Interpretable models for medical datasets by means of biclustering and association rules

Rosana Veroneze, Fernando J. Von Zuben
Laboratory of Bioinformatics and Bioinspired Computing (LBiC/DCA/FEEC/Unicamp)

Introduction: Veroneze & Von Zuben [1] recently proposed an enumerative biclustering algorithm to mine all maximal biclusters in mixed-attribute datasets. A mixed-attribute dataset is composed of numerical (discrete or continuous) and categorical (ordinal or nominal) attributes, being very common in medical data. As usual examples we may mention information on weight, height, age, gender, and location of pain. Notice that strictly numerical or categorical datasets are special cases of mixed-attribute datasets [1]. Alternative biclustering proposals to handle mixed-attributes datasets either do not simultaneously exhibit four key properties (which are all present in our approach), more specifically efficiency, completeness, correctness, and non-redundancy, or, when exhibiting the four key properties, the numerical attributes should pass mandatorily through discretization before the mining process, which inevitably promote information loss. Additionally, Veroneze & Von Zuben [1] presented the biclusters in a user-friendly and intuitive form, by automatically converting them to association rules [2], more specifically quantitative class association rules (QCARs). Here, we show how this proposal is valuable to yield a parsimonious set of relevant rules, automatically providing useful and interpretable models for medical datasets.

Materials and Methods: We used two datasets in our experiments. One of them is the Acute dataset, which contains 6 attributes of 120 patients. It has two decision variables that indicate the presence or absence of a disease of the urinary system, which are *inflammation of urinary bladder* (IUB) and *nephritis of renal pelvis origin* (NRP). The other dataset is the Heart dataset, which contains 13 attributes of 270 patients. The decision variable indicates the presence or absence of *heart disease*. Both datasets are publicly available at UCI Repository [3]. For more details about the datasets and the parameterization of the biclustering algorithm see [1]. The quality of the rules was measured by the metrics completenesses, confidence, lift and leverage [2].

Results: Table 1 shows some rules that we obtained for the Acute and Heart datasets. We refrained from presenting more rules due to space restriction.

Table 1: Some examples of rules for the Acute and Heart datasets.

#	Dataset	Rule	Comp	Conf	Lift	Lev
1	Acute (IUB)	urinePushing{yes}, micturitionPain{yes} ⇒ IUB{Yes}	0.83	1.00	2.03	0.21
2	Acute (NRP)	temperature[35.50,37.90], nausea{no} ⇒ NRP{No}	0.98	1.00	1.71	0.21
3	Acute (NRP)	temperature[39.40,41.50], lumbarPain{yes} ⇒ NRP{Yes}	0.71	1.00	2.40	0.20
4	Heart disease	sex{M}, chestPain{asymptom.}, vesselsColor{1} ⇒ HeartDisease{Yes}	0.23	0.97	2.17	0.06

Discussion: Rule #1 shows that 83% of the patients with IUB presented urine pushing and micturition pain. Rule #2 shows that 98% of the patients without NRP did not presented fever and nausea. Rule #3 shows that 71% of the patients with NRP presented fever and lumbar pain. Rule #4 shows that 23% of the patients with a heart disease were male, presented asymptomatic chest pain, and one major vessel colored by fluoroscopy.

Conclusion: The present work confirmed that QCARs directly extracted from biclusters are valuable and automatic means of providing useful and relevant interpretable models for medical datasets.

References: [1] Veroneze R & Von Zuben FJ, arXiv preprint arXiv:1710.03289, 2017; [2] Zaki MJ & Meira W, Cambridge University Press, 2014; [3] http://archive.ics.uci.edu/ml.