Data mining big data inpatient database using Cuckoo search

Uroš Mlakar
Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova 17, 2000 Maribor
uros.mlakar@um.si

Iztok Fister Jr.
Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova 17, 2000 Maribor
iztok.fister1@um.si

Monika Marković
Faculty of Medicine, University of Maribor
Taborska 6b, 2000 Maribor
monika.markovic@student.um.si

Iztok Fister
Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova 17, 2000 Maribor
iztok.fister@um.si

ABSTRACT

This paper investigates data mining in a medical dataset by using the stochastic population-based nature-inspired Cuckoo search algorithm. Particularly, association rules are mined by applying an objective function composed of support and confidence weighted by two parameters for controlling the importance of each measure. The rules are mined in a Nationwide Inpatient Sample dataset, which is a collection of discharge records of several hospitals in the USA. Only those records, where a patient was diagnosed with Type II diabetes mellitus were extracted for association rule mining. The results show that the found rules are simple, easy to understand and also interesting, as they were verified with actual clinical studies. The results obtained can be beneficial to either doctors or insurance companies.

Keywords

data mining, big data, association rule mining, cuckoo search

1. INTRODUCTION

With the increasing rate of data collected everyday, there is a need for automatic mining of useful information hidden within. But this may be a difficult task, since this data is either big in volume, has variety (different data sources or multiple data types), or is collected at a very fast pace (velocity). An example of such data are definitely the discharge records of hospital patients. There is a lot of hidden information within this data, such as interesting connections between apparently unrelated diseasepresenteds, or discovering interesting risk factors, that contribute to a particular disease (although not being directly related to the disease). Such knowledge would be beneficial to hospitals, and also to insurance compaines, which can make evidence based decisions, and can optimize, validate and refine the rules that govern their business [6]. This important hidden knowledge can be found with the help of data mining, with methods such as clustering, feature selection, association rule mining, and many more.

This paper is structured as follows. After the introduction, data mining methods are briefly discussed in Section 2, then the Cuckoo search algorithm and the Nationwide Inpatient Sample (NIS) dataset are presented in Sections 3 and 4. The preliminary results are presented in form of association rules in Section 5, then the paper is concluded with future directions in Section 6.

2. DATA MINING METHODS

Data mining is a computing process of discovering patterns in large datasets. The goal of data mining is to extract useful information from a dataset and transform it into an understandable structure, which may be used directly or processed further by another algorithm. There are several methods for which are used for data mining, such as cluster analysis [4], dimensionality reduction [8], association rule mining [9], etc. Association rule mining has gained a lot of attention for mining interesting patterns from large databases within the research community.

2.1 Association rule mining

Association rule mining (ARM) is a rule-based machine learning method for discovering interesting relations between attributes in large databases. ARM is used for identifying strong rules using measures of interestingness, where the most established method is the Apriori algorithm introduced by Agrawal et al. [1]. ARM can be mathematically expressed as follows. Let $I = \{i_i, i_2, \ldots, i_n\}$ be a set of attributes called items and $T = \{t_1, t_2, \ldots, t_m\}$ the a set of transactions (i.e. database). Each rule is defined as an implication $X \rightarrow Y$, where $X, Y \subseteq I$. X and Y are composed of two different set of items, which are also known as item-sets; X

DOI: https://doi.org/10.26493/978-961-7023-40-4.19-22

is also defined as the antecedent, while Y is called the consequent. To be able to select interesting rules from the set of all possible rules, various measures of interestingness are utilised and also constrained. The most used constraints are the minimum support and confidence, which are defined as follows:

$$Support(X \rightarrow Y) = \frac{|t \in T; X \subseteq t \wedge Y \subseteq t|}{|T|} \quad (1)$$

$$Confidence = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

Support is defined as a proportion of transactions containing X and Y, and the total number of transactions, while confidence is a proportion of transactions which contain X, and also contain Y.

Although being able to find interesting rules on smaller datasets, it faces computational problems when confronted with bigger datasets. To overcome this problem the research has gone in the direction of stochastic population-based nature-inspired algorithms, that treat the ARM as an optimization problem.

3. CUCKOO SEARCH

Cuckoo search (CS) is a stochastic population-based nature-inspired optimization algorithm proposed by Yang and Deb in 2009 [11]. It is classified as a Swarm Intelligence (SI) algorithm, since its mechanisms are inspired by the natural behaviour of some cuckoo species in nature. To be able to capture the behaviour of cuckoos and adapt it to be suitable for using as a computer optimization algorithm, the authors idealised three rules:

- A cuckoo lays only one egg, then dumps it into a randomly chosen nest,
- Nests that contain high-quality eggs, are carried over to the next generation,
- Any cuckoo egg may be discovered by the host bird with probability $p_a \in [0, 1]$. If an egg is discovered, the host bird may abandon the nest, and build a new one at a new location.

Each solution in population of the CS algorithm corresponds to a cuckoo nest, which represents the position of the egg within the search space, and can be mathematically expressed as follows:

$$\mathbf{x}_i^{(g)} = \{x_{i,j}^{(g)}\}, \quad \text{for} i = 1, \ldots, Np \quad \text{and} j = 1, \ldots, D, \quad (3)$$

where NP is the population size, and D the dimension of the optimization problem. In the CS algorithm, new solutions are created by exploitation of the current solutions as:

$$\mathbf{x}_i^{(g+1)} = \mathbf{x}_i^{(g)} + \alpha L(s, \lambda), \quad (4)$$

where

$$L(s, \lambda) = \frac{\lambda \Gamma(\lambda) \sin(\frac{\pi\lambda}{2})}{\pi} \frac{1}{s^{(1+\lambda)}}. \quad (5)$$

The term $L(s, \lambda)$ determines the characteristic scale and $\alpha > 0$ is the scaling factor of the step size s.

Table 1: Data elements considered from the NIS dataset.

Element Name	Element description
Age	Age of patient at admission (years)
Atype	Admission type
Died	Died during hospitalization
Female	Indicator of sex
Los	Length of stay
DX1	Principal diagnosis
DX$\{2-15\}$	Diagnoses$\{2-15\}$
PR1	Principal Procedure
PR$\{2-15\}$	Procedures$\{2-15\}$

3.1 Association rule mining using CS algorithm

Since the CS is used for ARM in this paper, the solution representation has to be adapted accordingly. There are two well established encodings available for representing rules for evolutionary algorithms (EA) and SI-based algorithms. The first is the Michigan encoding [5], where each solution represents a separate association rule. In the second, the Pittsburgh encoding [5], each solution represents a set of association rules. for the purpose of this study the Michigan encoding was used. Additionally a fitness evaluation function needs to be defined in order to find the most promising rules:

$$f(\mathbf{x}_i^{(g)}) = \frac{\beta Support(X \rightarrow Y) + \gamma Confidence(X \rightarrow Y)}{\beta + \gamma} \quad (6)$$

The fitness function $f(x)$ is defined as a weighted sum of support and confidence. The weights β and γ control the importance of both said measures. The user can set the values of these weights according to the importance of each measure in the domain of association rule mining. For the purpose of this study, the values of $\beta = \gamma = 1$.

4. NATIONWIDE INPATIENT SAMPLE DATASET

The Nationwide Inpatient Sample (NIS) dataset holds the records of hospital inpatient discharges, that date back to 1988, and is used for identifying, tracking and analysing trends in health care access, quality, and outcomes. It is a publicly available dataset, without any patient identifiers. It is worth noticing that it consists merely of US hospital discharges. It holds about 64 million records, with 126 clinical and non-clinical data elements. Only the elements listed in Table 1 were used in this study.

The DX1 and DX$\{2-15\}$ are the principal diagnosis and other diagnoses, respectively. The diagnoses are represented as codes by following the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Since the NIS dataset holds a lot of records, it is hard to find association rules by considering the whole dataset. The goal of this research is to uncover other risk factors for patients, who suffer from one particular disease. For this reason, we chose a disease with code '250.30', which is Type II diabetes mellitus (TIIDM), which is a heterogeneous group of disorders characterized by a variable degree of insulin resistance, impaired insulin secretion, and increased glucose production. There are many causes of which doctors and patients should be aware of and maybe even more of those

Table 2: Association rules as found by the Cuckoo search rule miner.

Antecedent	Consequent	Fitness value
$(DX = 201.23) \wedge (SEX = FEMALE)$	$(LOS =< 44)$	0.508
$(DX = 201.23) \wedge (SEX = FEMALE) \wedge (TYPE = URGENT)$	$(LOS =< 44)$	0.503
$(DX = 142.8) \wedge (AGE = 40 - 49)$	$(LOS =< 44)$	0.501
$(DX = 202.88)$	$(DIED = NO)$	0.501
$(DX = 070.22)$	$(DIED = NO)$	0.501
$(DX = 016.04) \wedge (AGE = 60 - 69)$	$(TYPE = EMERGENCY)$	0.5
$(DX = 036.3) \wedge (DIED = NO)$	$(SEX = MALE)$	0.5
$(DX = 197.8)$	$(DIED = NO)$	0.5
$(DX = 255.8) \wedge (DIED = NO)$	$(TYPE = EMERGENCY)$	0.5
$(DX = 206.00) \wedge (AGE = 80 - 89) \wedge (SEX = FEMALE)$	$(TYPE = EMERGENCY)$	0.5
$(DX = 010.04)$	$(SEX = FEMALE)$	0.5
$(DX = 012.33)$	$(LOS =< 44)$	0.5
$(DX = 201.66)$	$(DIED = NO) \wedge (LOS =< 44)$	0.5
$(DX = 079.88) \wedge (SEX = MALE) \wedge (DIED = NO)$	$(TYPE = EMERGENCY)$	0.5
$(DX = 211.7)$	$(SEX = FEMALE)$	0.5
$(DX = 010.03) \wedge (SEX = FEMALE) \wedge (DIED = YES)$	$(DX = 015.55)$	0.5
$(DX = 201.66) \wedge (SEX = MALE) \wedge (DIED = NO)$	$(LOS =< 44)$	0.5
$(DX = 171.4)$	$(SEX = FEMALE)$	0.5
$(DX = 085.5) \wedge (DIED = YES)$	$(LOS =< 44)$	0.5
$(DX = 232.8)$	$(DIED = NO)$	0.5

that we might not know [3]. The latter is also the reason for this study, and with this in mind, all records containing the disease with ICD-9-CM code '250.30' (Type II diabetes mellitus) were extracted from the whole NIS dataset to form a new smaller dataset.

5. RESULTS

In this section the results of association rule mining on the NIS dataset using the CS algorithm is presented. The results are reported in Table 2 in form of association rules. Additionally the fitness value of each rule is reported. Only the best 20 rules are reported in Table 2, but only the top five are additionally commented on. The average number of antecedent obtained in this study is 1.8, while the average number of consequent is 1.05. This is favourable for the user, since shorter association rules are easier to understand. It also worth emphasizing that all rules produces are somehow related to the TIIDM. The first two rules indicate that the TIIDM is involved with the pathogenesis of non-Hodgkin's sarcoma. This fact is supported by several studies in literature [2]. The third and fourth rules state that there is a connection of Maligant neoplasm of major salivary gland and other maligant lymphomas with TIIDM, which is supported by a study in [10] where an increased prevalence of diabetes was found in patients with salivary gland tumour. A chronic viral hepatitis B was found to be in connection with TIIDM in the fifth rule [7].

6. CONCLUSION

The CS algorithm was investigated as a association rule miner on a hospital discharge dataset. The CS produces rules, which are simple, easy to understand, and also interesting. The rules are found with the help of a objective function, which weighs the support and confidence of the rules. The weights control how each interestingness measure is important, and thus guides the search in the desired direction.

The obtained rules were compared with research in the field of Type II diabetes mellitus, where all results were confirmed to be reasonable and supported by a study.

In future we would like to use the CS algorithm for mining association rules for other diseases which occur commonly in the modern world.

7. REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] Chun Chao and John H. Page. Type 2 diabetes mellitus and risk of non-hodgkin lymphoma: A systematic review and meta-analysis. *American Journal of Epidemiology*, 168(5):471–480, 2008.

[3] Anthony S Fauci et al. *Harrison's principles of internal medicine*, volume 2. McGraw-Hill, Medical Publishing Division New York, 2008.

[4] Iztok jr. Fister. *Algoritmi računske inteligence za razvoj umetnega športnega trenerja*. PhD dissertation, University of Maribor, Faculty of Electrical Engineering and Computer Science, 2017.

[5] John H. Holland. Adaptation*. In Robert Rosen and Fred M. Snell, editors, *Progress in Theoretical Biology*, pages 263–293. Academic Press, 1976.

[6] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):51, 2011.

[7] Kar Neng Lai, Fernand Mac-Moune Lai, Nancy WY Leung, Stephen T Lo, and John S Tam. Hepatitis with isolated serum antibody to hepatitis b core antigen: a variant of non-a, non-b hepatitis? *American journal of clinical pathology*, 93(1):79–83, 1990.

[8] Uroš Mlakar, Iztok Fister, Janez Brest, and Božidar Potočnik. Multi-objective differential evolution for feature selection in facial expression recognition systems. *Expert Systems with Applications*, 89:129–137, 2017.

[9] Uroš Mlakar, Milan Zorman, Iztok Fister Jr, and Iztok Fister. Modified binary cuckoo search for association rule mining. *Journal of Intelligent & Fuzzy Systems*, 32(6):4319–4330, 2017.

[10] Zsuzsanna Suba, József Barabás, György Szabó, Daniel Takács, and Márta Ujpál. Increased prevalence

of diabetes and obesity in patients with salivary gland tumors. *Diabetes Care*, 28(1):228–228, 2004.

[11] Xin-She Yang and Suash Deb. Cuckoo search via lévy flights. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*, pages 210–214. IEEE, 2009.