Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

Article - e001

EFFECTS OF CONSTRAINT QUALITY AND DISTRIBUTION ON SEMI-SUPERVISED CLUSTERING PERFORMANCE

Tianshu Yang[1], **Nicolas Pasquier**[2]

[1,2]Laboratoire I3S, Université Côte d'Azur, CNRS, 2000 route des Lucioles, 06903 Sophia Antipolis, France

Received: 13/02/2026 **Revised**: 13/03/2026 **Accepted**: 18/03/2026

ABSTRACT

Semi-supervised clustering enhances traditional clustering algorithms by leveraging prior knowledge, such as class labels or pairwise constraints, to improve clustering quality. Despite numerous proposed methods, the effects of constraint set quality and distribution, particularly imbalanced or erroneous constraints, remain poorly understood. This paper investigates these issues through systematic simulations, evaluating six state-of-the-art semi-supervised clustering algorithms under imbalanced or incorrect constraint conditions. Experiments on multiple real-world and syn-thetic benchmark datasets reveal that must-link constraints generally provide greater benefit than cannot-link constraints, while relying solely on cannot-link constraints may reduce performance. Furthermore, incorrect constraints, especially erroneous must-link constraints, can significantly degrade clustering performance. These findings underscore the critical importance of carefully designing constraint sets to achieve reliable and effective semi-supervised clustering, providing clear guidance on the conditions under which different algorithmic approaches perform best.

KEYWORDS: Machine Learning, Data Mining and Knowledge Discovery, Semi-Supervised Clustering, Imbalanced Constraints, Inaccurate Constraints, Robustness Analysis

1. INTRODUCTION

The objective of clustering methods is to partition data into homogeneous groups, called clusters, such that the intra-cluster variability is minimal, while the inter-cluster variability is maximal [1, 2]. In other words, instances within the same cluster should exhibit higher levels of similarity to each other than to those in disparate clusters. Clustering methods are frequently utilized as unsupervised learning approaches for the analysis of datasets without supervisory information [3]. When some prior knowledge about the properties of the underlying data space clusters, such as limited instance class labels or constraints on clustering instances, is available, semi-supervised clustering can be utilized. Semi-supervised clustering, also known as constrained clustering, integrates prior knowledge to guide the

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

clustering process and recent studies introduced different approaches to effectively utilize available supervisory information and identify more significant clusters for users [4, 5].

Among the different types of prior knowledge considered in the literature, the most common form are must-link and cannot-link constraints between two instances [6]. A must-link constraint states that the two instances should be assigned to the same cluster. A cannot-link constraint states that the two instances should be assigned to different clusters. Different categories of pairwise constraints based semi-supervised clustering approaches can be distinguished. In the largest category, approaches extend classical clustering algorithms, typically the K-means algorithm, to integrate constraints [7, 6, 8, 9, 10, 11]. In another category, approaches deduce new metrics for the grouping of instances from pairwise constraints [12, 13, 14, 15] or use a declarative framework to model the problem [16, 17, 18, 19]. In a third category, approaches that combine several clustering algorithms are extended to integrate pairwise constraints into a collaborative process [20] or an ensemble process [21, 22, 23, 24]. The fourth category includes approaches that integrate pairwise constraints in a deep learning unsupervised process [25, 26].

Although a number of research works have proposed variants of semi-supervised clustering, the impact of unbalanced and erroneous constraints on semi-supervised clustering has been scarcely studied. In most studies, the input set of constraints is generated based on the ground-truth labeling of instances and is balanced in terms of cannot-link and must-link constraint numbers. The situation where there are incorrect constraints, also referred to as noisy or mis-specified constraints in the literature, and the distribution of constraints is imbalanced, which has a greater practical applicability [27, 28], is barely considered.

An incorrect must-link constraint erroneously states that two instances must be assigned to the same cluster, and an incorrect cannot-link constraint erroneously states that two instances must be assigned to different clusters. In numerous areas of study, experimental outcomes can be imprecise as a result of human errors, technical issues, or noise interference. These outcomes may subsequently result in erroneous constraints, regardless of whether they are generated automatically or specified by the user [28, 29, 30]. In biological studies involving experimental grouping of genes or proteins, it is frequently useful to evaluate associations between individual pairs, but it is not certain that the guidance produced by this method will be accurate [31, 32]. In multimedia resource classification applications, such as documents, images, and web pages, connectivity constraints are often generated from annotations and tags of resources that are automatically or manually defined. Such annotations and tags are prone to contain errors or misspecifications due to data and terminology ambiguity or polysemy, inadvertent human errors, or even deliberate mistakes by malicious users, leading to incorrect constraints [33, 28]. Recognizing and locating objects in a multidimensional space using a network of sensors, such as video cameras, is a key operation in many domains, such as surveillance applications, individual identification, autonomous navigation, and robot guidance, for example. Pairwise constraints, obtained automatically or from human feedback, that improve the performance of video object classification from video streams are also subject to association errors [34, 35].

Imbalanced constraints refer to the situation where the distribution of constraints is skewed; that is, the sets of must-link and cannot-link constraints are not equal in size. This situation can arise in certain applications if the classes are significantly imbalanced, or depending on the technical parameters of the application when the constraints are automatically

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

generated [36, 37, 38]. Class imbalance is prevalent in many application domains such as medical studies, fraud detection, spam filtering, credit risk assessment, finance, surveillance, and natural language processing, among others [39, 40, 41]. In extreme cases, such as paraphrase detection and open-domain question answering, the majority class can represent 99.99% of data instances [38] and sometimes even more [39, 40]. Classic supervised learning solutions, such as minority class oversampling or majority class undersampling, whose applicability depends on the amount of data available, do not resolve this issue for the generation of balanced pairwise constraints in semi-supervised learning [37]. In some applications, the must-link and cannot-link constraints are utilized at different levels or have different roles in the process [42]. This may result in employing imbalanced sets of constraints or potentially relying solely on one type of constraint. An example is resource classification applications, such as documents or web pages, using user-defined tags where each resource can belong to several groups or classes. Depending on the scope of the tagging and the user's interests, the tags can be used to identify resources relating to identical themes, but they cannot ensure the opposite [43].

Among the few articles dealing with this topic, [9] analyzes the complexity of their semi-supervised clustering approach in the case that the constraint set is skewed. [44] points out the impact of the constraint set on performance, demonstrating that some constraint sets can actually decrease algorithm accuracy, leading to further work that focuses on how to select informative and/or "easy" constraint sets [45]. Few publications discuss the impact of incorrect constraints, among which we can cite [28, 31, 29, 30], and the impact of an incorrect number k of clusters [30]. In addition, [27] and [46] report the behavior of semi-supervised clustering algorithms in the presence of incorrect constraints.

In this paper, we conduct a simulation-based study of the practical scenarios where the input, represented by pairwise constraints, is imbalanced or inaccurate. This implies that the constraint set deviates from uniformity or contains erroneous constraints. We analyze the robustness and accuracy of state-of-the-art semi-supervised clustering approaches to highlight the scenarios in which each approach is more suitable. The remainder of the paper is organized as follows. Section 2 introduces the state-of-the-art semi-supervised clustering approaches used in the article. In Section 3, we present the datasets, the strategy for generating imbalanced and incorrect constraint sets, and the experimental design. Section 4 reports the experimental results on five real-world benchmark datasets from the UCI repository and their analysis. Section 5 makes use of the Cassini benchmark dataset to visualize and analyze the impact of different constraint sets on the results of semi-supervised clustering approaches. Section 6 concludes the article with the essential findings of the study.

2. SEMI-SUPERVISED CLUSTERING APPROACHES

In this section, we present the six state-of-the-art semi-supervised clustering algorithms used in the experiments. These include COP-Kmeans [6], PC-Kmeans and MPC-Kmeans [8] variants of K-means that integrate constraints into the classical K-means algorithm, and RCA [47], MMC [12] and ITML [14] metric learning methods that use constraints to learn distance functions. We first briefly introduce the K-means algorithm.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

2.1 K-means algorithm

The K-means clustering algorithm [48] is commonly used to automatically divide a dataset into k clusters, where k is a user-define parameter. It begins with randomly selecting k instances as initial cluster centers, called centroids. The clusters are then iteratively refined through a two-stage process:

1. Each instance i is assigned to its nearest cluster centroid.
2. Each cluster centroid is updated as the average of its constituent instances.

The algorithm converges when there is no further change in the allocation of instances to clusters.

The choice of initial centroids has a significant influence on the convergence and final results of the algorithm. Poor initialization can lead K-means to converge to an undesirable local minimum, resulting in sub-optimal clusters. Initializing clusters using constraints, by ensuring that the initial centroids respect the constraints, is studied in [49]. For example, the initialization can prohibit multiple instances connected by must-link constraints from being used as initial centroids or select instances connected by cannot-link constraints as initial centroids. If the experimental results of using constraints for cluster initialization tend to show an improvement in outcomes, this effect may be attenuated in the case of incorrect constraints.

2.2 Constrained K-means

Constrained K-means clustering (COP-Kmeans) was proposed in [6] and is one of the most prominent constrained clustering algorithms. The idea is to ensure that none of the constraints is violated. An instance i is assigned directly to a cluster if the cluster contains an instance that has a must-link constraint with i. Otherwise, i will be assigned to the cluster with the closest centroid, excluding those clusters that contain an instance that has a cannot-link constraint with i. If a legal cluster cannot be found, then an empty partition is returned, meaning that the instance is excluded from the clustering result. The clusters are initialized using instances from the dataset that are randomly selected as cluster centroids, as in the K-means algorithm.

COP-Kmeans is a simple and efficient approach. However, since COP-Kmeans enforces the satisfaction of each constraint, an incorrect constraint will result in an inaccurate output. COP-Kmeans can also be sensitive to instance assignment order. Once an instance is assigned to a cluster, all the other instances that have a must-link constraint with it will be assigned to the same cluster. Thus, a different assignment order may end up in completely different clustering results for these instances. When an instance cannot be assigned to any cluster due to having a cannot-link constraint with instances in each cluster, COP-Kmeans returns an empty partition.

2.3 Pairwise Constrained K-means and Metric Pairwise Constrained K-means

Based on the idea of using constraints in the K-means algorithm, [8] proposed Pairwise Constrained K-means (PC-Kmeans) and Metric Pairwise Constrained K-means (MPC-Kmeans). PC-Kmeans utilizes constraints to seed the initial clusters and directs the cluster assignments to respect these constraints. The connected components, which consist of instances connected by must-link constraints, are taken as initial cluster centroids. The objective function is

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

formulated as the sum of the total squared distances between instances and their cluster centroids, and the penalty incurred by violating any constraints. During the cluster assignment step, an instance will be assigned to the cluster centroid that minimizes the objective function.

MPC-Kmeans involves cluster initialization, cluster assignment, and metric learning in a unified framework. The definition of the objective function and cluster initialization are the same as in PC-Kmeans. The distance metric is adapted by re-estimating the weight matrices during each iteration based on the current cluster assignments and constraint violations.

PC-Kmeans and MPC-Kmeans never return empty partitions as COP-Kmeans. However, as variants of the K-means algorithm, they can also be order-sensitive. The complexity of MPC-Kmeans is critical [46] as MPC-Kmeans updates the weight matrices during each iteration of the clustering process.

2.4 Relevant Components Analysis

Relevant Components Analysis (RCA) [47] is one of the earliest methods that integrates constraints in metric learning. It uses the notion of *chunklet*, which is essentially a connected component, to compute a global linear transformation to assign large weights to relevant dimensions and low weights to irrelevant dimensions [50]. This transformation is based on chunklet information only and does not use any cannot-link constraints. The learned metric is used in the K-means algorithm to calculate the distance of each instance from the centroid of each cluster, in order to assign that instance to the nearest centroid cluster. Clusters are randomly initialized as in the K-means algorithm.

Although the RCA approach was initially developed for image retrieval, it has since been used for many other applications in a large variety of fields.

2.5 Mahalanobis Metric for Clustering

Mahalanobis Metric for Clustering (MMC) [12] aims to minimize the sum of Mahalanobis distances between instances linked by must-link constraints, and at the same time enforce the distances between instances linked by cannot-link constraints to be greater than a constant (usually set to 1). This distance metric is trained using convex optimization, and thus the training process is free from local optimization. The learned metric is used in the K-means algorithm to calculate the distance between each instance and each cluster centroid. Initial centroids are randomly chosen by the K-means algorithm.

The computation of eigenvalues during the metric learning step can sometimes be time-consuming. Another restriction is its idealistic assumption that all clusters follow a unimodal distribution. In addition, MMC has been reported to have some uncertainty about the optimality of the final solution. In fact, the MMC gradient-based algorithm proposed requires the adjustment of several parameters and is not guaranteed to find the optimum solution without such an adjustment [13].

2.6 Information-Theoretic Metric Learning

Similar to MMC, the Information-Theoretic Metric Learning (ITML) approach [14] aims to learn an optimal Mahalanobis distance subject to constraints. It bijects the Mahalanobis distance to an equal mean multivariate Gaussian distribution. In this way, the problem is

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

translated into minimizing the differential relative entropy, also known as the Kullback-Leibler divergence, between two multivariate Gaussian distributions under constraints on the distance function. The problem is then expressed as a particular Bregman optimization problem by minimizing the LogDet divergence subject to linear constraints. As with other metric learning approaches, the learned metric is used in the K-means algorithm to calculate distances between instances and centroids, with the centroids initialized at random.

Unlike some other metric learning methods, in ITML, no eigenvalue computation or semi-definite programming is required. It can also handle a wide variety of constraints and can optionally incorporate prior knowledge about the distance function. However, a simple bijection between the Mahalanobis distance and the equal-mean multivariate Gaussian distribution over-simplifies the underlying metric structure. In practice, there will often be no feasible solution to the general ITML problem, particularly when the number of constraints is large, as reported in [51].

3. EXPERIMENTAL SETTING

The experimental evaluation is conducted using a selection of well-established real-world bench-marks from the UCI repository [52], specifically Iris, Wine, Seeds, Statlog, and Obesity, complemented by the two-dimensional Cassini synthetic dataset [53] to analyze the influence of constraint set properties on clustering outcomes. The selection of these specific benchmarks is motivated by their diverse topological properties, varying levels of class overlap, and differing feature spaces. While the Iris, Wine, and Seeds datasets represent classic low-to-medium dimensional problems with well-defined clusters, Statlog and Obesity introduce higher dimensionality and greater complexity in feature correlation. Furthermore, the inclusion of the synthetic Cassini dataset provides a rigorous test for the algorithms' ability to resolve non-convex structures and complex spatial boundaries in a controlled two-dimensional environment. The inclusion of the Statlog and Obesity datasets is particularly critical for evaluating algorithmic robustness under more challenging conditions. Unlike the well-separated classes found in simpler benchmarks, these datasets present significant class overlap and high-dimensional feature spaces. Specifically, the Obesity dataset introduces a mixture of categorical and numerical variables, reflecting the complexities of real-world health data where decision boundaries are often poorly defined.

Meanwhile, the Statlog benchmark—characterized by its multi-class nature and high degree of at-tribute correlation—serves to test the scalability and discriminative power of the semi-supervised constraints. These attributes ensure that the experimental results are not merely confined to idealized scenarios but are representative of performance in complex, noisy environments. The main characteristics of these datasets, the maximum number of constraints used in the experiments and the proportions of must-link and cannot-link constraints used in the experiments are presented in Table 1.

A primary rationale for selecting these specific datasets lies in their balanced class distributions. In the presence of a minority class, a significant discrepancy often arises between the structural density of the data and the predefined number of clusters, k. Such imbalances can introduce ambiguity during the parameterization of semi-supervised approaches, potentially con-founding the evaluation of the clustering process. By employing balanced datasets, we ensure that the experimental outcomes exclusively reflect the impact of constraint set proper-

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

ties—specifically their imbalance and inaccuracy—rather than being skewed by the underlying class distribution. This methodological choice allows for a more granular analysis of how varying levels of incorrect or disproportionate constraints affect the performance of each algorithm.

Ultimately, this rigorous selection of balanced benchmarks, encompassing both synthetic spatial challenges and high-dimensional real-world data, ensures the reproducibility and statistical reliability of our experimental framework. By isolating the influence of constraint quality from the inherent distribution of the data, we provide a controlled environment that accurately quantifies the robustness of each semi-supervised algorithm. This foundational consistency is essential for interpreting the performance trade-offs presented in the subsequent analysis.

Table 1: Benchmark dataset properties. For each experimental datasets, the number of classes, the number of variables, the number of instances, the maximum number of constraints used and the proportions of must-link and cannot-link constraints used in the experiments are shown.

Dataset	Number of classes	Number of variables	Number of instances	Maximum constraints	Proportion of constraints must-link/cannot-link (%)
Iris	3	4	150	210	0/0, 100/0, 0/100, 50/50
Wine	3	13	178	210	0/0, 100/0, 0/100, 50/50
Seeds	3	7	210	210	0/0, 100/0, 0/100, 50/50
Cassini	3	2	1,000	30	50/50, 50/50, 90/10, 10/90
Statlog	7	20	2,310	2,000	0/0, 100/0, 0/100, 50/50
Obesity	7	17	2,111	2,000	0/0, 100/0, 0/100, 50/50

The six semi-supervised clustering approaches presented in Section 2 are used for the experiments. Implementations of these semi-supervised clustering approaches are available in the *active-semi-supervised-clustering* [54] and *metric-learn* [55] Python packages. These approaches require an input number of clusters k as a parameter, for which the number of classes in the dataset is given. For the small Iris, Wine and Seeds datasets, the maximum number of iterations, defined by the *max_iter* parameter, is kept at its default value of 10 for the MPC-Kmeans approach and 100 for the other approaches. For the large Statlog and Obesity datasets, it is necessary to increase the maximum number of iterations up to 1,000 for all approaches excluding MPC-Kmeans to ensure convergence. The significant difference in the values of this parameter comes from the nature of the MPC-Kmeans approach, which is the only one that updates the weight matrices at each iteration, resulting in faster convergence.

The pseudocode of the algorithm implemented to generate the constraints is presented in Algorithm 1. The *update_closure()* function used in the algorithm is based on the *preprocess_constraints()* function of the *active-semi*-supervised-*clustering* and *metric-learn* Python packages. This function is used to update the set of constraints by adding new transitive constraints induced when a new must-link or cannot-link constraint is created. It uses a DFS (deep first search) approach to update the graph of constraints, where each node is an instance, and each edge represents a must-link or cannot-link constraint. This algorithm allows conflict-free constraint generation, that is, it ensures that a must-link constraint and a cannot-link constraint do not exist simultaneously between the same two instances.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

Algorithm 1: Generation of must-link and cannot-link constraint sets

Input: dataset *d*, number of true classes *k*, number of must-link constraints *mlink*, number of cannot-link constraints *clink*, number of incorrect must-link constraints *mnoise*, number of incorrect cannot-link constraints *cnoise*

Output: set of must-link constraints *ml_list*, set of cannot-link constraints *cl_list*

```
for each class in d do
    class_instances[i] ← list of instances of the class in d
end for
ml_list ← {}, cl_list ← {}
for j = 1 to mlink do                          // GENERATE MUST-LINK CONSTRAINTS
    ok ← false, p ← 0, q ← 0
    while ok = false do
        randomly choose a class number nclass in [1, k]
        randomly choose instance numbers p and q among instances[nclass]
        if p != q and (p, q) not in ml_list then
            add (p, q) to ml_list
            ml_list ← update_closure(ml_list)          // Add new transitive must-link constraints
            ok ← true
        end if
    end while
end for
for j = 1 to clink do                          // GENERATE CANNOT-LINK CONSTRAINTS
    ok ← false, p ← 0, q ← 0
    while ok = false do
        randomly choose class numbers nclass1 and nclass2 in [1, k] such that nclass1 != nclass2
        randomly choose instance number p among instances[nclass1]
        randomly choose instance number q among instances[nclass2]
        if (p, q) not in cl_list then
            add (p, q) to cl_list
            cl_list ← update_closure(cl_list)          // Add new transitive cannot-link constraints
            ok ← true
        end if
    end while
end for
for j = 1 to mnoise do                 // GENERATE INCORRECT MUST-LINK CONSTRAINTS
    ok ← false, p ← 0, q ← 0
    while ok = false do
        randomly choose class numbers nclass1 and nclass2 in [1, k] such that nclass1 != nclass2
        randomly choose instance number p among instances[nclass1]
        randomly choose instance number q among instances[nclass2]
        if (p, q) not in cl_list and (p, q) not in ml_list then
            add (p, q) to ml_list
            ml_list ← update_closure(ml_list)      // Add new transitive incorrect must-link constraints
            ok ← true
        end if
    end while
end for
for j = 1 to cnoise do                 // GENERATE INCORRECT CANNOT-LINK CONSTRAINTS
    ok ← false, p ← 0, q ← 0
    while ok = false do
        randomly choose a class number nclass in [1, k]
        randomly choose instance numbers p and q among instances[nclass]
        if p != q and (p, q) not in ml_list and (p, q) not in cl_list then
            add (p, q) to cl_list
            cl_list ← update_closure(cl_list)      // Add new transitive incorrect cannot-link constraints
```

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

51: *ok* ← true
52: **end if**
53: **end while**
54: **end for**

The semi-supervised clustering results are evaluated using the standard NMI index score [56]. The NMI score shown in the curves corresponds to the average of the scores obtained for each approach and constraint set tested. To account for the inherent sensitivity of K-means-based algorithms to initial centroid placement, 50 independent trials were conducted for each approach and constraint set, with results subsequently averaged to ensure statistical robustness.

The experiments were carried out on a Dell server with 32 Intel Xeon E5-4620 processors clocked at 2.20 GHz, each with 8 cores. The server, running the Linux CentOS operating system, has 529 GB of RAM and 6.4 TB of hard drives configured in RAID 0 for performance.

4. EXPERIMENTAL RESULTS

The experimental results regarding the impact of imbalanced or incorrect pairwise constraints on the performance of semi-supervised clustering approaches are reported in this section. Building upon the diverse structural and dimensional characteristics of the benchmark datasets used, this section presents a comparative analysis of the experimental outcomes. By subjecting the six semi-supervised algorithms to varying constraint densities across these datasets, we aim to delineate the specific conditions under which each approach maintains its predictive accuracy and structural integrity. Section 4.1 tackles the problem of imbalanced constraint sets in terms of the number of cannot-link and must-link constraints. In Section 4.2, we address the problem of incorrect cannot-link and must-link constraints in the constraint sets.

4.1 Impact of Imbalanced Constraint Sets

During this experiment, we analyze the impact of highly imbalanced constraint sets on semi-supervised clustering approaches. For the small Iris, Wine, and Seeds datasets, the total number of constraints ranges from 0 to 210. For the large Statlog and Obesity datasets, the total number of constraints ranges from 0 to 2000. For each number of constraints, 30 different sets of constraints with different proportions of must-link and cannot-link constraints are generated. The generated constraint sets contain 50% must-link constraints and 50% cannot-link constraints, only must-link constraints, or only cannot-link constraints, to separately investigate their impact on performance.

The results of COP-Kmeans, PC-Kmeans, MPC-Kmeans, RCA, MMC and ITML are presented in Figure 1 for the Iris dataset, in Figure 2 for the Wine dataset, in Figure 3 for the Seeds dataset, in Figure 4 for the Statlog dataset and in Figure 5 for the Obesity dataset. The horizontal axis shows the number of total pairwise constraints used during the run and the vertical axis shows the average NMI index score of the output clustering solution over all trials for each approach.

The vertical axis is normalized to the [0.0,1.0] range for all figures. The blue curve corresponds to the result of unsupervised clustering in which no constraint is used. The orange

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

curve corresponds to the NMI index score of each approach in the case where the number of must-link constraints is equal to the number of cannot-link constraints. The green and red curves illustrate the NMI evaluation for each approach when the constraint set consists exclusively of must-link or cannot-link constraints, respectively. The red curve is not illustrated for the RCA approach because this approach does not utilize cannot-link constraints.

For the Wine and Seeds datasets, the MMC approach does not find the optimal final solution, demonstrating the statement in [13] that the MMC approach requires adjustment of several parameters and is not guaranteed to find the optimum without such adjustment. For the large Statlog and Obesity datasets, the use of constraints with the MMC approach does not improve the results either, and even degrades them when only must-link constraints are provided for Obesity with an average decrease in NMI of 14.7% compared with using no constraints.

When only must-link constraints are provided, the ITML approach fails to find a feasible solution for the Iris, Statlog and Obesity datasets for the largest numbers of constraints, as presented by the green curve. This corresponds to the report in [51], stating that there will often be no feasible solution to the general ITML problem in practice, in particular when the number of constraints is large with regard to the number of instances.

The COP-Kmeans, PC-Kmeans and MPC-Kmeans approaches behave very similarly for all five datasets. For the large Statlog and obesity datasets, must-link constraints significantly im-prove the results if their number is large enough, while cannot-link constraints have no significant impact on the results, regardless of their number. Using the example of the Statlog dataset with the maximal number of must-link constraints, the increase in NMI is 17.1% for COP-Kmeans, 37.4% for PC-Kmeans and 36.1% for MPC-Kmeans compared with using no constraints. For the small Iris, Wine and Seeds datasets, generally similar conclusions can be made, even if we can observe some variations, such as the weak or even negative impact of must-link constraints for COP-Kmeans with the Wine and Seeds datasets, and the negative impact of increasing the number of cannot-link constraints for MPC-Kmeans with the Iris dataset.

The RCA approach, which utilizes only must-link constraints, rapidly attains its optimal outcome as soon as a sufficient number of constraints, somewhat small relative to the size of the dataset, are provided. This minimal number of must-link constraints to attain maximal NMI is 90 for Iris, 125 for Wine, 90 for Seeds and 600 for Statlog and 600 for Obesity. This result remains constant as the number of constraints increases. This observation can be made for all five datasets. RCA gives the best results overall for all five datasets when a small set of must-link constraints only is provided as input.

Comparing the blue curve with the orange, green, and red curves, we can clearly see the negative effect issue in several situations. This is clearly visible for MMC when using must-link or cannot-link constraints with the Wine and Seeds datasets, and when using only must-link constraints with the Obesity dataset, for example. The average decrease in NMI for the Wine dataset is 11.1% for only cannot-link constraints, 15.3% for half cannot-link and must-link constraints, and 32.1% for only must-link constraints, for example. This confirms that the use of pairwise constraints as supervised information in the clustering process can sometimes lead to worse performance than using no constraint. This can occur with semi-supervised clustering approaches in particular when the number of constraints is small as illustrated by MPC-Kmeans with the Wine, Seeds and Obesity datasets for example.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

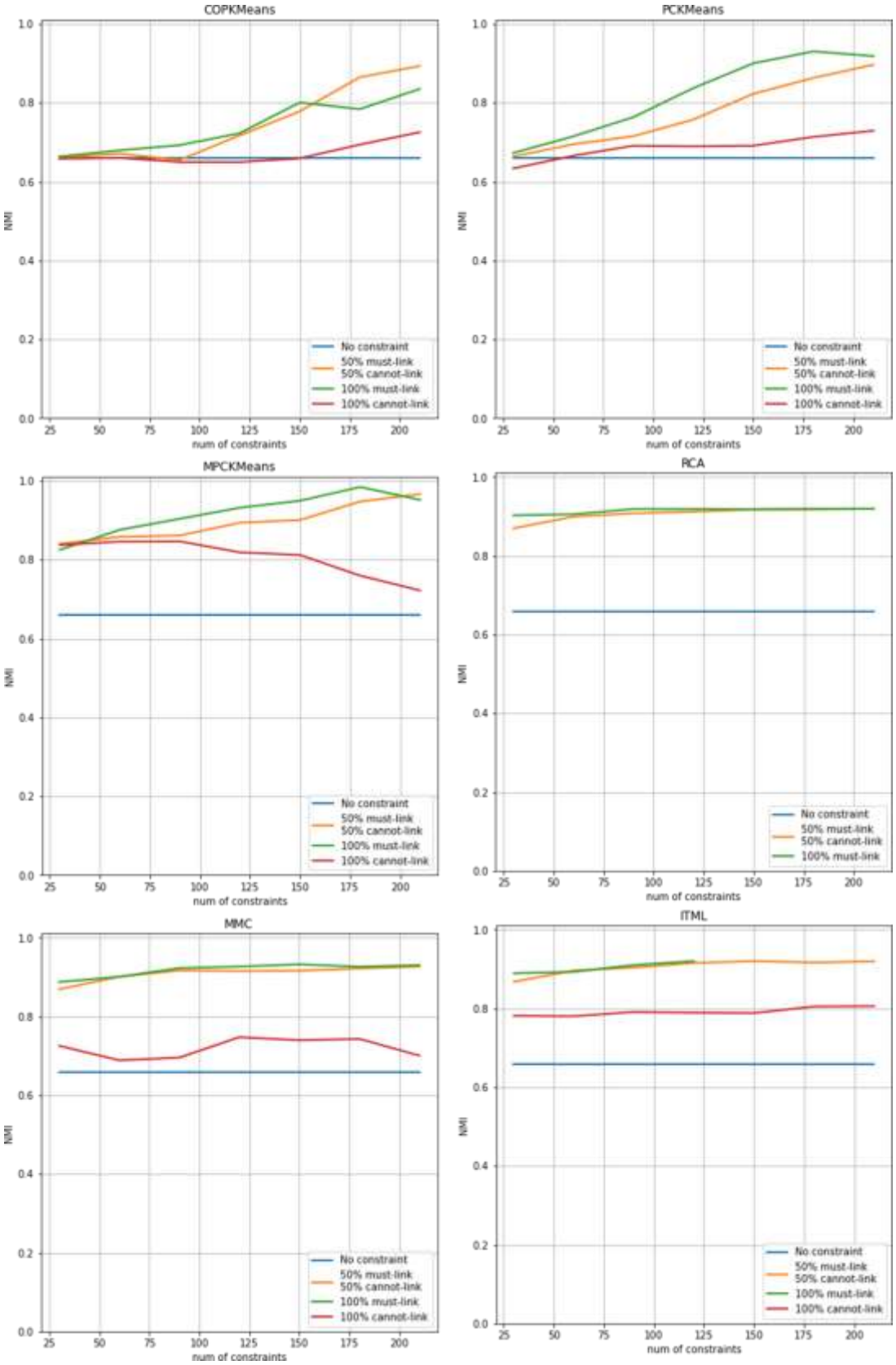

Figure 1: Performance of semi-supervised clustering approaches with imbalanced constraint sets for the Iris dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

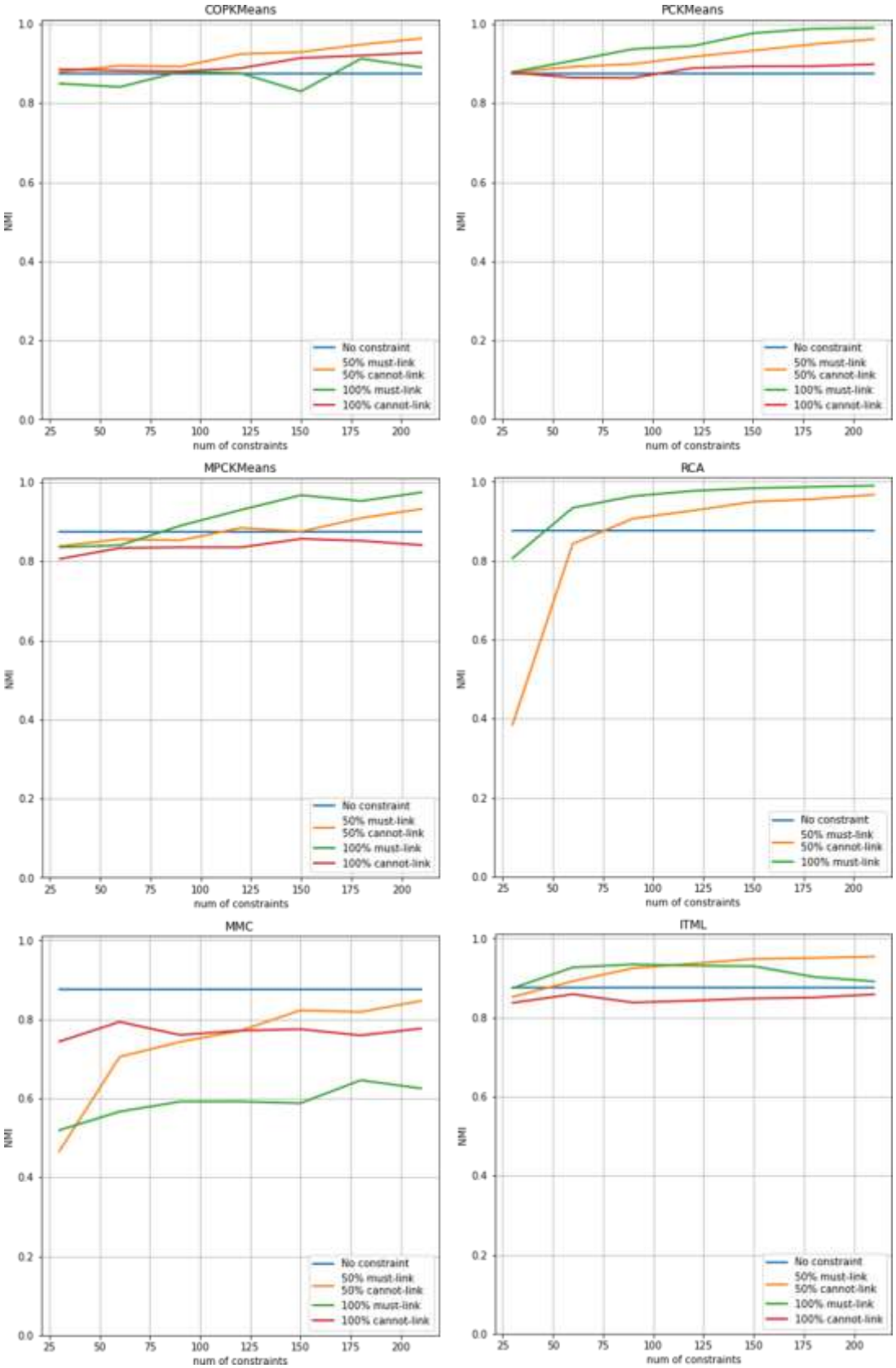

Figure 2: Performance of semi-supervised clustering approaches with imbalanced constraint sets for the Wine dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

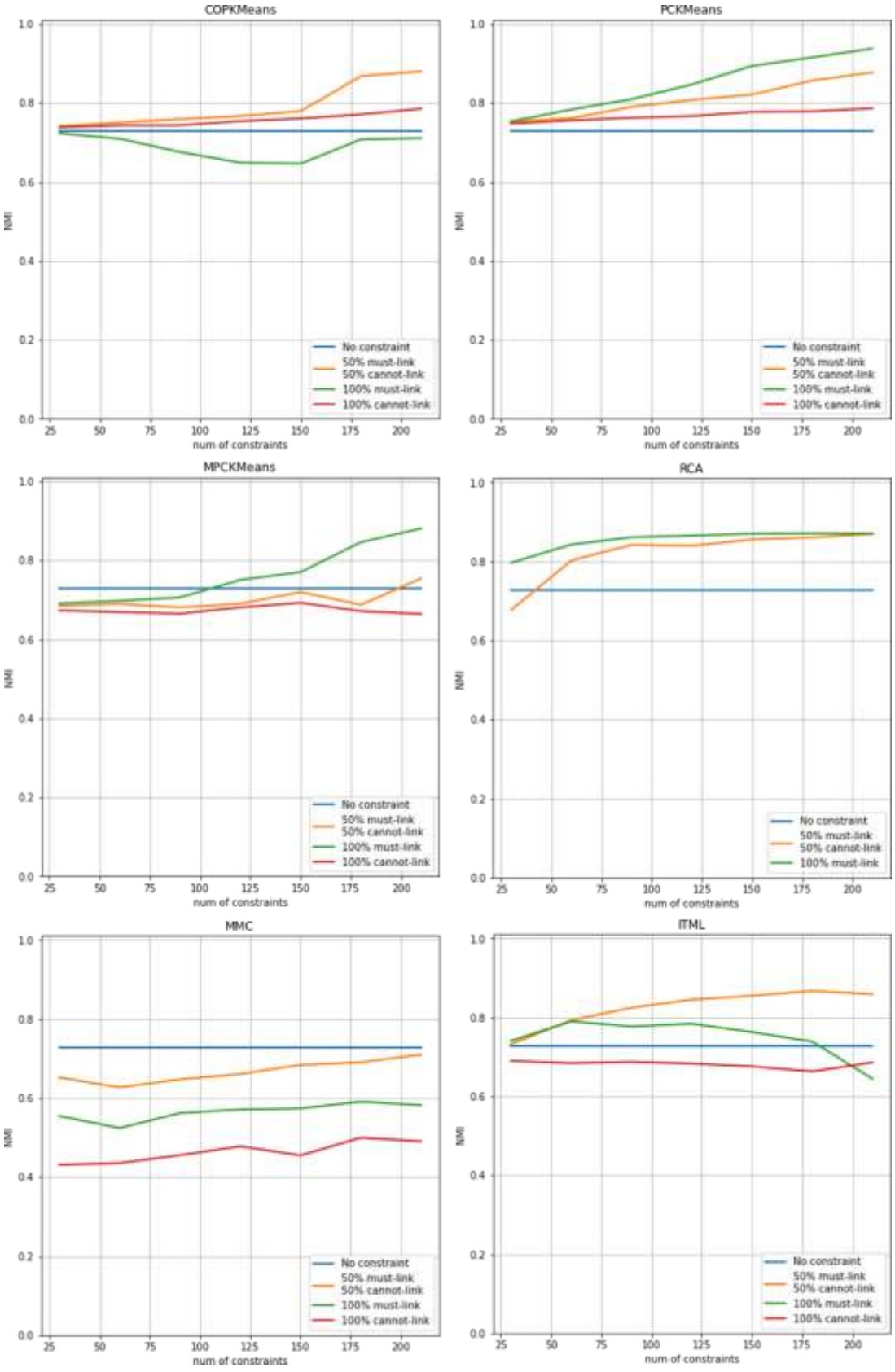

Figure 3: Performance of semi-supervised clustering approaches with imbalanced constraint sets for the Seeds dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

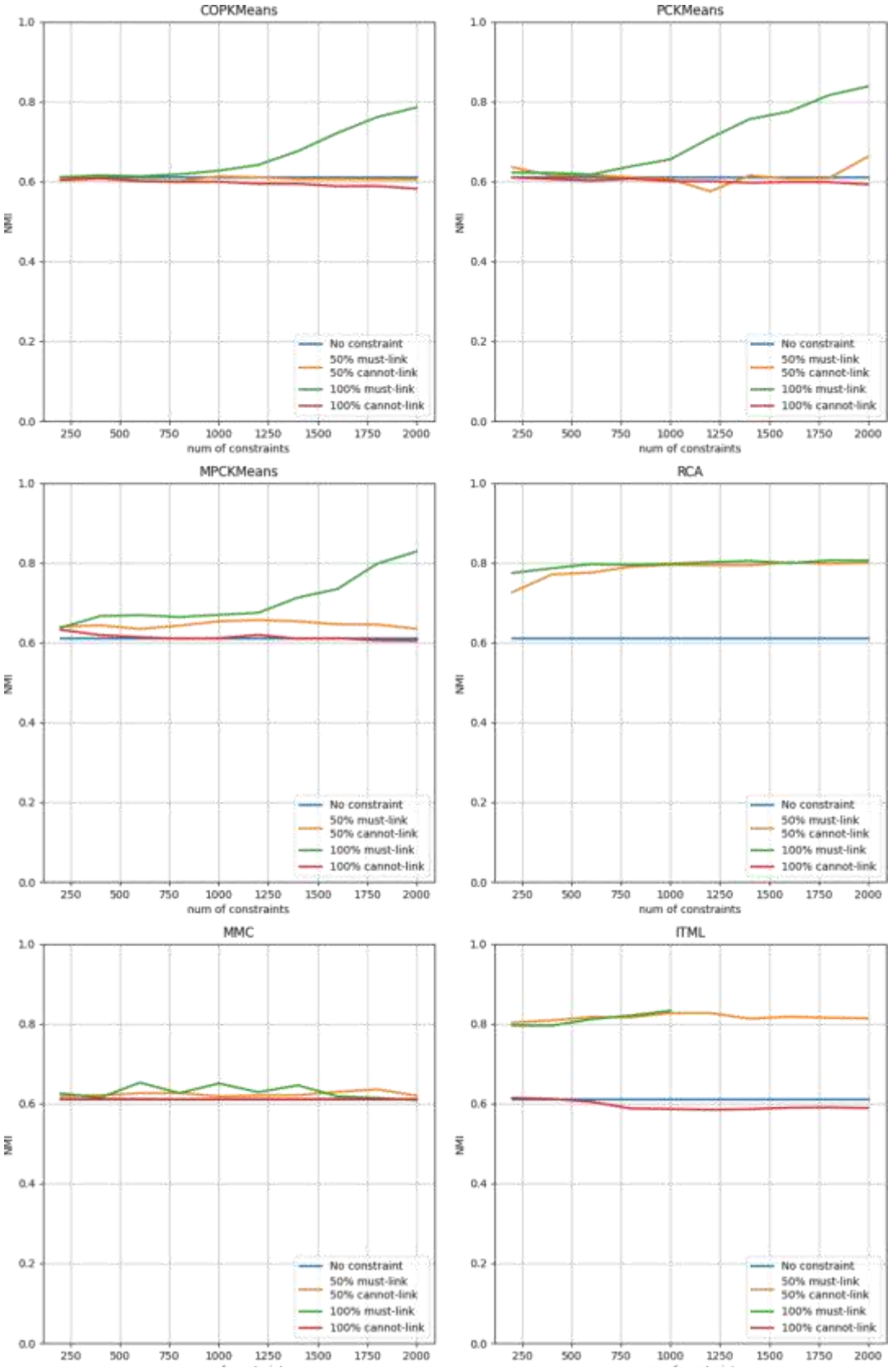

Figure 4: Performance of semi-supervised clustering approaches with imbalanced constraint sets for the Statlog dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

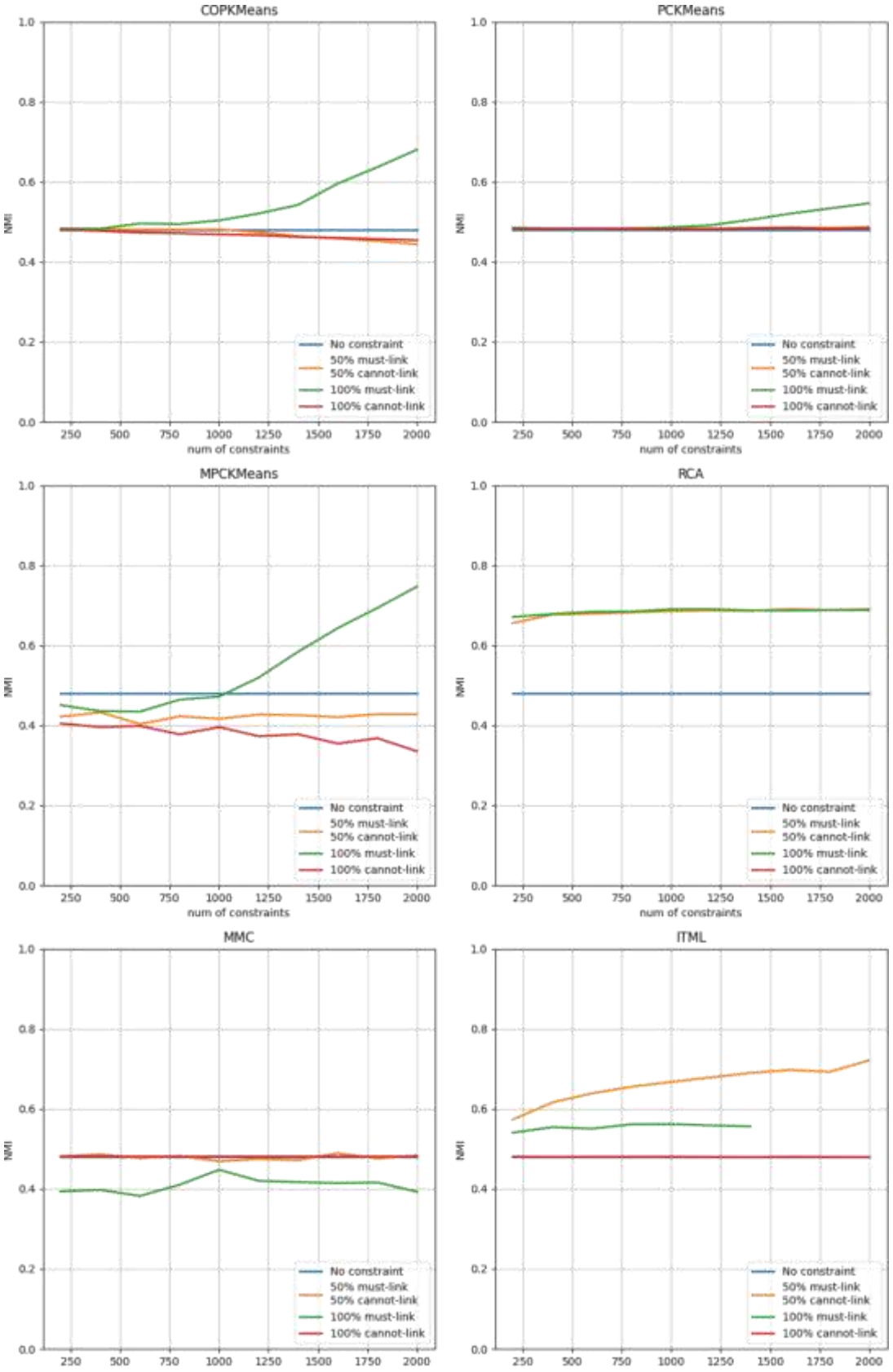

Figure 5: Performance of semi-supervised clustering approaches with imbalanced constraint sets for the Obesity dataset.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

The Seeds dataset reflects well this difficulty with an average decrease in NMI of 8.9% for half cannot-link and must-link constraints, 22.8% for only must-link constraints, and 35.5% for only cannot-link constraints compared with using no constraints. The CGCA and K-Means clustering algorithms tested produced similar results, and the greatest principal component visualization of the data shows a distribution of instances of the three classes in three distinct convex zones of similar sizes in the data space which makes it a suitable data space for the K-Means algorithm. must-link constraints, 22.8% for only must-link constraints, and 35.5% for only cannot-link constraints compared with using no constraints. The CGCA and K-Means clustering algorithms tested produced similar results, and the greatest principal component visualization of the data shows a distribution of instances of the three classes in three distinct convex zones of similar sizes in the data space which makes it a suitable data space for the K-Means algorithm.

Overall, the semi-supervised clustering approaches benefit more from must-link constraints than cannot-link constraints. A pure cannot-link constraint set, as demonstrated by the red curve, sometimes leads to a decrease in performance, especially for metric learning methods MPC-Kmeans, MMC and ITML. In contrast, must-link constraints have a significant positive impact on performance. Most approaches achieve their best performance with only a set of must-link constraints, as illustrated by the green curve. The COP-Kmeans, ITML and MMC approaches are shown to have the ability of making use of cannot-link constraints, since their orange curve has better performance than the green curve in several cases.

The metric learning methods RCA, MMC and ITML converge faster than the COP-Kmeans, PC-Kmeans and MPC-Kmeans K-means variants, as the curves quickly reach peak performance when the size of the constraint set is small. The performance then remains stable, even as the number of constraints increases. However, K-means variants, especially MPC-Kmeans and PC-Kmeans, seem to have a higher NMI index score when the number of constraints is large enough. Among all metric learning methods, the RCA approach generally has the best performance, even if only must-link constraints are used.

For all five datasets, the highest NMI scores are obtained with the PC-Kmeans approach, for Wine, Seeds and Statlog, and the MPC-Kmeans approach, for Iris and Obesity, with the highest numbers of must-link constraints.

4.2 Impact of Incorrect Constraint Sets

In this experiment, the impact of incorrect constraints on semi-supervised clustering approaches is analyzed. For the small Iris, Wine and Seeds datasets, the total number of pairwise constraints is fixed to 100, including 50 must-link constraints and 50 cannot-link constraints. For the large Statlog and Obesity datasets, the total number of pairwise constraints is fixed to 2000, including 1000 must-link constraints and 1000 cannot-link constraints. This guarantees that the semi-supervised clustering approaches have sufficient supervised information to generate an initial clustering solution. The number of incorrect constraints ranges from 0 to 20 for Iris, Wine and Seeds and from 0 to 200 for Statlog and Obesity. For each number of incorrect constraints, 30 different sets of constraints were generated. The generated incorrect constraints may exist equally in must-link constraints and cannot-link constraints, or only in must-link constraints, or only in cannot-link constraints.

The results of COP-Kmeans, PC-Kmeans, MPC-Kmeans, RCA, MMC and ITML are presented in Figure 6 for the Iris dataset, in Figure 7 for the Wine dataset, in Figure 8 for the

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

Seeds dataset, in Figure 9 for the Statlog dataset and in Figure 10 for the Obesity dataset. The horizontal axis shows the total number of incorrect constraints used during the run, and the vertical axis shows the average NMI index score of the output clustering solution over all trials for each approach. The vertical axis is normalized to the [0.0,1.0] range for all plots. The blue curve corresponds to the NMI index score of each approach in the case where the number of incorrect must-link constraints is equal to the number of incorrect cannot-link constraints. The orange and green curves, respectively, represent the NMI evaluation of each approach in the situation where incorrect constraints exist only within must-link constraints or only within cannot-link constraints. The green curve is not illustrated for the RCA approach as this approach does not utilize cannot-link constraints.

For COP-Kmeans, PC-Kmeans and MPC-Kmeans variants of K-means, a similar behavior can be observed when faced with incorrect constraints: Performance decreases linearly as the number of incorrect constraints increases. This decrease is more accentuated for must-link constraints than for cannot-link constraints, as shown by the orange curve of must-link constraints that is below the green curve of cannot-link constraints. However, this decrease is lower, and even null for the Statlog and Obesity datasets, in the case of the PC-Kmeans approach for both types of constraints. Using the example of the small Seeds dataset, for the maximal number of incorrect constraints, that is 20, the decrease in NMI is 34.8% for only cannot-link constraints, 38% for only must-link constraints, and 33.6% for half cannot-link and must-link constraints compared with using no constraints for the COP-Kmeans approach. For the MPC-Kmeans approach, the decrease in NMI is 33.6% for only cannot-link constraints, 41.8% for only must-link constraints, and 38.6% for half cannot-link and must-link constraints compared with using no constraints. For the large Statlog dataset and the maximal number of incorrect constraints, that is 200, the decrease in NMI is 26% for only cannot-link constraints, 51.4% for only must-link constraints and 42.9% for half cannot-link and must-link constraints compared with using no constraints for the COP-Kmeans approach. For the MPC-Kmeans approach, the decrease in NMI is 33.5% for only cannot-link constraints, 57.3% for only must-link constraints, and 46.1% for half cannot-link and must-link constraints compared with using no constraints.

The COP-Kmeans and MPC-Kmeans approaches fail to find a feasible solution for the Obesity dataset for the highest numbers of incorrect must-link constraints. There is no convergence for numbers of incorrect must-link constraints greater than 120 for COP-Kmeans and 160 for MPC-Kmeans for this dataset.

For RCA, MMC and ITML metric learning approaches, performance also decreases as the number of incorrect must-link constraints increases. However, this decrease is visibly lower for Seeds and Statlog than for the other three datasets, suggesting that intrinsic data properties, such as cluster separability for example, can influence the impact of incorrect must-link constraints. The RCA approach fails to find a feasible solution for the Obesity dataset for the highest numbers of incorrect must-link constraints, that is beyond 160 constraints. The MMC and ITML approaches have good robustness against incorrect cannot-link constraints, as illustrated by the green curve which remains constant for all five datasets.

Overall, the PC-Kmeans approach is the most robust in the presence of incorrect must-link constraints. It also offers excellent robustness in the presence of incorrect cannot-link constraints. The ITML approach is the most robust in the presence of incorrect cannot-link constraints, in addition to providing the best clustering results with the highest NMI score for

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

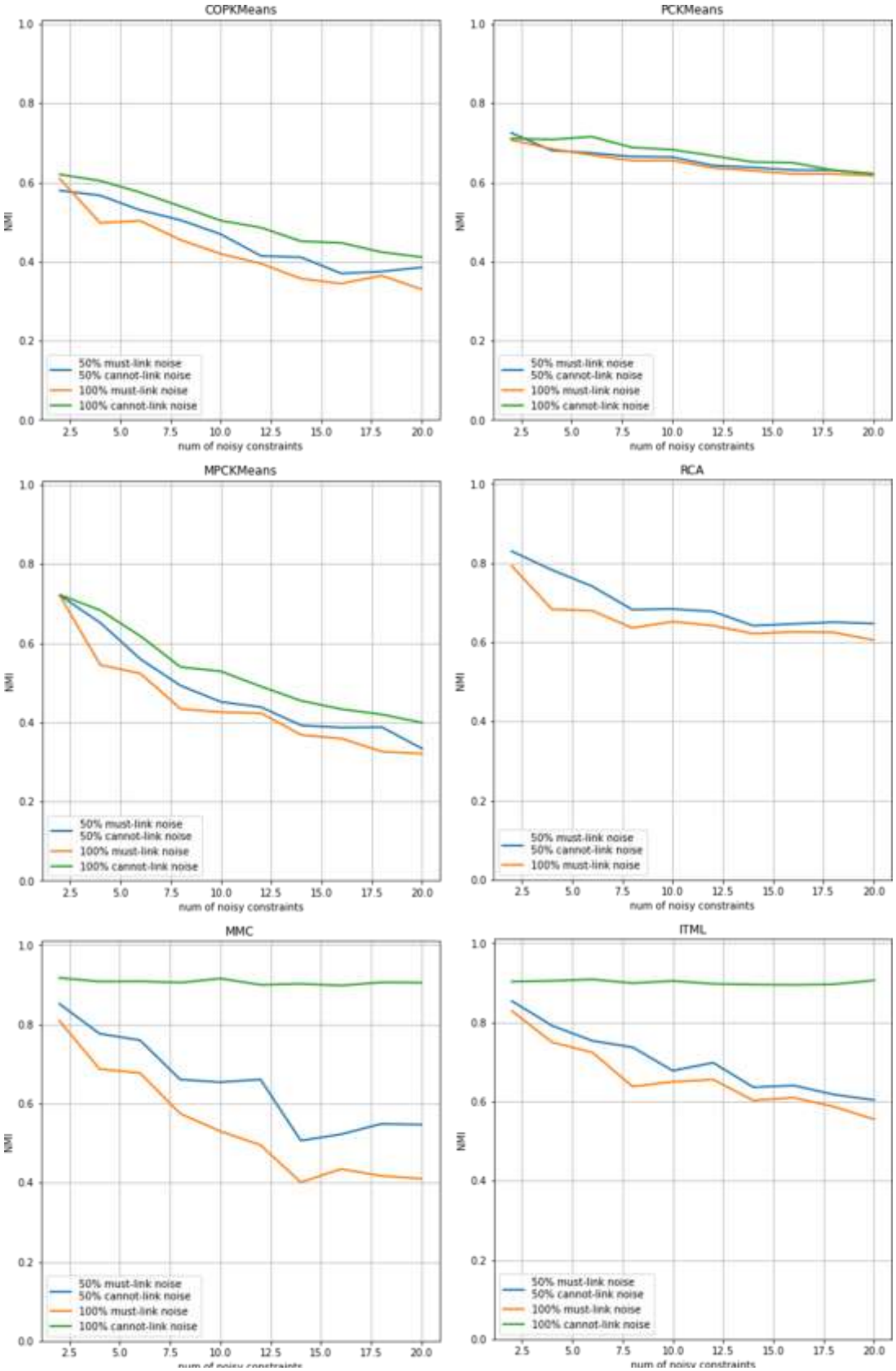

Figure 6: Performance of semi-supervised clustering approaches with incorrect constraint sets for the Iris dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

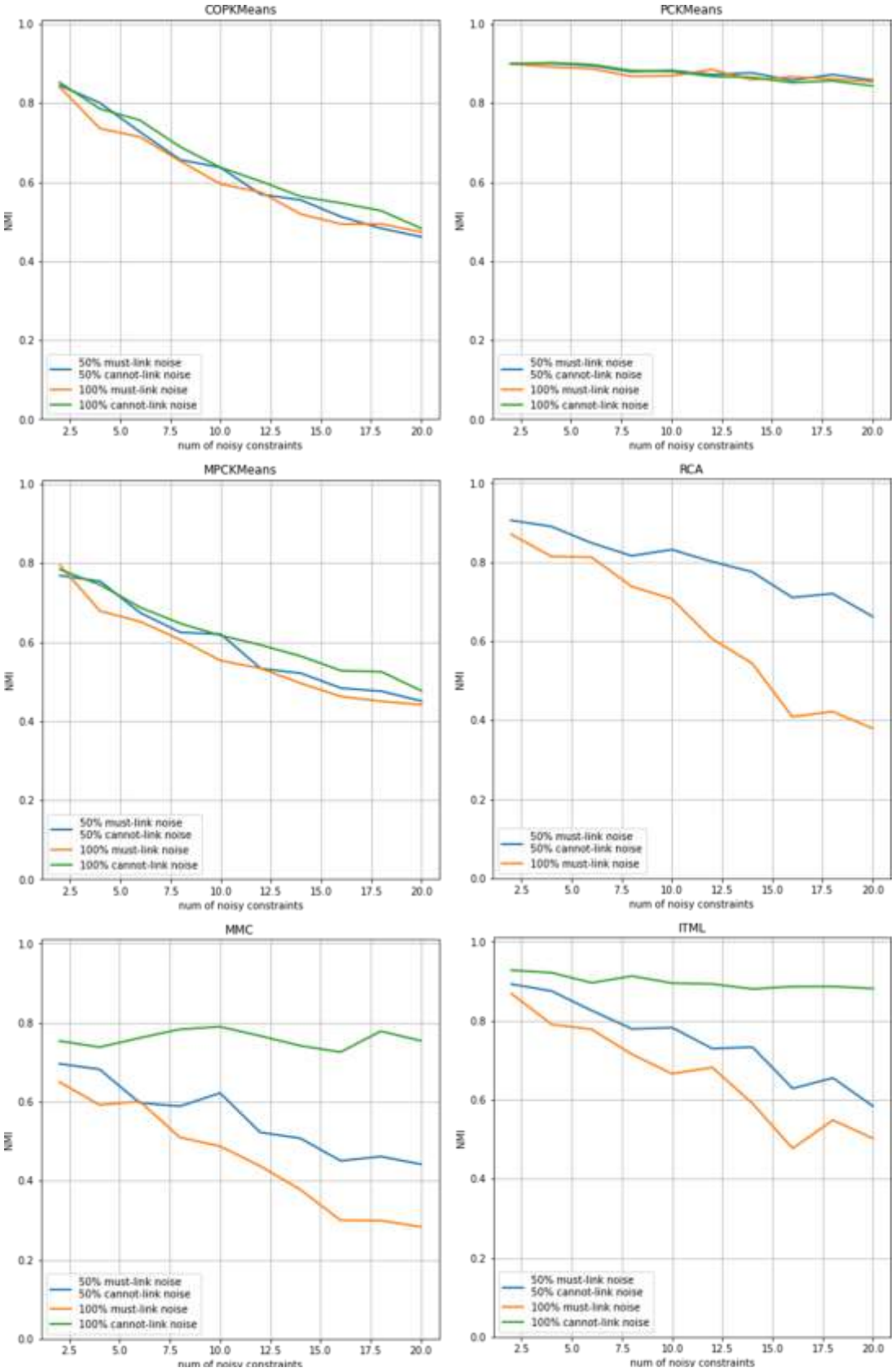

Figure 7: Performance of semi-supervised clustering approaches with incorrect constraint sets for the Iris dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

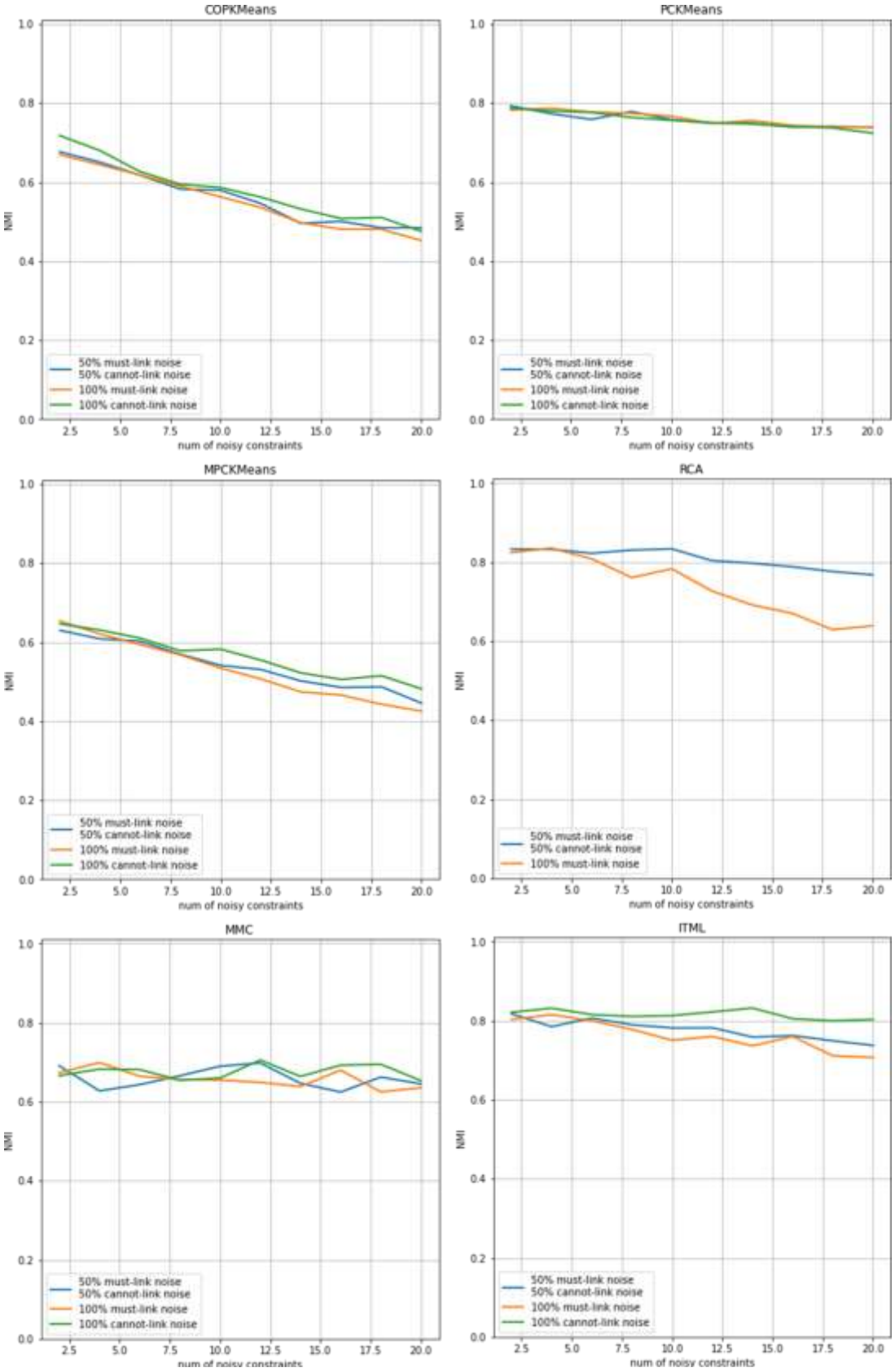

Figure 8: Performance of semi-supervised clustering approaches with incorrect constraint sets for the Seeds dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

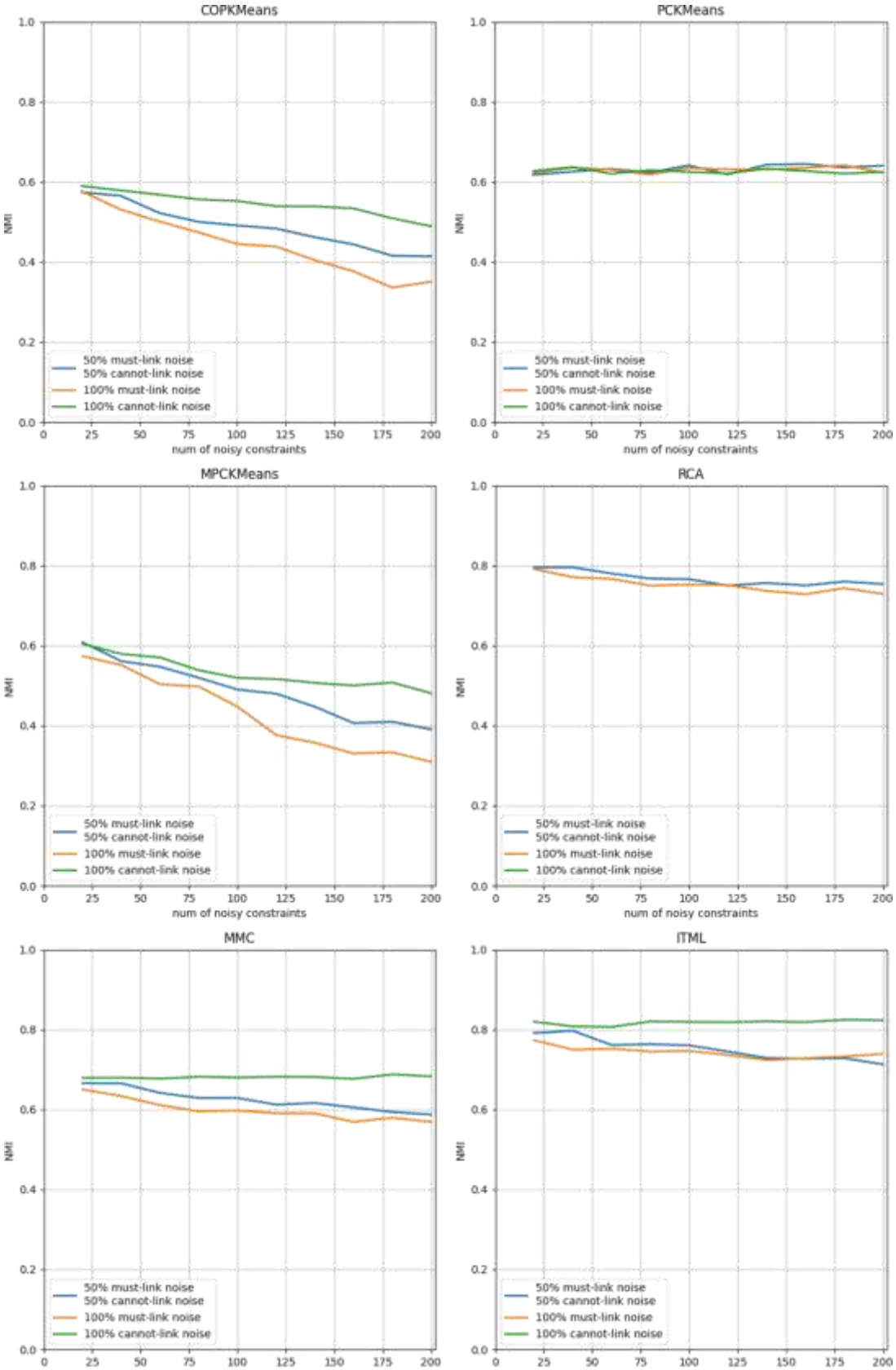

Figure 9: Performance of semi-supervised clustering approaches with incorrect constraint sets for the Statlog dataset.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

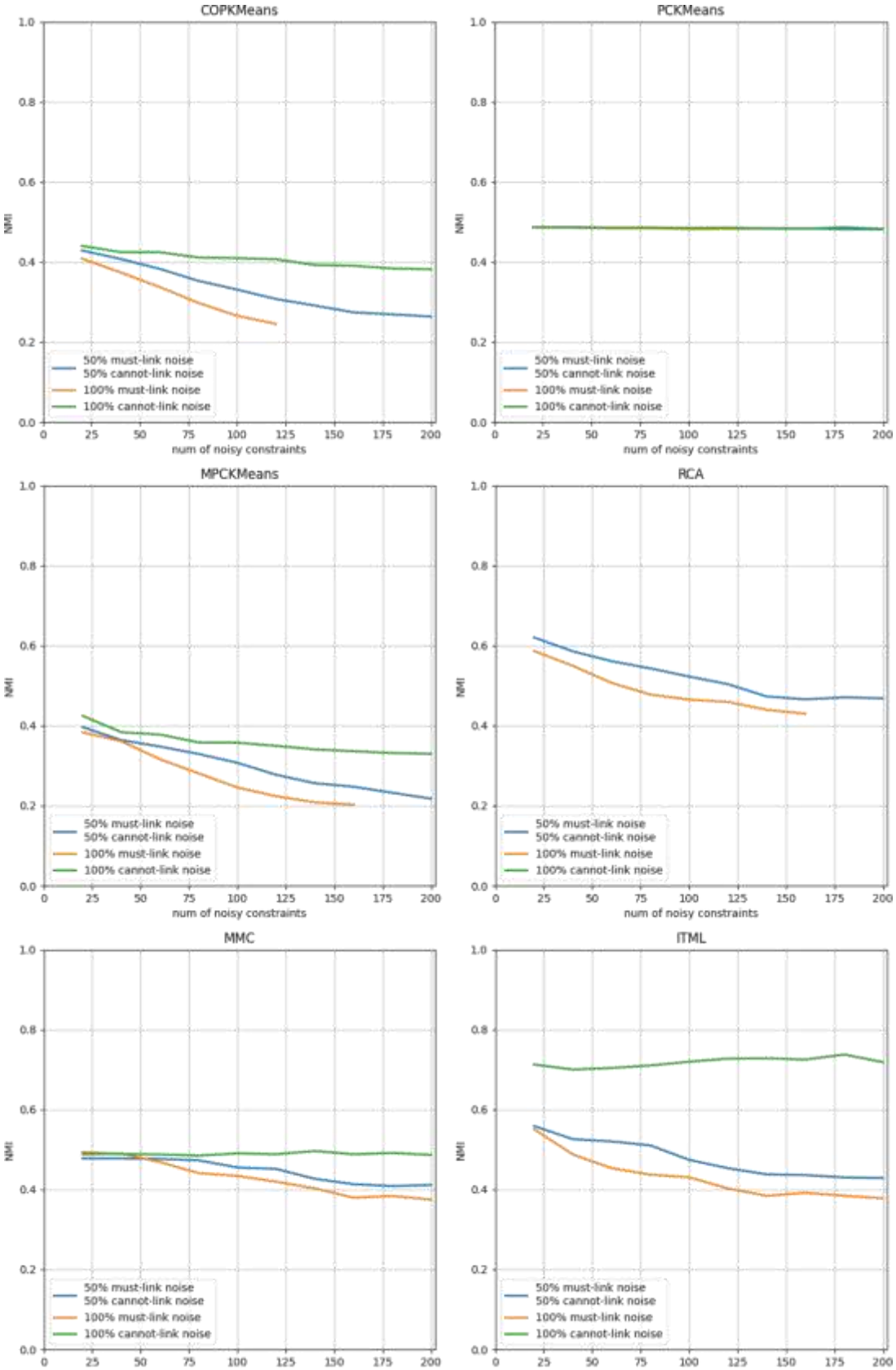

Figure 10: Performance of semi-supervised clustering approaches with incorrect constraint sets for the Obesity dataset.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

all five datasets. According to the analysis in the previous experiment, the tested semi-supervised clustering approaches are more affected and benefit more from must-link constraints. This finding is con-firmed in this experiment, as we can observe from the orange curves, since there is a global significant decrease in performance when inaccuracy exists in must-link constraints.

This experiment illustrates the weakness of the tested approaches in stability in the presence of inaccuracy in constraints, especially when there are incorrect constraints among must-link constraints.

5. VISUALIZATION AND ANALYSIS

We use the two-dimensional Cassini dataset to visualize and analyze the impact of must-link and cannot-link constraints using four different balanced and imbalanced constraint sets. This dataset features three well-defined instance groups: A core circular cluster flanked by two non-convex, curvilinear clusters. Such geometrical configurations are widely regarded as significant benchmarks for evaluating clustering performance, as documented in [57]. The presence of curvilinear structures poses a significant challenge for the K-means algorithm; due to its inherent bias toward generating convex partitions, it is unable to accurately resolve the three underlying clusters. To assess the comparative performance of the COP-Kmeans, PC-Kmeans, MPC-Kmeans, RCA, MMC and ITML semi-supervised approaches, this dataset is employed to visualize and analyze the influence exerted by different sets of constraints on the resulting clusters.

We selected four different sets of constraints to evaluate their impact and compare the final clustering result of each semi-supervised clustering approach. The first one is a randomly selected small set of constraints, with three must-link constraints and three cannot-link constraints that were randomly selected. The second is a representative small set of constraints, where three representative must-link constraints and three cannot-link constraints were manually selected. These manually selected must-link and cannot-link constraints were chosen according to the position of the instances involved regarding the distribution of the clusters they belong to in the data space of the Cassini dataset. These constraints were chosen to be representative of the three clusters which are distributed vertically and whose top and bottom clusters have a horizontal elongated shape. The third is an imbalanced randomly selected large set of constraints, with thirty cannot-link constraints and three must-link constraints. The fourth is an imbalanced randomly selected large set of constraints, with three cannot-link constraints and thirty must-link constraints.

To visualize and analyze the impact of the four constraint sets, the pairwise constraints used and the output clustering solutions for the unsupervised K-means algorithm and the six semi-supervised clustering approaches are plotted in two-dimensional graphical representations. These graphical representations are shown in Figure 11 for the first set of constraints, in Figure 12 for the second set of constraints, in Figure 13 for the third set of constraints and in Figure 14 for the fourth set of constraints. The horizontal and vertical axes depict the two-dimensional data space. The cannot-link constraints used are represented by red lines, and the must-link constraints used are represented by green lines. The color of each point indicates the cluster to which the corresponding instance is assigned in the output clustering solution.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

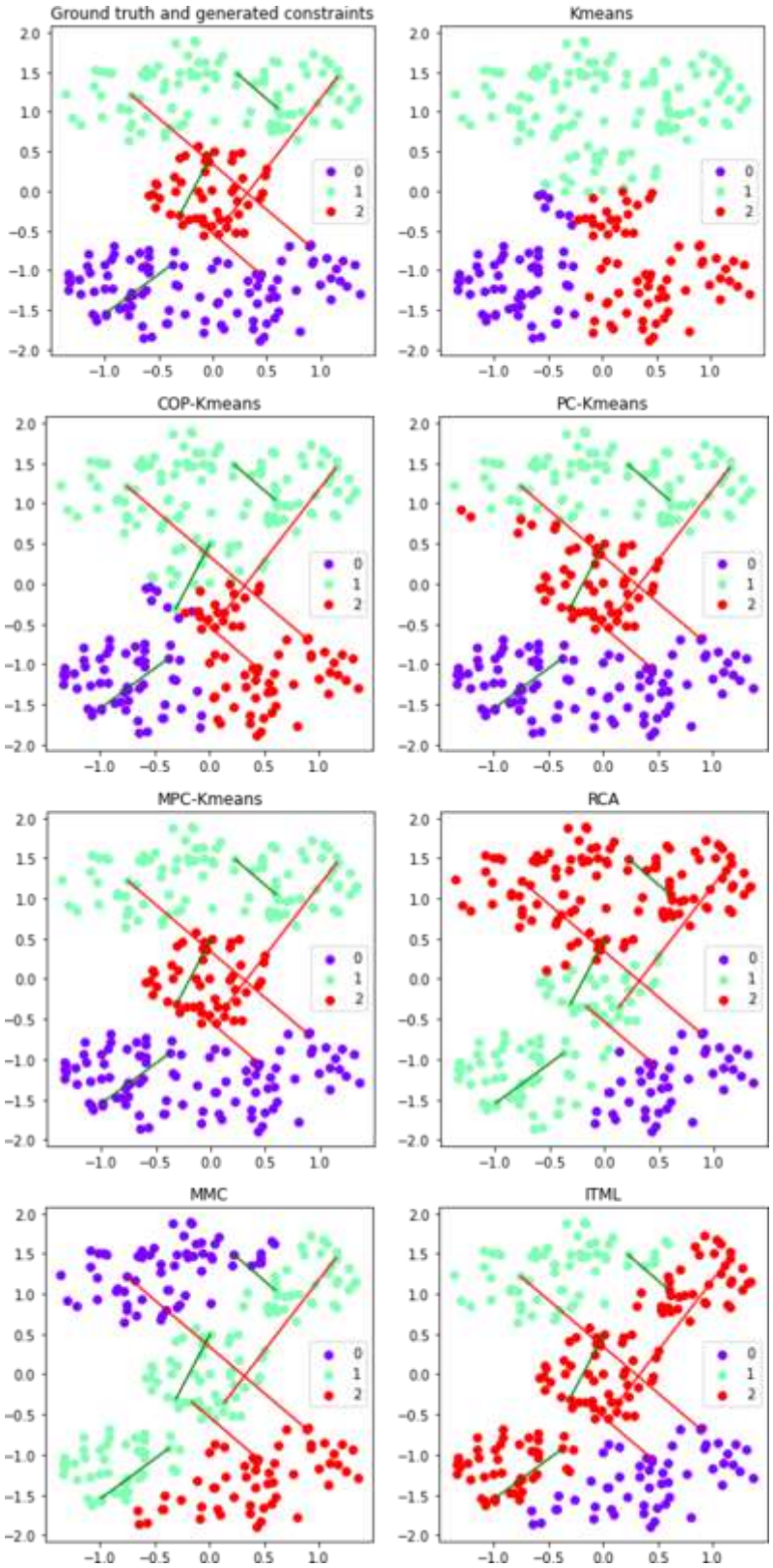

Figure 11: Performance of semi-supervised clustering approaches for the Cassini dataset with randomly selected constraints. The three cannot-link and the three must-link constraints are represented by red and green lines respectively.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

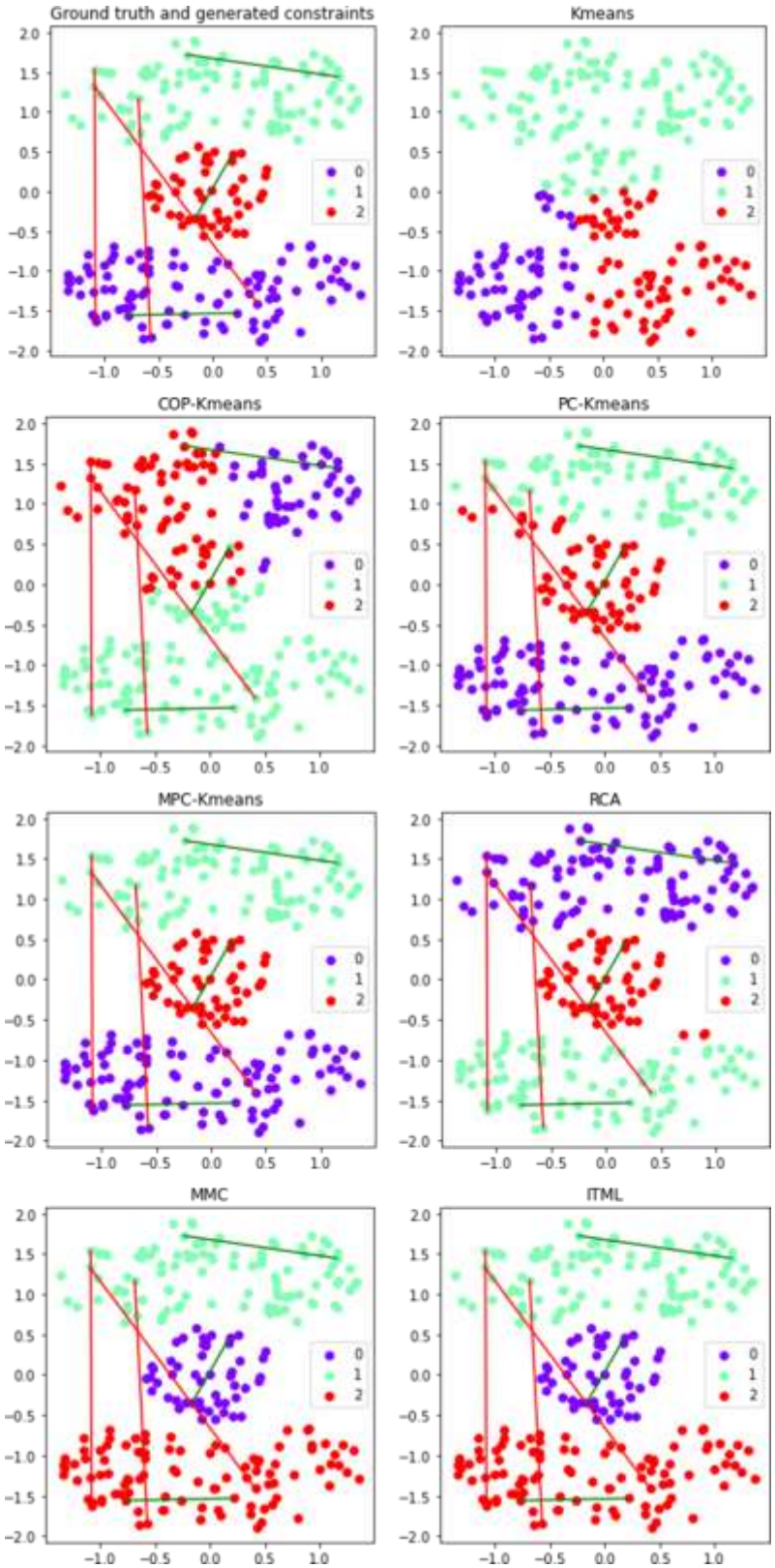

Figure 12; Performance of semi-supervised clustering approaches for the Cassini dataset with manually selected representative constraints. The three cannot-link and the three must-link constraints are represented by red and green lines respectively.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

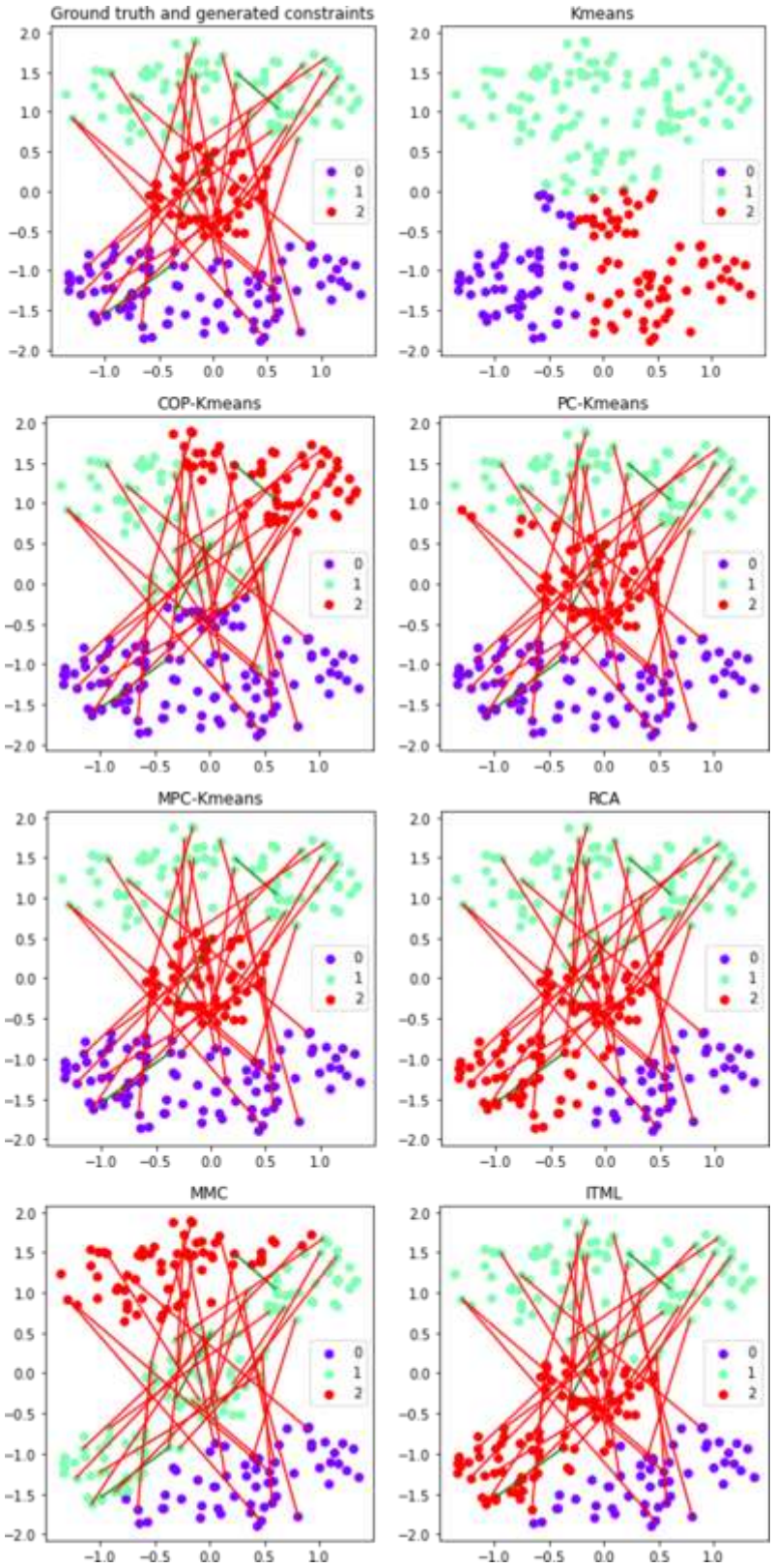

Figure 13: Performance of semi-supervised clustering approaches for the Cassini dataset with more cannot-link constraints than must-link constraints. The thirty cannot-link and the three must-link constraints are represented by red and green lines respectively.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

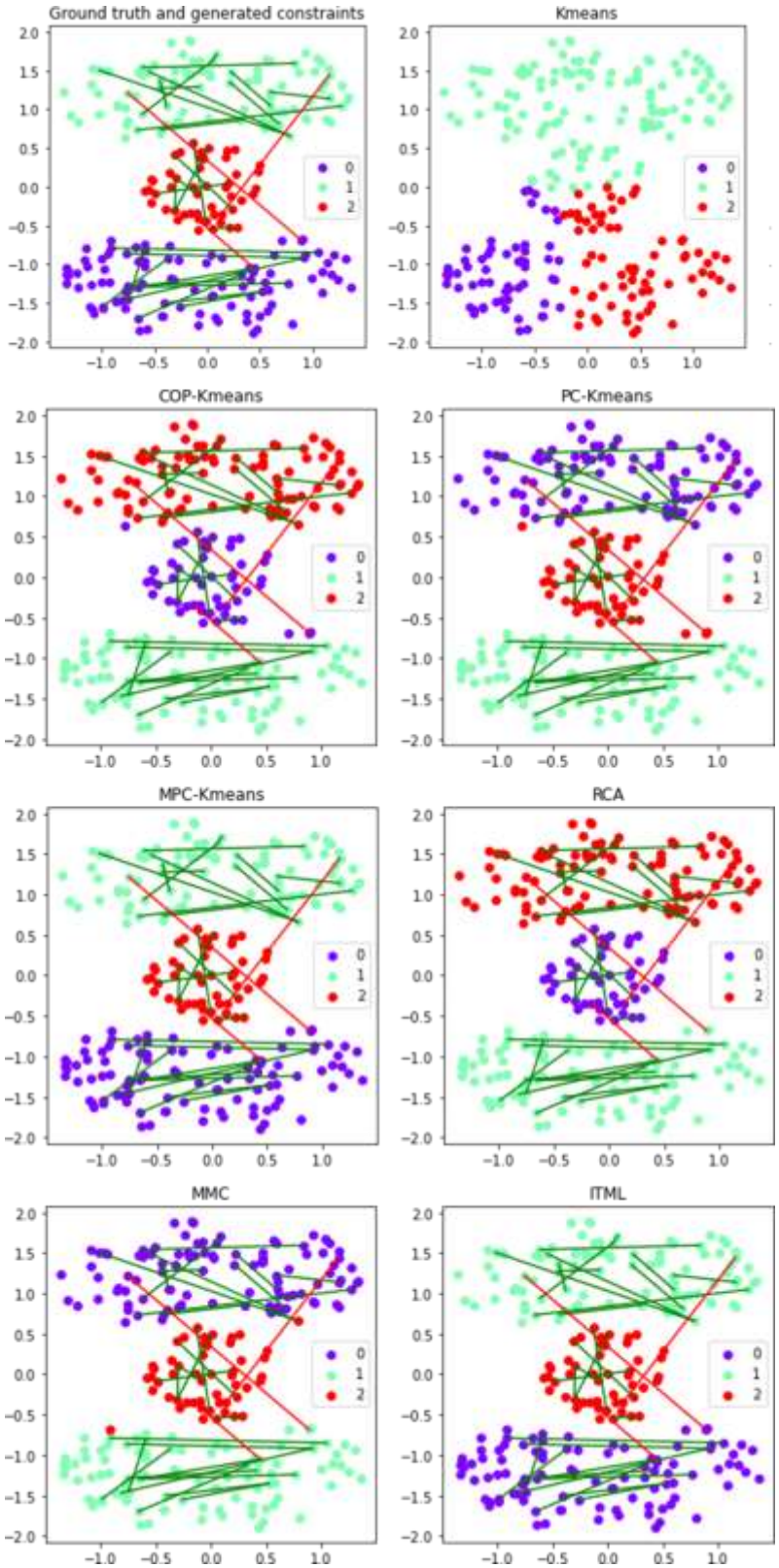

Figure 14: Performance of semi-supervised clustering approaches for the Cassini dataset with more must-link constraints than cannot-link constraints. The three cannot-link and the thirty must-link constraints are represented by red and green lines respectively.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

When comparing the results for the first and second constraint sets, we can see that metric learning approaches are strongly affected if the constraint set does not effectively express the cluster distribution. The reason is that the metric learning approaches aim to learn an adapted metric based on constraints. With the representative small constraint set, the must-link constraints within the two curvilinear clusters help the metric learning approaches to "shrink" the horizontal distance, and the cannot-link constraints allow for enlarging the vertical distance. Therefore, metric learning approaches achieve high quality clustering. However, based on the randomly selected small constraint set, the metric learning approaches adapt the metric to reduce the diagonal distance from the lower left corner to the upper right corner, which does not reflect well the data structure and results in poor performance. This also implies the potential reason for the negative effect issue. As the metric learning approaches strongly rely on the representative must-link constraints, non-representative constraints in the small constraint set may cause these approaches to fall into a local optima and give worse performance than using no constraints. On the other hand, the K-means variants integrate constraints to improve the clustering process of K-means. Thus, the quality of constraint sets has less influence on their performance than on metric learning approaches. They also have fewer negative effect issues.

As we can see from the result for the third constraint set, increasing the size of cannot-link constraints does not help much in improving the performance. With the same randomly selected must-link constraints as in the first constraint set, even when the number of cannot-link constraints is increased to thirty, the semi-supervised clustering approaches still fail to give good performance. In contrast, when the number of must-link constraints increases, the performance of all approaches is significantly improved. This confirms the finding outlined in Section 4.1 that the constraints of must-linking have an effective positive impact on performance. A presumed reason for this is that, as the number of must-link constraints increases, the probability that representative must-link constraints are included in the constraint set also increases.

6. CONCLUSION

In this paper, we examine the impact of imbalanced and incorrect constraint sets on six state-of-the-art semi-supervised clustering approaches. Experiments conducted on several UCI bench-mark datasets (Iris, Seeds, Wine, Statlog, and Obesity), complemented by visualizations and analyses on the representative Cassini dataset, provide several important insights. Metric learning–based approaches generally exhibit faster convergence than K-means–based variants. How-ever, when a sufficiently large number of constraints is available, K-means variants may achieve superior clustering performance. The results further indicate that semi-supervised clustering methods tend to benefit more from must-link constraints than from cannot-link constraints, suggesting that must-link constraints often exert a stronger positive influence on clustering quality. In contrast, relying exclusively on cannot-link constraints may in some cases lead to performance degradation. Moreover, incorrect constraints, particularly incorrect must-link constraints, can adversely affect clustering performance. Overall, these findings highlight the critical role played by both the quality and the balance of must-link and cannot-link constraints in ensuring reliable performance of semi-supervised clustering methods, and emphasize the importance of carefully constructing constraint sets in practical applications.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

DATA AVAILABILITY STATEMENT

The data presented in this study are available in UCI Machine Learning Repository at https://archive.ics.uci.edu reference number [52], and in mlbench: Machine Learning Benchmark Problems at https://CRAN.R-project.org/package=mlbench, reference number [53].

ACKNOWLEDGMENTS

The authors declare that no financial or institutional support was received for this research.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

REFERENCES

1. Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005. https://doi.org/10.1109/TNN.2005.845141.
2. Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015. https://doi.org/10.1007/s40745-015-0040-1.
3. Kiri L. Wagstaff. Value, cost, and sharing: Open issues in constrained clustering. In *Proceedings of the International Workshop on Knowledge Discovery in Inductive Databases*, pages 1–10. Springer International Publishing, 2006. https://doi.org/10.1007/978-3-540-75549-4_1.
4. Derya Dinler and Mustafa Kemal Tural. A survey of constrained clustering. In *Unsupervised Learning Algorithms*, pages 207–235. Springer International Publishing, 2016. https://doi. org/10.1007/978-3-319-24211-8_9.
5. Pierre Gançarski, Bruno Crémilleux, Germain Forestier, Thomas Lampert, et al. Constrained clustering: Current and new trends. In *A Guided Tour of Artificial Intelligence Research*, pages 447–484. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-06167-8_14.
6. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl, et al. Constrained K-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, volume 1, pages 577–584. Morgan Kaufmann Publishers Inc., 2001. https://dl.acm.org/doi/10.5555/645530.655669.
7. Kiri Wagstaff and Claire Cardie. Clustering with instance-level constraints. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1103–1110. Morgan Kaufmann Publishers Inc., 2000. https://doi.org/10.5555/645529.658275.
8. Mikhail Bilenko, Sugato Basu, and Raymond J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st*

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

International Conference on Machine learning, page 11. Association for Computing Machinery, 2004. https://doi. org/10.1145/1015330.1015360.

9. Ian Davidson and S.S. Ravi. Clustering with constraints: Feasibility issues and the K-means algorithm. In *Proceedings of the SIAM International Conference on Data Mining*, pages 138–149. SIAM, 2005. https://doi.org/10.1137/1.9781611972757.13.

10. Dan Pelleg and Dorit Baras. K-means with large and noisy constraint sets. In *Proceedings of the European Conference on Machine Learning*, pages 674–682. Springer International Publishing, 2007. https://doi.org/10.1007/978-3-540-74958-5_67.

11. Mohadeseh Ganji, James Bailey, and Peter J Stuckey. Lagrangian constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 288–296. SIAM, 2016. https://doi.org/10.1137/1.9781611974348.33.

12. Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 521–528. MIT Press, 2002. https://doi.org/10.5555/2968618.2968683.

13. Aharon Bar-Hillel, Tomer Hertz, Noam Shental, Daphna Weinshall, and Greg Ridgeway. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6):937–965, 2005. https://dl.acm.org/doi/10.5555/1046920.1088704.

14. Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216. Association for Computing Machinery, 2007. https://doi.org/10.1145/1273496.1273523.

15. Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang. An efficient sparse metric learning in high-dimensional space via ℓ1-penalized log-determinant regularization. In *Proceedings of the 26th International Conference on Machine Learning*, pages 841–848. Association for Computing Machinery, 2009. https://doi.org/10.1145/1553374.1553482.

16. Ian Davidson, SS Ravi, and Leonid Shamis. A SAT-based framework for efficient constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 94–105. SIAM, 2010. https://doi.org/10.1137/1.9781611972801.9.

17. Tias Guns, Siegfried Nijssen, and Luc De Raedt. k-pattern set mining under constraints. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):402–418, 2011. https://doi.org/10.1109/TKDE.2011.204.

18. Abdelkader Ouali, Samir Loudni, Yahia Lebbah, Patrice Boizumault, Albrecht Zimmermann, and Lakhdar Loukil. Efficiently finding conceptual clustering models with integer linear programming. In *Proceedings of the 25th International Joint Conferences on Artificial Intelligence*, pages 647–654. AAAI Press, 2016. https://doi.org/10.5555/3060621.3060712.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

19. Sean Gilpin and Ian Davidson. A flexible ILP formulation for hierarchical clustering. *Artificial Intelligence*, 244:95–109, 2017. https://doi.org/10.1016/j.artint.2015.05.009.

20. Germain Forestier, Pierre Gançarski, and Cédric Wemmert. Collaborative clustering with background knowledge. *Data & Knowledge Engineering*, 69(2):211–228, 2010. https://doi. org/10.1016/j.datak.2009.10.004.

21. Muna Al-Razgan and Carlotta Domeniconi. Clustering ensembles with active constraints. In *Applications of Supervised and Unsupervised Ensemble Methods*, pages 175–189. Springer International Publishing, 2009. https://doi.org/10.1007/978-3-642-03999-7_10.

22. Ashraf Mohammed Iqbal, Abidalrahman Moh'd, and Zahoor Khan. Semi-supervised clustering ensemble by voting. *arXiv preprint arXiv:1208.4138*, pages 1–5, 2012. https://doi.org/10.48550/arXiv.1208.4138.

23. Wenchao Xiao, Yan Yang, Hongjun Wang, Tianrui Li, and Huanlai Xing. Semi-supervised hierarchical clustering ensemble and its application. *Neurocomputing*, 173:1362–1376, 2016. https://doi.org/10.1016/j.neucom.2015.09.009.

24. Tianshu Yang, Nicolas Pasquier, and Frédéric Precioso. Semi-supervised consensus clustering based on closed patterns. *Knowledge-Based Systems*, 235:107599, 2022. https://doi.org/10.1016/j.knosys.2021.107599.

25. Yazhou Ren, Kangrong Hu, Xinyi Dai, Lili Pan, Steven CH Hoi, and Zenglin Xu. Semi-supervised deep embedded clustering. *Neurocomputing*, 325:121–130, 2019. https://doi.org/ 10.1016/j.neucom.2018.10.016.

26. Hongjing Zhang, Tianyang Zhan, Sugato Basu, and Ian Davidson. A framework for deep constrained clustering. *Data Mining and Knowledge Discovery*, pages 1–28, 2021. https://doi.org/10.1007/s10618-020-00734-4.

27. M Eduardo Ares, Javier Parapar, and Álvaro Barreiro. An experimental study of constrained clustering effectiveness in presence of erroneous constraints. *Information Processing & Management*, 48(3):537–551, 2012. https://doi.org/10.1016/j.ipm.2011.08.006.

28. Xiatian Zhu, Chen Change Loy, and Shaogang Gong. Constrained clustering: Effective constraint propagation with imperfect oracles. In *Proceedings of the 13th International Conference on Data Mining*, pages 1307–1312. IEEE, 2013. https://doi.org/10.1109/ICDM.2013.45.

29. Blaine Nelson and Ira Cohen. Revisiting probabilistic models for clustering with pair-wise constraints. In *Proceedings of the 24th International Conference on Machine Learning*, pages 673–680. Association for Computing Machinery, 2007. https://doi.org/10.1145/1273496.1273581.

30. Hongjing Zhang, Sugato Basu, and Ian Davidson. A framework for deep constrained clustering-algorithms and advances. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 57–

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

72. Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-46150-8_4.

31. Tom Coleman, James Saunderson, and Anthony Wirth. Spectral clustering with inconsistent advice. In *Proceedings of the 25th International Conference on Machine Learning*, pages 152–159. Association for Computing Machinery, 2008. https://doi.org/10.1145/1390156.1390176.

32. Erliang Zeng, Chengyong Yang, Tao Li, and Giri Narasimhan. On the effectiveness of constraints sets in clustering genes. In *Proceedings of the 7th International Symposium on BioInformatics and BioEngineering*, pages 79–86. IEEE, 2007. https://doi.org/10.1109/BIBE.2007.4375548.

33. M. Eduardo Ares, Javier Parapar, and Alvaro Barreiro. Improving text clustering with social tagging. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 430–433, 2021. https://doi.org/10.1609/icwsm.v5i1.14148.

34. Ronglu Yan, J. Zhang, Jiangshuai Yang, and Alexander Hauptmann. *Learning with pairwise constraints for video object classification*, pages 397–430. Chapman and Hall/CRC, 2008. https://doi.org/10.1201/9781584889977.

35. Dan Pelleg and Dorit Baras. K-means with large and noisy constraint sets. In *Proceedings of the European Conference on Machine* Learning, pages 674–682. Springer, 2007. https://doi.org/10.1007/978-3-540-74958-5_67.

36. Chen Gong, Keren Fu, Qiang Wu, Enmei Tu, and Jie Yang. Semi-supervised classification with pairwise constraints. *Neurocomputing*, 139:130–137, 2014. https://doi.org/10.1016/j.neucom.2014.02.053.

37. Hui Liu, Yuheng Jia, Junhui Hou, and Qingfu Zhang. Imbalance-aware pairwise constraint propagation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1605–1613. ACM, 2019. https://doi.org/10.1145/3343031.3350968.

38. Stephen Mussmann, Robin Jia, and Percy Liang. On the importance of adaptive data collection for extremely imbalanced pairwise tasks. In *Empirical Methods in Natural Language Processing*, pages 3400–3413. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.305.

39. Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

40. Bartosz Krawczyk. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. https://doi.org/10.1007/s13748-016-0094-0.

41. S. Layaq and Dr. B. Manjula. A recapitulation of imbalanced data. *International Journal of Innovative Technology and Exploring Engineering*, 9(3):452–455, 2020. https://doi.org/10.35940/ijitee.b8120.019320.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

42. Eric Bair. Semi-supervised clustering methods. *Wiley interdisciplinary reviews. Computational statistics*, 5(5):349–361, 2013. https://doi.org/10.1002/wics.1270.

43. Korinna Bade and Andreas Nurnberger. Personalized hierarchical clustering. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 181–187, 2006. https://doi.org/10.1109/WI.2006.131.

44. Ian Davidson, Kiri L Wagstaff, and Sugato Basu. Measuring constraint-set utility for partitional clustering algorithms. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pages 115–126. Springer International Publishing, 2006. https://doi.org/10.1007/11871637_15.

45. Ian Davidson and SS Ravi. Identifying and generating easy sets of constraints for clustering. In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 1, pages 336–341. AAAI Press, 2006. https://doi.org/10.5555/1597538.1597593.

46. Thiago F. Covoes, Eduardo R. Hruschka, and Joydeep Ghosh. A study of K-means-based algorithms for constrained clustering. *Intelligent Data Analysis*, 17(3):485–505, 2013. https://doi.org/10.5555/2595566.2595574.

47. Noam Shental, Tomer Hertz, Daphna Weinshall, and Misha Pavel. Adjustment learning and relevant component analysis. In *Proceedings of the European Conference on Computer Vision*, pages 776–790. Springer International Publishing, 2002. https://doi.org/10.1007/3-540-47979-1_52.

48. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967. https://scispace.com/pdf/some-methods-for-classification-and-analysis-of-multivariate-4pswti19oz.pdf.

49. Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the 19th International Conference on Machine Learning*, pages 27–34. Morgan Kaufmann Publishers Inc., 2002. https://dl.acm.org/doi/10.5555/645531.656012.

50. Liu Yang and Rong Jin. *Distance Metric Learning: A Comprehensive Survey*. Michigan State University, 2006. https://api.semanticscholar.org/CorpusID:850937.

51. Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012. https://doi.org/10.1561/2200000019.

52. Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu.

53. Friedrich Leisch and Evgenia Dimitriadou. mlbench: Machine learning benchmark problems. Available online: https://CRAN.R-project.org/package=mlbench.

54. Jakub Svehla. Active semi-supervised clustering, 2020. Available online: https://github.com/datamole-ai/active-semi-supervised-clustering.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/67e51f07

55. William de Vazelhes, C. J. Carey, Yuan Tang, Nathalie Vauquier, and Aurélien Bellet. metric-learn: Metric Learning Algorithms in Python. *Journal of Machine Learning Research*, 21(138):1–6, 2020. https://doi.org/10.48550/arXiv.1908.04710.

56. Tarald Kvålseth. On Normalized Mutual Information: Measure derivations and properties. *Entropy*, 19(11):631, 2017. https://doi.org/10.3390/e19110631.

57. Christian Wiwie, Jan Baumbach, and Richard Röttger. Comparing the performance of biomedical clustering methods. *Nature Methods*, 12(11):1033–1038, 2015. https://doi.org/10.1038/nmeth.3583.