On the evaluation of graph construction methods for semi-supervised transductive classification

Leonardo Macedo Freire[1], Jefferson Tales Oliva[2], Valéria de Carvalho Santos[1], Vander Luis de Souza Freitas[1], Jadson Castro Gertrudes[1]

[1] Postgraduate Program in Computer Science, Federal University of Ouro Preto.
leonardo.macedo@aluno.ufop.edu.br
[2] Federal University of Technology – Paraná (UTFPR), Pato Branco, Paraná, Brazil

Abstract. Semi-supervised learning addresses critical challenges in machine learning when labeled data is scarce but unlabeled data is abundant. However, ensuring effective label propagation in transductive semi-supervised classification algorithms is challenging due to the significant influence of the underlying graph's topological structure. This article systematically investigates this problem by evaluating various graph construction methods alongside traditional approaches, including the novel application of the HDBSCAN*-derived Mutual Reachability Minimum Spanning Tree (MST_R) and the Disparity Filter (DF). We employed the Local and Global Consistency (LGC) algorithm for label propagation and assessed performance using macro-averaged F-measure and statistical tests (Friedman and Nemenyi). Our experiments show that the DF demonstrates compelling performance, showing no statistical difference from top-ranked methods like SYM_FKNN. On the other hand, MST_R exhibited limitations that could be verified in future studies. This comprehensive analysis, further supported by Spearman's rank correlation with graph and data properties, provides important insights into optimal graph selection for robust semi-supervised classification workflows.

CCS Concepts: • **Computing methodologies → Semi-supervised learning settings**.

Keywords: density-based clustering, disparity filter, graph construction, semi-supervised classification, transductive classification.

1. INTRODUCTION

Semi-supervised learning addresses scenarios where obtaining labeled data for training a learning model is costly, finding practical value in tasks such as speech recognition and spam filtering [Chapelle et al. 2006]. It combines supervised and unsupervised learning strategies to take advantage of labeled and unlabeled data [Chapelle et al. 2006]. Semi-supervised classification algorithms, including inductive and transductive types, are generally extensions of traditional classification algorithms that utilize both labeled and unlabeled data during their learning process. In particular, transductive algorithms, the focus of this article, commonly use graphs to propagate labels from labeled to unlabeled data, relying on assumptions such as the graph assumption, where class labels are considered "smooth" in relation to the graph, exhibiting slow variation, such that labels tend to be the same for points connected by a strong edge [Gertrudes et al. 2018]. Consequently, how the graph is presented to the transductive semi-supervised classification algorithm can directly influence its performance.

Regarding the problem of graph construction for transductive algorithms, [de Sousa et al. 2013] and [Berton et al. 2018] presented studies comparing the influence of various graph construction methods on semi-supervised classification. Their observations confirmed that different graph structures affect

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The following Brazilian research agencies have also supported this work: CNPq and Minas Gerais Research Foundation (FAPEMIG - grant APQ-02036/24).

the performance of semi-supervised tasks. However, over the years, new graph construction methods have emerged that warrant evaluation in the context of transductive semi-supervised classification. One such example is the mutual reachability minimum spanning tree (MST_R), a graph structure that can be efficiently computed and is employed by the Hierarchical DBSCAN* (HDBSCAN*) framework proposed by [Campello et al. 2015] for unsupervised learning (data clustering) and semi-supervised learning for semi-supervised classification [Gertrudes et al. 2018].

Network science analysis offers backbone extraction techniques, such as the disparity filter (DF) [Serrano et al. 2009], to simplify densely connected weighted graphs [Menczer et al. 2020]. The idea is to prune less significant edges while preserving topologically important connections. The DF evaluates edge significance relative to each vertex's local weight distribution through statistical hypothesis testing, providing an alternative approach to graph construction.

Therefore, this article evaluates the performance of new approaches for graph construction methods in transductive semi-supervised classification algorithms. For this, we systematically compared graph construction models, comparing well-established methods from the area with the MST_R and the DF. The results suggest DF as a good choice for semi-supervised classification.

2. BACKGROUND

2.1 Semi-supervised Classification Framework

The semi-supervised classification framework operates on a dataset ($\mathbf{X}_n$) composed of labeled ($\mathbf{X}_l$) and unlabeled ($\mathbf{X}_u$) subsets, where $n = l + u$ is the dataset size. Its objective involves combining information from both to enhance classification performance, proving particularly valuable when labeled instances are limited ($l \ll n$) [Zhu 2005].

Semi-supervised learning algorithms can be categorized into two principal paradigms. Inductive approaches incorporate $\mathbf{X}_l$ and $\mathbf{X}_u$ during its training, resulting in classifiers capable of generalizing to novel instances. On the other hand, transductive methods specifically target optimal label assignments from $\mathbf{X}_l$ to $\mathbf{X}_u$, aiming to boost performance on the particular dataset at hand [Zhu 2005; Chapelle et al. 2006]. Graph-based formulations is one of these transductive approaches, and have demonstrated notable success by explicitly modeling the data's underlying manifold geometry, which enables more effective label propagation through the graph structure.

2.2 Graph Construction algorithms

Graph-based semi-supervised algorithms begin by constructing a graph from the combined training data $\mathbf{X}_n$. Effective label propagation from labeled instances $\mathbf{X}_l$ to unlabeled instances $\mathbf{X}_u$ requires that all vertices remain connected within the graph structure, ensuring information flow across both data subsets.

k-nearest neighbors (kNN) methods: the kNN method generates an adjacency matrix $\mathbf{A}$ from a dissimilarity matrix $\mathbf{\Psi}$ (e.g., Euclidean distance derived from $\mathbf{X}$). For each instance $\mathbf{x}_i$, its neighborhood comprises the k closest instances in $\mathbf{\Psi}$, establishing the k-neighborhood $\mathcal{N}_k(\mathbf{x}_i)$. The adjacency matrix entries are defined as $a_{ij} = 1$ if $\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)$, and 0 otherwise [de Sousa et al. 2013]. This construction yields asymmetric adjacency matrices due to non-reciprocal neighborhood relationships: $\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)$ does not imply $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j)$.

There are three well-established symmetrization strategies to address this asymmetry: Symmetric kNN (SYM_KNN) converts directed edges to undirected connections when either $\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)$ or $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j)$. This produces a binary symmetric matrix $\hat{\mathbf{A}} = \max\{\mathbf{A}, \mathbf{A}^T\}$ with $\hat{\mathbf{A}} \in \mathbb{B}^{n \times n}$. Symmetry-Favored kNN (SYM_FKNN) [Liu and Chang 2009] generates weighted undirected edges: bidirectional connections receive weight 2 while unidirectional links receive weight 1. The weighted adjacency matrix

$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{A}^T$ yields $\hat{\mathbf{A}} \in \mathbb{N}^{n\times n}$. Mutual kNN (MUT_KNN) retains only reciprocal edges, requiring both $\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)$ and $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j)$. The binary symmetric matrix $\hat{\mathbf{A}} = \min\{\mathbf{A}, \mathbf{A}^T\}$ ($\hat{\mathbf{A}} \in \mathbb{B}^{n\times n}$) may contain isolated vertices, which are connected to their nearest neighbors despite the reciprocity condition following [de Sousa et al. 2013].

The M_KNN [Ozaki et al. 2011] method enhances the MUT_KNN approach by integrating a Maximum Spanning Tree (MST), specifically addressing connectivity limitations inherent to mutual k-nearest neighbor graphs. This hybrid framework operates within a similarity matrix paradigm, where the MST strategically incorporates high-similarity edges to guarantee a single connected component. M_KNN eliminates isolated vertices in the symmetrized adjacency matrix $\hat{\mathbf{A}} = \min\{\mathbf{A}, \mathbf{A}^T\}$ ($\hat{\mathbf{A}} \in \mathbb{B}^{n\times n}$), ensuring robust connectivity essential for effective label propagation. The MST integration thus transforms MUT_KNN from a locally constrained method into a globally coherent graph construction technique.

Sequential KNN (S_KNN) constructs quasi-regular graphs through incremental edge addition. The algorithm processes vertices from $k = 1$ to $k_{\max}$, connecting each vertex to neighbors with degree $< k$ based on a relevance criterion. The ordering criterion employs the closeness (Cl') centrality measure [Vega-Oliveros et al. 2014], a multidimensional adaptation of standard closeness centrality defined as $Cl' = \frac{N}{\sum_{j\neq i} |\mathbf{x}_i - \mathbf{x}_j|}$, where N denotes vertex count and $|\cdot|$ represents a distance metric (e.g., Euclidean distance). Lower Cl' values indicate peripheral vertices (greater average distances), while higher values indicate central vertices. S_KNN prioritizes vertices with minimal Cl' values, enabling peripheral vertices to establish connections before central ones, preventing hub formation.

Disparity Filter: The disparity filter (DF) [Serrano et al. 2009] method prunes less significant edges by evaluating edge significance relative to each vertex's local weight distribution through statistical hypothesis testing. For an edge $\langle i, j\rangle$ connecting vertices v_i and v_j with weight w_{ij}, the disparity filter computes the probability p_{ij} under the null hypothesis of random weight distribution by $p_{ij} = (1 - w_{ij}/s_i)^{k_i - 1}$, where $s_i = \sum_j w_{ij}$ denotes the strength (total edge weight) of vertex v_i, and k_i its degree. Edges are retained if $p_{ij} < \alpha$ for a chosen significance level α, which controls backbone sparsity: lower α values yield sparser networks. Since each edge connects two vertices, dual probabilities p_{ij} and p_{ji} are computed. Our implementation retains edges when $\min\{p_{ij}, p_{ji}\} < \alpha$, producing sparse backbones while maintaining essential connectivity.

Mutual reachability Minimum spanning tree: The MST$_R$ is an important structure derived from the HDBSCAN* [Campello et al. 2015], representing the hierarchy of DBSCAN* clusterings. This hierarchy effectively encapsulates all possible DBSCAN* solutions across a continuous range of ϵ values ($\epsilon \in [0, \infty]$) for a given fixed m_{pts} (minimum number of points to consider a point $\mathbf{x}_i$ dense). The construction of this hierarchy is performed upon a proximity graph in a transformed space, defined by the *mutual reachability distance* (d_{mreach}). The d_{mreach} quantifies the minimum distance at which each pair of points can be considered density-connected. Within the scope of the present article, the MST$_R$ is explicitly built in this mutual reachability space to delineate essential connectivity paths among objects. We consider that this specific connectivity representation is sufficient for effective use on the label propagation step in semi-supervised classification algorithms. Consequently, our methodology involves constructing a symmetric adjacency matrix $\hat{\mathbf{A}}$ where connections are established only between MST$_R$ neighbors, assigned a weight of 1.

2.3 Label propagation

Once the graph structure is obtained from the graph construction algorithm, the label propagation procedure is performed, which involves gradually propagating labels from the labeled subset ($\mathbf{X}_l$) to the unlabeled subset ($\mathbf{X}_u$) using the graph's connections.

The Local and Global Consistency (LGC) [Zhou et al. 2003] algorithm is a well-established method. The LGC computes a labeling matrix $\mathbf{F} \in \mathbb{R}^{n\times c}$ by minimizing Equation (1), given a regularization parameter $\mu > 0$ and a label matrix $\mathbf{Y} \in \mathbb{B}^{n \times c}$:

$$\mathbf{F} = \arg \min_{\mathbf{F} \in \mathbb{R}^{n \times c}} \quad \mathrm{tr}(\mathbf{F}^T \mathbf{L} \mathbf{F} + \mu(\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})), \tag{1}$$

where $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ is the normalized graph Laplacian, $\mathbf{I}_n$ is the identity matrix, $\mathbf{D} = \mathrm{diag}(s_i)$ is the degree matrix with $s_i = \sum_j W_{ij}$, and $\mathbf{W}$ represents the affinity matrix. The closed-form solution for $\mathbf{F}$ is given by $\mathbf{F} = (\mathbf{I}_n + \mathbf{L}/\mu)^{-1}\mathbf{Y}$.

3. MATERIAL AND METHODS

Methodology overview: Our systematic evaluation framework, illustrated in Figure 1, contains four phases: (i) dataset characterization, (ii) graph construction, (iii) label propagation, and (iv) comparative analysis. We also extracted clustering and graph-theoretic measures from both the dataset and the graph structure during execution.

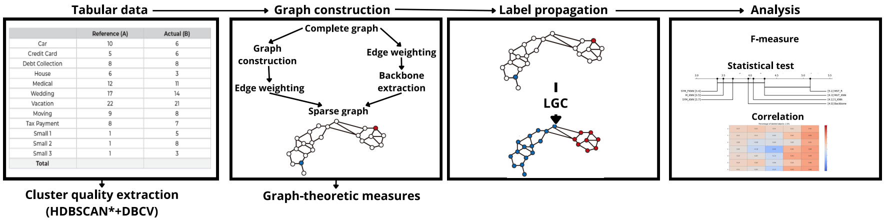

Fig. 1. Overview of the systematic methodology of our work.

Dataset characterization: Our study employs a curated collection of 31 publicly available datasets, detailed in Table I, containing diverse domains such as biology, text, and chemistry. This dataset collection, previously used by [Gertrudes et al. 2018], was adopted in its preprocessed form.

We randomly selected labeled instances from each dataset to simulate semi-supervised learning conditions. Each trial included at least one labeled example per class, generating 20 distinct variants of each dataset for every predefined labeling ratio. These variants ensure that any algorithm could be benefited just by chance. To investigate the impact of label availability on model performance, we evaluated multiple proportions of labeled data: 2%, 5%, 8%, and 10%, the same ratio used in [Gertrudes et al. 2018].

Graph construction: We evaluated seven graph construction methods described in Section 2. For MST$_R$, MUT_KNN, M_KNN, SYM_FKNN, SYM_KNN, and S_KNN, we tested the neighborhood size parameter (m_{pts} or k) across values in $2, 4, 6, 8, 10, 12, 14, 16$. For DF, we examined three threshold values $\alpha = \{0.01, 0.05, 0.1\}$. Following [de Sousa et al. 2013], edge weights were computed using the radial basis function kernel (RBF kernel): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\Psi(i,j)^2}{2\sigma^2}\right)$. The bandwidth parameter σ was set as $\sigma = \frac{\sum_{i=1}^{n} \Psi(i, i_k)}{3n}$, where $\Psi(i, i_k)$ denotes the distance between instance $\mathbf{x}_i$ and its k-th nearest neighbor.

Cluster quality and graph-theoretic measures: Following the sparse graph construction or backbone extraction (Second box in Figure 1), we computed several graph-theoretic measures to characterize the resulting graphs and validate findings. Specifically, we quantified the following properties: number of vertices, average degree, maximum degree, density, and graph diameter. We also computed the Density-Based Cluster Validity (DBCV) index [Moulavi et al. 2014] to assess the quality of density-based clustering solutions. This involved applying HDBSCAN* with its default parameters ($m_{pts} = 5$), from which the DBCV index was then derived for the resulting clustering.

Table I. List of real datasets collected to perform the semi-supervised classification experiments.

Dataset		#obj	#att	#cl	#Distance
ACE ECFP4	[Sutherland et al. 2004]	114	1,025	2	Tanimoto
ACE ECFP6	[Sutherland et al. 2004]	114	1,025	2	Tanimoto
Analcatdata authorship	[Vanschoren et al. 2014]	841	70	4	Cosine
Armstrong-v1	[de Souto et al. 2008]	72	1,082	2	Cosine
Articles-1442-5	[Naldi et al. 2011]	253	4636	5	Cosine
Articles-1442-80	[Naldi et al. 2011]	253	388	5	Cosine
Auto price	[Vanschoren et al. 2014]	159	16	2	Euclidean
Bank note–Authentication	[Vanschoren et al. 2014]	1,372	5	2	Euclidean
Cardiotocography	[Vanschoren et al. 2014]	2,126	36	10	Euclidean
Chowdary	[de Souto et al. 2008]	104	183	2	Cosine
Chcase Geyser1	[Vanschoren et al. 2014]	222	2	2	Euclidean
COX2 ECFP6	[Sutherland et al. 2004]	322	1,025	2	Tanimoto
DHFR ECFP4	[Sutherland et al. 2004]	397	1,025	2	Tanimoto
DHFR ECFP6	[Sutherland et al. 2004]	397	1,025	2	Tanimoto
Diggle table	[Vanschoren et al. 2014]	310	8	9	Euclidean
Fontaine ECFP4	[Fontaine et al. 2005]	435	1,024	2	Tanimoto
Fontaine ECFP6	[Fontaine et al. 2005]	435	1,024	2	Tanimoto
Gordon	[de Souto et al. 2008]	181	1,627	2	Cosine
Iris	[Lichman 2013]	150	5	3	Euclidean
M1 ECFP4	[Gaulton et al. 2017]	769	1,025	2	Tanimoto
M1 ECFP6	[Gaulton et al. 2017]	769	1,025	2	Tanimoto
Mfeat-factors	[Vanschoren et al. 2014]	2,000	216	10	Euclidean
Mfeat-Karhunen	[Vanschoren et al. 2014]	2,000	65	10	Euclidean
Seeds	[Lichman 2013]	210	8	3	Euclidean
Segmentation	[Vanschoren et al. 2014]	2,100	20	7	Euclidean
Semeion	[Vanschoren et al. 2014]	1,593	256	10	Cosine
Stock	[Vanschoren et al. 2014]	950	10	2	Euclidean
Transplant	[Vanschoren et al. 2014]	131	4	2	Euclidean
WDBC	[Lichman 2013]	569	32	2	Euclidean
Wine	[Lichman 2013]	178	13	3	Euclidean
Yeast galactose	[Yeung et al. 2003]	205	81	4	Euclidean

Label propagation: We employed the Local and Global Consistency (LGC) algorithm for the label propagation phase with its default parameter setting ($\mu = 0.9$) as the parameter set by [Ozaki et al. 2011] during their experiments. We limited our analysis to this single propagation algorithm, as our primary focus was to assess and compare the performance of different graph construction methods rather than evaluate label propagation techniques.

Analysis: We report the macro-averaged F-measure, computed as the harmonic mean between precision and recall for each class, then averaged across all classes [Sokolova and Lapalme 2009]. For this evaluation, we excluded instances whose labels served as initial points for label propagation during the semi-supervised learning step.

For statistical analysis, we employed the two-step procedure proposed by Demšar [Demšar 2006]. Initially, the Friedman test [Friedman 1937] was applied to verify whether significant differences existed among the graph construction algorithms across all datasets. Upon rejection of the null hypothesis (absence of significant algorithmic difference) at a significance level of 5% ($\alpha = 0.05$), the Nemenyi post-hoc test [Nemenyi 1963] was conducted to identify specific algorithm pairs exhibiting significant differences. The Nemenyi test establishes a significant performance difference if the average ranks of two algorithms diverge by at least the critical distance (CD) value.

To investigate the relationships between F-measure performance in the label propagation step and (i) dataset properties, as well as (ii) structural characteristics of graphs generated by construction methods, we applied Spearman's rank correlation coefficients. This analysis reveals distinct behavioral patterns across graph construction algorithms.

4. RESULTS AND DISCUSSION

Table II reports the mean and standard deviation of F-measure values from 20 experimental trials for each dataset, along with the optimal neighborhood size parameter (k). Although focusing on the 2% labeled data scenario, these findings generalize to other labeling percentages (5%, 8%, and 10%), exhibiting qualitatively similar patterns across all conditions. The observed F-measure range for 2%

labeled data was $[0.53, 0.99]$, with some datasets showing improved performance at higher labeling percentages.

Table II. The average and standard deviation of F-measure values across 20 experimental repetitions for each dataset and the optimal neighborhood size parameter (k) yielded the highest performance. 2% of labeled data.

Dataset	DF	M_KNN	MST$_R$	S_KNN	MUT_KNN	SYM_FKNN	SYM_KNN
ACE ECFP4	**0.72 (6) ± 0.08**	0.71 (12) ± 0.12	0.66 (4) ± 0.12	0.67 (12) ± 0.1	0.71 (12) ± 0.11	0.67 (16) ± 0.11	0.64 (4) ± 0.19
ACE ECFP6	**0.72 (6) ± 0.03**	0.7 (14) ± 0.11	0.68 (4) ± 0.12	0.6 (8) ± 0.1	0.7 (16) ± 0.12	0.66 (8) ± 0.09	0.66 (8) ± 0.1
ANALCATDATA AUTHORSHIP-458	0.98 (8) ± 0.01	0.98 (10) ± 0.01	0.97 (4) ± 0.02	0.98 (12) ± 0.01	0.98 (16) ± 0.01	0.98 (10) ± 0.01	**0.99 (12) ± 0.01**
ARMSTRONG2002-v1	0.78 (4) ± 0.08	0.81 (8) ± 0.1	0.8 (4) ± 0.14	0.79 (10) ± 0.07	0.8 (16) ± 0.12	0.81 (4) ± 0.09	**0.85 (4) ± 0.11**
ARTICLES-1442-5	0.96 (6) ± 0.04	0.95 (8) ± 0.05	0.98 (14) ± 0.03	**0.99 (16) ± 0.0**	0.9 (14) ± 0.12	**0.99 (14) ± 0.01**	0.98 (16) ± 0.02
ARTICLES-1442-80	0.96 (6) ± 0.04	0.96 (8) ± 0.04	**0.99 (14) ± 0.01**	0.98 (16) ± 0.01	0.94 (14) ± 0.07	0.98 (14) ± 0.01	0.97 (16) ± 0.01
AUTO PRICE	0.79 (14) ± 0.1	0.81 (14) ± 0.11	0.78 (10) ± 0.1	**0.83 (10) ± 0.09**	0.76 (16) ± 0.12	0.82 (16) ± 0.09	0.82 (16) ± 0.09
BANKNOTE-AUTHENTICATION	0.96 (14) ± 0.01	**0.99 (10) ± 0.0**	**0.99 (4) ± 0.0**	0.98 (10) ± 0.01	0.97 (10) ± 0.02	0.98 (10) ± 0.01	0.98 (10) ± 0.01
CARDIOTOCOGRAPHY	0.98 (14) ± 0.0	**0.99 (12)** ± 0.01	0.98 (10) ± 0.01	**0.99 (16)** ± 0.0	0.98 (16) ± 0.0	**0.99 (14) ± 0.0**	**0.99 (14) ± 0.0**
CHOWDARY 2006	0.94 (16) ± 0.11	0.97 (10) ± 0.04	0.95 (6) ± 0.08	**0.98 (8) ± 0.01**	0.97 (14) ± 0.01	0.97 (12) ± 0.05	**0.98 (12) ± 0.0**
CHSCASE GEYSER-1	0.7 (12) ± 0.13	**0.72 (4) ± 0.17**	0.71 (10) ± 0.15	0.68 (10) ± 0.12	**0.72 (12) ± 0.17**	**0.72 (16) ± 0.16**	0.68 (6) ± 0.12
COX2 ECFP6	0.56 (8) ± 0.1	0.55 (14) ± 0.1	0.53 (12) ± 0.13	**0.57 (16) ± 0.08**	0.54 (16) ± 0.1	0.56 (12) ± 0.09	0.56 (14) ± 0.09
DHFR ECFP4	0.59 (16) ± 0.07	**0.66 (12) ± 0.07**	0.58 (4) ± 0.09	0.6 (16) ± 0.08	0.61 (12) ± 0.09	0.61 (16) ± 0.07	0.61 (16) ± 0.09
DHFR ECFP6	0.58 (16) ± 0.09	0.63 (16) ± 0.08	0.56 (4) ± 0.1	0.6 (14) ± 0.09	**0.65 (16) ± 0.09**	0.59 (16) ± 0.09	0.59 (16) ± 0.07
DIGGLE TABLE	0.94 (16) ± 0.05	**0.97 (4) ± 0.03**	0.94 (4) ± 0.05	0.92 (10) ± 0.04	0.95 (16) ± 0.04	0.92 (12) ± 0.04	0.91 (10) ± 0.04
FONTAINE ECFP4	**0.81 (8) ± 0.09**	0.77 (6) ± 0.09	0.74 (4) ± 0.15	0.75 (16) ± 0.08	0.78 (16) ± 0.1	0.77 (10) ± 0.11	0.78 (16) ± 0.1
FONTAINE ECFP6	**0.85 (6) ± 0.11**	0.78 (14) ± 0.08	0.79 (4) ± 0.09	0.79 (8) ± 0.1	0.76 (14) ± 0.1	0.78 (8) ± 0.1	0.76 (16) ± 0.14
GORDON 2002	0.95 (8) ± 0.01	**0.97 (14) ± 0.01**	0.93 (10) ± 0.16	0.96 (12) ± 0.01	0.92 (16) ± 0.01	**0.97 (14) ± 0.02**	0.96 (16) ± 0.02
IRIS	0.83 (16) ± 0.13	**0.86 (14) ± 0.11**	**0.86 (8) ± 0.15**	**0.86 (12) ± 0.11**	0.86 (14) ± 0.11	0.85 (4) ± 0.12	0.85 (4) ± 0.12
M1 ECFP4	0.78 (10) ± 0.04	0.77 (16) ± 0.04	0.73 (4) ± 0.04	0.76 (8) ± 0.05	**0.81 (16) ± 0.04**	0.76 (6) ± 0.04	0.76 (4) ± 0.03
M1 ECFP6	**0.76 (8) ± 0.04**	**0.76 (16) ± 0.04**	0.73 (4) ± 0.02	0.74 (10) ± 0.03	**0.76 (16) ± 0.04**	0.74 (14) ± 0.03	0.75 (4) ± 0.04
MFEAT-FACTORS	**0.89 (4)** ± 0.03	0.87 (16) ± 0.04	0.87 (4) ± 0.03	0.87 (14) ± 0.04	0.87 (16) ± 0.04	0.88 (12) ± 0.03	0.88 (6) ± 0.04
MFEAT-KARHUNEN	0.89 (4) ± 0.03	**0.90 (16) ± 0.01**	0.87 (4) ± 0.04	0.88 (10) ± 0.04	0.89 (16) ± 0.04	0.89 (8) ± 0.03	0.89 (8) ± 0.04
SEEDS	0.84 (14) ± 0.06	0.83 (14) ± 0.06	**0.85 (6) ± 0.09**	0.84 (16) ± 0.06	0.84 (16) ± 0.06	0.84 (12) ± 0.06	**0.85 (14) ± 0.06**
SEGMENTATION-NORMCOLS	0.79 (16) ± 0.03	0.79 (14) ± 0.03	0.78 (4) ± 0.04	**0.8 (16) ± 0.03**	0.77 (14) ± 0.03	**0.80 (16) ± 0.04**	**0.80 (14) ± 0.03**
SEMEION	0.69 (6) ± 0.04	0.72 (16) ± 0.04	0.67 (4) ± 0.06	0.67 (12) ± 0.04	**0.73 (14) ± 0.03**	0.71 (8) ± 0.03	0.7 (10) ± 0.04
STOCK	**0.87 (16) ± 0.03**	0.85 (16) ± 0.04	0.84 (6) ± 0.04	0.82 (10) ± 0.05	0.85 (12) ± 0.04	**0.87 (16)** ± 0.03	0.84 (4) ± 0.04
TRANSPLANT	0.87 (6) ± 0.07	0.86 (10) ± 0.1	0.78 (6) ± 0.16	0.86 (12) ± 0.08	0.87 (16) ± 0.08	**0.89 (10) ± 0.08**	0.88 (10) ± 0.1
WDBC	0.92 (12) ± 0.02	0.92 (10) ± 0.04	0.92 (4) ± 0.04	0.92 (12) ± 0.02	0.91 (10) ± 0.04	**0.93 (16) ± 0.02**	0.91 (8) ± 0.04
WINE-187	0.82 (16) ± 0.14	0.82 (16) ± 0.15	0.82 (16) ± 0.14	**0.84 (16) ± 0.13**	0.81 (16) ± 0.15	0.83 (16) ± 0.14	**0.84 (16) ± 0.13**
YEAST GALACTOSE	0.97 (14) ± 0.02	**0.98 (16) ± 0.02**	0.93 (6) ± 0.05	0.96 (16) ± 0.04	**0.98 (16) ± 0.03**	**0.98 (14)** ± 0.01	**0.98 (16)** ± 0.01

The distribution of best-performing methods in Table II indicates an interesting balance among algorithms; each method (excluding MST$_R$) achieved optimal F-measure in approximately eight datasets on average. This equilibrium suggests that no single method universally dominates, emphasizing the necessity of including all algorithms in comparative analyses. We employed the Friedman test for comparison of multiple graph construction methods over multiple datasets [Demšar 2006] and obtained a p-value lower then 0.05, indicating statistical difference between pair of algorithms. To access these statistical significance, we followed by Nemenyi post-hoc analysis.

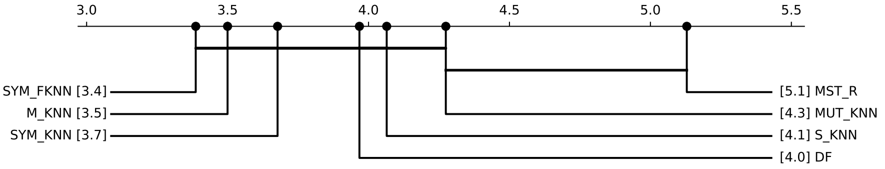

Fig. 2. Nemenyi post-hoc test results for 2% labeled data.

Figure 2 summarizes the average ranks of graph construction methods across all datasets, utilizing 2% labeled data for label propagation. The relative ranking of methods remained consistent across all tested labeling percentages (5%, 8%, and 10%), with only minor variations in critical distance thresholds. The Critical Distance (CD) diagram identifies SYM_FKNN as the top-performing method, although it shows no statistically significant difference ($\alpha = 0.05$) from M_KNN, SYM_KNN, DF, S_KNN, and MUT_KNN. This outcome positions SYM_FKNN as a robust baseline for semi-supervised learning workflows, indicating its strong initial performance without precluding the other methods from comprehensive semi-supervised analysis. On the other hand, MST$_R$ demonstrated the weakest performance, ranking significantly lower than most methods. This discrepancy may arise from MST$_R$'s sensitivity to datasets lacking clear cluster structures, a hypothesis further investigated in subsequent sections through cluster quality and graph-theoretic properties.

Table III. Spearman coefficient ($p_{\text{value}} < 0.05$) obtained from comparisons between the cluster quality and graph-theoretic properties, and the F-measure results presented in Table II for 2% of labeled data.

	DF	**MST$_R$**	**MUT_KNN**	**M_KNN**	**SYM_FKNN**	**SYM_KNN**	**S_KNN**
Max Degree				-0.42			
Diameter				0.36	0.43		0.43
DBCV	0.42	0.50	0.42	0.48	0.48	0.46	0.48

Although its original conception was from outside the area of semi-supervised learning, the Disparity Filter (DF) demonstrated competing performance in the semi-supervised task, exhibiting no statistical difference from SYM_FKNN. This outcome shows its potential as a valuable addition to the state-of-the-art shelf algorithms for graph construction in semi-supervised learning.

Table III presents the significant ($p_{value} < 0.05$) Spearman's rank correlation coefficients between the F-measure average performance of the label propagation step (results summarized in Table II) and (i) dataset properties and (ii) structural characteristics of graphs generated by construction methods. For all methods, we observed moderate correlations ($\rho \in [0.4, 0.5]$) between F-measure and the Density-Based Cluster Validity (DBCV) index. This suggests that these methods perform effectively on datasets possessing well-defined cluster structures, aligning with the cluster assumption principle in semi-supervised learning [Chapelle et al. 2006], which posits that decision boundaries should reside in low-density regions. M_KNN also exhibits a negative moderate correlation ($\rho = -0.42$) with maximum vertex degree. This implies that while inherent cluster structure enhances its performance, the presence of hub nodes (vertices with exceptionally high degrees) negatively impacts label propagation efficacy. [Berton et al. 2018] corroborate such findings, who attributed similar topology-sensitive behavior to neighborhood saturation in dense graphs. S_KNN and SYM_FKNN also demonstrate positive correlations with graph diameter ($\rho = 0.43$), which indicates robust performance under conditions of both sparse connectivity (where larger diameters mitigate over-smoothing) and inherent cluster structures. Such versatility suggests that they adapt effectively to diverse topological regimes, balancing local and global graph properties.

5. CONCLUSIONS

We investigated the influence of various graph construction methods on the performance of transductive semi-supervised classification algorithms. Our findings show that the choice of graph structure indeed impacts the efficacy of label propagation. Specifically, methods such as the MST$_R$ and the DF were explored in this context as novel graph construction approaches. While SYM_FKNN consistently exhibited robust performance, often ranking among the top methods, the DF, despite not originally for semi-supervised learning, demonstrated competing results, showing no statistical difference from the leading methods. On the other hand, MST$_R$ displayed lower performance, particularly its sensitivity to datasets lacking well-defined cluster structures. Spearman's correlation between the F-measure and the DBCV index may support one direction of this behavior.

Future work aims to investigate the performance of MST$_R$, particularly its observed sensitivity to dataset structure. One possible outcome would be exploring modifications or adaptive strategies to enhance its robustness for semi-supervised classification tasks. Besides, a more exhaustive evaluation of the Disparity Filter's applicability across an even broader spectrum of datasets and problem domains would be important to ascertain its generalizability capability. Beyond these specific methods, other backbone extraction techniques could be investigated in graph construction methods for transductive semi-supervised learning.

REFERENCES

Berton, L., de Andrade Lopes, A., and Vega-Oliveros, D. A. A comparison of graph construction methods for semi-supervised learning. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018.* IEEE, pp. 1–8, 2018.

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM TKDD* 10 (1): 1–51, 2015.

Chapelle, O., Scholkopf, B., and Zien, A. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006.

de Sousa, C. A. R., Rezende, S. O., and Batista, G. E. A. P. A. Influence of graph construction on semi-supervised learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezný (Eds.). Lecture Notes in Computer Science, vol. 8190. Springer, pp. 160–175, 2013.

de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B., and Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* vol. 9, pp. 1–14, 2008.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *JMLR* vol. 7, pp. 1–30, 2006.

Fontaine, F., Pastor, M., Zamora, I., and Sanz, F. Anchor-grind: Filling the gap between standard 3d qsar and the grid-independent descriptors. *Journal of Medicinal Chemistry* 48 (7): 2687–2694, 2005. PMID: 15801859.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* 32 (200): 675–701, 1937.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo-Meullenet, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* 45 (Database-Issue): D945–D954, 2017.

Gertrudes, J. C., Zimek, A., Sander, J., and Campello, R. J. G. B. A unified framework of density-based clustering for semi-supervised classification. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management, SSDBM 2018, Bozen-Bolzano, Italy, July 09-11, 2018*, D. Sacharidis, J. Gamper, and M. H. Böhlen (Eds.). ACM, pp. 11:1–11:12, 2018.

Lichman, M. UCI machine learning repository, 2013.

Liu, W. and Chang, S. Robust multi-class transductive learning with graphs. In *Proc. IEEE CVPR*. pp. 381–388, 2009.

Menczer, F., Fortunato, S., and Davis, C. A. *A first course in network science.* Cambridge University Press, 2020.

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., and Sander, J. Density-based clustering validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, M. J. Zaki, Z. Obradovic, P. Tan, A. Banerjee, C. Kamath, and S. Parthasarathy (Eds.). SIAM, pp. 839–847, 2014.

Naldi, M. C., Campello, R. J. G. B., Hruschka, E. R., and Carvalho, A. C. P. L. F. Efficiency issues of evolutionary k-means. *Appl.iedSoft Computing* 11 (2): 1938–1952, 2011.

Nemenyi, P. B. *Distribution-free multiple comparisons.* Ph.D. thesis, Princeton University, 1963.

Ozaki, K., Shimbo, M., Komachi, M., and Matsumoto, Y. Using the mutual k-nearest neighbor graphs for semi-supervised classification on natural language data. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, S. Goldwater and C. D. Manning (Eds.). ACL, pp. 154–162, 2011.

Serrano, M. Á., Boguñá, M., and Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences* 106 (16): 6483–6488, 2009.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45 (4): 427–437, 2009.

Sutherland, J. J., O'Brien, L. A., and Weaver, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med.Chem.* 47 (22): 5541–5554, 2004.

Vanschoren, J., Van Rijn, J. N., Bischl, B., and Torgo, L. Openml: networked science in machine learning. *ACM SIGKDD Expl. Newsletter* 15 (2): 49–60, 2014.

Vega-Oliveros, D. A., Berton, L., Eberle, A. M., de Andrade Lopes, A., and Zhao, L. Regular graph construction for semi-supervised learning. In *Journal of physics: Conference series.* Vol. 490. IOP Publishing, pp. 012022, 2014.

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. Clustering gene-expression data with repeated measurements. *Genome Biol* 4 (5): R34, 2003.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, S. Thrun, L. K. Saul, and B. Schölkopf (Eds.). MIT Press, pp. 321–328, 2003.

Zhu, X. Semi-supervised learning literature survey — tr1530. Tech. rep., University of Wisconsin, Madison, 2005.