COMBINED UNSUPERVISED AND SEMI-SUPERVISED LEARNING FOR DATA CLASSIFICATION

MLSP2016
IEEE International Workshop on
MACHINE LEARNING FOR SIGNAL PROCESSING
September 13-16, 2016 **Vietri sul Mare, Salerno, Italy**

Fabricio Aparecido Breve, Daniel Carlos Guimarães Pedronette
State University of São Paulo (UNESP), Rio Claro, SP, Brazil
fabricio@rc.unesp.br ; daniel@rc.unesp.br

unesp

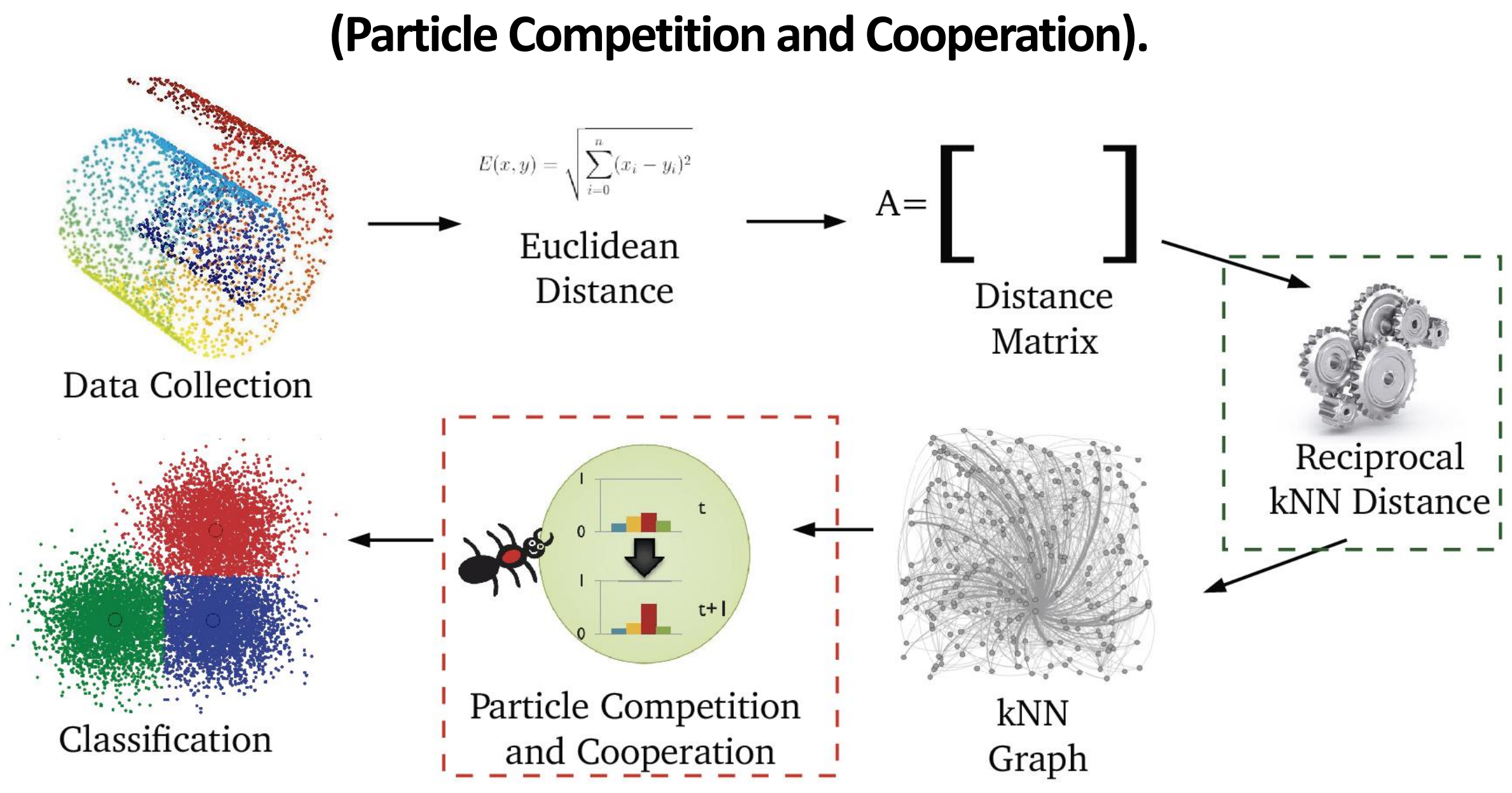
Abstract - Semi-supervised learning methods exploit both labeled and unlabeled data items in their training process, requiring only a small subset of labeled items. Although capable of drastically reducing the costs of labeling process, such methods are directly dependent on the effectiveness of distance measures used for building the kNN graph. On the other hand, unsupervised distance learning approaches aims at capturing and exploiting the dataset structure in order to compute a more effective distance measure, without the need of any labeled data. In this paper, we propose a combined approach which employs both unsupervised and semi-supervised learning paradigms. An unsupervised distance learning procedure is performed as a pre-processing step for improving the kNN graph effectiveness. Based on the more effective graph, a semi-supervised learning method is used for classification. The proposed Combined Unsupervised and Semi-Supervised Learning (CUSSL) approach is based on very recent methods. The Reciprocal kNN Distance is used for unsupervised distance learning tasks and the semi-supervised learning classification is performed by Particle Competition and Cooperation (PCC). Experimental results conducted in six public datasets demonstrated that the combined approach can achieve effective results, boosting the accuracy of classification tasks.

Reciprocal kNN Distance

Reciprocal kNN Distance, was proposed to provide a more effective distance measure in image retrieval scenarios. It takes into account the intrinsic dataset structure by analyzing the reciprocal references at top rank positions. The modelling in terms of rank information enables its use in many other scenarios, specially in cases which require the computation of k nearest neighbors, as the PCC.

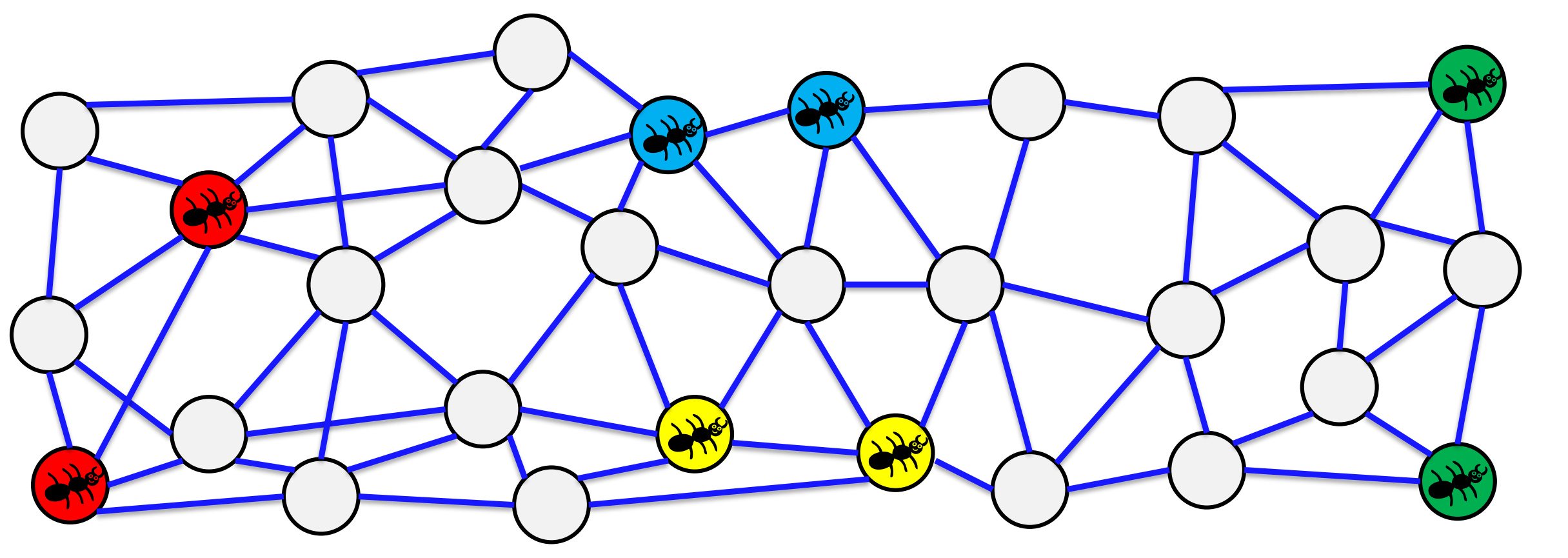
The Reciprocal kNN Distance between two data items $x_q, x_i \in \mathcal{X}$ is computed based on the number of reciprocal neighbors at top positions of ranked lists $\tau_q, \tau_i \in \mathrm{T}$. Additionally, a weight for each pair of reciprocal neighbors is considered, proportionally to their position in the ranked lists τ_q and τ_i.

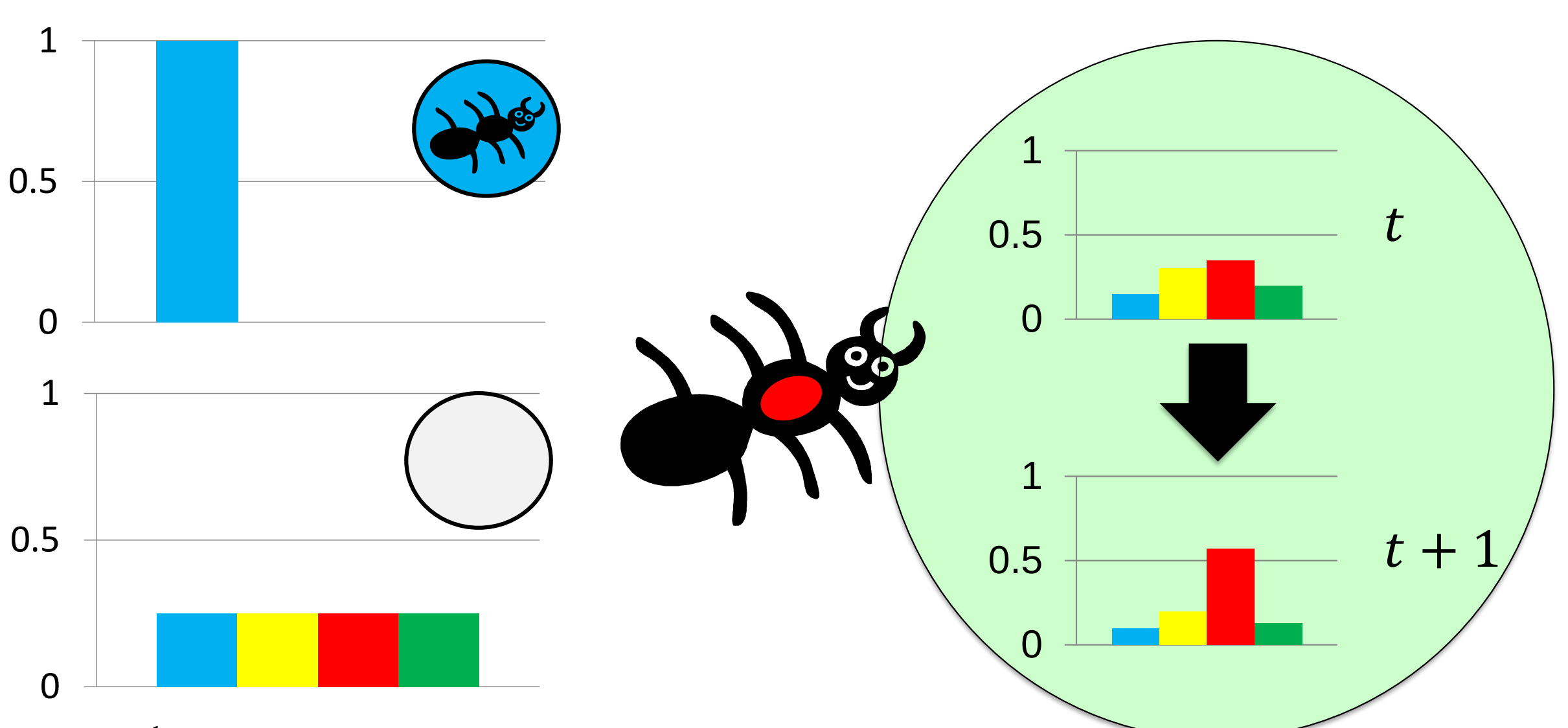
Overall work-flow of combined method: in green dashed box, the unsupervised distance learning method (Reciprocal kNN Distance); in red dashed box, the semisupervised learning method used for classification (Particle Competition and Cooperation).



Particle Competition and Cooperation

A particle is generated for each labeled node of the graph. Particles initial position are set to their corresponding nodes. Particles with same label play for the same team and cooperate with each other. Particles with different labels compete against each other. They increase the domination level of their respective team in the nodes they visit. At the same time, they decrease the domination levels of other teams.



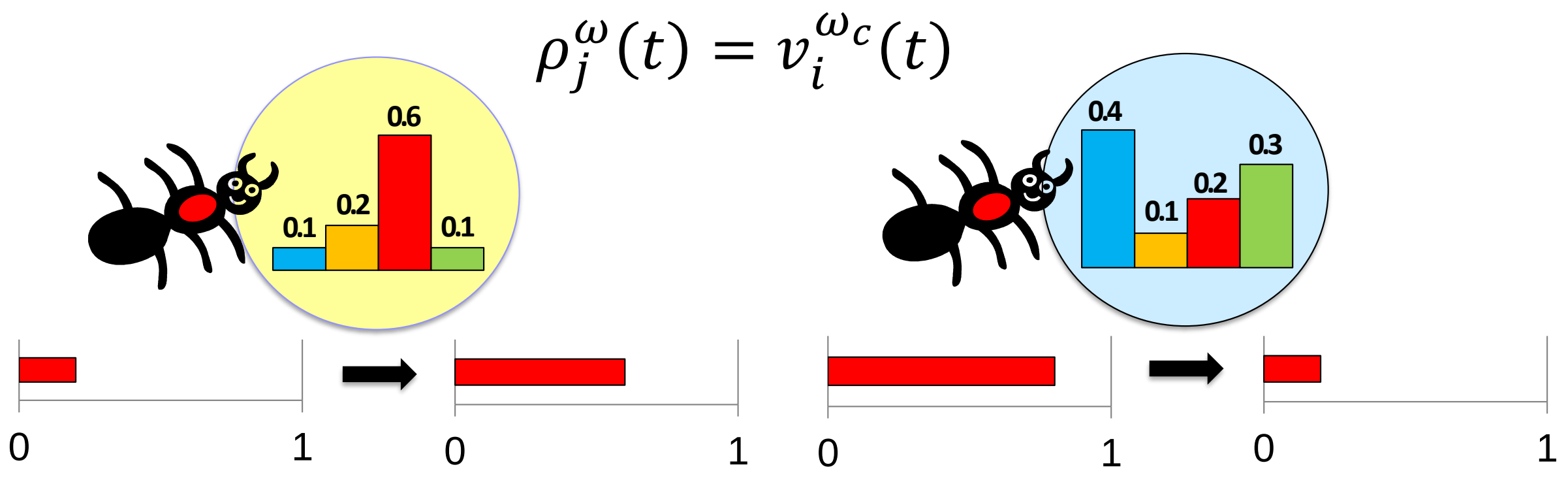


$$v_i^{\omega_\ell} = \begin{cases} 1 & \text{if } y_i = \ell \\ 0 & \text{if } y_i \neq \ell \text{ e } y_i \in L \\ 1/c & \text{if } y_i = \emptyset \end{cases}$$

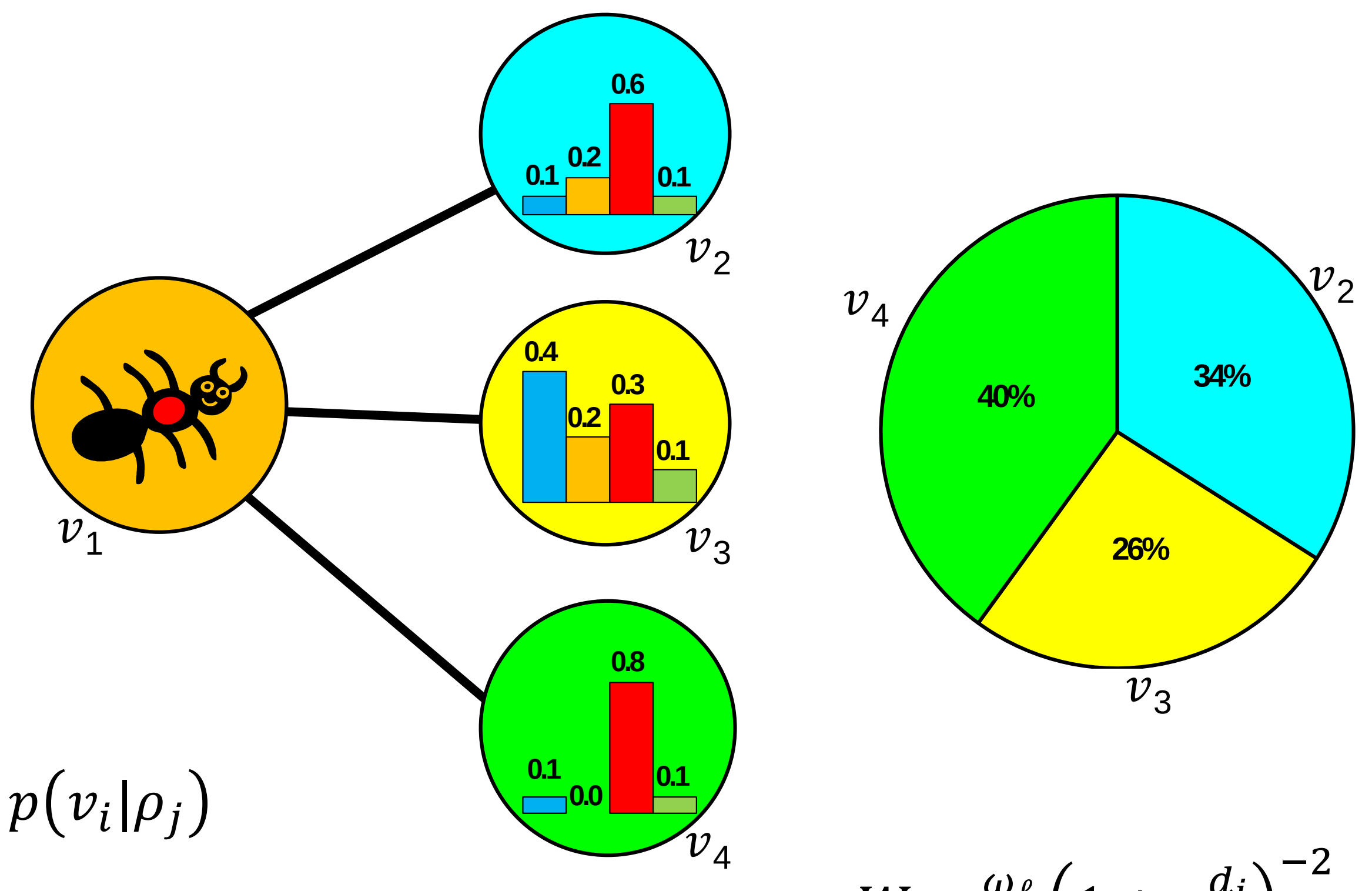
$$v_i^{\omega_\ell}(t+1) = \begin{cases} \max\left\{0, v_i^{\omega_\ell}(t) - \frac{\Delta_v \rho_j^{\omega}(t)}{c-1}\right\} & \text{if } y_i = 0 \text{ e } \ell \neq \rho_j^f \\ v_i^{\omega_\ell}(t) + \sum_{q\neq\ell} v_i^{\omega_q}(t) - v_i^{\omega_q}(t+1) & \text{if } y_i = 0 \text{ e } \ell = \rho_j^f \\ v_i^{\omega_\ell}(t) & \text{if } y_i \neq 0 \end{cases}$$

A particle gets strong when it visits a node being dominated by its own team, but it gets weak when it selects a node being dominated by another team.

$$\rho_j^{\omega}(t) = v_i^{\omega_c}(t)$$



Each particles randomly chooses a neighbor to visit at each iteration. Nodes which are already dominated by the particle team and closer to the particle initial node have higher probabilities of being chosen.



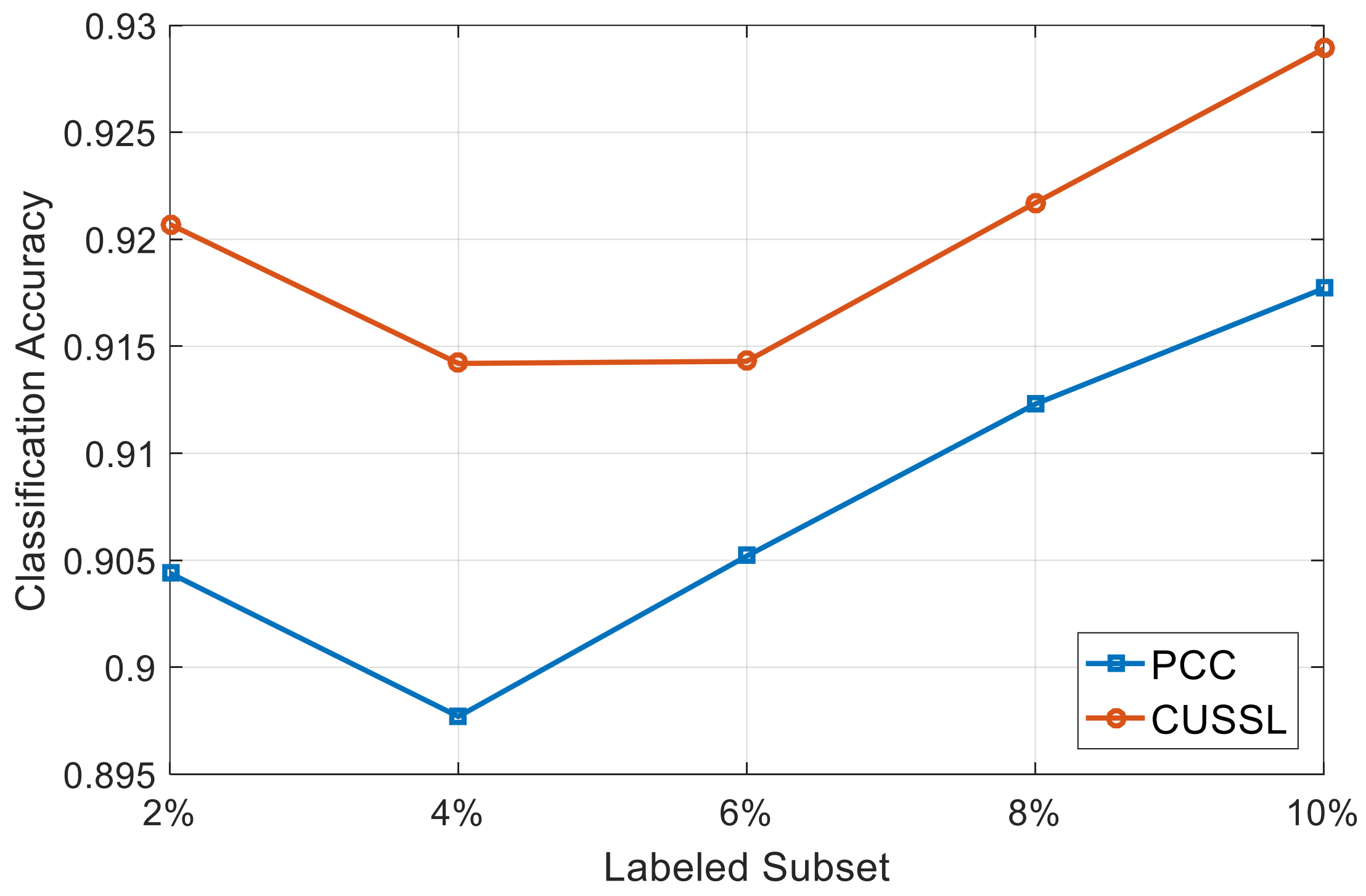
$$p(v_i|\rho_j) = (1 - p_{grd})\frac{W_{qi}}{\sum_{\mu=1}^{n} W_{q\mu}} + p_{grd}\frac{W_{qi} v_i^{\omega_\ell}\left(1+\rho_j^{d_i}\right)^{-2}}{\sum_{\mu=1}^{n} W_{q\mu}\, v_\mu^{\omega_\ell}\left(1+\rho_j^{d_\mu}\right)^{-2}}$$

Computer Simulations

Computer simulation using some artificial and real-world data sets are presented in order to show the effectiveness of the proposed method. For each data set, we applied both the original PCC method and the proposed method (CUSSL). For PCC, the parameter k defines the size of k-neighborhood (amount of nearest neighbors) for the graph construction using the Euclidean distance (L2). For CUSSL, two parameters are considered: (i) the size of k-neighborhood used by the unsupervised distance learning algorithm, referred in this section as k_r; and (ii) the size of k-neighborhood for the graph construction, referred in this section as k_n.

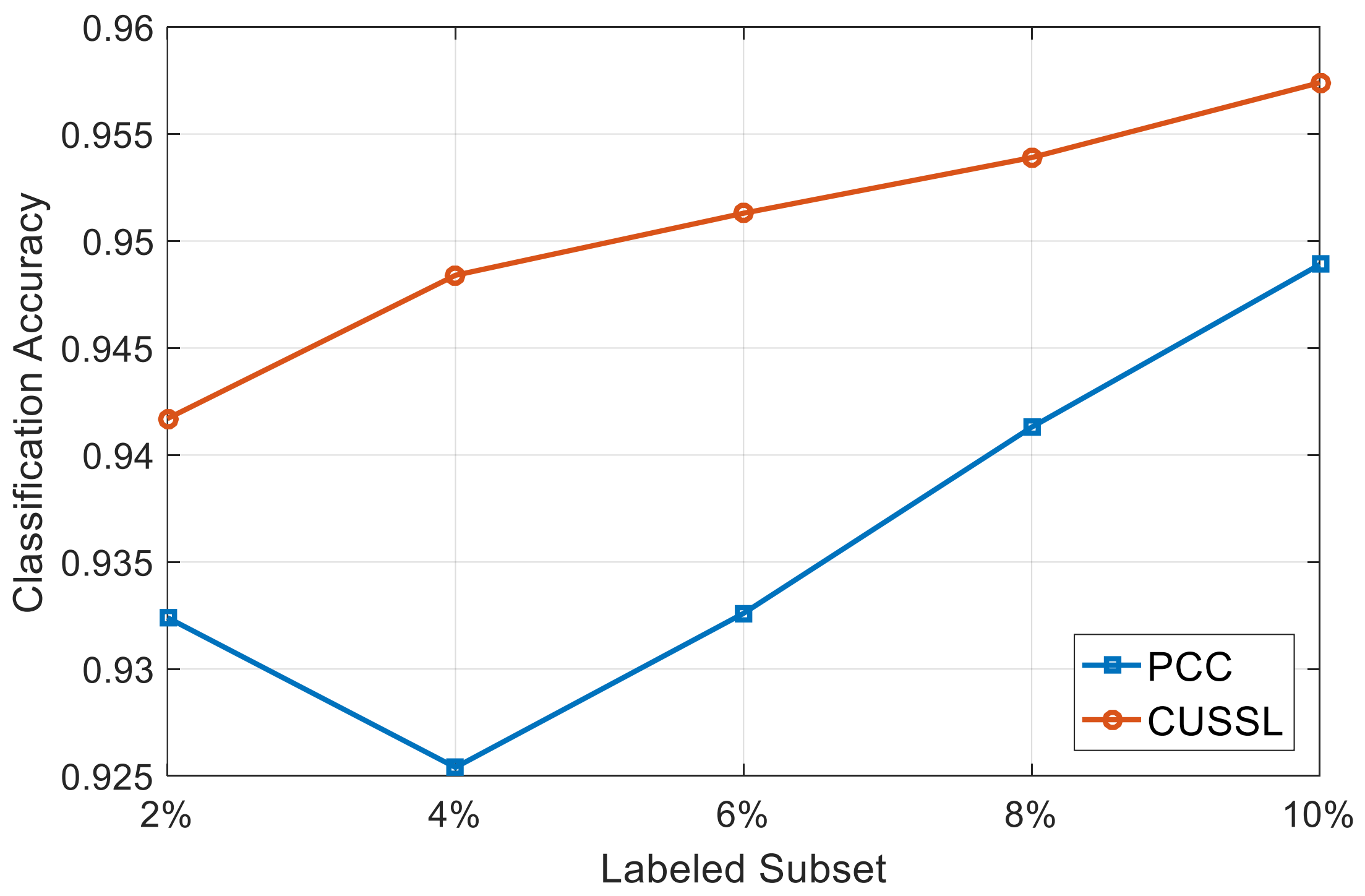
Classification Accuracy on the Iris data set with 2% to 10% labeled samples

Labeled	2%	4%	6%	8%	10%
PCC	90.44%	89.77%	90.52%	91.23%	91.77%
CUSSL	**92.07%**	**91.42%**	**91.43%**	**92.17%**	**92.89%**



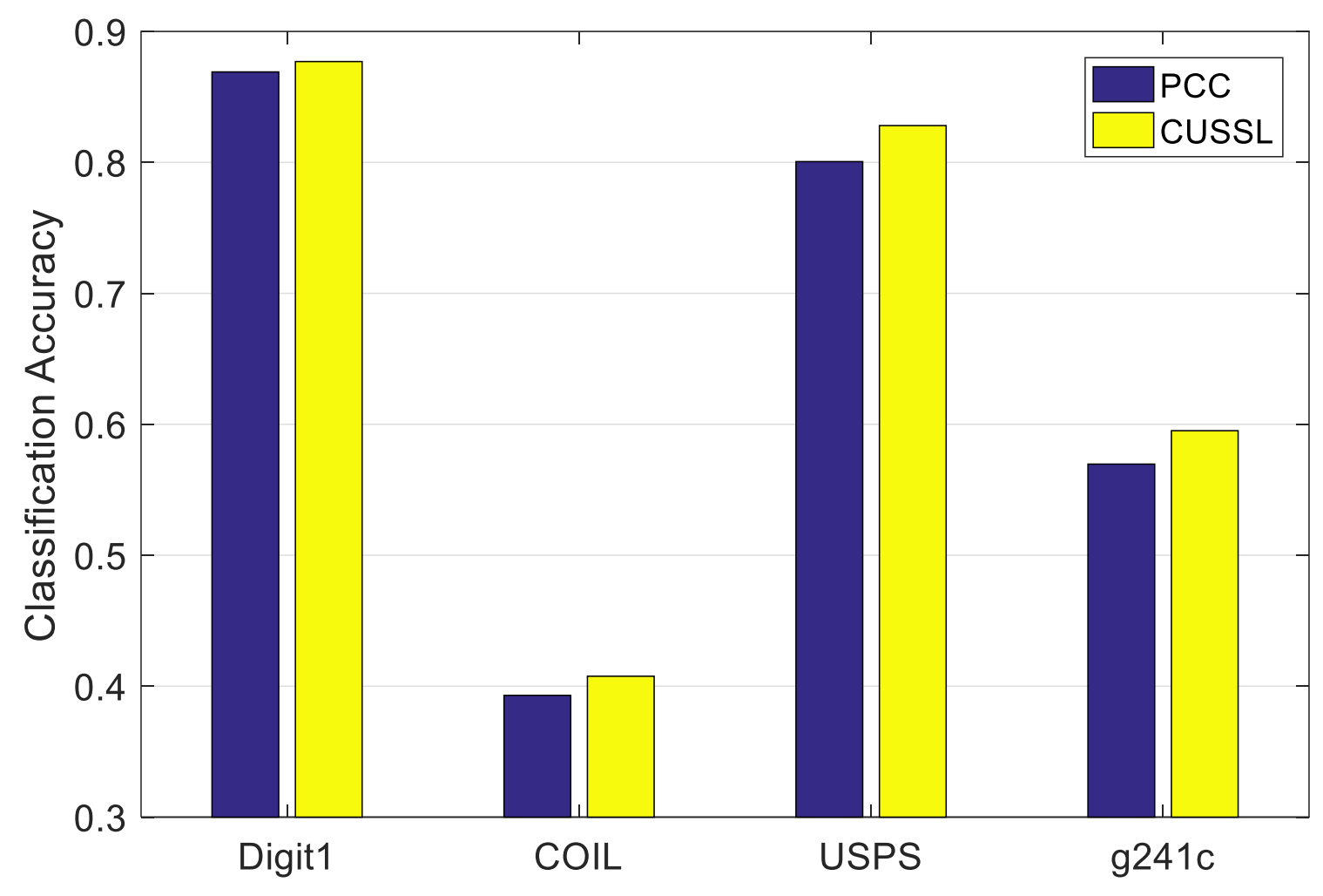
Classification Accuracy on the Wine data set with 2% to 10% labeled samples

Labeled	2%	4%	6%	8%	10%
PCC	93.24%	92.54%	93.26%	94.13%	94.89%
CUSSL	**94.17%**	**94.84%**	**95.13%**	**95.39%**	**95.74%**

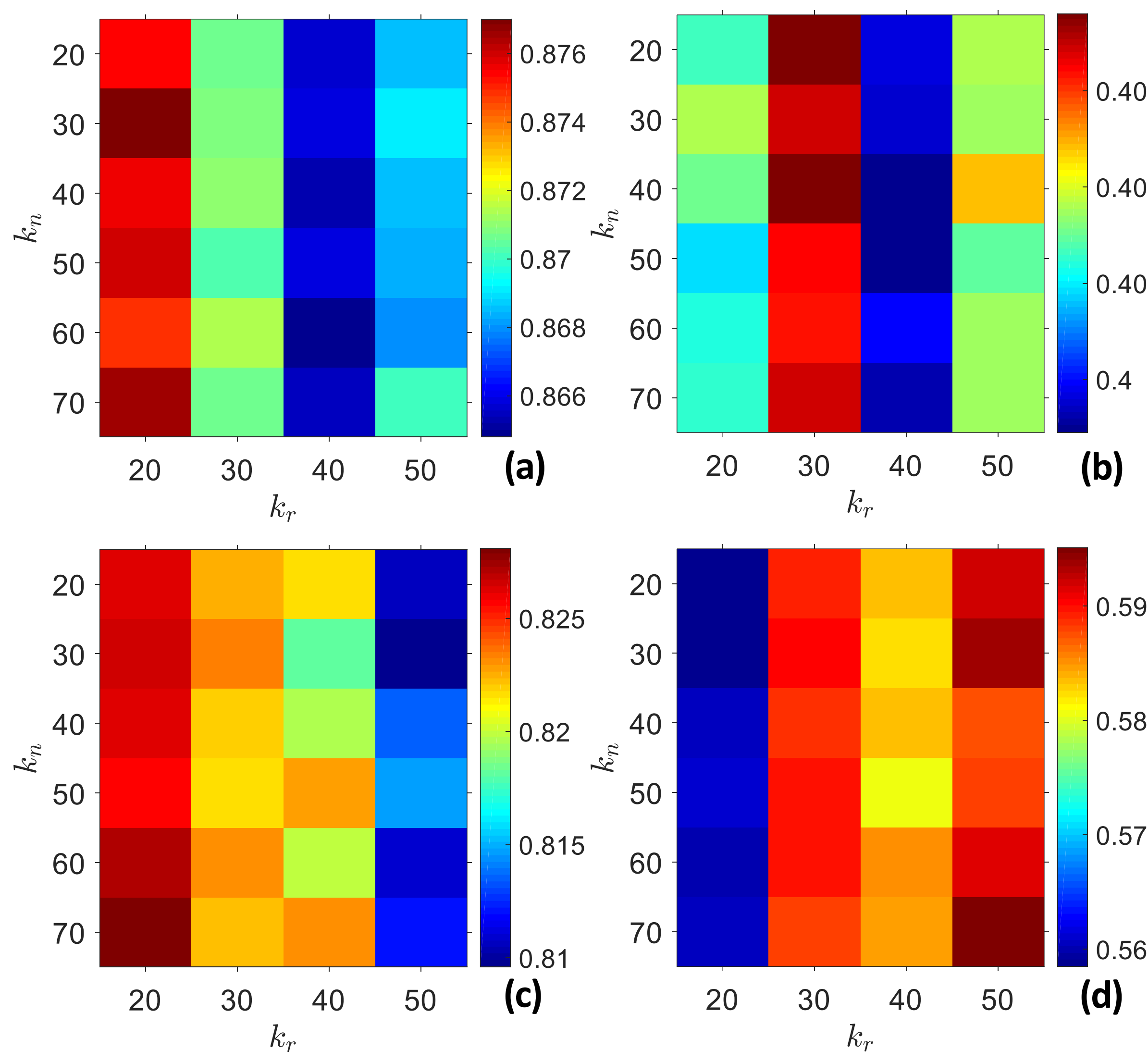


Classification Accuracy on Digit1, COIL, USPS, and g241c data sets with 10 labeled samples

Dataset	Digit1	COIL	USPS	G241c	Mean
PCC	86.91%	39.30%	80.06%	56.96%	65.77%
CUSSL	**87.70%**	**40.76%**	**82.81%**	**59.51%**	**67.28%**



Classification accuracy achieved by CUSSL with different combinations of k_r and k_n on (a) Digit1, (b) COIL, (c) USPS, and (d) g241c data sets with 10 labeled samples



Classification Accuracy on Digit1, COIL, USPS, and g241c data sets with 100 labeled samples

Dataset	Digit1	COIL	USPS	G241c	Mean
PCC	97.31%	74.21%	93.59%	**73.87%**	84.73%
CUSSL	**97.48%**	**76.87%**	**95.19%**	73.85%	**85.00%**

Conclusion

The proposed approach performs an unsupervised distance learning step, without the need of any labeled or training data, through the Reciprocal kNN Distance. The objective consists in exploiting the intrinsic dataset structure for improving the distance among data items. Subsequently, a k-NN graph is computed based on the learned distance and used as input for a semisupervised learning step. The semi-supervised learning step is based on the Particle Competition and Cooperation approach. It combines both labeled and unlabeled data items in its training process.

An experimental evaluation was conducted considering six public datasets, including artificial and real-world data sets. The computer simulations also considered various different size of labeled sets used in the training procedure. The vast majority of experimental results demonstrated the benefits of the combined approach, CUSSL, in comparison to the original PCC method.

Acknowledgments

The authors would like to thank the São Paulo Research Foundation - FAPESP (grants 2011/17396-9 and 2013/08645-0) and the National Counsel of Technological and Scientific Development - CNPq (grants 475717/2013-9) for the financial support.

FAPESP
CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico