Proc. of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023)
19-21 July 2023, Tenerife, Canary Islands, Spain

Classification based on Association Rules: Complexity and Interestingness Guided Algorithm

Carlos Molina
Dept. of Software Engineering
University of Granada
Granada, Spain
carlosmo@ugr.es (ORCID: 0000-0002-7281-3065)

Belén Prados-Suárez
Dept. of Software Engineering
University of Granada
Granada, Spain
belenps@ugr.es (ORCID: 0000-0002-3980-102X)

***Abstract*—Classification based on association rules has proved able to get high precision and deal with data noise in real scenarios. However, very low support thresholds are used to get these results, generating a great number of rules. In this process, interesting and non-interesting itemsets are considered for subsequent rule generation. In this paper, we propose a new method to identify not only frequent but also interesting itemsets, including an *interestingness measure* when pruning candidate itemsets. With this approach, we reduce the number of association rules in the final classifier and increase precision. We combine it with an iterative approach, reducing the execution time and complexity of the final rule set. We have tested the proposal using a COVID19 patient database.**

***Index Terms*—CBA, Complexity Reduction, Classification, Interest measure**

I. Introduction

Classification methods have been used in several contexts like credit card marketing [1], pattern extraction in industry [2], or security issues [3]. We can find several heuristic classification methods based on genetic algorithms (e.g. [4]), genetic network programming (e.g. [3]), or even greedy approaches (e.g. [5]). But in most cases, the approaches fail when applied in real scenarios due to bad tolerance to noisy data or overfitting problems (see [6], [7] for more details). In these cases, classification based on association rules (CAR) has yielded better results and several classifiers have been proposed following this idea (e.g [1], [8]–[10]). However, most of these proposals are based on generating a huge number of frequent itemsets to generate the rules. This increases execution time and handling all these itemsets can be a problem. Several proposals use special structures to deal with the itemsets (e.g. tree [11], production functions [12], special sets [6], etc.), aiming to reduce the time required to calculate the support.

The proposed solutions increase the speed of support calculation but do not reduce the search space, so the final number of generated rules is high. The main resulting issue is that the user then has problems to understand the full system.

What we propose in this paper is to control the number of generated rules to obtain the simplest rule set, without generating all frequent itemsets. In this process, we try to assemble the simplest rules as possible (minimising the amount of items in the antecedent). This way, the entire system will be more understandable for the user. To reduce the search space, we have introduced an *interestingness measure* together with the support so that we do not need to use a very low support threshold to retrieve all interesting itemsets. This way, we reduce the number of itemsets for rule generation. Combining this reduction with an iterative approach like in [13], we can reduce the execution time and complexity of the final rule set. In this iterative approach, we do not generate all frequent itemsets (of all sizes). This method is presented in Section II as the starting point of our proposal. The following sections present the adaptation proposed to consider interesting itemsets. In Section III, we present the COVID19 patient datacube used to test the effect of the interestingness measure in the process. Finally, we set out our conclusions and final remarks in Section IV.

II. COGiARE: Algorithm for Interpretable Associations

For our proposal, we have used *COGiARE* [13] as our base algorithm. This method extracts association rules for classification in a iterative way (Figure 1 shows this process). Instead of generating all the strong association rules and selecting the rules later, the method first obtains size 2 association rules (only one antecedent and the class as consequent) and chooses rules to make a classification. Then, size 3 itemsets are calculated and we get a set of classification rules considering the size 2 and 3 rules. The quality of this new rule set is compared with that obtained in the previous iteration. If we find an improvement in quality, we choose this new rule set and continue with the size 4 itemsets. The process is repeated either until we can get a classification rule set that improves the one chosen in the previous iteration for one size of itemsets or we can not generate any more rules. A detailed explanation can be found in [13] and an application to hospital management in [14].

The algorithm *COGiARE* considers the support when selecting frequent itemsets to build the rules. The quality of the rules

979-8-3503-2297-2/23/$31.00

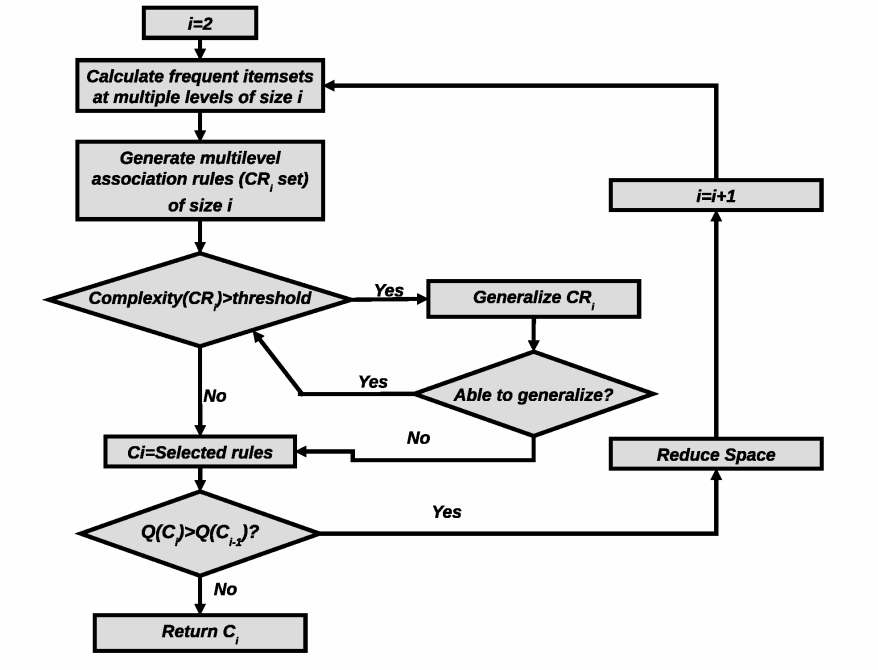

Fig. 1. COGIARE algorithm.

is based on the *certainty factor (CF)* proposed in [15]. The use of this measure avoids some of the well-known problems of Confidence (see [16] for more details). Below we introduce a variation to consider not only frequent but *interesting* itemsets for rule generation.

A. Interesting Itemsets

The concept of *interesting itemset* is different from that of *frequent itemset*, which refers to facts that occur together in a large number of studied cases. It does not, however, allow there to be itemsets that are highly related but which are not frequent, as in the case of rare diseases. The purpose of *interesting itemsets* is to model not only what is relevant because it is frequent but also what is relevant because it is highly related, even though its overall frequency is low. A number of interestingness measures have been published (e.g. lift [17]), but as these are applied during rule generation, interesting itemsets can be lost if they are neither frequent nor considered for later rule generation.

In order to model this concept, we have defined the following measure:

Definition 1: The *interest* of the itemset $\{i_1, ..., i_n\}$ is the function *In*, defined as follows:

$$In(\{i_1, .., i_n\}) = \frac{f1 - f2}{f2} \in [-1, +1] \quad (1)$$

where

$$f1 = \frac{sup(\{i_1, ..., i_n\})}{min\{sup(i_1), ..., sup(i_n)\}} \quad (2)$$

$$f2 = max\{sup(i_1), ..., sup(i_n)\} \quad (3)$$

where $sup(i)$ is the $support$ of itemset i.

If $In(I) > 0$, this means that the probability of the items in I appearing together is higher than the expected probability of them being independent. Conversely, $In(I) < 0$ means the opposite: the probability of the items appearing together is lower than the probability of the items being independent. In our case, values greater than 0 mean that the items are interesting. In the algorithm, we will use $threshold_{In}$ to indicate the minimum value to consider an itemset for rule generation.

This way, we consider the itemsets as candidates to generate rules if the support exceeds a threshold ($threshold_{Sup}$) or if the interest exceeds another threshold ($threshold_{In}$).

III. Experiments

A. Datacube: COVID19 patient data

In order to test the proposed algorithm, we used data from a COVID-19 patient database [18]. This dataset contains information about 2,547 patients (Figure 2).

We built a simple datacube considering 11 dimensions with hierarchies:

- *Patient*: including information on each patient's sex and age.
- *Outcomes*: including information regarding patient outcomes (values: *ICU*, *Home* and *Dead*).
- *Diagnosis*: with ICD-10 codes [19] of the patient's hospital diagnosis.
- *Procedure*: using the ICD-10-PCS [19] code for the applied procedures.
- *Drug*: all the drugs given to each patient. The drugs are coded using the Anatomical Therapeutic Chemical (ATC)

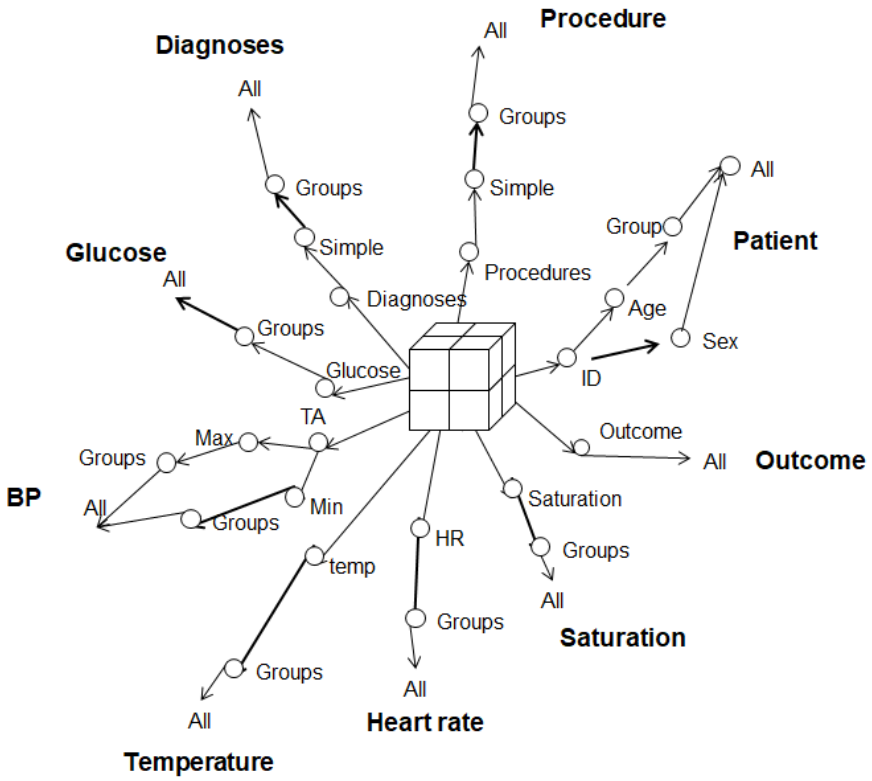

Fig. 2. COVID19 DataCube.

classification system [20]. For the drugs, we define a hierarchy considering the anatomic level and the therapeutic group.

- *BP*: data regarding blood pressure (BP), with minimum and maximum BP at different times. We defined fuzzy labels for these values to establish when blood pressure is *low*, *normal* or *high*.
- *Heart rate*: grouping the values using fuzzy labels similar to BP (blood pressure).
- *Glucose*: measuring blood glucose.
- *Temperature*: body temperature to identify fever.
- *Saturation*: information on oxygen saturation in blood.

In this datacube, we want to classify what the outcome was for each patient (*Outcome* dimension at *base* level with the values *Dead*, *Home*, *ICU*).

B. Tests

To test the effect of considering *interesting itemsets*, we have performed several experiments changing only the threshold $threshold_{In}$ applied to itemset candidates to be used in rule generation. In all the tests, we used $threshold_{Support} = 0.01$ and $threshold_{CF} = 0.2$. For $threshold_{In}$ we used the values 1.0 (without considering the interestingness measure and only looking for frequent itemsets), 0.8, 0.6, 0.4, 0.2 and 0.0. We have limited the size of the itemsets to 4 (3 iterations of the algorithm) so we can compare and show the results. We expected to get better results for low values of the interestingness threshold ($trheshold_{In}$).

We have applied a 10-cross-fold validation to test the quality of the generated rule set. The results of the executions are shown in Table I. We explain the meaning of the columns below:

- *FI*: average number of *frequent itemsets* generated in each iteration.
- *AR gen*: average number of *association rules* generated.
- *Rules*: average number of rules in the rule set obtained after each iteration.
- *Prec.*: average *precision* of the rule set given by the method as a final result.

As expected, there are some differences regarding the number of itemsets generated (column *FI*). When $threshold_{In}$ is low (considering more interesting but not frequent itemsets) the method considers more itemsets, so it generates more rules (column *AR gen*). Looking at the size of the final rule set for classification (column *Rules*), we can see there are small differences that have an effect on classification precision (column *Prec.*). As we can see, considering *interesting itemsets* gives a better classification. The improvement does not mean using more rules but better ones, as the results using low values for $threshold_{In}$ achieve higher precision while using fewer rules.

TABLE I
RESULTS OF THE EXPERIMENTS

	Size 2			Size 3			Size 4			
$threshold_{In}$	FI	AR gen	Rules	FI	AR gen	Rules	FI	AR gen	Rules	Prec.
0.0	429	80	55	2301	707	279	4680	2080	465	0.802
0.2	382	76	50	2224	637	228	4679	2009	438	0.797
0.4	358	76	50	2195	610	228	4678	1982	435	0.786
0.6	329	76	50	2170	592	227	4676	1962	480	0.775
0.8	313	76	50	2166	588	227	4676	1958	480	0.775
1.0	300	76	50	2162	584	227	4676	1954	469	0.771

IV. CONCLUSIONS

In this paper, we have presented a classification method based on association rules that considers not only frequent itemsets but also interesting itemsets. We have tested the effect on classification in a COVID19 patient database. The results show that the use of interesting itemsets improves classification precision without a big increase in the number of itemsets and rules.

In future work, we need to compare the proposed method against other well-known proposals, and with other much more complex databases, to test the actual improvement.

ACKNOWLEDGMENT

This publication is part of the Grant PID2021-126363NB-I00 funded by MCIN/AEI/10.13039/501100011033, "ERDF A way of making Europe" and B-TIC-744-UGR20 *ADIM: Accesibilidad de Datos para Investigación Médica* research project of the Junta de Andalucía.

REFERENCES

[1] Y. li Zhu, Y.-F. Wang, S.-P. Wang, and X. juan Guo, "Mining association rules based on cloud model and application in credit card marketing," in *Wearable Computing Systems (APWCS), 2010 Asia-Pacific Conference on*, 2010, pp. 165–168.

[2] J. w. Lee and C. Giraud-Carrier, "Results on mining nhanes data: A case study in evidence-based medicine," *Computers in Biology and Medicine*, vol. 43, no. 5, pp. 493–, Jun. 2013. [Online]. Available: http://search.proquest.com/docview/1324273684?accountid=14542

[3] S. Mabu, C. Chen, N. Lu, K. Shimada, and K. Hirasawa, "An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 1, pp. 130–139, 2011.

[4] Y. ling Tang and F. Qin, "Research on web association rules mining structure with genetic algorithm," in *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, 2010, pp. 3311–3314.

[5] C.-M. Wu, Y.-F. Huang, and J.-Y. Chen, "Privacy preserving association rules by using greedy approach," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol. 4, 2009, pp. 61–65.

[6] L. T. Nguyen, B. Vo, T.-P. Hong, and H. C. Thanh, "Car-miner: An efficient algorithm for mining class-association rules," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2305 – 2311, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412011591

[7] S. Ghosh, S. Biswas, D. Sarkar, and P. Sarkar, "Association rule mining algorithms and genetic algorithm: A comparative study," in *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, 2012, pp. 202–205.

[8] R. Sumithra and S. Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery," in *Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on*, 2010, pp. 1–5.

[9] H. Qinglan and D. Longzhen, "Multi-level association rule mining based on clustering partition," in *Intelligent System Design and Engineering Applications (ISDEA), 2013 Third International Conference on*, 2013, pp. 982–985.

[10] X. S. Xiong, L. Fan, and Z. Lei, "Ontology-based association rule quality evaluation using information theory," in *Computational and Information Sciences (ICCIS), 2010 International Conference on*, 2010, pp. 170–173.

[11] B. Vo and B. Le, "A novel classification algorithm based on association rule mining." in *LNAI 5465 Pacific rim knowledge acquisition workshop, held with PRICAI 2008, Viet Nam.*, 2008, pp. pp. 61–75.

[12] Y. Yusof and M. Refai, "Mmcar: Modified multi-class classification based on association rule," in *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*, 2012, pp. 6–11.

[13] C. Molina, B. Prados-Suarez, D. Sanchez, and A. Vila, "Complexity reduction in classification based on asocciations over datacubes," in *XVII Congreso Español sobre Tecnologias y Logica Fuzzy*.

[14] M. Prados De Reyes, C. Molina, B. Prados, and C. Peña-Yañez, "Interpretable associations over datacubes: application to hospital managerial decision making," in *e-Health–For Continuity of Care*. IOS Press, 2014, pp. 131–135.

[15] E. Shortliffe and B. Buchanan, "A model of inexact reasoning in medicine," *Math. Biosci.*, vol. 23, pp. 351–379, 1975.

[16] M. Delgado, N. Marin, D. Sanchez, and M. Vila, "Fuzzy association rules: General model and applications," *IEEE transactions on Fuzzy Systems*, vol. 11, no. 2, pp. 214–225, 2003.

[17] N. Hussein, A. Alashqur, and B. Sowan, "Using the interestingness measure lift to generate association rules," *Journal of Advanced Computer Science & Technology*, vol. 4, no. 1, p. 156, 2015.

[18] MH Hospitales. (2021) Covid Data Save Lives. https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version.

[19] World Health Organization. Icd-10 version:2019. [Online]. Available: https://icd.who.int/browse10/2019

[20] WHO Collaborating Centre for Drug Statistics Methodology. Anatomical, therapeutic, chemical classification system. [Online]. Available: https://www.whocc.no/atc_ddd_index/