Assignment 3: Pattern mining with PBC4cip

Diaz-Moraila J.E.[1], Cortes-Guzman, M.[2], and Hinojosa-Cavada, C.[3]

[1] A00828174@itesm.mx
[2] A01270966@itesm.mx
[3] A01137566@itesm.mx

Abstract. In the following work we present the construction of several classification models using PBC4cip weka package with various hyperparameters to extract contrast patterns (CP) that describe the best 100 universities of the QS World University Rankings 2019. Since PBC4cip is a tree-based classifier these classification models we built using 1000 trees and exploring the hyperparameters such as random features, depth of trees, minumum object by leaf and random subspace. We depict an analysis of the obtained results, reporting how different hypterparameters changes the contrast patters found by PBC4cip.

1 Introduction

White-box classification algorithms plays a key role in the field of artificial intelligence, since its useful to know why a given algorithm is taking a given decision, contrast patterns (CPs) were a major element of the white-box classification algorithms partly due to the advantage of them being able to explain classification results in a language that is easy to understand for an expert. CP-based classifiers, when using contrast patterns extracted by miners based on decision trees, attain accuracies comparable with other state-of-the-art classifiers [1] .

In the present work we explore different hypterparameters of the PCB4cip weka package to build 16 different models from which the top 4 contrast patterns are extracted per model (Each of these patterns must obey the property $s1\_r - s2\_r > 0.3$; where $s1\_r$ and $s2\_r$ are the supports of the pattern in the class of the universities ranked from 1 to r and ranked from r+1 to 200, respectively) .

The features of the dataset are described in the following list:

- intColYYYY: (International Collaboration) Percent of publications reporting international collaboration for the year YYYY.
- acaColYYYY: (Academic-Corporate Collaboration) Percent of publications reporting collaborations from other educational institutions for the year YYYY.
- pubYYYY: (Scholarly Output) Number of articles published in the year YYYY.
- citYYYY: Number of citations received in the year YYYY.
- pubTCPYYYY: (Outputs in Top 10% Citation Percentiles) Number of publications in the top 10% of the most-cited publications in the year YYYY.

– pubTJPYYYY: (Publications in Top 10% Journal Percentiles) Number of publications in the top 10% of the most-cited journals in the year YYYY.
– citPPYYYY: Ratio of citations per publication for the year YYYY.
– authorsYYYY: Number of authors publishing in the year YYYY.
– citPAYYYY: Ratio of citations per author for the year YYYY.
– pubPAYYYY: Ratio of publications per author for the year YYYY.
– h5Index: The h-index computed over the last five years.
– fwCitImpYYYY: (Field-Weighted Citation Impact) Number of citations received by an entity's publications compared with the average number of citations received by all other similar publications in the data universe for the year YYYY. The fwCitImpYYYY value indicates if the entity's publications have been cited exactly the same, more, or less than would be expected based on the global average for similar publications. For example, a value of 2.11 means 111% more than the world average, a value of 0.87 means 13% less than the world average, and a value equal to 1 means that it was exactly as expected.

2 Experiment runs

For the first experiment, a parameter of 1000 trees in Random Forest (RF) was used in the univariate PBC4cip framework in Weka, all other parameters were left as defaults. We obtained the following interesting patterns:

Table 1. Top patterns found with default settings

Pattern	Relative support for top universities
citPA2016 > 14.79 cit2016 > 117378.00 pubTCP2018 > 19.10 pubPA2016 > 0.95	0.4
citPA2016 > 14.79 cit2018 > 31738.50 pubPA2018 > 0.98 citPP2017 > 8.65 h5Index > 165.00	0.39
pub2017 > 7940.00 pubTCP2017 > 21.50 pubPA2016 > 0.95	0.37
h5Index > 172.50 cit2016 > 115133.00 citPP2018 > 4.25 pubPA2016 > 0.95	0.36

In Table 1, the second column depicts the difference between the support of class 1 (top universities) and class 2 (the rest), so the patterns presented at the

top patterns that support good universities relative to bad ones. It is possible to observe that citations per author in 2016 is an important feature as it appears in two different patterns with the same value. This is also the case for total citations in 2016 and the h-index. Remarkably, the number of publications per author per year is always present in the patterns (even if it is for different years) which implies that this particular metric is very important for the success of a university.

For the second experiments we varied the number of random features parameter in PBC4cip while keeping the same number of trees (1000). The resulting interesting patterns are as follows:

Table 2. Top patterns found with varying the number of random features

Random Features	Pattern	Relative Support
5	pub2016 >7543.00 AND citPP2017 >8.65 AND intCol2018 >36.45	0.37
	pub2016 >7543.00 AND fwCitImp2016 >1.94 AND pubPA2017 >0.97	0.36
	pubTCP2018 >19.45 AND cit2016 >117378.00 AND intCol2018 >36.45	0.36
	fwCitImp2016 >1.94 AND cit2016 >117378.00 AND pubPA2018 >0.98	0.36
10	h5Index >172.50 AND cit2018 >38284.00 AND citPP2018 >4.25 AND pubPA2016 >0.95	0.36
	pub2017 >7940.00 AND citPA2017 >9.68	0.35
	authors2017 >7050.50 AND citPA2017 >9.69 AND cit2018 >38603.00	0.34
	cit2016 >107650.00 AND intCol2016 >37.45 AND h5Index >165.00	0.34
15	pub2017 >7940.00 AND citPA2017 >9.68	0.35
	intCol2016 >37.45 AND cit2016 >117378.00	0.34
	cit2016 >107650.00 AND intCol2016 >37.45 AND h5Index >165.00	0.34
	pub2016 >7543.00 AND citPA2017 >9.68	0.34

When comparing Table 1 with Table 2, it is easy to see that moving the random features parameter from $log_2(36 + 1)$ leads to overall smaller relative supports for the top universities class in the found top patterns although the change is not that big. Some new features appear which were not observed in the model with default parameters such as international collaboration and number of publishing authors.

For the third experiment, the depth of trees parameter was varied keeping the same number of trees that we have used so far.

Table 3. Top patterns found with varying depth of trees

Depth of Trees	Pattern	Relative Support
2	cit2016 >107650.00	0.48
	cit2017 >75794.50	0.46
	cit2018 >35720.00	0.45
	pub2016 >7543.00	0.44
3	cit2016 >107650.00 AND citPP2017 >6.40	0.49
	cit2016 >107650.00 AND acaCol2018 >2.20	0.49
	cit2016 >107650.00 AND pubTCP2017 >13.90	0.49
	cit2016 >107650.00 AND pubTJP2017 >30.15	0.49
4	authors2017 >7050.50 AND acaCol2016 >2.20 AND cit2016 >107115.50	0.48
	authors2016 >6535.50 AND cit2016 >107115.50 AND acaCol2018 >2.20	0.48
	authors2017 >7050.50 AND cit2016 >107115.50 AND pubTJP2018 >28.10	0.48
	authors2018 >6722.50 AND fwCitImp2016 >1.94 AND pubPA2017 >0.96	0.4

Comparing Table 1 with Table 3 reveals that limiting the depth of tree results in slightly better support values overall that distinguish top universities from others. It is interesting that particularly for depth = 2, only singleton propositions are found and 3 of them are for the total number of citations received in different years which indicates that this very simple metric alone is important.

For experiment number 4 and while keeping the same number of trees again, we varied the minimum objects by leaf parameter while keeping all other parameters as defaults.

Table 4. Top patterns found with varying the minimun objects by leaf

Min Object By Leaf	Pattern	Relative Support
3	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2016 >0.95 AND pubTCP2018 >19.10 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >2.90 AND intCol2016 >35.70 AND cit2016 >117378.00	0.37
	cit2018 >35720.00 AND pubTCP2017 >21.50 AND pubPA2017 >0.96 AND pub2018 >8184.50	0.36
4	cit2018 >35720.00 AND pubPA2016 >0.95 AND pubTCP2017 >21.50 AND pub2016 >7472.50	0.37
	pub2016 >7543.00 AND citPP2017 >8.65 AND intCol2018 >36.45	0.37
	pubTCP2016 >13.65 AND intCol2016 >35.70 AND cit2016 >117378.00	0.37
	citPA2016 >14.79 AND cit2018 >31738.50 AND intCol2016 >35.55 AND cit2016 >117378.00	0.36
5	h5Index >172.50 AND pub2018 >7036.50 AND citPP2018 >4.25 AND pubPA2018 >0.98	0.35
	pubTJP2018 >38.70 AND cit2016 >117378.00 AND intCol2016 >35.70	0.35
	citPA2017 >10.10 AND cit2017 >79138.00	0.35
	pub2017 >7940.00 AND citPA2017 >9.68	0.35

We noticed that varying the number of minimum objects by leaf did not produce great changes compared to running the algorithm with default parameters. Some patterns observed with defaults seem to also be present in essence in Table 4, but with slight variations.

For the fifth experiment, we maintained the number of trees in 1000 and this time we varied the size of the random subspace metaclassifier in the PBC4cip univariate framework.

Table 5. Top patterns found with varying the size of the random subspace

Random Subspace	Pattern	Relative Support
0.2	citPA2016 >14.79 AND cit2016 >117378.00 AND pubTCP2018 >19.10 AND pubPA2016 >0.95	0.4
	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2018 >0.98 AND citPP2017 >8.65 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >4.05 AND cit2016 >117378.00 AND citPA2017 >9.69	0.36
0.4	citPA2016 >14.79 AND cit2016 >117378.00 AND pubTCP2018 >19.10 AND pubPA2016 >0.95	0.4
	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2018 >0.98 AND citPP2017 >8.65 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >4.05 AND cit2016 >117378.00 AND citPA2017 >9.69	0.36
0.6	citPA2016 >14.79 AND cit2016 >117378.00 AND pubTCP2018 >19.10 AND pubPA2016 >0.95	0.4
	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2018 >0.98 AND citPP2017 >8.65 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >4.05 AND cit2016 >117378.00 AND citPA2017 >9.69	0.36

In table 5, we obtained the same patterns with the same supports for all the size of random subspace parameters we tried (size of bag for RF). This means that independently of this parameter, the patterns can be discovered by PBC4cip which in turn implies that they are in fact decisive patterns at the moment of telling a top university from a not so good one.

For the sixth experiment, we utilized different feature selection algorithms in the PBC4cip framework and obtained again the top 4 patterns per algorithm while keeping all settings as default.

Table 6. Top patterns found with varying the attribute selection algorithm

Attribute Selection Algorithm	Pattern	Relative Support
CfsSubsetEval	cit2017 >75794.50 AND cit2018 >43571.50	0.33
	pub2016 >7543.00 AND cit2018 >43571.50	0.33
	pub2016 >7472.50 AND cit2018 >43571.50	0.33
	cit2016 >107650.00 AND pubTJP2017 >30.15 AND cit2018 >43571.50	0.33
InfoGainAttributeEval	citPA2016 >14.79 AND cit2016 >117378.00 AND pubTCP2018 >19.10 AND pubPA2016 >0.95	0.4
	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2018 >0.98 AND citPP2017 >8.65 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >4.05 AND cit2016 >117378.00 AND citPA2017 >9.69	0.36
ClassifierAttributeEval	citPA2016 >14.79 AND cit2016 >117378.00 AND pubTCP2018 >19.10 AND pubPA2016 >0.95	0.4
	citPA2016 >14.79 AND cit2018 >31738.50 AND pubPA2018 >0.98 AND citPP2017 >8.65 AND h5Index >165.00	0.39
	pub2017 >7940.00 AND pubTCP2017 >21.50 AND pubPA2016 >0.95	0.37
	citPP2018 >4.05 AND cit2016 >117378.00 AND citPA2017 >9.69	0.36

Notable things for Table 6 include that using the Correlation-based Feature Subset Selection (CfsSubsetEval) results in noticeably lower relative supports that using no feature selection with default settings. Both the attribute selection based on information gain (InfoGainAttributeEval) and on a classifier (ClassifierAttributeEval) yield the same pattern supports to those observed without any feature selection algorithm (in fact they yield exactly the same patterns between them too).

The pattern with the greatest support for all random subspace parameters, and two of the feature selection algorithms (and for default settings) includes many types of features such as total citations, citations per author, publications per author, citations per publication. Some of these features are definitely related to h-index which also appears in the same pattern predicate.

We also wanted to point out that one attribute, the academic-corporate collaboration, never appeared in any of the top 4 patters for any of the models

we created with PBC4cip. This is interesting as this indicates that universities shouldn't be prioritizing this activity and would probably be better off dedicating resources to other areas if they want to increase their university ranking. All other attributes appear at least once in the patterns presented in this document, but particular features with low counts include Number of publications in the top 10% of the most-cited journals (4 times appeared), and Field-Weighted Citation Impact (3 times appeared).

References

1. Leonardo Cañete-Sifuentes, Raúl Monroy, Miguel Angel Medina-Pérez, Octavio Loyola-González, and Francisco Vera Voronisky. Classification based on multivariate contrast patterns. *IEEE Access*, 7:55744–55762, 2019.