Cluster Validation Using VIC

Edoardo Bucheli-Susarrey, Hugo Velasco-Romero, and Sebastián Loredo-Gomez

Tecnologico de Monterrey

Abstract. VIC is a Cluster Validation Technique that uses an ensemble of classifiers to evaluate a clustering for some given data set. This work is a technical report discussing our implementation of VIC and the results obtained over 50 different clusterings of a Scientometrics data set containing the Top 200 Universities world-wide according to QS.

Keywords: Clustering · Cluster Validation · VIC · QS Ranking

1 Introduction

The analysis of a certain partition or clustering of a data set can be a difficult task when there are no labels in the data set. This is a very common occurrence in real world applications of clustering algorithms. Some methods have been proposed that estimate the distance and separability of some clustering such as the Calinsky Harabaz Score [2], the Silhouette Score [5] and the Davies Bouldin Score [3]. This report presents experiments using the VIC method proposed in [4] which uses an ensemble of classifiers to generate a validity index for a given clustering of the data.

This report is tied to a GitHub repository by the name Cluster-Validation-With-VIC [1].

2 The VIC Algorithm

The Culster Validity Index using Supervised Classifiers (VIC) uses as its name implies a set of supervised classifiers where the clusters in some given partition are used as the label for each object in the data set. The motivation behind this idea is that a good partition of a data set consists of clusters that are compact and separated enough for an artificial expert to recognize. To correct for the bias that a given classifier might induce, an ensemble of classifiers is used instead.

There are three main components to VIC: (1) A clustering algorithm Ω (2) A data set D and (3) a set of classifiers Ψ. Using Ω we can generate a partition of D. Next, we use the obtained partition and add them as labels for each element in D and for each classifier in Ψ we generate and test models using 10-Fold Cross Validation and report the average AUC for each classifier.

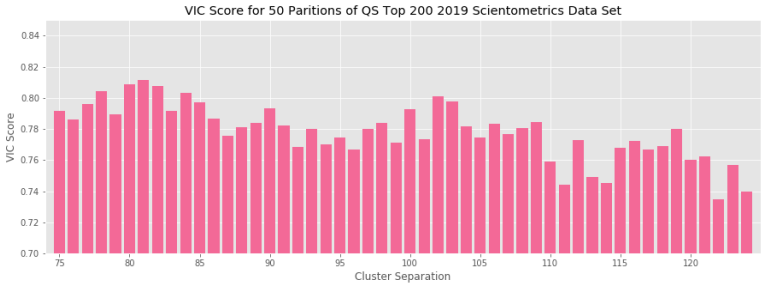

Fig. 1: VIC Score for each partition used

3 QS 2019 Top 200 Scientometrics Data Set

To test the algorithm, we generated a data set with data from Scopus and Scival for each of the Top 200 Universities world wide according to the QS 2019 ranking. The data set includes variables with information from a 5 year period related to scientific publications for each institution.

There is a total of 550 variables out of which most are numeric. Categorical Variables were separated into binary dummy variables which resulted in 757 total attributes. For additional information regarding the data set, please check the report included in the previously mentioned GitHub repository.

Instead of using some clustering algorithm, we have created the partitions by separating the universities into two clusters according to the QS 2019 Ranking. Each partition splits the data using a different rank, any school above that rank belongs to the first cluster and its complement belongs to the second cluster. We iterate over ranks 75 to 124 to obtain 50 different partitions. The goal is to find which is the best way to split the data set.

4 Implementation

Our implementation has been coded in Python and it uses 6 Classifiers listed below,

1. Random Forest
2. Support Vector Machine
3. Naive Bayes
4. Linear Discriminant Analysis
5. Gradient Boosting
6. Logistic Regression

The implementation has been made so it is easy to add and test new classifiers with the set of partitions. For more information on that, please refer to the repository.

5 Results

Figure 1 shows the overall VIC Score for each of the 50 partitions used. In it we can see that the best partition found was with a separation at rank 81 with an AUC score of 0.8112. Instead of resembling a normal distribution there seems to be several local optimal partitions to the data.

To give ourselves a better idea of how each classifier informed the final VIC Score, Figure 2 presents the average AUC for each classifier and each partition. Random Forest seems to be the most similar followed by Gradient Boosting. The worst performing model was the Support Vector Machine which appears to get stuck making one prediction every time. Naive Bayes presents a peak at 81 but it's not the global optima. The results seem to point out that since there are a lot of variables in the model, classifiers than emphasise attribute selection seem to perform the best for the data set.

Fig. 2: Average AUC per clustering for all classifiers

Figure 3 presents the frequency with which each classifier obtained the highest score for a given partition. In it we can see that the classifier with the most high scores was the Random Forest Classifier with almost 40 wins followed by Gradient Boosting with a little more than 11 wins and Naive Bayes with only 1

win. This results further indicate that the SVM, LDA and Logistic Regression Classifiers may not be suitable for this data set.

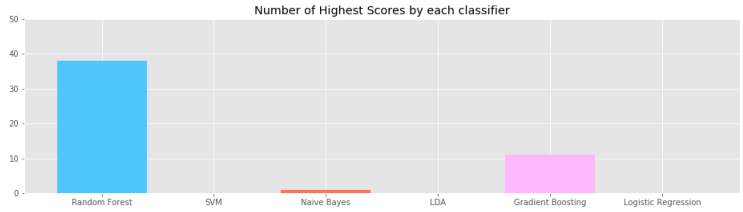

Fig. 3: Frequency of highest score for each classifier

Finally, to compare the obtained results we have tested the partitions with other Clustering scores presented in figure 4. Two out of three clustering scores seem to agree with VIC. Interestingly, the Silhouette Score seems similar to the results provided by the Naive Bayes Classifier in VIC.

6 Conclusions

We have presented a brief experiment to analyze the best partition for a set of ordered data. Rather than depending on some clustering, we used the previously known order (QS ranking) to facilitate analysis by separating the data set into two classes rather than having to describe each element by itself. This will prove useful for further experiments when we try to define the characteristics that make a certain University appear in the highest rankings of publications such as QS.

We have achieved this using VIC and validated our results with other previously known methods for unsupervised cluster validation. We have also seen which type of classifiers work best and to improve our confidence in the results we could either use more classifiers that perform attribute selection or try the parametric models using fewer previously selected attributes.

References

1. Bucheli, E., Velasco, H., Loredo, S.: Cluster-validation-with-vic (2019), gitHub Repository, https://github.com/ebucheli/Cluster-Validation-With-VIC
2. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974). https://doi.org/10.1080/03610927408827101, https://www.tandfonline.com/doi/abs/10.1080/03610927408827101
3. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (April 1979). https://doi.org/10.1109/TPAMI.1979.4766909

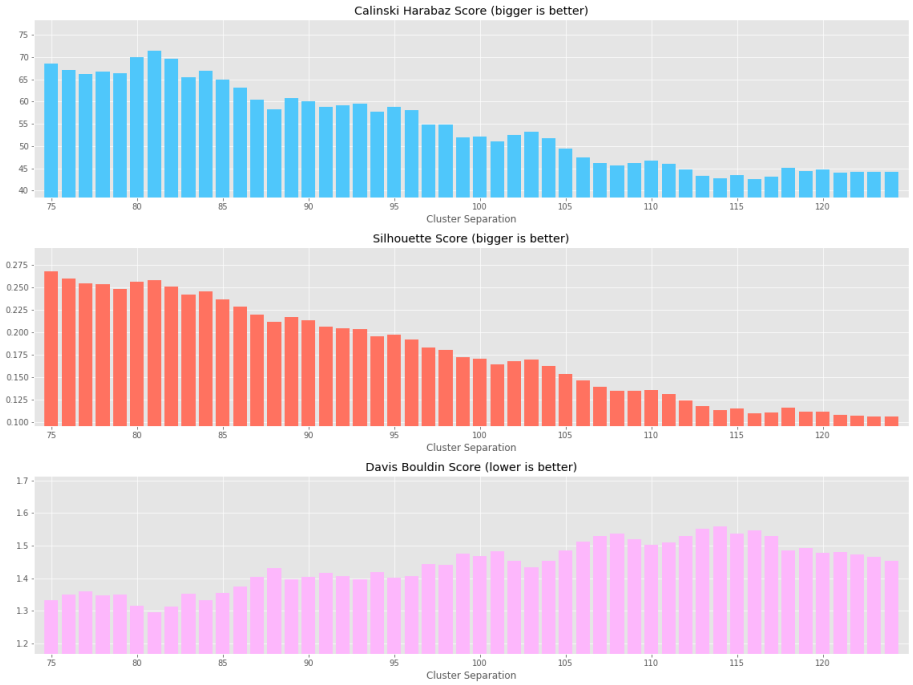

Fig. 4: Results for other clustering metrics

4. Rodriguez, J., Medina-Perez, M.A., Gutierrez-Rodriguez, A.E., Monroy, R., Terashima-Marin, H.: Cluster validation using an ensemble of supervised classifiers. Knowledge-Based Systems **145**, 134–144 (2018), https://0-doi-org.millenium.itesm.mx/10.1016/j.knosys.2018.01.010
5. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53 – 65 (1987). https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7, http://www.sciencedirect.com/science/article/pii/0377042787901257