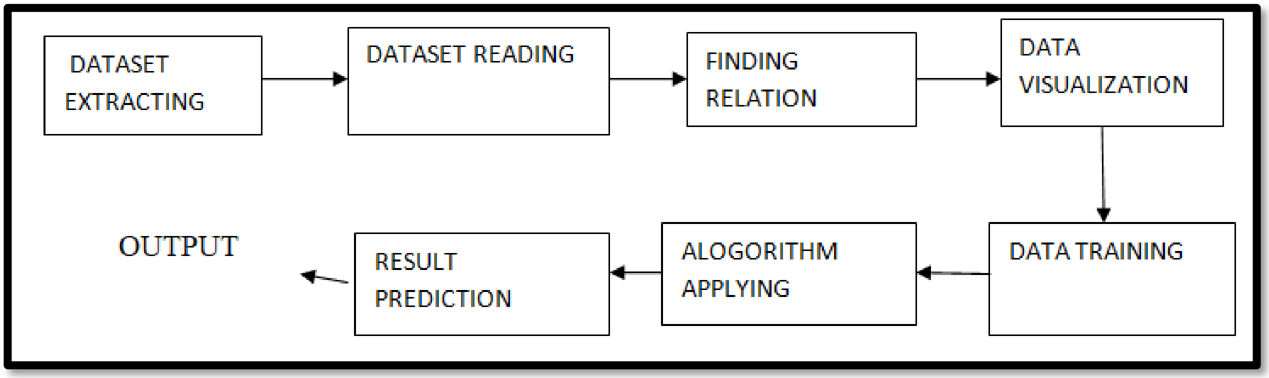
Component Diagram:



Activity Diagram:



Module description



Dataset Extraction:

Data extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured. Data extraction makes it possible to consolidate, process, and refine data so that it can be stored in a centralized location in order to be transformed. These locations may be on-site, cloud-based, or a hybrid of the two.

Finding Relation

Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection, filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable.

Data Visualisation

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career. Whether you work in finance, marketing, tech, design, or anything else, you need to visualize data. That fact showcases the importance of data visualization.

Data Training

In machine learning, training data is the data you use to train a machine learning algorithm or model. Training data requires some human involvement to analyze or process the data for machine learning use with supervised learning, people are involved in choosing the data features to be used for the model.

REFERENCES

[1]. U.S. Government, U.S. Centers for Medicare & Medicaid Services. The Official
[2]. U.S. Government Site for Medicare. https://www.medicare.gov/. Accessed 21 Jan 2017.
[3]. Feldstein M. Balancing the goals of health care provision and financing. Health Affairs. 2006;25(6):1603–11.
[4]. Administration for Community Living: Profile ofolder Americans. 2015. http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/.
[5]. Accessed 2015.
[6]. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, Horst C, Li Z, Matyasz T, Reynolds A. Factors associated with increases in us health care spending, 1996–2013. Jama. 2017;318(17):1668–78.
[7]. Aetna. The facts about rising health care costs. http://www.aetna.com/health- reform- connection/aetnasvision/facts-about-costs.html. Accessed 2015.