# Complex Network Measures for Data Classification

Murillo G. Carneiro, Barbara C. Gama and Otavio S. Ribeiro
*Faculty of Computing*
*Federal University of Uberlândia*
Uberlândia, Brazil
{mgcarneiro, barbaracgama, otaviosoares}@ufu.br

***Abstract*—Complex networks have become an increasingly relevant research topic in machine learning, with many learning systems in the literature successfully exploring complex network properties and measures. In data classification, the use of complex networks allows the detection of structural and topological patterns related, for example, to the formation pattern of the input data. Some measures of complex networks have already been used in this sense. However, a systematic study capable of characterizing such measures in the context of data classification is lacking in the literature. In this work, we evaluate comparatively the predictive performance of some measures. Specifically, eight complex network measures were selected from the literature, namely: assortativity, average local clustering coefficient, average degree, betweenness, average shortest path length, closeness, global clustering coefficient and eigenvector centrality. For our analyses, both artificial and real-world data sets were considered. The results show that measures such as average shortest path and assortativity, besides presenting high predictive capability, are also more robust to the variation of the network structure. In summary, this research paves a way to support other related works in selecting more appropriate complex network measures for data classification.**

***Index Terms*—Complex networks, data classification, network measures, high-level classification, network science.**

## I. Introduction

Although data classification is a well-known machine learning task, some studies have shown that several classification techniques, such as decision trees, neural networks, support vector machines, etc., have troubles in detecting semantic patterns of the input data, like the pattern formation [1]–[3], for example. The reason pointed out by these studies is that such techniques perform data classification based only on the physical attributes of the data (e.g., distance, similarity or distribution). By the contrary, the analysis of structural and topological properties of the data has been demonstrated to be an efficient tool to uncover such patterns, especially when considering measures of complex networks [4]–[7].

Learning methods based on complex networks have been the subject of recent studies related especially to two major topics: the construction of the network representation from feature vector data, and the exploration of complex network measures to take advantage of the structural, dynamical and topological properties of the data. Regarding the construction, most studies build up the graph by using methods based on k-nearest neighbors heuristics, while others use density-based methods, like the $\epsilon$-radius neighborhood, or even the combination of both heuristics [8]–[13]. On the other hand, there are also works focused on the construction of optimized graphs, which are capable of achieve a better predictive performance [14], [15], although demanding a higher computational cost.

Complex network measures have been explored in different ways by current classification techniques. Consequently, new data classification concepts emerged, such as pattern compliance and characterization of importance. The former, originally proposed in [1], aims to classify each test item into the network class in which its insertion causes the least variation of the network measures. The latter, originally proposed in [3], considers an importance score derived from the network connectivity patterns in order to assign to the test item the class label to which it gets the highest importance score.

Despite the literature contains an uncountable number of network measures, very few have been investigated in the context of data classification. Examples include assortativity, average local clustering coefficient, average degree, closeness, eigenvector centrality and pagerank [1]–[3], [15]–[18]. In addition, to the best of our knowledge, there is no study in the literature which evaluates comparatively the predictive performance of such measures in separate, which means that the probable contribution of each network measure to the overall classification process is indefinite. To address that, we select relevant and potential network measures taking into account the data classification context, adapt the technique based on pattern compliance to analyze the measures individually, and design an experimental setup to evaluate the network measures in both artificial and real data sets. Thereby, the hypothesis investigated in this work states that certain network measures present better predictive accuracy than others.

In a few words, the proposed technique builds up a network from feature vector data using several graph construction methods based on the k-nearest neighbors and $\epsilon$-radius neighborhood heuristics [11]. A network measure is then calculated over each network subgraph (corresponding to a given class). In the following, each test item is virtually inserted into the network, the measure is recalculated, and the subgraph label in which the insertion caused the least variation of the measure is assigned to the test item. To be specific, eight network measures are evaluated in this paper, namely: assortativity, average local clustering coefficient, average degree, betweenness, average shortest path, closeness, global clustering coefficient and eigenvector centrality.

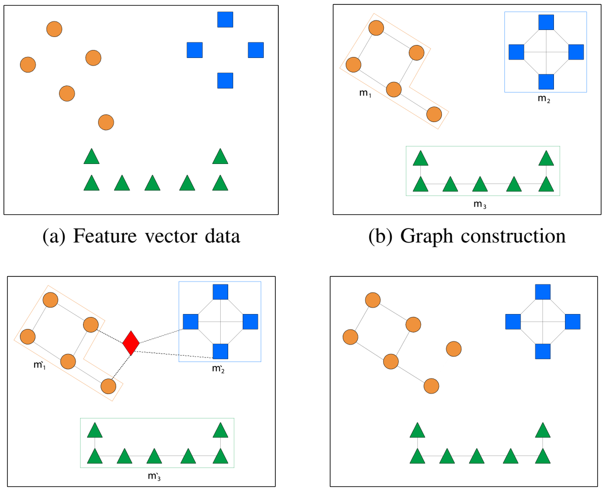
Thanks to the Brazilian National Council for Scientific and Technological Development - CNPq (grant number 439556/2018-0).

In relation to the experiments, we evaluate the network measures in terms of predictive capability and robustness considering artificial data sets with low and high noise levels as well as eight real-world data sets. The results pointed out the average shortest path length as the most accurate and robust measure, able to achieve better predictive performance in data sets with higher noise levels and also less influenced by the variation of the graph structure. They also suggested that the average degree, which is a commonly adopted measure, is neither robust nor achieve top results in any data set.

The remainder of this paper is organized as follows. Sect. II describes the high-level classification model; Sect. III presents the results obtained by the network measures in artificial and real data sets; Sect. IV shows the application of the measures recommended by our study to improve the performance of traditional classifiers; and Sect. V ends the paper.

## II. Model Description

The evaluation of the network measures is conducted through of the pattern compliance technique proposed in [1]. As illustrated by Fig. 1, the technique is composed by two major phases. In the training phase, a graph construction method is responsible to represent the training data (usually in the form of feature vector) into a network, in which the structural patterns associated to each network class, countoured in Fig. 1(b), are then calculated by a given measure $m$. In the test phase, a new data item is virtually inserted into the network and classified where its insertion causes the smallest variation of such a network measure. In the following we describe in detail each one of these phases as well as the graph construction heuristics and the network measures considered in this study.



(a) Feature vector data

(b) Graph construction

(c) Analysis of measure variation

(d) Classification

Fig. 1: Overview about the classification model based on complex network measures.

### A. Training Phase

Let $\mathcal{X}$ denote a set of training data, in which each data item is represented by $(\mathbf{x}_i, y_i)$, $i = \{1, \ldots, n\}$, with $\mathbf{x}_i = \{x_1, \ldots, x_d\}$ representing a d-dimensional vector of features and $y_i \in \mathcal{Y} = \{1, \ldots, L\}$ a class label associated with the i-th data item. A supervised data classification task aims to learn a mapping function $f\colon \mathbf{x} \to y$ in order to associate a class label to any new (test) item $\mathbf{x}_u \notin \mathcal{X}$ in which $y_u$ is unknown.

Particularly, as previous works in the literature, we assume that the network representation may reveal high-level patterns about the input data which can be uncovered by complex network measures. As most data sets are in the feature vector form, we provide an additional step to convert such data into networks. Therefore, our network-based classifier is given by $f\colon g(\mathbf{x}) \to y$, where $g(\cdot)$ is a function to build up a network $\mathcal{G} = \{V, E\}$ from the feature vector data. In such a network, each labeled data item $\mathbf{x}_i$ is associated to a vertex $v_i \in V$ and each edge $e_{ij} \in E$ refers to a link between vertices $v_i$ and $v_j$ based on an affinity criterion derived from a graph construction method. In this paper, we assume $\mathcal{G}$ is an undirected and unweighted graph with neither multiple edges nor self-loops.

Here we consider the following graph construction methods as the function $g$: k-nearest neighbors (kNN), selective kNN (S-kNN), mutual kNN (MkNN), selective MkNN (S-MkNN) as well as their combinations with $\epsilon$-radius neighborhood ($\epsilon$N). All methods based on k-nearest neighbors consider as affinity criterion the proximity of the data items and their class label. In common, there is an edge $e_{ij}$ if both vertices $v_i$ and $v_j$ have the same class label and a particular k-nearest neighbors relationship. In the kNN and S-kNN graphs, this relationship means that $v_i$ belongs to the k-nearest neighbors of $v_j$ or vice-versa [11], while in MkNN and S-MkNN such a relationship is required to be mutual, which means that $v_i$ and $v_j$ must belong to the k-nearest neighbors of each other in order to create a link [10], [14]. On the other hand, the difference between the selective and non-selective methods lies on the treatment of pairs of vertices which satisfy a given k-nearest neighbor relationship, but have different class labels. In such cases, the kNN and MkNN graphs simply do not create that connection, while the S-kNN and S-MkNN avoid such a situation by looking for the nearest neighbors only among vertices of the same class [11], [14]. Regarding the $\epsilon$N graph, it is a density-based method which creates an edge $e_{ij}$ if both vertices $v_i$ and $v_j$ have the same class label and are close at a distance less than or equal to $\epsilon$ [11]. Notice that its combination with any other nearest neighbors methods (e.g., kNN+$\epsilon$N, S-kNN+$\epsilon$N, MkNN+$\epsilon$N and S-MkNN+$\epsilon$N) is nothing more than the maximization of both adjacency matrices.

With the graph at hand, the patterns associated to each class $l \in \mathcal{Y}$ are calculated by a network measure $m$ over each subgraph $\mathcal{G}^{(l)} \subset \mathcal{G}$, i.e., $m(\mathcal{G}^{(l)})$. In this paper, we evaluate eight network measures which are further presented.

### B. Test Phase

In this phase, a given test item $\mathbf{x}_u$ is mapped as a vertex $v_u$ which is virtually connected, according to the graph construction heuristic, to its neighbors vertices of the training data. In the following, for each subgraph $\mathcal{G}^{(l)} \in \mathcal{G}$ in which $v_u$ has been connected, denoted by $\mathcal{G}^{(l)}_{\mathbf{x}_u}$, we calculate the variation of the network measure $m$ as follows:

$$\Delta\mathcal{G}^{(l)}_{\mathbf{x}_u} = \left[ \frac{\left| m(\mathcal{G}^{(l)}) - m(\mathcal{G}^{(l)}_{\mathbf{x}_u}) \right|}{\sum_{c \in \mathcal{Y}'} \left| m(\mathcal{G}^{(c)}) - m(\mathcal{G}^{(c)}_{\mathbf{x}_u}) \right|} \right]^{-1} \quad , \qquad (1)$$

in which $m(\mathcal{G}^{(l)})$ and $m(\mathcal{G}^{(l)}_{\mathbf{x}_u})$ denote the results of applying a given network measure $m$ respectively before and after the insertion of the test item; the denominator term is for normalization and considers the variations obtained to every subgraph (class label) in which $v_u$ was virtually connected (denoted by $\mathcal{Y}'$); and the inverse function is to change the magnitude order in a way that small variations receive high scores and big variations low ones. Notice that each subgraph examined by our formulation corresponds to a unique and separate class label, and that $c, l \in \mathcal{Y}'$. Otherwise, in the case of $v_u$ has not been connected with any vertex of a given subgraph, the probability associated to the classification of $\mathbf{x}_u$ in that corresponding class label is automatically defined as zero. Intuitively, this means that the test item does not conform to the pattern related to that class.

In the following, the measure variation is post-processed in order to take into account imbalanced classes:

$$\mathcal{H}^{(l)}_{\mathbf{x}_u} = \frac{\Delta\mathcal{G}^{(l)}_{\mathbf{x}_u} p^{(l)}}{\sum_{c \in \mathcal{Y}'} \Delta\mathcal{G}^{(c)}_{\mathbf{x}_u} p^{(c)}} \ , \qquad (2)$$

where $p^{(l)}$ means the proportion of data items belonging to the class $l \in \mathcal{Y}'$ in the whole training data, and the denominator is for normalization matters.

*1) High-level classification:* The classification based only in the associations provided by the complex network measures is performed by assigning to the test item $\mathbf{x}_u$ the label with highest probability score, i.e., in which its insertion caused the smallest variation of the network measure:

$$f(\mathbf{x}_u) = \underset{l \in \mathcal{Y}}{\arg\max}\ \mathcal{H}^{(l)}_{\mathbf{x}_u} \ . \qquad (3)$$

Notice that such a formulation is employed in most experiments of this paper in order to provide a soundness comparative evaluation about the predictive power of each network measure investigated.

*2) Hybrid classication:* We also evaluate the combination of the high-level associations provided from complex networks with those produced by traditional classification techniques. Such techniques are known to provide low-level associations as they perform data classification essentially based on the physical features of the data (e.g., proximity or distribution). In this sense, some works have evidenced the ability of the network measures to improve the predictive performance of such techniques by also taking into account topological and structural features of the networked data. Let $\mathcal{C}$ be the associations provided by a given low-level classifier (e.g., naive bayes or support vector machine), the combination with the high-level classifier is given by:

$$\mathcal{M}^{(l)}_{\mathbf{x}_u} = \lambda\mathcal{H}^{(l)}_{\mathbf{x}_u} + (1-\lambda)\mathcal{C}^{(l)}_{\mathbf{x}_u} \ , \qquad (4)$$

in which $\lambda$ is a parameter that represents the linear convex combination between the high and low-level associations. The final classification uses the same procedure as denoted in (3), but now changing the term $\mathcal{H}^{(l)}_{\mathbf{x}_u}$ by $\mathcal{M}^{(l)}_{\mathbf{x}_u}$.

### C. Complex Network Measures

In the high-level classification, structural and topological information are captured from the networked data through of complex network measures. Despite the literature contains a wide range of network measures, very few have been investigated in such a context. The network measures evaluated in this study are listed by Tab. I. In the table, the column "Refs." indicates related works that employed the corresponding measure for data classification.

## III. Experimental Results

In this section we present the experimental results obtained by evaluating the eight network measures over real and artificial data sets. In the following experiments, we adopt a 10-fold stratified cross-validation process averaged over three executions, totaling 30 runs. Graph methods based on the k-nearest neighbors heuristic have their parameter $k$ selected over the following range $\{1, 2, \ldots, 15\}$, while those based on the $\epsilon$-radius neighborhood have their parameter $\epsilon$ selected over the following range $\{0.1\overline{d}, 0.2\overline{d}, \ldots, 0.5\overline{d}\}$, with $\overline{d}$ representing the average distance among the input data. We also use the Euclidean distance as the dissimilarity metric for every graph method.

### A. Results on real-world data

Table II presents the real-world data sets considered here, which are available in [25]. The selection was made to encompass diversity from domain as well as number of samples, features and classes. We evaluate the predictive performance of the network measures in terms of predictive capability and robustness. With the former, we analyze the potential of the network measures to detect patterns through of the networked data. By analyzing the predictive capability we are interested in answering "how good a given network measure may be in the classification task?". With the latter, we investigate the robustness of the network measures in relation to the variation of the graph structure. Thus, we are looking on to answer "how straightforward is to achieve that predictive capability?".

*1) Predictive capability analysis:* Table III shows the best predictive accuracy obtained by each network measure over the eight data sets using the graph construction methods presented before. In the table, boldfaced results indicate the top predictive accuracy achieved for a given data set and underlined ones indicate the graph construction method which provided better results for each network measure. Notice that the reference performance (with $k = 1$) is presented in the "Baseline" column and, with exception of the Dig. data set, the network measures are able to improve the baseline in any other data set, with such an improvement varying from 1% (Bal.) to 7.4% (Eco.) of accuracy. Interestingly, five different

TABLE I: Complex network measures evaluated in this study.

| Abbrev | Network Measure | Definition | Refs. |
|---|---|---|---|
| ASS | Assortativity [19] | $\frac{\lvert E\rvert^{-1}\sum_u i_u - [\lvert E\rvert^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}{\lvert E\rvert^{-1}\sum_u \frac{1}{2}(i_u^2 + k_u^2) - [\lvert E\rvert^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}$ | [1], [15]–[18] |
| ALC | Avg. Local Clustering Coefficient [20] | $\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert e_{us}\rvert}{N_i\,(N_i - 1)}$ | [1], [2], [15]–[18] |
| ADG | Avg. Degree | $\frac{1}{n}\sum_{i=1}^{n} k_i$ | [1], [2], [16]–[18] |
| BET | Betweenness [21] | $\frac{1}{n}\sum_{i=1}^{n}\sum_{u,v\,\in\,V-\{i\}}\frac{\eta_{uv}^{i}}{\eta_{uv}}$ | - |
| ASP | Avg. Shortest Path Length [4] | $\frac{1}{n(n-1)}\sum_{i\neq j} d(i,j)$ | [17] |
| CLO | Closeness [22] | $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-1}{\sum_{j=1}^{n} d(i,j)}\right)$ | [2], [15] |
| GCC | Global Clustering Coefficient [23] | $\frac{\text{number of closed triplets}}{\text{number of open and closed triplets}}$ | [17] |
| ECE | Eigenvector Centrality [24] | $Ax = \lambda x, \qquad \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j$ | [2] |

TABLE II: Meta-description of the real data sets in terms of number of instances, features and classes.

| Abbrev. | Name | #Inst. | #Feat. | #Class |
|---|---|---|---|---|
| App. | Appendicitis | 106 | 7 | 2 |
| Bal. | Balance | 625 | 4 | 3 |
| Dig. | Digits | 5620 | 64 | 10 |
| Eco. | Ecoli | 336 | 7 | 8 |
| Gla. | Glass | 214 | 9 | 7 |
| Iris | Iris | 150 | 4 | 3 |
| Son. | Sonar | 208 | 4 | 3 |
| Thy. | Thyroid | 215 | 5 | 3 |

network measures and eight different graph construction methods achieved the best predictive results in at least one data set, which emphasizes the ability of complex network measures to capture a considerable range of structural and topological properties of the data. In the table, one can see that the network measures CLO and ASP achieved the best results to three data sets each, and ASS was able to outperform the baseline in seven of eight data sets. Also in the table, the S-kNN+$\epsilon$N graph contributed in three of the best results. Regarding the configuration of network measure and graph construction, ASS returned five of its best results with a kNN graph, GCC did the same with the S-kNN one, and ASP returned six and five of its best results respectively with S-kNN+$\epsilon$N and S-kNN.

In order to analyze statistically the predictive capability of the network measures, we conduct the Wilcoxon Signed Ranks test [26] by comparing the accuracy results of any pair of (measure, graph) against each other. As our simulations have eight network measures and eight graph methods, we have a total of 64 models, thus 63 comparisons for each model. At a confidence level of 95% ($\alpha = 0.05$), Tables V and VI present respectively the top 10 models with more favorable ("wins") and unfavorable ("losses") significant differences. In Table V, one can see five different network measures, with CLO occupying five positions in the rank; in addition, there are seven of eight graph construction methods, emphasizing the importance of considering several graphs in our experiments. Otherwise, six network measures are present in Table VI, all them working on graphs composed by the $\epsilon$N method. A probable reason of such a drawback in the $\epsilon$N-based methods can be seen in the results of some data sets in Table III (e.g, Gla.), in which the predictive accuracy of the network measures is largely decreased (worse than the baseline). This suggest that such $\epsilon$N-based methods are missing out local and relevant topological features.

*2) Robustness analysis:* In the following we analyze the robustness of the network measures regarding the variation of the graph structure. This is a relevant analysis as it may reveals how sensible (or not) a given measure is in function of the parameter choice. Table IV shows the predictive performance of the network measures as an average of their results for each value of $k$ considering the graph construction methods. As one can see, ASP has a notable performance here, much better than any other network measure. The measure achieves the best results for the eight data sets under analysis. In the case of "Eco." data set, for example, ASP provided better averaged results than the best results of any other measure presented in Tab. III. Interestingly, ASP and GCC obtained their best results always with the S-kNN or S-kNN+$\epsilon$N graphs. Otherwise, ADG prefers the MkNN or M-kNN+$\epsilon$N graphs.

We conduct the Wilcoxon test in order to analyze statistically the results in terms of robustness. Taking a confidence level of 95% again, Tables VII and VIII present respectively the top 10 models with more favorable ("wins") and unfavorable ("losses") significant differences. In Table VII, one can see three different network measures (ASP, ASS and ALC), with ASP achieving an outstanding number of 62 wins. Such a result emphasizes the salient features of that network measure which, besides the high predictive capability (boldfaced in Table V), has a robust performance (boldfaced in Table VII).

TABLE III: Predictive capability of the network measures in function of different graph construction methods. Best local results are underlined and best global results are boldfaced.

| Data | Graph | ASS | ALC | ADG | BET | ASP | CLO | GCC | ECE | Baseline |
|---|---|---|---|---|---|---|---|---|---|---|
| App. | kNN | 82.7±5.2 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 80.6±7.6 | 79.6±7.6 | **83.4±5.0** | 79.6±7.6 |
| | kNN+$\epsilon$N | 82.7±5.2 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 80.6±7.6 | 79.6±7.6 | **83.4±5.0** | |
| | S-kNN | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | **83.4±7.8** | 82.1±6.9 | 79.6±7.6 | 82.1±6.6 | |
| | S-kNN+$\epsilon$N | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | **83.4±7.8** | 82.1±6.9 | 79.6±7.6 | 82.1±6.6 | |
| | MkNN | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 80.9±6.1 | |
| | MkNN+$\epsilon$N | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 81.5±5.8 | |
| | S-MkNN | 79.6±7.6 | 80.8±7.4 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | |
| | S-MkNN+$\epsilon$N | 80.8±6.8 | 80.2±7.7 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | 82.7±4.6 | 79.6±7.6 | 79.6±7.6 | |
| Bal. | kNN | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 95.8±2.0 | 94.9±2.3 |
| | kNN+$\epsilon$N | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.9±2.2 | |
| | S-kNN | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 95.6±2.4 | 94.9±2.3 | |
| | S-kNN+$\epsilon$N | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.7±2.3 | 95.0±2.3 | |
| | MkNN | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 96.1±2.1 | 94.9±2.3 | 94.9±2.3 | |
| | MkNN+$\epsilon$N | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | **96.2±2.2** | 95.0±2.3 | 95.0±2.3 | |
| | S-MkNN | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 94.9±2.3 | 96.1±2.1 | 94.9±2.3 | 94.9±2.3 | |
| | S-MkNN+$\epsilon$N | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | 95.0±2.3 | **96.2±2.1** | 95.0±2.3 | 95.0±2.3 | |
| Dig. | kNN | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 |
| | kNN+$\epsilon$N | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | S-kNN | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | S-kNN+$\epsilon$N | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | MkNN | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | MkNN+$\epsilon$N | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | S-MkNN | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| | S-MkNN+$\epsilon$N | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | 98.8±0.3 | |
| Eco. | kNN | 82.9±5.0 | 81.3±5.3 | 80.6±4.6 | 80.6±4.6 | 80.6±4.6 | 84.4±5.7 | 80.6±4.6 | 83.7±5.5 | 80.6±4.6 |
| | kNN+$\epsilon$N | 82.9±5.0 | 81.3±5.3 | 80.6±4.6 | 80.6±4.6 | 80.6±4.6 | 84.4±5.7 | 80.6±4.6 | 83.7±5.5 | |
| | S-kNN | 80.6±4.6 | 82.7±6.0 | 82.2±5.4 | 84.3±5.9 | **88.0±4.9** | 84.1±5.2 | 81.0±5.4 | 82.5±5.5 | |
| | S-kNN+$\epsilon$N | 80.6±4.6 | 82.7±6.0 | 82.2±5.4 | 84.3±5.9 | **88.0±4.9** | 84.1±5.2 | 81.0±5.4 | 82.5±5.5 | |
| | MkNN | 81.1±5.8 | 80.7±4.7 | 81.5±4.0 | 80.6±4.6 | 81.4±5.6 | 82.8±4.9 | 80.9±5.2 | 82.1±5.8 | |
| | MkNN+$\epsilon$N | 81.1±5.8 | 80.7±4.7 | 81.5±4.0 | 80.6±4.6 | 81.4±5.6 | 82.8±4.9 | 80.9±5.2 | 81.2±5.9 | |
| | S-MkNN | 82.8±6.2 | 80.7±4.9 | 82.2±5.4 | 81.3±5.9 | 82.0±5.0 | 82.8±4.9 | 81.8±5.5 | 82.9±5.5 | |
| | S-MkNN+$\epsilon$N | 82.8±6.2 | 80.7±4.9 | 82.2±5.4 | 81.3±5.9 | 82.0±5.0 | 82.8±4.9 | 81.8±5.5 | 83.5±5.2 | |
| Gla. | kNN | 72.9±9.6 | 72.9±9.6 | 73.1±10.7 | 72.9±9.6 | 72.9±9.6 | 75.2±10.3 | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 |
| | kNN+$\epsilon$N | 71.6±10.1 | 65.8±11.0 | 66.9±10.0 | 66.3±10.7 | 68.0±11.1 | 65.9±9.7 | 67.5±10.5 | 65.3±11.1 | |
| | S-kNN | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 | 73.1±10.8 | 73.5±11.2 | 72.9±9.6 | |
| | S-kNN+$\epsilon$N | 72.3±10.4 | 66.0±12.9 | 65.3±9.8 | 67.2±10.9 | 68.6±10.8 | 65.8±9.7 | 68.3±11.3 | 67.8±9.9 | |
| | MkNN | 73.3±10.7 | 73.7±9.7 | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 | 75.2±9.7 | 73.5±10.6 | 73.2±9.1 | |
| | MkNN+$\epsilon$N | 70.2±10.9 | 66.1±10.5 | 67.1±9.5 | 65.8±11.0 | 67.6±11.2 | 67.0±9.6 | 66.9±10.7 | 65.7±9.5 | |
| | S-MkNN | 73.1±11.6 | 74.1±10.2 | 72.9±9.6 | 72.9±9.6 | 72.9±9.6 | **75.3±9.6** | 73.5±9.6 | 72.9±9.6 | |
| | S-MkNN+$\epsilon$N | 71.1±10.9 | 69.5±12.2 | 66.1±10.3 | 65.7±11.4 | 69.2±10.7 | 66.7±9.5 | 68.6±10.8 | 67.8±9.2 | |
| Iris | kNN | **97.8±4.0** | 96.2±4.8 | 96.7±4.5 | 97.6±4.0 | 96.7±5.4 | 97.3±4.4 | 96.2±4.8 | **97.8±4.0** | 96.2±4.8 |
| | kNN+$\epsilon$N | 96.2±4.8 | 94.9±5.1 | 96.4±4.8 | 95.3±5.7 | 96.7±3.8 | 97.1±4.1 | 94.4±5.7 | 95.3±4.9 | |
| | S-kNN | 96.7±4.5 | 96.2±4.8 | 96.2±4.8 | 97.6±4.0 | 97.3±3.7 | 96.2±4.8 | 96.9±4.5 | 96.7±4.5 | |
| | S-kNN+$\epsilon$N | 94.7±5.6 | 95.6±5.0 | 96.2±5.1 | 96.9±4.8 | 97.3±3.7 | 95.3±5.7 | 95.6±5.3 | 95.6±5.0 | |
| | MkNN | 96.7±4.1 | 96.2±4.8 | 96.2±4.8 | 96.2±4.8 | 96.2±4.8 | **97.8±4.0** | 96.2±4.8 | 96.9±4.5 | |
| | MkNN+$\epsilon$N | 96.7±4.1 | 94.2±5.6 | 97.1±4.1 | 95.3±4.9 | 96.2±4.1 | 97.1±4.1 | 94.2±5.6 | 95.3±4.9 | |
| | S-MkNN | 96.2±4.8 | 96.2±4.8 | 96.2±4.8 | 96.2±4.8 | 96.2±4.8 | **97.8±4.0** | 96.2±4.8 | 96.7±4.1 | |
| | S-MkNN+$\epsilon$N | 96.7±4.1 | 96.2±4.4 | 97.1±4.1 | 96.0±4.7 | 97.1±3.3 | 96.9±4.5 | 94.4±5.5 | 95.3±4.9 | |
| Son. | kNN | 83.2±6.7 | 82.8±6.1 | 83.8±5.8 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 |
| | kNN+$\epsilon$N | 83.2±6.7 | 82.8±6.1 | 83.8±5.8 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | |
| | S-kNN | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 83.0±6.5 | 82.8±6.1 | 84.4±5.5 | 83.0±5.9 | |
| | S-kNN+$\epsilon$N | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 82.8±6.1 | 83.0±6.5 | 82.8±6.1 | 84.4±5.5 | 83.1±6.3 | |
| | MkNN | 82.8±6.1 | 83.6±6.2 | 82.8±6.1 | 83.6±6.6 | 83.0±6.8 | 82.8±6.1 | 83.5±6.5 | 82.8±6.5 | |
| | MkNN+$\epsilon$N | 82.8±6.1 | 83.6±6.2 | 82.8±6.1 | 83.8±6.5 | 83.0±6.8 | 82.8±6.1 | 83.5±6.5 | 83.0±6.1 | |
| | S-MkNN | 82.8±6.1 | 84.6±6.9 | 82.8±6.1 | 84.2±6.4 | 83.9±6.9 | 82.8±6.1 | 84.6±6.9 | 82.8±6.1 | |
| | S-MkNN+$\epsilon$N | 82.8±6.1 | 84.6±6.9 | 82.8±6.1 | 84.6±6.9 | 83.9±6.9 | 82.8±6.1 | **84.7±7.2** | 83.5±5.8 | |
| Thy. | kNN | 96.3±4.1 | 95.6±5.0 | 95.6±5.0 | 96.1±5.6 | 96.1±4.6 | 95.6±5.0 | 95.6±5.0 | 95.6±5.0 | 95.6±5.0 |
| | kNN+$\epsilon$N | 95.6±4.9 | 95.0±4.5 | 96.1±4.4 | 96.1±4.4 | 96.3±5.1 | 96.1±4.4 | 95.0±4.2 | 94.9±4.9 | |
| | S-kNN | 95.6±5.0 | 95.8±4.9 | 95.6±5.0 | 96.1±5.6 | 96.7±3.8 | 95.6±5.0 | 96.4±4.3 | 95.6±5.0 | |
| | S-kNN+$\epsilon$N | 95.2±4.8 | 95.5±5.8 | 96.1±4.4 | 96.1±4.4 | **96.9±4.9** | 95.0±5.1 | 95.8±4.1 | 95.3±4.9 | |
| | MkNN | 96.1±4.9 | 95.6±5.0 | 95.6±5.0 | 95.9±5.1 | 96.3±4.8 | 95.6±5.0 | 95.6±5.0 | 95.6±5.0 | |
| | MkNN+$\epsilon$N | 94.8±4.3 | 94.7±5.0 | 96.1±5.6 | 96.1±4.4 | 94.7±5.0 | 96.6±5.1 | 94.9±5.1 | 95.5±4.7 | |
| | S-MkNN | 96.3±3.3 | 95.6±5.0 | 95.6±5.0 | 96.4±4.8 | 95.6±5.0 | 95.6±5.0 | 95.6±5.0 | 95.6±5.0 | |
| | S-MkNN+$\epsilon$N | 95.8±4.5 | 94.7±5.0 | 96.1±5.6 | 96.1±4.4 | 95.4±5.5 | 96.6±5.1 | 94.9±5.1 | 95.3±4.7 | |

Another measure which deserves attention is the ASS that was also present (boldfaced) in the top positions of both Tables V and VII. Indeed, with exception of ASP.S-kNN, ASP.S-kNN+$\epsilon N$ and ASS.kNN, no other measure was able to achieve the top positions of both analysis. Otherwise, three network measures (ADG, CLO and ECE) are present in Table VIII. Despite CLO and ECE presented good predictive capability in Table V, they seem to be very sensitive to the variation of the graph structure. The reason of that may lies in the nature of such measures which are based on centrality heuristics. In this sense, an alternative design of the pattern compliance strategy may be desirable for such measures, or even their investigation in a more related context, like the classification via characterization of importance.

### B. Results on artificial data

The artificial data sets considered in this work are presented by Fig. 2. The Moons1 data set is exhibited by Fig. 2(a) and the Gauss2 by Fig. 2(b). Notice that every data set represents a binary classification problem and denotes a configuration with low and high noise levels.

The results obtained by the eight network measures are presented by Fig. 3, in which each subfigure denotes their predictive accuracy over a given data set using the kNN and S-kNN graphs. In the Moons1 data set, one can see that there is a small difference between the results provided through of both graphs. In addition, the network measures in the Moons1

TABLE IV: Robustness analysis of the network measures in function of different graph construction methods. Best local results are underlined and best global results are boldfaced.

| Data | Graph | ASS | ALC | ADG | BET | ASP | CLO | GCC | ECE |
|---|---|---|---|---|---|---|---|---|---|
| App. | kNN | 64.5±12.2 | 70.5±5.3 | 51.8±17.5 | 50.9±16.5 | 71.0±3.5 | 51.8±17.4 | 62.9±10.3 | 54.2±18.8 |
| | kNN+ϵN | 67.7±2.8 | 70.9±3.4 | 51.8±17.5 | 50.9±16.5 | 71.0±3.5 | 51.8±17.4 | 62.9±10.3 | 54.2±18.8 |
| | S-kNN | 56.4±16.4 | 61.1±12.0 | 52.0±17.6 | 51.6±17.0 | **80.6±2.0** | 52.0±17.5 | 68.3±7.4 | 54.7±19.7 |
| | S-kNN+ϵN | 59.6±8.9 | 61.1±12.0 | 52.0±17.6 | 51.6±17.0 | **80.6±2.0** | 52.0±17.5 | 68.3±7.4 | 54.7±19.7 |
| | MkNN | 61.5±10.5 | 62.4±7.2 | 53.6±19.6 | 51.2±16.5 | 53.8±14.0 | 52.7±18.5 | 59.6±10.1 | 55.9±17.9 |
| | MkNN+ϵN | 66.8±3.0 | 73.3±1.8 | 53.6±19.6 | 51.2±16.5 | 62.2±4.9 | 52.7±18.5 | 63.9±3.7 | 56.2±18.1 |
| | S-MkNN | 65.2±9.1 | 58.6±13.6 | 52.3±17.7 | 51.9±17.6 | 61.9±9.2 | 53.4±17.7 | 59.4±13.2 | 57.2±17.8 |
| | S-MkNN+ϵN | 74.8±2.0 | 58.8±11.2 | 52.3±17.7 | 51.9±17.5 | 61.9±9.1 | 53.4±17.7 | 60.3±6.4 | 57.6±18.2 |
| Bal. | kNN | 80.8±7.3 | 78.1±7.6 | 71.5±10.3 | 72.1±11.7 | 85.1±3.6 | 71.4±11.2 | 78.3±8.4 | 72.9±11.7 |
| | kNN+ϵN | 80.9±7.2 | 78.1±7.7 | 71.6±10.4 | 72.2±11.7 | 85.0±3.6 | 71.5±11.2 | 78.4±8.5 | 73.1±11.7 |
| | S-kNN | 76.0±9.7 | 79.4±7.1 | 68.3±12.0 | 73.0±11.3 | **89.8±1.8** | 70.6±10.9 | 80.0±7.9 | 72.2±11.9 |
| | S-kNN+ϵN | 76.0±9.8 | 79.4±7.2 | 68.6±12.6 | 73.0±11.4 | **89.8±1.8** | 70.7±11.0 | 79.9±8.1 | 72.3±12.0 |
| | MkNN | 79.4±8.3 | 77.5±7.6 | 71.8±10.7 | 72.2±11.4 | 80.7±5.5 | 71.7±11.8 | 76.5±8.6 | 72.2±11.3 |
| | MkNN+ϵN | 79.4±8.2 | 77.7±7.6 | 71.8±10.6 | 72.3±11.4 | 80.8±5.5 | 71.7±11.8 | 76.4±8.7 | 72.4±11.4 |
| | S-MkNN | 77.6±9.7 | 77.7±7.6 | 69.8±11.4 | 72.0±11.1 | 85.4±3.2 | 71.5±11.7 | 77.0±8.1 | 71.6±11.8 |
| | S-MkNN+ϵN | 77.6±9.7 | 77.7±7.6 | 70.0±11.6 | 72.0±11.1 | 85.3±3.3 | 71.5±11.8 | 77.0±8.2 | 71.5±11.9 |
| Dig. | kNN | 94.8±2.8 | 95.1±2.1 | 92.7±3.9 | 93.2±4.2 | 96.9±0.9 | 92.9±3.8 | 94.1±3.1 | 92.7±4.0 |
| | kNN+ϵN | 94.8±2.8 | 95.1±2.1 | 92.8±4.0 | 93.2±4.2 | 96.9±0.9 | 92.9±3.8 | 94.1±3.1 | 92.7±4.0 |
| | S-kNN | 94.4±3.1 | 94.8±2.3 | 92.6±3.9 | 93.2±4.2 | **97.0±0.8** | 92.6±4.0 | 94.4±3.0 | 92.7±4.0 |
| | S-kNN+ϵN | 94.4±3.1 | 94.8±2.3 | 92.7±4.0 | 93.2±4.2 | **97.0±0.8** | 92.6±4.0 | 94.4±3.0 | 92.7±4.0 |
| | MkNN | 94.2±3.3 | 94.2±2.7 | 92.9±4.0 | 93.1±4.0 | 93.9±3.1 | 93.4±3.6 | 93.4±3.5 | 93.0±3.8 |
| | MkNN+ϵN | 94.2±3.3 | 94.2±2.7 | 93.2±4.1 | 93.1±4.0 | 93.9±3.1 | 93.4±3.6 | 93.4±3.5 | 93.0±3.9 |
| | S-MkNN | 94.3±3.3 | 94.3±2.6 | 92.8±4.0 | 93.1±4.0 | 94.0±3.0 | 93.3±3.6 | 93.4±3.5 | 93.0±3.8 |
| | S-MkNN+ϵN | 94.3±3.3 | 94.3±2.6 | 93.1±4.1 | 93.1±4.0 | 94.0±3.0 | 93.3±3.6 | 93.4±3.5 | 93.0±3.9 |
| Eco. | kNN | 74.7±5.9 | 68.6±5.8 | 55.2±15.0 | 63.2±12.3 | 73.2±3.0 | 59.4±13.1 | 58.8±12.0 | 56.2±16.7 |
| | kNN+ϵN | 74.7±5.9 | 70.8±1.4 | 55.2±15.0 | 63.2±12.3 | 73.2±3.0 | 59.4±13.1 | 58.8±12.0 | 56.2±16.7 |
| | S-kNN | 63.6±10.6 | 70.4±5.3 | 54.2±15.3 | 63.9±12.7 | **85.8±1.8** | 57.4±13.1 | 62.3±12.6 | 55.2±16.3 |
| | S-kNN+ϵN | 63.6±10.6 | 70.4±5.3 | 54.2±15.3 | 63.9±12.7 | **85.8±1.8** | 57.4±13.1 | 62.3±12.6 | 55.2±16.3 |
| | MkNN | 69.7±6.7 | 63.9±8.2 | 55.5±15.3 | 60.7±12.5 | 63.7±9.0 | 58.0±15.1 | 57.7±12.8 | 58.0±15.5 |
| | MkNN+ϵN | 71.2±1.6 | 69.1±2.6 | 55.5±15.3 | 60.7±12.5 | 63.7±9.0 | 58.0±15.1 | 57.7±12.8 | 57.9±15.4 |
| | S-MkNN | 68.0±8.7 | 71.5±3.9 | 54.4±15.7 | 62.1±11.8 | 70.5±5.2 | 58.0±13.9 | 60.6±11.2 | 57.1±15.7 |
| | S-MkNN+ϵN | 68.0±8.7 | 71.5±3.9 | 54.4±15.7 | 62.1±11.8 | 70.5±5.2 | 58.0±13.9 | 60.6±11.2 | 57.1±15.7 |
| Gla. | kNN | 55.2±10.6 | 58.9±6.3 | 45.8±14.5 | 47.6±14.1 | 58.9±6.8 | 46.5±15.7 | 53.6±10.1 | 44.8±15.6 |
| | kNN+ϵN | 54.3±9.6 | 57.2±4.2 | 43.8±12.1 | 45.8±12.0 | 57.1±5.0 | 44.4±13.0 | 52.1±8.3 | 42.8±13.3 |
| | S-kNN | 52.7±11.4 | 60.5±6.3 | 43.3±15.1 | 47.8±14.3 | **65.1±2.8** | 44.5±15.2 | 54.7±11.0 | 46.7±15.5 |
| | S-kNN+ϵN | 51.5±10.3 | 58.5±4.4 | 41.5±13.0 | 45.8±12.0 | 63.9±1.5 | 42.3±12.6 | 53.3±9.2 | 44.7±13.3 |
| | MkNN | 54.8±10.1 | 55.5±8.3 | 48.0±15.2 | 47.9±14.5 | 52.0±11.2 | 50.8±13.2 | 52.2±10.7 | 47.2±14.2 |
| | MkNN+ϵN | 54.4±8.7 | 53.8±5.8 | 46.1±13.1 | 45.9±12.2 | 50.4±9.5 | 48.4±10.3 | 50.6±8.7 | 45.2±11.9 |
| | S-MkNN | 55.7±11.3 | 55.1±9.3 | 45.3±15.0 | 47.8±14.6 | 58.9±6.2 | 53.5±10.6 | 54.7±9.9 | 49.3±13.2 |
| | S-MkNN+ϵN | 54.7±10.4 | 53.9±7.8 | 43.5±13.1 | 45.9±12.4 | 57.8±5.2 | 51.2±7.8 | 53.6±8.4 | 46.9±11.4 |
| Iris | kNN | 92.1±3.1 | 90.4±3.2 | 84.5±7.4 | 86.0±8.0 | 94.0±1.9 | 85.4±7.3 | 86.8±6.4 | 84.5±7.6 |
| | kNN+ϵN | 91.7±2.7 | 90.3±2.8 | 84.3±7.3 | 85.6±7.5 | 96.4±0.1 | 85.2±7.2 | 87.3±0.0 | 84.1±7.0 |
| | S-kNN | 88.4±4.7 | 89.3±4.8 | 84.4±7.4 | 85.5±8.4 | 96.5±0.5 | 84.5±7.4 | 88.2±6.5 | 84.5±7.4 |
| | S-kNN+ϵN | 88.0±4.2 | 89.1±4.5 | 84.3±7.3 | 85.4±8.2 | **97.0±0.4** | 84.2±7.0 | 87.8±6.0 | 84.2±6.9 |
| | MkNN | 88.8±5.2 | 88.2±4.5 | 84.7±7.7 | 85.0±7.4 | 89.4±3.8 | 86.3±6.5 | 86.2±6.3 | 85.1±7.2 |
| | MkNN+ϵN | 88.3±4.8 | 87.9±4.0 | 84.7±7.8 | 84.9±7.1 | 96.2±0.0 | 86.5±6.7 | 87.3±0.0 | 84.5±7.0 |
| | S-MkNN | 88.0±5.6 | 90.2±4.0 | 84.7±7.7 | 85.0±7.8 | 92.9±2.0 | 85.9±6.8 | 87.6±5.3 | 85.4±7.1 |
| | S-MkNN+ϵN | 87.7±5.3 | 89.8±3.6 | 84.7±7.8 | 84.9±7.8 | 96.2±0.0 | 86.0±7.0 | 87.6±5.1 | 84.7±6.9 |
| Son. | kNN | 64.4±9.9 | 62.0±11.0 | 54.7±16.3 | 55.5±17.7 | 67.1±8.4 | 56.2±15.1 | 64.8±9.8 | 53.9±16.2 |
| | kNN+ϵN | 64.4±9.9 | 62.0±11.0 | 54.7±16.4 | 55.5±17.7 | 67.2±8.4 | 56.2±15.1 | 64.8±9.8 | 53.9±16.2 |
| | S-kNN | 64.6±10.8 | 63.0±9.8 | 52.5±15.9 | 55.9±18.0 | **68.9±7.6** | 53.1±15.6 | 68.7±9.7 | 54.5±16.3 |
| | S-kNN+ϵN | 64.6±10.8 | 63.0±9.8 | 52.6±16.0 | 55.9±18.1 | **68.9±7.6** | 53.1±15.6 | 68.7±9.7 | 54.5±16.3 |
| | MkNN | 59.1±15.2 | 66.3±8.1 | 56.9±15.5 | 58.7±14.2 | 58.9±13.3 | 56.0±13.7 | 58.9±13.3 | 55.5±15.3 |
| | MkNN+ϵN | 59.3±15.4 | 66.4±8.1 | 57.0±15.7 | 58.7±14.3 | 59.0±6.9 | 56.0±13.7 | 58.9±13.4 | 55.3±15.0 |
| | S-MkNN | 64.2±11.6 | 62.8±11.1 | 53.6±15.9 | 55.2±17.1 | 61.9±11.7 | 55.8±13.8 | 64.9±10.8 | 55.8±14.5 |
| | S-MkNN+ϵN | 64.2±11.6 | 62.9±11.2 | 53.9±16.2 | 55.3±17.2 | 61.9±11.7 | 55.9±13.8 | 65.0±10.9 | 56.0±14.7 |
| Thy. | kNN | 92.1±3.3 | 92.8±2.2 | 86.7±6.3 | 93.4±1.9 | 95.3±0.7 | 88.9±4.2 | 89.4±5.3 | 86.5±7.0 |
| | kNN+ϵN | 91.4±2.8 | 92.7±2.1 | 86.9±6.6 | 93.5±2.0 | 95.3±0.7 | 89.0±4.4 | 89.5±5.3 | 86.2±6.7 |
| | S-kNN | 89.7±4.7 | 92.9±2.6 | 86.5±6.5 | 93.6±1.8 | 95.3±0.9 | 88.7±4.0 | 89.7±6.0 | 86.6±7.1 |
| | S-kNN+ϵN | 89.3±4.6 | 92.9±2.6 | 86.7±6.8 | 93.6±1.8 | **95.6±0.7** | 88.7±4.0 | 89.5±5.8 | 86.4±6.9 |
| | MkNN | 88.1±6.8 | 91.1±3.0 | 88.0±5.4 | 92.4±3.3 | 90.8±3.9 | 89.6±4.1 | 87.5±5.5 | 87.8±5.8 |
| | MkNN+ϵN | 87.7±6.4 | 91.2±2.8 | 88.8±1.8 | 92.6±3.4 | 90.8±3.7 | 90.2±4.6 | 87.8±5.7 | 87.2±6.0 |
| | S-MkNN | 90.2±5.2 | 91.8±2.4 | 86.5±6.6 | 93.2±2.2 | 93.0±1.6 | 87.3±5.8 | 89.0±4.8 | 86.5±7.0 |
| | S-MkNN+ϵN | 89.6±4.8 | 91.8±2.3 | 86.9±3.7 | 93.4±2.2 | 93.8±0.8 | 87.8±6.3 | 89.3±5.0 | 86.4±7.1 |

TABLE V: Predictive capability analysis of the top 10 measures with favorable significant differences (wins) according to the statistical test of Wilcoxon.

| Pos | Measure.Graph | #Wins ↑ | #Draws | #Losses |
|---|---|---|---|---|
| 1 | **ASP.S-kNN** | 44 | 19 | 0 |
| 2 | CLO.MkNN | 41 | 22 | 0 |
| | CLO.S-MkNN | 41 | 22 | 0 |
| 4 | **ASS.kNN** | 34 | 29 | 0 |
| | ECE.kNN | 34 | 29 | 0 |
| 6 | ASP.S-kNN+e | 21 | 42 | 0 |
| 7 | GCC.S-kNN | 19 | 44 | 0 |
| | CLO.kNN | 19 | 44 | 0 |
| 9 | CLO.MkNN+e | 11 | 52 | 0 |
| | CLO.S-MkNN+e | 11 | 52 | 0 |

TABLE VI: Predictive capability analysis of the top 10 measures with unfavorable significant differences (losses) according to the statistical test of Wilcoxon.

| Pos | Measure.Graph | #Wins | #Draws | #Losses↑ |
|---|---|---|---|---|
| 1 | GCC.kNN+e | 0 | 23 | 40 |
| 2 | ALC.kNN+e | 0 | 24 | 39 |
| 3 | ALC.MkNN+e | 0 | 29 | 34 |
| 4 | ASS.S-kNN+e | 0 | 35 | 28 |
| 5 | GCC.MkNN+e | 0 | 36 | 27 |
| 6 | BET.kNN+e | 0 | 43 | 20 |
| 7 | ASP.MkNN+e | 0 | 46 | 17 |
| 8 | ALC.S-kNN+e | 1 | 47 | 15 |
| 9 | ECE.MkNN+e | 0 | 49 | 14 |
| 10 | BET.MkNN+e | 0 | 51 | 12 |
| | ASS.MkNN+e | 0 | 51 | 12 |

TABLE VII: Robustness analysis of the top 10 measures with favorable significant differences (wins) according to the statistical test of Wilcoxon.

| Pos | Measure.Graph | #Wins ↑ | #Draws | #Losses |
|---|---|---|---|---|
| 1 | **ASP.S-kNN** | 62 | 1 | 0 |
|  | ASP.S-kNN+e | 62 | 1 | 0 |
| 3 | ASP.kNN | 60 | 1 | 2 |
|  | ASP.kNN+e | 60 | 1 | 2 |
| 5 | **ASS.kNN** | 47 | 12 | 4 |
|  | ASS.kNN+e | 47 | 12 | 4 |
| 7 | ALC.kNN+e | 44 | 15 | 4 |
| 8 | ALC.kNN | 43 | 16 | 4 |
| 9 | ASP.S-MkNN+e | 41 | 18 | 4 |
| 10 | ALC.MkNN+e | 40 | 19 | 4 |
|  | ALC.S-kNN+e | 40 | 19 | 4 |
|  | ALC.S-kNN | 40 | 19 | 4 |
|  | ASP.S-MkNN | 40 | 19 | 4 |

TABLE VIII: Robustness analysis of the top 10 measures with unfavorable significant differences (losses) according to the statistical test of Wilcoxon.

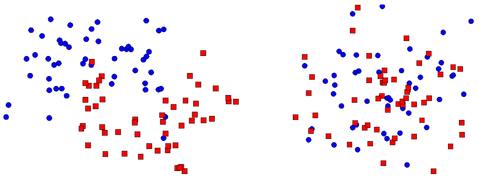
| Pos | Measure.Graph | #Wins | #Draws | #Losses↑ |
|---|---|---|---|---|
| 1 | ADG.S-kNN+e | 0 | 1 | 62 |
| 2 | ADG.S-kNN | 0 | 3 | 60 |
| 3 | ADG.S-MkNN | 2 | 9 | 52 |
|  | ADG.kNN+e | 2 | 9 | 52 |
|  | ADG.S-MkNN+e | 2 | 9 | 52 |
| 6 | ADG.kNN | 2 | 10 | 51 |
| 7 | CLO.S-kNN+e | 1 | 14 | 48 |
| 8 | CLO.S-kNN | 2 | 13 | 48 |
| 9 | ECE.S-kNN+e | 2 | 21 | 40 |
| 10 | ECE.kNN+e | 1 | 23 | 39 |

data set requires smaller values of $k$ to provide the best results than in the Gauss2 data set. This could suggest that bigger values of $k$ may be required as the noise levels get higher. Generally speaking, most of the network measures achieve their best results with small values of $k$, which means that their predictive power is directly associated to the highest affinities among the data items. Otherwise, ASP has its predictive capability less sensible to the changes in the graph structure, which makes it a robust measure.

In summary, the experiments with artificial data confirm some insights obtained with the real ones. They showed, for example, that ASP, ASS and ALC are more robust to the variation of the graph structure; that measures like BET and CLO present good predictive capability in specific situations, such as in the case of graphs constructed with small values of $k$; and that measures like ADG seems to have limitations in terms of both predictive capability and robustness.

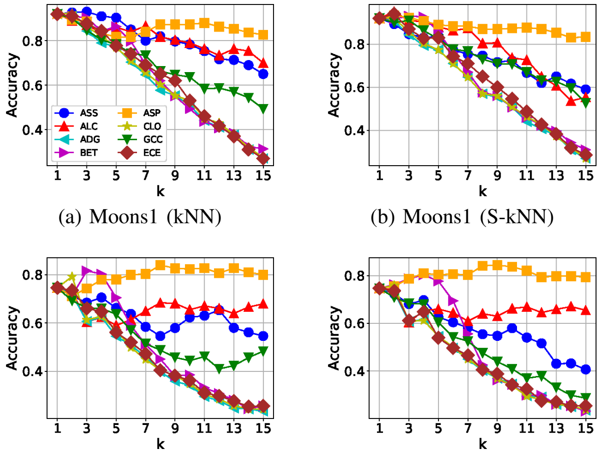
## IV. Hybrid classification analysis

Our last experiment consists of analyzing the combination of low-level classifiers with the high-level analysis provided by the network measures investigated. Two widely known low-level techniques have been adopted here: the Naive Bayes (NB) which is a classical technique and the Support Vector Machine (SVM) which is a state-of-the-art technique for a wide range of classification problems. Regarding the parameters of both techniques, in the NB we assume the likelihood to be Gaussian; and in the SVM we adopted the radial basis function (rbf) as kernel and selected the cost and kernel parameters respectively over the range $\{2^{-4}, 2^{-2}, \ldots, 2^{12}\}$ and $\{2^{-6}, 2^{-4}, \ldots, 2^{10}\}$. For the high-level associations, we selected the recommended network measures according to our results discussed before, i.e., ASS with the kNN graph (ASS.kNN) and ASP with the S-kNN one (ASP.S-kNN). Table IX presents the results obtained by the hybrid classification method. The *Networkless* rows denote results obtained by traditional classifiers when $\lambda = 0$, i.e., without the high-level associations provided by the complex network measures. One can see in the table that both network measures ASS and ASP are able to improve the predictive performance of the traditional classifiers. For example, ASP improved the NB results in every data set under analysis, while ASS improved the SVM results over seven of eight data sets.



(a) Moons1 (b) Gauss.2

Fig. 2: Artificial data sets with lower and higher noise levels.



(a) Moons1 (kNN) (b) Moons1 (S-kNN)

(c) Gauss.2 (kNN) (d) Gauss.2 (S-kNN)

Fig. 3: Comparative evaluation of the network measures on the artificial data sets.

## V. Conclusions

In this paper we evaluate comparatively the contribution of eight well-known complex network measures for data classification. Our hypothesis was that certain complex network measures present better predictive performance than others. To the best of our knowledge, this is the first study focused on the individual characterization of the measures in that context.

TABLE IX: Predictive accuracy of the hybrid classification model considering the NB and SVM low-level classifiers, and ASS.kNN and ASP.S-kNN high-level ones. The *Networkless* rows refer to classification results provided only by the low-level techniques, i.e., without the network measures.

| Data | $\Delta G$ | NB | SVM |
|---|---|---|---|
| App. | *Networkless* | 80.3±12.0 | 81.7±7.5 |
| | ASS.kNN | 83.1±5.6 ($\lambda = 0.8$) | **83.6±7.0** ($\lambda = 0.4$) |
| | ASP.S-kNN | 82.4±10.1 ($\lambda = 0.9$) | 82.7±6.6 ($\lambda = 1$) |
| Bal. | *Networkless* | 90.0±2.2 | 99.0±1.3 |
| | ASS.kNN | 96.3±2.7 ($\lambda = 0.3$) | **99.6±0.9** ($\lambda = 0.2$) |
| | ASP.S-kNN | 95.7±2.3 ($\lambda = 0.3$) | **99.6±0.8** ($\lambda = 0.2$) |
| Dig. | *Networkless* | 79.8±2.5 | **99.3±0.2** |
| | ASS.kNN | 98.6±0.5 ($\lambda = 0.9$) | 99.3±0.2 ($\lambda = 0.1$) |
| | ASP.S-kNN | 98.7±3.7 ($\lambda = 0.6$) | 99.3±0.2 ($\lambda = 0.1$) |
| Eco.. | *Networkless* | 85.9±6.2 | 88.1±5.4 |
| | ASS.kNN | 86.8±6.1 ($\lambda = 0.2$) | 88.3±5.2 ($\lambda = 0.2$) |
| | ASP.S-kNN | 88.2±5.8 ($\lambda = 0.8$) | **89.1±5.5** ($\lambda = 0.3$) |
| Gla. | *Networkless* | 44.5±8.9 | 71.9±8.5 |
| | ASS.kNN | 73.2±8.6 ($\lambda = 0.9$) | **74.8±8.7** ($\lambda = 0.3$) |
| | ASP.S-kNN | 73.3±8.7 ($\lambda = 0.6$) | **74.8±9.8** ($\lambda = 0.3$) |
| Iris | *Networkless* | 97.1±4.4 | 94.4±12.5 |
| | ASS.kNN | 98.0±3.9 ($\lambda = 0.5$) | **98.2±3.8** ($\lambda = 0.4$) |
| | ASP.S-kNN | 97.3±4.4 ($\lambda = 0.8$) | 97.3±4.4 ($\lambda = 0.4$) |
| Son. | *Networkless* | 66.2±10.8 | 74.6±8.9 |
| | ASS.kNN | **84.4±7.5** ($\lambda = 0.5$) | 83.2±6.0 ($\lambda = 0.4$) |
| | ASP.S-kNN | 84.3±7.4 ($\lambda = 0.5$) | 83.2±6.9 ($\lambda = 0.4$) |
| Thy. | *Networkless* | 97.1±3.3 | 97.4±3.7 |
| | ASS.kNN | 96.9±3.3 ($\lambda = 0.1$) | **97.5±3.7** ($\lambda = 0.1$) |
| | ASP.S-kNN | 97.4±3.1 ($\lambda = 0.5$) | 97.4±3.7 ($\lambda = 0.3$) |

To analyze the measures, we considered several graph construction methods based on k-nearest neighbors and $\epsilon$-radius neighborhood heuristics, adapted a high-level classification technique based on pattern compliance, and designed an experimental setup to apply the network measures in both artificial and real data sets. The measures were evaluated in terms of predictive capability and robustness.

The results revealed some network measures with great potential, which were able to achieve high predictive capability and also robustness to the change of the graph structure. The recommended network measures are the average shortest path (ASP), which had outstanding performance, and assortativity (ASS), which presented competitive and straightforward results. We also found that ASS often provides its best results with the kNN graph, while ASP are usually better with the S-kNN and S-kNN+$\epsilon$N graphs.

On the other hand, despite the average degree (ADG) is a commonly adopted measure in the literature, our study shows that it is neither robust nor achieve top predictive results in any data set. Therefore, it does not seem very appropriate to the high-level classification via pattern compliance. Indeed, other network measures based on centrality also share some of these limitations. Forthcoming works will investigate such measures in the context of the high-level classification via characterization of importance.

## References

[1] T. C. Silva and L. Zhao, "Network-based high level data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, p. 954–970, 2012.

[2] T. H. Cupertino, M. G. Carneiro, Q. Zheng, J. Zhang, and L. Zhao, "A scheme for high level data classification using random walk and network measures," *Expert Systems with Applications*, vol. 92, pp. 289–303, 2018.

[3] M. G. Carneiro and L. Zhao, "Organizational data classification based on the importance concept of complex networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3361–3373, 2018.

[4] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[5] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006.

[7] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, "Characterization of complex networks: A survey of measurements," *Advances in physics*, vol. 56, no. 1, pp. 167–242, 2007.

[8] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[9] T. Jebara, J. Wang, and S.-F. Chang, "Graph construction and b-matching for semi-supervised learning," in *International Conference on Machine Learning*, 2009, pp. 441–448.

[10] K. Ozaki, M. Shimbo, M. Komachi, and Y. Matsumoto, "Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data," in *ACL Conference on Computational Natural Language Learning*, 2011, pp. 154–162.

[11] M. G. Carneiro and L. Zhao, "Analysis of graph construction methods in supervised data classification," in *Brazilian Conference on Intelligent Systems*. IEEE, 2018, pp. 390–395.

[12] J. R. Bertini Jr, L. Zhao, R. Motta, and A. de Andrade Lopes, "A nonparametric classification method based on k-associated graphs," *Information Sciences*, vol. 181, no. 24, pp. 5435–5456, 2011.

[13] B. Araujo and L. Zhao, "Data heterogeneity consideration in semi-supervised learning," *Expert Systems with Applications*, vol. 45, pp. 234–247, 2016.

[14] M. G. Carneiro, R. Cheng, L. Zhao, and Y. Jin, "Particle swarm optimization for network-based data classification," *Neural Networks*, vol. 110, pp. 243–255, 2019.

[15] M. G. Carneiro, L. Zhao, R. Cheng, and Y. Jin, "Network structural optimization based on swarm intelligence for highlevel classification," in *International Joint Conference on Neural Networks*. IEEE, 2016, pp. 3737–3744.

[16] M. G. Carneiro, J. L. G. Rosa, A. A. Lopes, and L. Zhao, "Network-based data classification: combining k-associated optimal graphs and high-level prediction," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, pp. 1–14, 2014.

[17] T. Colliri, D. Ji, H. Pan, and L. Zhao, "A network-based high level data classification technique," in *International Joint Conference on Neural Networks*. IEEE, 2018, pp. 1–8.

[18] V. H. Resende and M. G. Carneiro, "High-level classification for multi-label learning," in *2020 International Joint Conference on Neural Networks*. IEEE, 2020, pp. 1–8.

[19] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.

[20] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[21] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of mathematical sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[22] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[23] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[24] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[25] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[26] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.