Graph-based Clustering Index for Topology and Classification Analysis

Jose G. Fernandes * **Frederico Coelho** ** **Antonio P. Braga** **

* *Graduate Program in Electrical Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG (e-mail: josegeraldof@ufmg.br)*
** *Department of Eletronic and Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG (e-mail: fredgfc,apbraga@ufmg.br)*

Abstract: This work proposes a quality index based on a graph of distance relationships, the Gabriel Graph, to analyze the performance of classification models. The results have shown a consistent correlation between the quality index and the model's predictions, both at an individual and global level. These findings emphasize the importance of considering the data space to explain and optimize the performance of classification models.

Keywords: Clustering Index; Supervised Learning; Gabriel Graph; Machine Learning; Local Clustering.

1. INTRODUCTION

In classification problems, assessing the amount of overlapping between opposite classes, resulted either by embedded learning or input space handcrafted feature representation, may help to explain the performance of classifiers and to gain insights into the problem. In fact, the amount of overlapping may be seen as an indirect measure of how difficult it is to fit the data with an inductive learning model. It reflects also how representative are features either learned with embedded mapping or by direct input space representation. Also, regularized models are likely to yield a higher degree of overlapping in feature space in order to trade-off learning set error and model complexity (Geman et al., 1992; Tikhonov and Arsenin, 1977). So, understanding the degree of overlapping may uncover interesting properties of the learned model and of the data itself.

Data set superposition is frequently estimated considering global statistical figures, such as Silhouette (Rousseeuw, 1987), Davies-Bouldin (Davies and Bouldin, 1979) and Calinski-Harabasz (Caliński and Harabasz, 1974). Such measures, however, are global and not sensitive to local properties of the distribution, such as overlapping in the separation region of the data. In this paper, the properties of a planar graph (Gabriel and Sokal, 1969) are explored in order to obtain a quality index that reflects how a sample is related to its graph neighborhood, considering class labels. Such representation differs from global statistical measures since the graph construction rule preserves locality properties of the data, so graph adjacency reflects local spatial relations of the data. For instance, a pattern in the margin region has most of its neighbors also in the margin.

The analysis is accomplished for a collection of tabular datasets from the UCI repository (Dua and Graff, 2017). These datasets cover a wide range of domains and have been widely used in various applications, making them suitable for evaluating the proposed quality index. A greater correlation was observed between the proposed index and the average accuracy of a supervised classification model. This relationship was also verified at the individual level, for each sample, with a negative correlation of misclassification.

2. LITERATURE REVIEW

2.1 Gabriel Graph

The Gabriel Graph (Gabriel and Sokal, 1969) is a construction based on a set of points $\boldsymbol{\mathcal{S}}$, which defines the vertices of the graph, and another set of edges $\boldsymbol{\mathcal{E}}$ such that two points $\boldsymbol{x_i}, \boldsymbol{x_j}$ are adjacent and define an edge if there are no other points in $\boldsymbol{\mathcal{S}}$ within the hypersphere with diameter defined as the distance between the two points $\boldsymbol{x_i}, \boldsymbol{x_j}$, as in Equation 1 and represented schematically in Figure 1.

$$\begin{gathered}(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{\mathcal{E}} \leftrightarrow \\ \delta^2(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq [\delta^2(\boldsymbol{x_i}, \boldsymbol{x_k}) + \delta^2(\boldsymbol{x_j}, \boldsymbol{x_k})] \; \forall \boldsymbol{x_k} \in \mathcal{S}\end{gathered} \tag{1}$$

Where δ is a distance metric used in the construction, commonly, but not restricted, defined as the Euclidean distance.

The construction of this graph is a technique from Computational Geometry and was borrowed for learning problems as a way to express the neighborhood characteristics of the data structure.

2.2 Overlap Filter

Overlap filters based on the Gabriel Graph have been adopted as a form of regularization in the Support Edge Classifier (CLAS) (Torres et al., 2015).

Based on the class difference between the vertex and the samples connected by its edges, a measure Q is defined to represent the quality of that sample, as shown in Equation 2, where V represents the number of edges and

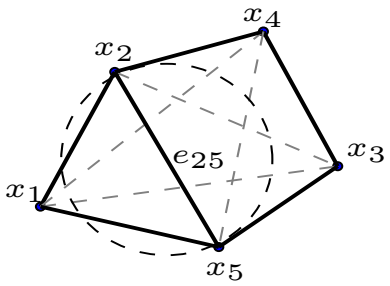

Figure 1. Schematic representation of a Gabriel graph construction. Solid lines represent those edges that are included in the graph.

V_{eq} represents only the edges from samples of the same class.

$$Q(\boldsymbol{x}_i) = \frac{V_{eq}(\boldsymbol{x}_i)}{V(\boldsymbol{x}_i)} \tag{2}$$

Samples in the boundary and overlap region will have their quality affected by the degree of overlapping. Therefore, a simple filter is applied to discard samples with quality below a threshold. A dynamic threshold TC is adopted, representing the average quality of samples from a specific class. The effect of removing noisy patterns in the margin region is to transform the problem into a linearly separable one, so that it becomes treatable by the linear separator. The regularization parameter C, which determines the error tolerance in the margin SVMs Hearst et al. (1998), is given, or estimated with ad-hoc methods such as cross-validation. Similarly, in the graph approach, the vertex degree, which is an implicit property of the Gabriel graph, is adopted in order to eliminate the uncertainties in the margin region. Such graph property, however, does not require a parameter to be set in advance, since it is an inherent property of the graph and the data set structure.

In this work, however, it is of interest to use this indicator, the vertex quality Q, as a way to evaluate the space of a dataset and explain the prediction of a supervised classifier.

2.3 Clustering Indices

For space evaluation, it is common to perform comparative analysis based on classification performance (Fehervari et al., 2019; Han et al., 2015; Moutafis et al., 2016). Although this approach directly aligns with the problem of interest, it is important to consider a large and diverse number of datasets to eliminate any objective bias.

Often, the cleanliness of the data representation space is as important as the final predictor's outcome. In this regard, it is suggested to use clustering quality indices to evaluate this perspective.

Calinski-Harabasz (Caliński and Harabasz, 1974) is an index that quantifies how far the samples are from the cluster centroids, the cohesion variable, concerning the distance of these centroids to the global centroid, the separability variable. Davies-Bouldin (Davies and Bouldin, 1979) is similar in its formulation but differs in the separability variable, where it computes a function of the distances between centroids. Finally, the Silhouette (Rousseeuw, 1987) computes cohesion as the sum of distances from all points within the same cluster, while separability is the sum of distances to the nearest neighboring cluster.

3. METHODOLOGY

3.1 Datasets

To evaluate whether the quality index can explain the performance of classification models, a set of 12 datasets from the UCI repository (Dua and Graff, 2017) is selected for all tests. These datasets are divided into 10 partitions for k-fold cross-validation (Kohavi et al., 1995).

To ensure the reliability of the results, the selected datasets are widely used in similar applications. As preprocessing steps, all attributes are normalized, and categorical variables are converted to numerical values. All datasets are binary classification problems. The following is a description of the selected datasets: *Statlog Australian Credit Approval* (australian); *Breast Cancer Wisconsin* (breastcancer); *Breast Cancer Hess Probes* (breastHess); *Liver Disorders* (bupa); *Climate Model Simulation Crashes* (climate); *Pima Indian Diabetes* (diabetes); *Statlog German Credit Data* (german); *Gene Expression* (golub); *Haberman's Survival* (haberman); *Statlog Heart Disease* (heart); *Indian Liver Patient* (ILPD); *Parkinson's* (parkinsons).

Table 1 presents the main characteristics of this dataset selection. Note the high diversity of the problems to ensure the generalization of the method.

Table 1. Characteristics of the selected datasets.

Dataset	Samples	Features
australian	690	14
breastcancer	683	6
breastHess	133	30
bupa	345	6
climate	540	18
diabetes	768	8
german	1000	24
golub	72	50
haberman	306	3
heart	270	13
ILPD	579	10
parkinsons	195	22

3.2 Classification

The classification problem follows the standard procedure. For each fixed dataset and fold, the model is trained on the training set and tested on the separate test set.

A Support Vector Machine (SVM) with RBF kernel and a Multi-Layer Perceptron (MLP) are applied for classification. The performance is evaluated using the average accuracy with cross-validation across the folds. No hyperparameter tuning was performed as the proposed approach is agnostic to the classification model.

3.3 Quality Index

To leverage the heuristics of the overlap filter in CLAS as a supervised clustering index, we condition Equation 2 on the distance between samples, as shown in Equation 3. Note that $V_{eq}(\boldsymbol{x}_i) = 1 - \sum_j |y_i - y_j|$, but each term in the summation is weighted by a decreasing function of the neighbors' distances.

$$q(\boldsymbol{x}_i) = 1 - \frac{\sum_j |y_i - y_j| \exp(-\delta(\boldsymbol{x}_i, \boldsymbol{x}_j))}{V(\boldsymbol{x}_i)} \quad (3)$$

To evaluate the explanation power of this index q with respect to model performance, the Pearson correlation coefficient is calculated in two separate experiments.

First, a micro analysis is conducted for individual sample prediction. The correlation ρ between the sample's quality index $q(\boldsymbol{x}_i)$ and the classifier's error on that sample $|y_i - f(\boldsymbol{x}_i)|$ is computed. It is expected that a lower quality index corresponds to a sample that is deeply overlapped in the opposite class region, resulting in a higher expected error. The reported result for each dataset d is the average between all folds and all $n(d)$ test samples as in Equation 4.

$$\text{micro}(d) = K^{-1} \sum_{k}^{K} n(d)^{-1} \sum_{i}^{n(d)} \rho(q(\boldsymbol{x}_i), |y_i - f(\boldsymbol{x}_i)|) \quad (4)$$

Second, a macro analysis is performed to assess the dataset globally. The correlation between the average index $\bar{q}$ of the samples in the test fold and its accuracy is calculated, as in Equation 5. Once again, a high average index indicates well-separated classes and an easier classification problem, resulting in higher accuracy.

$$\text{macro}(d) = K^{-1} \sum_{k}^{K} \rho(\bar{q}_k, \text{acc}_k) \quad (5)$$

4. RESULTS AND DISCUSSION

Table 2 presents the results obtained for the average accuracy of SVM and MLP, as well as the average quality index $\bar{q}$ for all datasets.

Table 2. Accuracy of the models and average quality index for each dataset.

Dataset	SVM	MLP	$\bar{q}$
australian	0.85	0.85	0.90
breastcancer	0.97	0.97	0.98
breastHess	0.78	0.77	0.89
bupa	0.69	0.73	0.65
climate	0.93	0.95	0.99
diabetes	0.76	0.75	0.77
german	0.77	0.73	0.94
golub	0.74	0.69	0.94
haberman	0.73	0.73	0.68
heart	0.81	0.92	0.92
ILPD	0.72	0.72	0.72
parkinsons	0.89	0.93	0.93

There is a wide range of performance observed across these datasets, indicated by the index as well, with some datasets being more separable than others. Additionally, the models performed similarly.

To investigate the correlation between these variables, Tables 3 and 4 show the results of the correlation analysis in both micro and macro modes. In the micro analysis, the correlation between the sample's quality index and the corresponding prediction error is reported. In the macro analysis, the correlation between the average index of the test folds and their accuracies is examined.

Table 3. Correlations calculated for individual samples and globally based on SVM predictions.

Dataset	Micro	Macro
australian	−0.64	0.86
breastcancer	−0.80	0.29
breastHess	−0.77	0.57
bupa	−0.52	0.46
climate	−0.86	0.86
diabetes	−0.69	0.79
german	−0.58	0.52
golub	−0.80	0.77
haberman	−0.65	0.38
heart	−0.66	0.71
ILPD	−0.51	0.69
parkinsons	−0.63	0.34
Mean	−0.68	0.60

Table 4. Correlations calculated for individual samples and globally based on MLP predictions.

Dataset	Micro	Macro
australian	−0.64	0.83
breastcancer	−0.78	0.43
breastHess	−0.65	0.86
bupa	−0.44	−0.03
climate	−0.60	0.38
diabetes	−0.65	0.84
german	−0.50	0.55
golub	−0.74	0.74
haberman	−0.66	0.71
heart	−0.65	0.58
ILPD	−0.54	0.44
parkinsons	−0.53	0.55
Mean	−0.62	0.57

Confirming the hypothesis, a consistent negative correlation was observed between the individual quality index and the classifier's error. These samples have a neighborhood with samples from a different class, which makes it difficult to model a separation surface in that region.

In the macro analysis, a consistent positive correlation was also observed, except for one experiment with the bupa dataset and the MLP model, which had a correlation close to zero. Since this phenomenon was not observed in the SVM experiment, it is possible that this can be explained by a training failure. It is worth noting that the bupa dataset has the lowest average quality index among the compiled datasets.

The same experiments were also carried out with the clustering indices for comparison purposes. Table 5 shows the average score for all folds in cross validation.

The DB and CH indices do not have an individualized score per sample, for this reason the micro analysis was performed only with the Silhouette. Tables 6 and 7 show the results of the macro correlation analysis for SVM and MLP models.

The proposed index showed a greater correlation with accuracy in most datasets, with a large difference in the average correlation between these. Note that the DB index has a negative correlation as it is reversed, a lower value means that the clustering is better.

Regarding the micro analysis only the Silhouette index was evaluated. Table 8 shows the results for both models.

Table 5. Average clustering index for each dataset.

Dataset	silh	DB	CH
australian	0.18	1.99	16.22
breastcancer	0.57	0.78	92.86
breastHess	0.06	1.49	2.54
bupa	0.01	5.09	1.38
climate	0.02	2.82	1.63
diabetes	0.10	3.12	6.77
german	0.03	4.90	3.35
golub	0.15	1.34	2.53
haberman	0.00	5.61	0.88
heart	0.12	2.38	4.64
ILPD	0.01	5.57	1.40
parkinsons	0.14	1.62	4.10

Table 6. Global correlations calculated based on SVM predictions for each clustering index.

Dataset	silh	DB	CH
australian	0.87	−0.87	0.84
breastcancer	0.12	−0.17	0.10
breastHess	0.69	−0.39	0.37
bupa	0.02	−0.67	0.42
climate	−0.12	−0.65	−0.66
diabetes	0.76	−0.52	0.50
german	0.33	−0.80	0.42
golub	0.53	−0.48	0.55
haberman	0.64	0.13	0.20
heart	0.81	−0.75	0.81
ILPD	−0.00	−0.66	0.18
parkinsons	0.44	−0.62	0.03
Mean	0.42	−0.54	0.31

Table 7. Global correlations calculated based on MLP predictions for each clustering index.

Dataset	silh	DB	CH
australian	0.62	−0.63	0.58
breastcancer	0.37	−0.63	0.62
breastHess	0.48	−0.62	0.35
bupa	0.00	−0.48	0.09
climate	−0.53	−0.28	−0.33
diabetes	0.52	−0.42	0.47
german	0.30	−0.58	0.39
golub	0.76	−0.60	0.82
haberman	0.14	0.16	−0.11
heart	0.71	−0.66	0.68
ILPD	−0.12	−0.55	0.06
parkinsons	−0.48	0.25	0.02
Mean	0.23	−0.42	0.30

Again, the negative correlation was lower than with the proposed index.

Table 8. Correlations calculated for individual samples from Silhoute score.

Dataset	SVM	MLP
australian	−0.82	−0.65
breastcancer	−0.40	−0.44
breastHess	−0.24	−0.27
bupa	−0.14	−0.19
climate	0.20	−0.07
diabetes	−0.72	−0.58
german	−0.44	−0.37
golub	−0.84	−0.71
haberman	−0.47	−0.58
heart	−0.75	−0.53
ILPD	−0.04	−0.30
parkinsons	−0.26	−0.25
Mean	−0.41	−0.41

In fact, these results are expected since the clustering indices are proposed to evaluate a Gaussian cohesion of point clouds, and do not deal with non-linear margins in space. The quality index from the graph is able to ignore this restriction as it evaluates essentially local relations of the points.

5. CONCLUSION

This study presented a quality index for clustering based on a graph of distance relations, Gabriel Graph, in an attempt to explain the performance of classification models. The analysis evaluated both individual aspects, from samples, and global aspects, from datasets.

A sample with a low quality index has its neighborhood populated by samples from a different class and is likely to have high overlap with that class region. As a result, it is less likely that the model has extended its separation surface to include that sample, resulting in a classification error.

The global analysis assesses the separability of the clouds of samples from opposing classes in the space defined by the distance metric. The greater the overlap between these clouds relative to the overall set, the lower the average quality index of that set. This overlap also poses a challenge for any classification model.

A consistent correlation was observed between the quality index and the model's prediction both individually and globally. This correlation is more significant with the proposed index compared to other clustering indices in literature.

Overlapping in embedded projections may also reflect smoothness of the output function and convey the effects of regularization. Considering the quality index as an objective function in further works may also pave the way to representation learning according to a prior adjacency matrix, not necessarily given by a graph structure.

ACKNOWLEDGMENTS

This work was made possible by the availability of the datasets provided by the UCI Machine Learning Repository Dua and Graff (2017).

REFERENCES

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1–27.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

Dua, D. and Graff, C. (2017). UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Fehervari, I., Ravichandran, A., and Appalaraju, S. (2019). Unbiased evaluation of deep metric learning algorithms. *arXiv preprint arXiv:1911.12528*.

Gabriel, K.R. and Sokal, R.R. (1969). A new statistical approach to geographic variation analysis. *Systematic zoology*, 18(3), 259–278.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1), 1–58.

Han, X., Leung, T., Jia, Y., Sukthankar, R., and Berg, A.C. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3279–3286.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 18–28.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, 1137–1145. Montreal, Canada.

Moutafis, P., Leng, M., and Kakadiaris, I.A. (2016). An overview and empirical comparison of distance metric learning methods. *IEEE transactions on cybernetics*, 47(3), 612–625.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.

Tikhonov, A.N. and Arsenin, V.I. (1977). *Solutions of Ill-posed Problems: Andrey N. Tikhonov and Vasiliy Y. Arsenin. Translation Editor Fritz John*. Wiley.

Torres, L., Castro, C., Coelho, F., Sill Torres, F., and Braga, A. (2015). Distance-based large margin classifier suitable for integrated circuit implementation. *Electronics Letters*, 51(24), 1967–1969.