International Journal of Biomedical Data Mining
ISSN: 2090-4924
Almadhoun and El-Halees, Int J Biomed Data Min 2017, 6:2
DOI: 10.4172/2090-4924.1000129
Review Article
OMICS International

Different Mining Techniques for Health Care Data Case Study of Urine Analysis Test

Mohamed D Almadhoun[1]* and Alaa M El-Halees[2]

[1]*Department of IT, University College of Applied Sciences, Gaza, Palestine*
[2]*Department of Computer Science, The Islamic University of Gaza, Gaza, Palestine*

Abstract

To make huge amounts of data that is produced by health care information systems useful and important to the potential, we apply knowledge discovery. This study considers urine analysis test results as an input data to different data mining techniques in order to discover the hidden and meaningful patterns in data. It also shows results of evaluation and analysis for data patterns.

Data mining techniques were, 1) Classification to support functionality for alerting about new instance that does not match the predicted value by classification model, 2) Association rules to inform about relations and changeability between elements, 3) Clustering to categorize patients into separate groups to give an indication about how to deal with each patient, and 4) Outlier analysis to discover the most sick patients or unfamiliar cases that need a special care.

Resulting knowledge were novel, actionable, understandable, and valid. This was stated by applying two methods of evaluation, first was a survey about results and was filled by medical specialists, second was by cross validation and T-Test.

Keywords: Health care information system; Data mining; Knowledge discovery in database; Urine analysis

Introduction

Data mining is an approach to extract patterns from large data sets and deduce knowledge insights from patterns [1]. Data mining can find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the data owner [2]. In this study we will use data mining to find relationships and rules, and the development trends of medical information resources [3].

The healthcare field faces strong pressures to reduce costs while increasing quality of services delivered, so healthcare information systems should be utilized for decision support and knowledge management [1]. After applying data mining we can identify significant clinical variables with respect to a diagnosis or therapy plan [4].

A urinalysis is a group of manual and/or automated qualitative and semi-quantitative tests performed on a urine sample. Purpose of this analysis is to apply general health screening to detect renal and metabolic diseases, diagnosis of diseases or disorders of the kidneys or urinary tract, monitoring of patients with diabetes. Routine urinalysis consists of three testing groups: physical characteristics (as measuring the color, transparency (clarity), and specific gravity of a urine sample), biochemical tests (are performed using dry reagent strips, often called dipsticks, the person performing the test dips the strip into the urine), and microscopic evaluation (measures presence of bacteria and white blood cells, the presence of cellular casts, and identifies both normal and abnormal crystals) [5].

This paper follows the process of knowledge discovery with all steps starting from data gathering, followed by data cleaning, and aggregation to make data ready to be utilized for data visualization and data mining, reaching to the evaluation and knowledge representation. Data mining phase will be shown in details and knowledge represented from each data mining model.

Four data mining models used: first was classification by decision tree, and rule induction, second was association rules by FP-Growth approach, third clustering by k-Means approach, and finally outlier analysis by distance based approach.

Related Works

Ramachandran et al. [4] introduced a readmission risk profiling by building a predictive scoring model that enables the assessment of re-hospitalization risk for patients. Longer the readmission gap, lower the hospitalization costs. Challenge was to determine if data mining techniques could generate a scoring model that could be used to reduce the "early returnees". The approach was to build a classification model to identify a readmission risk score that reflects the expected relative length of the readmission gap for an individual patient.

Gosain and Kumar [2] developed prototype/approach that is specially designed to monitor the Human immune-deficiency virus (HIV) patients that receive antiretroviral therapy (ART) to investigate the association between HIV and ART. For data mining, decision tree is first applied to the database for identifying patterns with relatively high support and confidence. Further Association rule is applied to predict the changes which are occurring frequently among people.

Ben-Chang et al. [6] analyzed World Health Organisation's (WHO) Health for All (HFA) database. Data was extracted for 39 European countries, careful initial selection of attributes and associated pre-processing. The empirical studies were carried out primarily using the Kohonen Self Organising Map (SOM) neural network technique.

***Corresponding author:** Mohamed D Almadhoun, Department of IT, University College of Applied Sciences, Gaza, Palestine, Tel: 00970599239967; E-mail: mdmadhoun@ucas.edu.ps.

Received July 14, 2017; **Accepted** August 01, 2017; **Published** August 07, 2017

Citation: Almadhoun MD, El-Halees AM (2017) Different Mining Techniques for Health Care Data Case Study of Urine Analysis Test. Int J Biomed Data Min 6: 129. doi: 10.4172/2090-4924.1000129

Work resulted in the identification of six groups. Characteristics of the groups ranged from the lowest life expectancy, coupled with the highest probability of dying before five and infant mortality rate, to the highest mean life expectancy, coupled with the lowest probability of dying before five, and infant mortality rate.

Discussion

Data resource

Dataset was created by a health care information system in a medical analysis lab located in Gaza; it's for urine analysis test results with 607 instances.

Dataset elements are numerous, general like name, age, sex, doctor, and notes, and special elements like color, volume, appearance, sp. gravity, pH, protein, glucose, ketones, blood, bilirubin, nitrite, RBC, WBC, epithcells, crystals, casts, mucus threads, and bacteria.

Data mining process and tasks

Figure 1 lists data mining process steps. First phase of data mining process was business understanding of how urine analysis can be applied and what benefits of this test. This information was gotten from specialists in the field of medical analysis.

Second phase data understanding achieved by collecting information about each special elements, understanding how these values were gotten and what test steps used to get each value, and what normal and abnormal values of each element.

By the aid of rapid-miner open source software third phase data preparation started by selecting the interesting elements, filtering instances with noisy incorrect values, removing instances that contain null values, and finally an auto useless element remover operator was used to filter out the most interesting attributes. Resulting attributes from data preparation were appearance, SP gravity, pH, blood, RBCs, WBCs, epithelial cells, crystals, mucus threads, and bacteria.

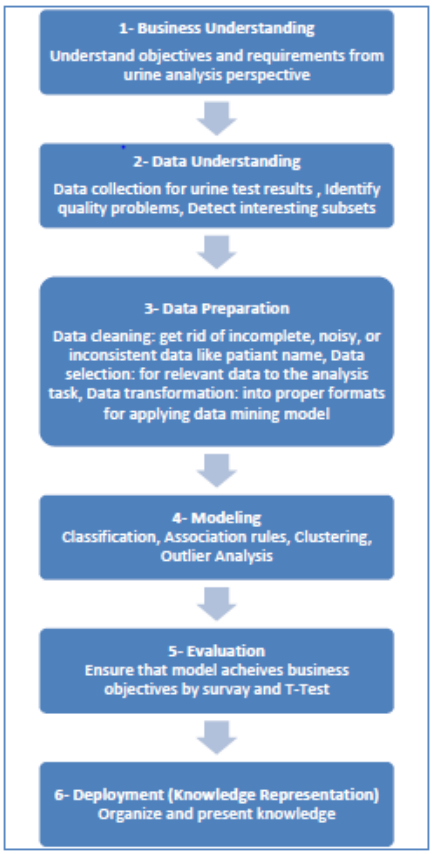

Figure 1: Data mining process steps and tasks.

By repeated experiments, it was inferred that it's impossible to get valuable models with using all special elements, so useless remover was used to get more accurate results.

Fourth phase modelling was applied in four different tasks, each of which informs about different knowledge. They were applied using rapid-miner software.

First task was classification using two techniques: First is decision tree which is a tree-structured plan of a set of attributes having several possible alternative branches of values in order to predict the class label which was the element 'appearance'. Second is rule induction which is a classifier that extracts a set of rules that show relationships between elements of dataset and class label which is the 'appearance' [7]. Classification can be useful in the field of urine analysis where health care information systems can be improved to support functionality for alerting about new instance that does not match the predicted value by classification model. Figure 2 shows a branch of resulting decision tree when WBCs are many. Table 1 shows some of resulting rules by rule induction model. Decision tree model was rich of knowledge; pruning was used to get rid of useless branches to avoid over fitting. Some of medical knowledge inferred from decision tree was like: lots of WBCs makes appearance turbid, finding RBCs with calcium oxalate crystals makes appearance turbid despite of little WBCs, finding crystals of type triple phosphate or amorphous phosphate makes appearance turbid, finding amorphous urate with little RBCs makes appearance clear, and turbid if RBCs were more than 3, present mucus threads make appearance turbid, and female sex has a high probability of turbid appearance. Decision tree model algorithm seeks to build tree with least levels and this depends on calculating information gain that increases

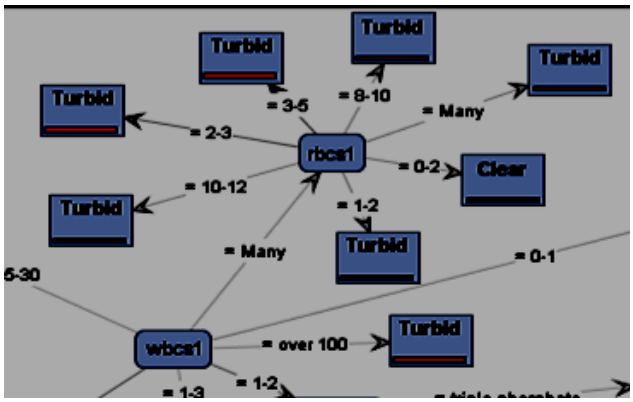

Figure 2: Part of resulting decision tree.

Rule Model
if wbcs1 = 2-3 and crystals1 = Absent then Clear (20 / 1 / 1)
if wbcs1 = 4-6 and ph1 = Acidic 6 then Turbid (0 / 0 / 3)
if wbcs1 = 6-8 and mucusthreads1 = Absent then Turbid (0 / 0 / 4)
if wbcs1 = 6-8 and rbcs1 = 2-3 then Turbid (0 / 0 / 3)
else Clear (14 / 2 / 13)
Correct: 495 out of 605 training examples.

Table 1: Some resulting rules from rule induction.

with the average purity of subsets that an element produces [7], so the chosen root was WBCs element.

Second task of data mining modelling phase was creating association rules by FP-Growth operator. This model is used to discover interesting relationships between disjoint items of dataset. Items can be defined as different values of elements. Relationship can be presented in a pattern of the form x - >y where x and y are item sets [7]. Association rules can be useful in the field of urine analysis by informing about those relations between elements and give useful medical knowledge about changeability relations between elements as a medical guidance of appreciating patients' situation. Figure 3 shows some resulting rules by association rules model with their evaluation factors support, confidence, and lift. Association rules were numerous, knowledge were extracted carefully according to medical background in that field and good consideration for data mining factors which are support, confidence, and lift. So association rules added that most tests that contain lots of epithelial cells are for females, turbid appearance of urine sample has more chance to appear in female tests, turbid appearance probability increases in case of finding calcium oxalate crystals, present mucus threads, and female sex, big chance of making appearance is clear occur when bacteria is absent, blood is nill, mucus threads absent, crystals absent, epithelial cells absent, and pH is acidic.

Third task was clustering by k-Means operator. This model is used to group data instances with similar characteristics or features together into a separate cluster [7]. Clustering helps to categorize patients into separate groups so any patients can be categorized into some category and this gives an indication to how to deal with that patient. Figure 4 shows how points are distributed between clusters. Clusters shown in Figure 5 were distributed up to their counts as shown in Figure 5, It's obvious that cluster_0 is the majority (376 items), and by studying different samples from this cluster it was noted that the cluster_0 has no problems in blood, RBCs, and WBCs. This cluster is considered the most healthy, or less sick. As for cluster_1 (160 items) it was noted that most of points were female, and have small problems in RBC, and WBCs. As for cluster_3 (56 items) it was noted that most of points have big problems in blood, RBCs, and WBCs. But for cluster_2 (31 items), it is the most sick, they have different big problems in blood, RBCs, WBCs, crystals, mucus threads. Fourth task was outlier detection using distance based approach which computes distance between every pair of data points and distinguishes points with neighbour's number less than a constant value within some distance as outliers [8]. Outlier analysis helps to get the most sick patients or unfamiliar cases that need a special care. Figure 6 shows the region of detected outliers. The statistical graph in Figure 7 shows that most outliers were from the group who has turbid appearance in the urine sample. Also from studying outlier points, it was found that outliers are patients with big problems that are not public and rarely occur and most of them were from cluster_2. Two methods of evaluation were followed, First of them was by medical section, where a urine medical specialists evaluated the resulting knowledge using evaluation survey. Evaluation survey contained a list of mined results with three available answers: Familiar, Unfamiliar, and Wrong. 70% of them were familiar, 16% unfamiliar and 14% wrong as they stated.

Second evaluation method held by using rapid-miner software by applying cross validation and T-Test operator to compare results between classification models.

Premises	Conclusion
mucusthreads1 = Absent, epithcells1 = Moderate	sex = Female
mucusthreads1 = Absent, epithcells1 = Many	sex = Female
epithcells1 = Moderate	sex = Female
epithcells1 = Many	blood1 = Nil, sex = Female
epithcells1 = Many	sex = Female

Premises	Support	Confidence	Lift
mucusthreads1 = Absent, epithcells1 = Moderate	0.129	0.930	1.492
mucusthreads1 = Absent, epithcells1 = Many	0.111	0.986	1.581
epithcells1 = Moderate	0.205	0.934	1.498
epithcells1 = Many	0.144	0.840	1.762
epithcells1 = Many	0.168	0.981	1.573

Figure 3: Some resulting rules from association rules model.

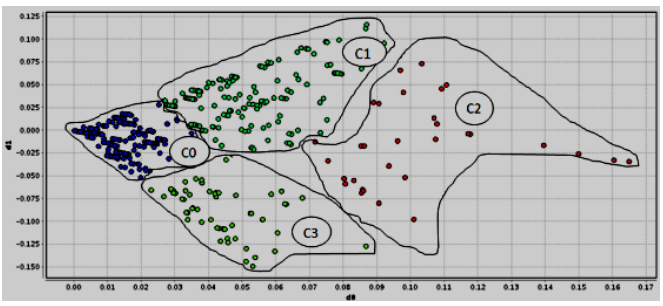

Figure 4: Resulting clusters.

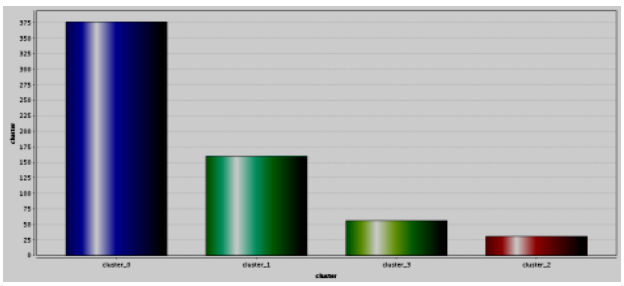

Figure 5: Clusters counts.

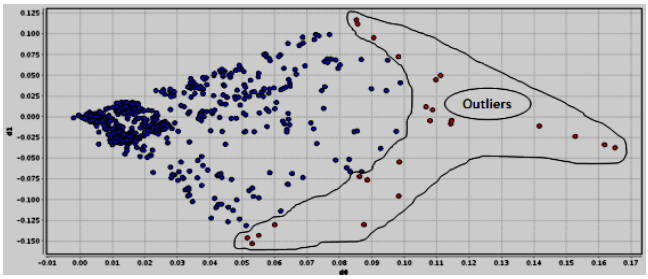

Figure 6: Detected outliers.

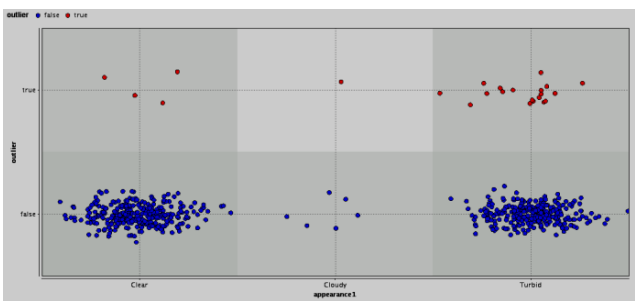

Figure 7: Outliers distribution between label classes.

Accuracy of constructed decision tree model was 70.33%, and accuracy of rule induction model was 74.58%.

To overcome the problem of low accuracy percentage, the strongest rules and branches were selected to extract knowledge.

Conclusion

Health care information systems can be exploited well if a good data mining process was applied on it with the guidance of medical specialists. This paper followed all data mining process steps to produce trusted results.

Main goal of data mining is to announce novel, actionable, understandable, and valid patterns. Up to evaluation steps followed by the rapid-miner and medical specialists, results were valid, novel, useful, and understandable. Best models were association rules and decision tree as they produced a rich accepted knowledge.

Most important knowledge informed medical specialists that females have some special abnormalities such as lots of epithelial cells and big chance of turbidity. It discovered also lots of WBCs or RBCs with calcium oxalate makes sample turbid, and finding some of triple phosphate or amorphous phosphate causes turbidity.

Future work can be considered if another datasets were available to re-apply models and get more accurate and actionable patterns. Recommendation to health care software developers is to add the functionalities of alerting data entry users about new misclassified instances and how far his entry data are, this will help to decrease errors in medical analysis tests and increase trust in medical results.

References

1. Kraft M, Desouza K, Androwich I (2002) Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population. Proceedings of the 36th Hawaii International Conference on System Sciences.
2. Gosain A, Kumar A (2009) Analysis of Health Care Data Using Different Data Mining Techniques. Intelligent Agent & Multi-Agent Systems 2009, IAMA.
3. Ang Q, Wang W, Liu Z, Li K (2010) Explored Research on Data Pre-processing and Mining Technology for Clinical Data Applications.
4. Ramachandran S, Erraguntla M, Mayer R, Benjamin P (2007) Data Mining in Military Health Systems -Clinical and Administrative Applications. Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering.
5. http://www.surgeryencyclopedia.com/
6. Lloyd-Williams M (1998) Case Studies in the Data Mining Approach to Health Information Analysis.
7. Han J, Kamber M (2001) Data Mining: Concepts and Techniques. The Morgan Kaufmann.
8. Ben-Chang S, Yi-Ting D, Ya-Wun J, Ting –Wei C (2009) Data Mining Technology for Applying of RFID and Health Examination.