IEEE World Conference on Computational Intelligence – WCCI 2010
International Joint Conference on Neural Networks – IJCNN 2010

Semi-Supervised Learning from Imperfect Data through Particle Cooperation and Competition

Fabricio A. Breve[1] fabricio@icmc.usp.br
Liang Zhao[1] zhao@icmc.usp.br
Marcos G. Quiles[2] quiles@unifesp.br

[1] Department of Computer Science, Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), São Carlos, SP, Brazil

[2] Department of Science and Technology (DCT), Federal University of São Paulo (Unifesp), São José dos Campos, SP, Brazil

Outline

- Learning from Imperfect Data
- Proposed Method
  - Nodes and Particles Dynamics
  - Random-Deterministic Walk
- Computer Simulations
- Conclusions

Learning from Imperfect Data

- In Supervised Learning
  - Quality of the training data is very important
  - Most algorithms assume that the input label information is completely reliable
  - In practice mislabeled samples are common in data sets.

Learning from Imperfect Data

- In Semi-Supervised learning
  - Problem is more critical
    - Small subset of labeled data
    - Errors are easier to be propagated to a large portion of the data set
  - Besides its importance and vast influence on classification, it gets little attention from researchers

[33] P. Hartono and S. Hashimoto, "**Learning from imperfect data**," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 353–363, 2007.

[34] D. K. Slonim, **"***Learning from imperfect data in theory and practice*," Cambridge, MA, USA, Tech. Rep., 1996.

[35] T. Krishnan, "**Efficiency of learning with imperfect supervision**," *Pattern Recogn.*, vol. 21, no. 2, pp. 183–188, 1988.

[36] M.-R. Amini and P. Gallinari, "**Semi-supervised learning with an imperfect supervisor**," *Knowl. Inf. Syst.*, vol. 8, no. 4, pp. 385–413, 2005.

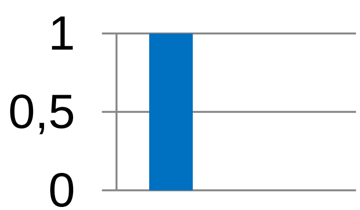
[37] ——, "**Semi-supervised learning with explicit misclassification modeling**," in IJCAI'03: *Proceedings of the 18th international joint conference on Artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 555–560.

Proposed Method

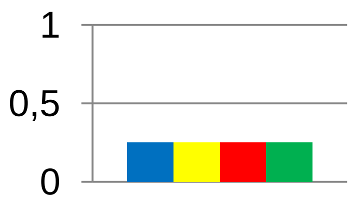
- Particles competition and cooperation in networks
  - Cooperation among particles from the same team (label / class)
  - Competition for possession of nodes of the network
- Each team of particles…
  - Tries to dominate as many nodes as possible in a cooperative way
  - Prevents intrusion of particles from other teams

Initial Configuration

- A particle is generated for each labeled node of the network
  - The node will be called that particle's *home node*
- Particles initial position are set to their respective home nodes.
- Particles with same label play for the same team
- Nodes have a domination vector
  - Labeled nodes have ownership set to their respective teams.
  - Unlabeled nodes have levels set equally for each team



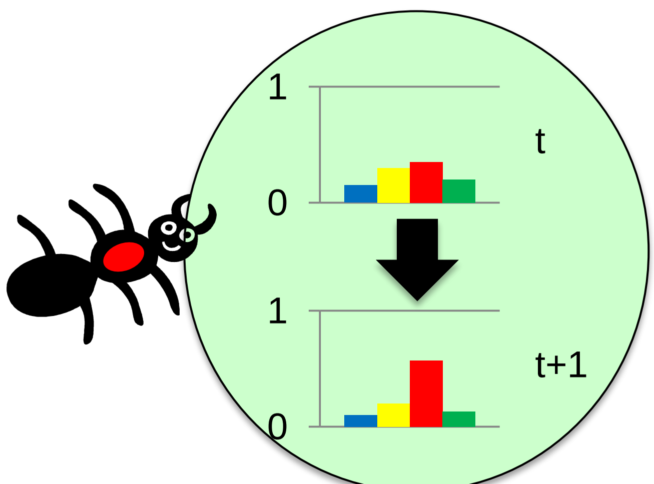
Ex: [1 0 0 0]
(4 classes, node labeled as class A)



Ex: [0.25 0.25 0.25 0.25]
(4 classes, unlabeled node)

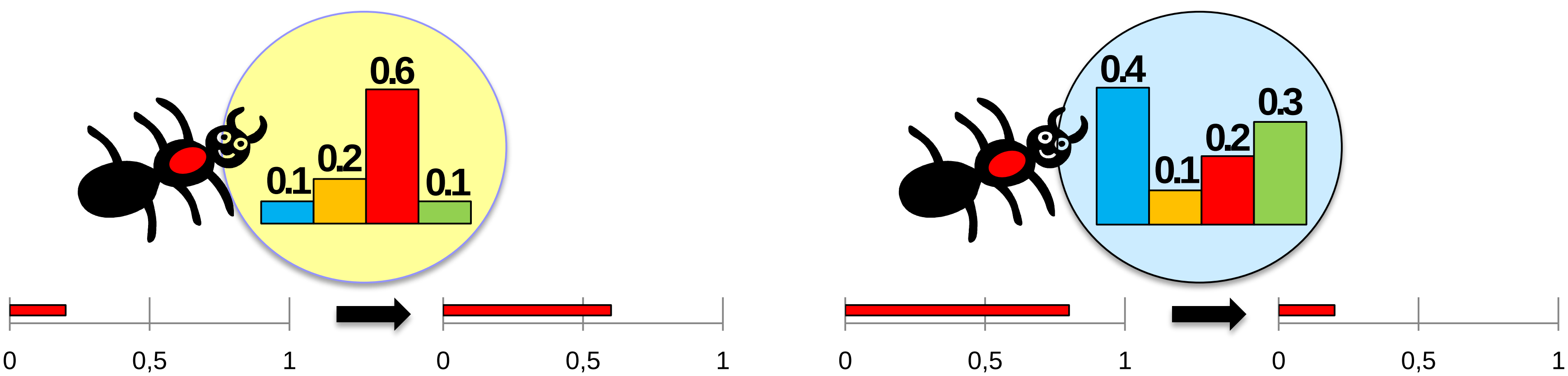
Node Dynamics

- When a particle selects a neighbor to visit:
  - It decreases the domination level of the other teams
  - It increases the domination level of its own team
  - Exception:
    - Labeled nodes domination levels are fixed



Particle Dynamics

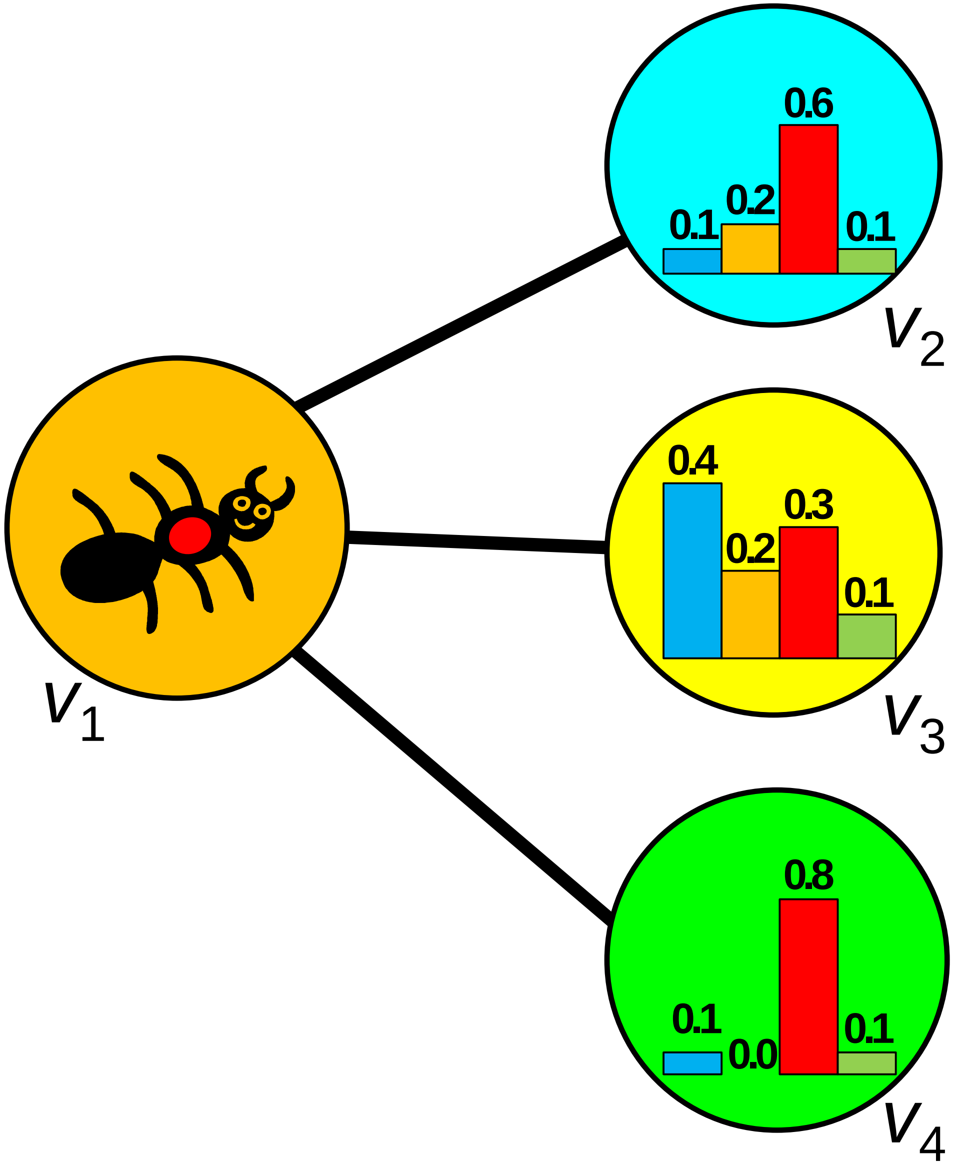
- A particle gets:
  - stronger when it selects a node being dominated by its team
  - weaker when it selects node dominated by other teams



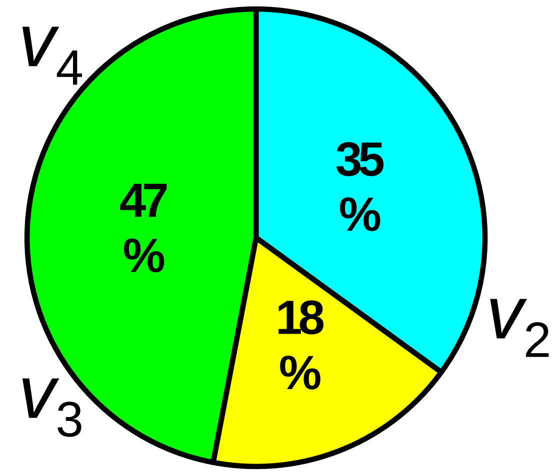
Random-Deterministic Walk

- Random walk
  - The particle randomly chooses any neighbor to visit with no concern about domination levels or distance
- Deterministic walk
  - The particle will prefer visiting nodes that its team already dominates and nodes that are closer to their home nodes

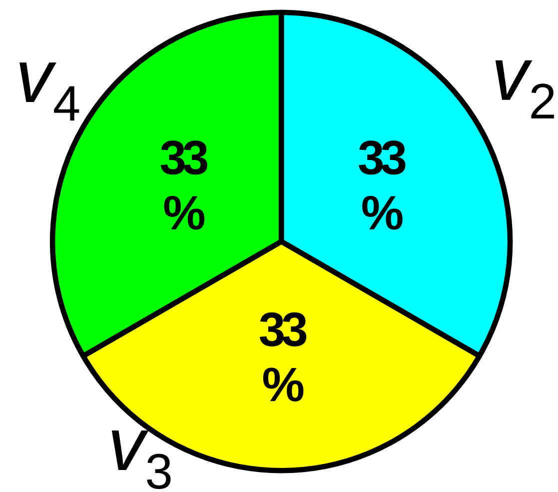
The particles must exhibit both movements in order to achieve an equilibrium between exploratory and defensive behavior



Deterministic Moving Probabilities

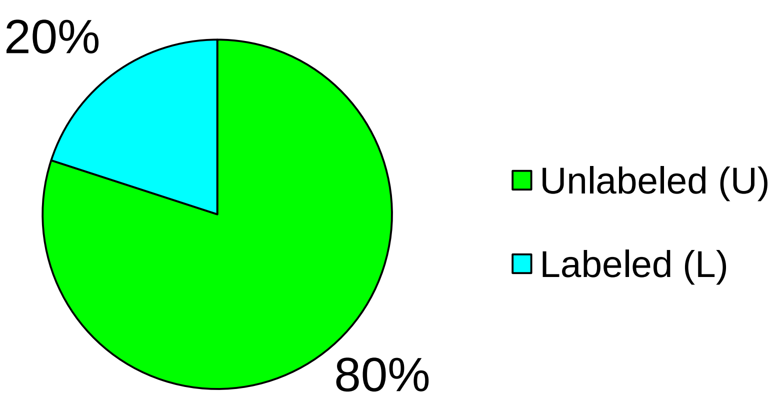


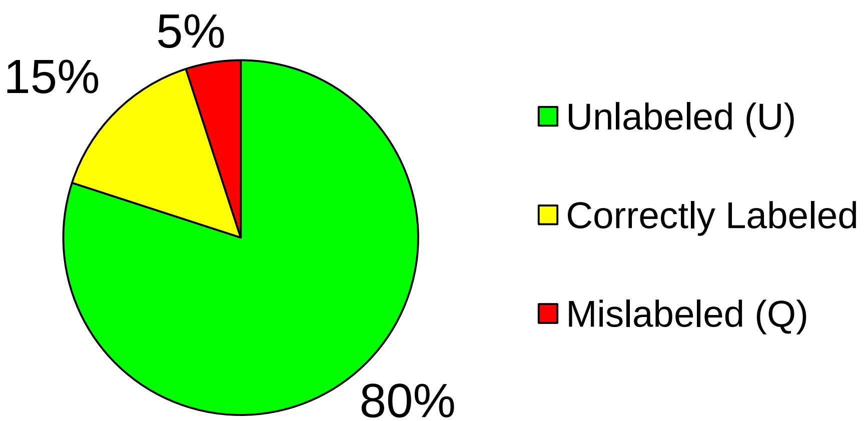
Random Moving Probabilities



Computer Simulations

- Network are generated with:
  - Different sizes and mixtures
  - Elements divided into 4 classes
  - Set of nodes N
  - Labeled subset $L \subset N$
  - Mislabeled subset $Q \subset \mathrm{L} \subset N$





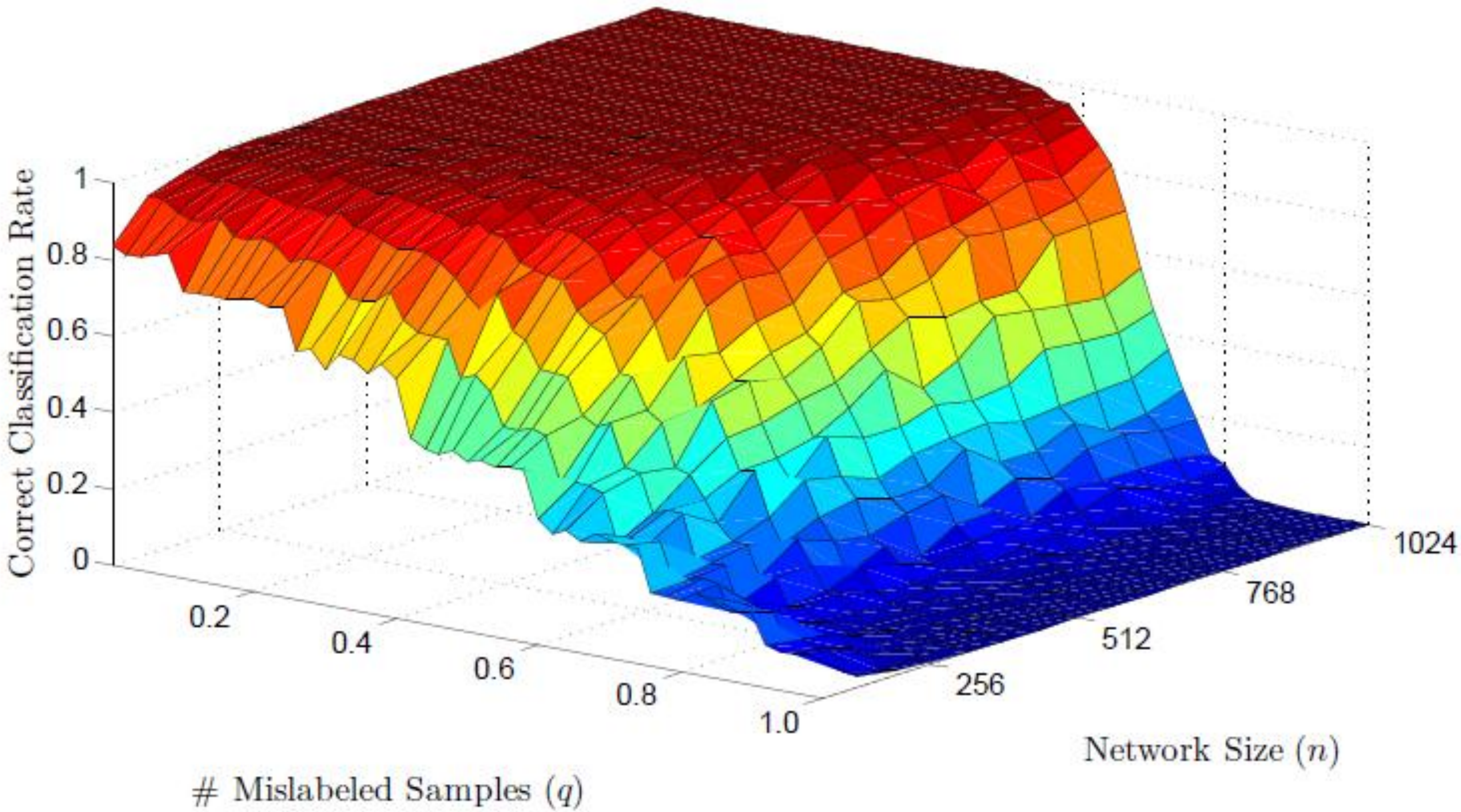


Fig. 3: Correct Classification Rate with different network sizes and mislabeled subset sizes. $z_{out}/\langle k \rangle = 0.25$, $l/n = 0.1$.

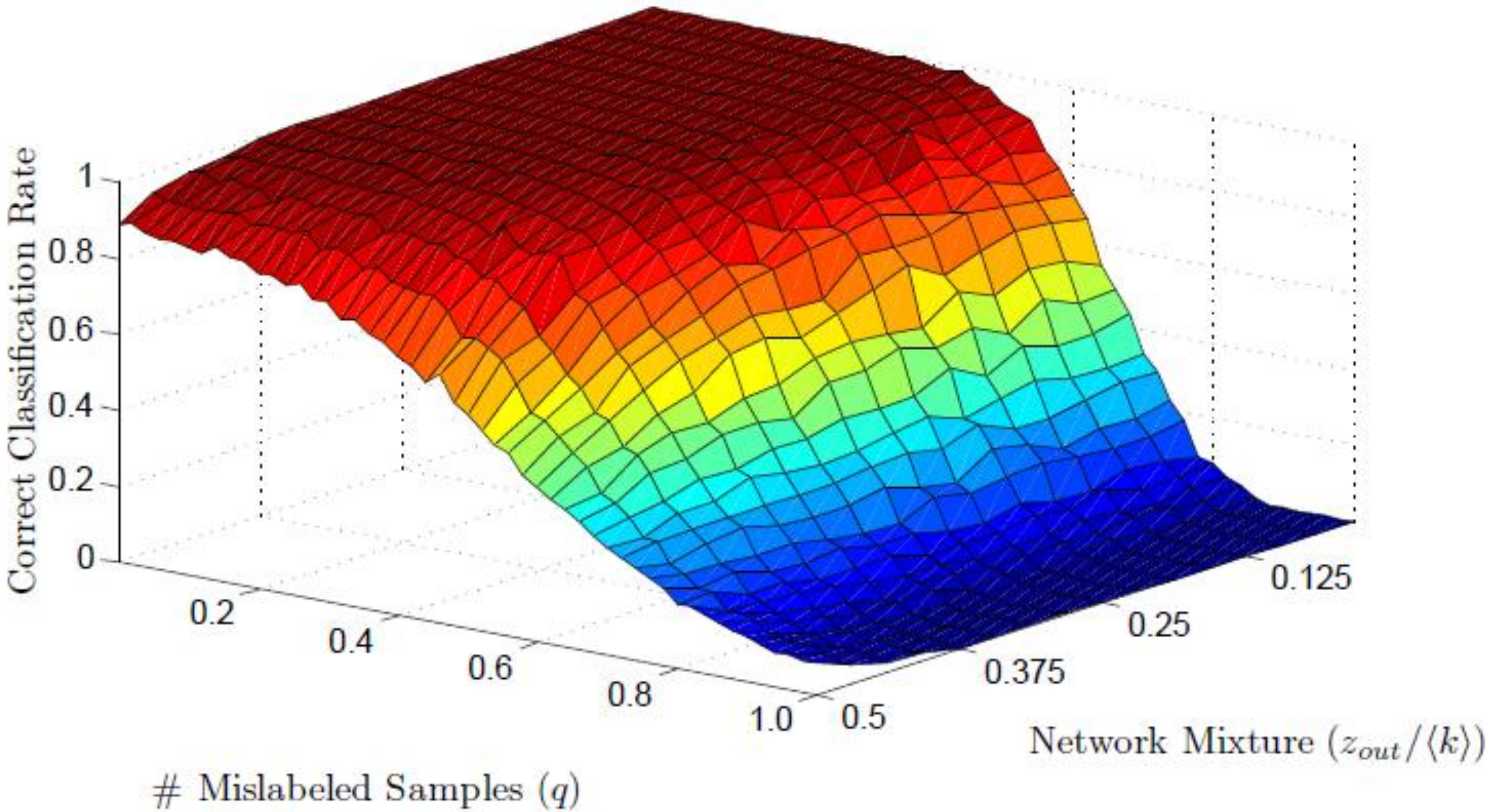


Fig. 4: Correct Classification Rate with different network mixtures and mislabeled subset sizes. $n = 512$, $l = 64$.

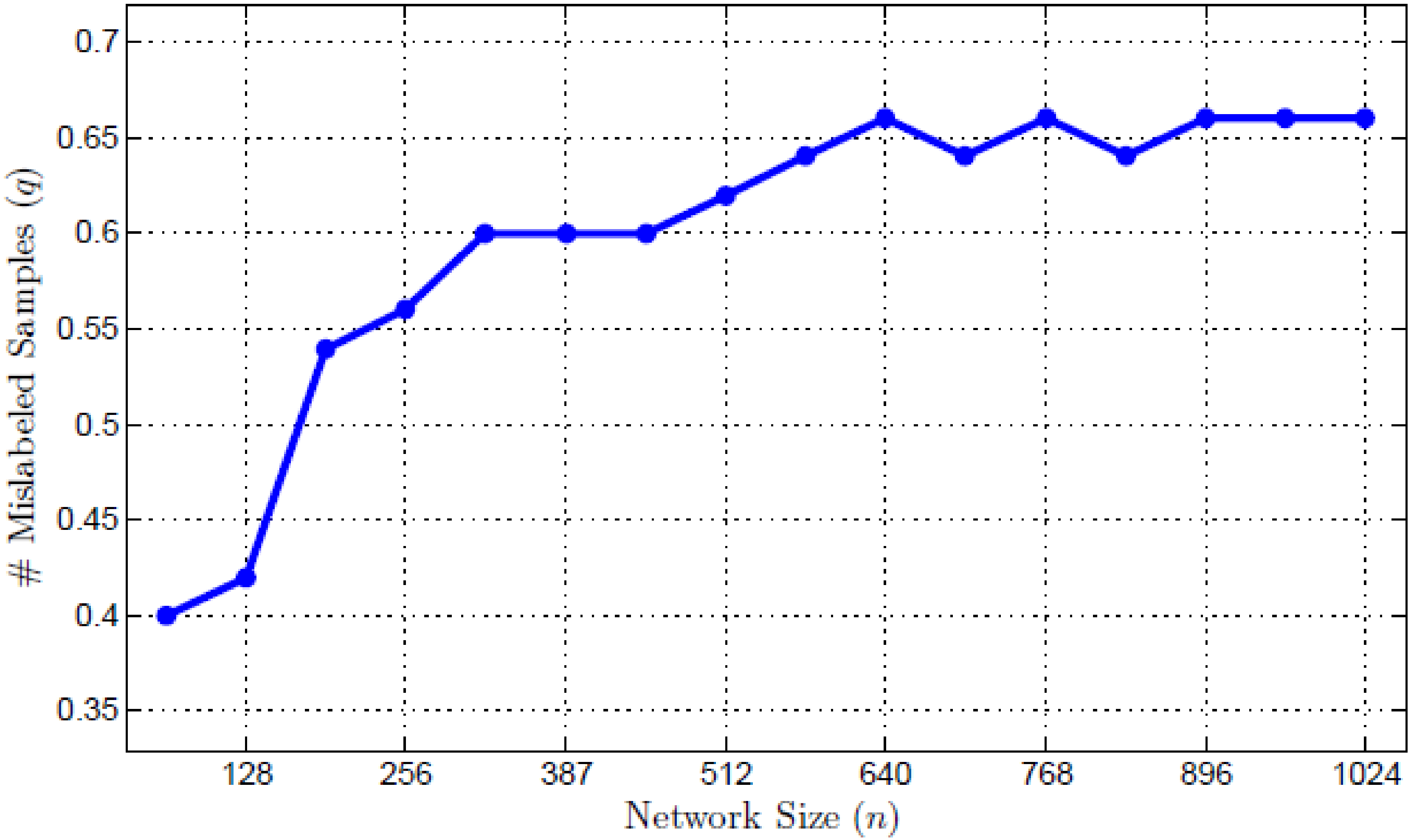


Fig. 5: The first critical point in the mislabeled samples curves with different network sizes. $z_{out}/\langle k \rangle = 512$, $l/n = 0.1$.

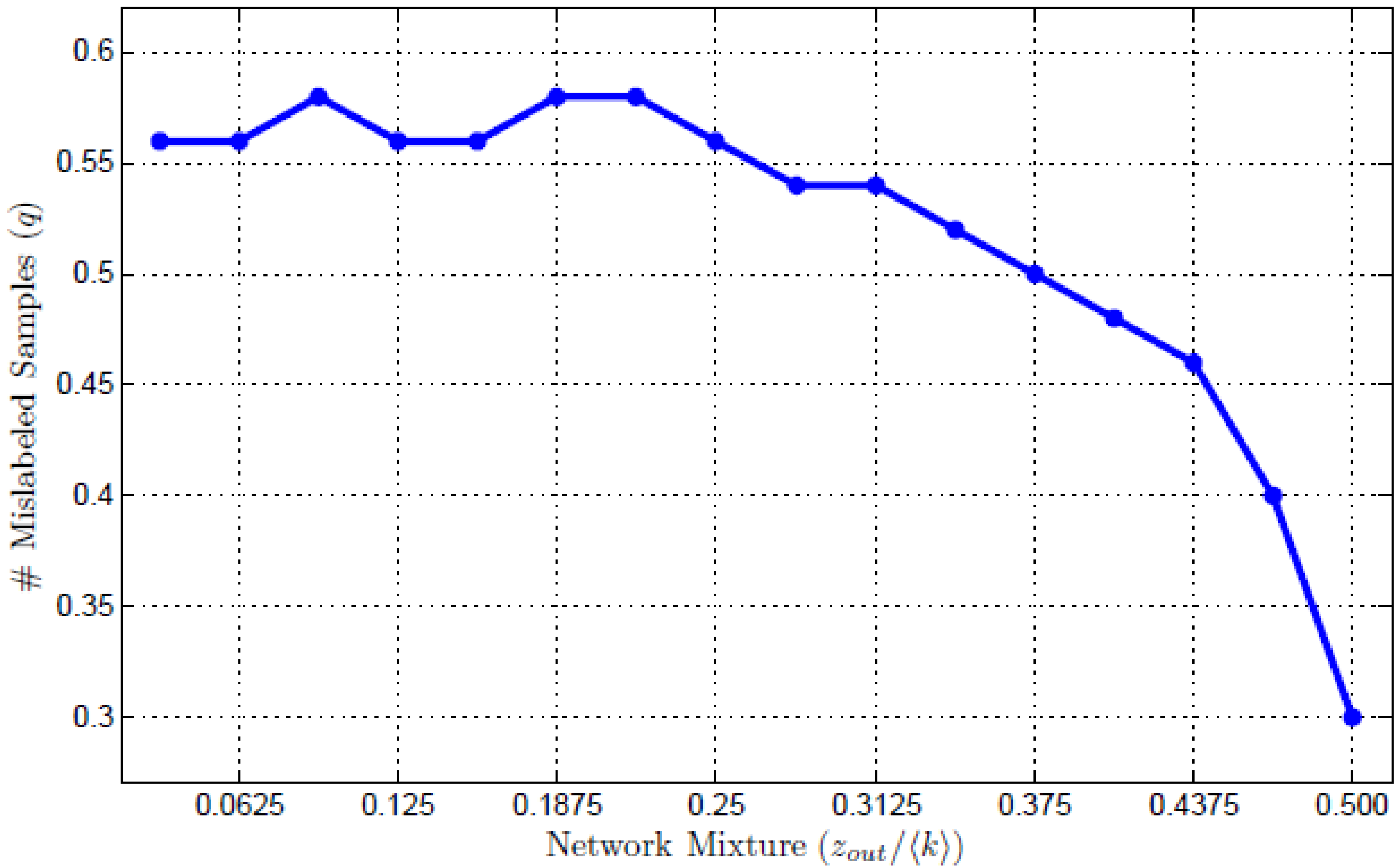


Fig. 6: The first critical point in the mislabeled samples curves with different network mixtures. $n = 512$, $l = 64$.

TABLE I: Maximum mislabeled subset size for different network sizes (n). $z_{out}/\langle k \rangle = 0.250$, $l/n = 0.1$.

n	Correct Classification Rate $> 90\%$	Correct Classification Rate $> 80\%$	n	Correct Classification Rate $> 90\%$	Correct Classification Rate $> 80\%$
64	-	8%	576	52%	58%
128	10%	26%	640	54%	58%
192	26%	44%	704	56%	60%
256	40%	48%	768	56%	60%
320	44%	48%	832	56%	60%
384	48%	52%	896	58%	62%
448	48%	56%	960	58%	62%
512	52%	56%	1024	60%	64%

TABLE II: Maximum mislabeled subset size for different network mixtures ($z_{out}/\langle k \rangle$). $n = 512$, $l = 64$.

$z_{out}/\langle k \rangle$	Correct Classification Rate $> 90\%$	$> 80\%$	$z_{out}/\langle k \rangle$	Correct Classification Rate $> 90\%$	$> 80\%$
0.0313	52%	56%	0.2813	50%	54%
0.0625	52%	56%	0.3125	48%	54%
0.0938	50%	58%	0.3438	48%	52%
0.1250	52%	56%	0.3750	46%	50%
0.1563	52%	56%	0.4063	40%	48%
0.1875	52%	58%	0.4375	32%	44%
0.2188	50%	56%	0.4688	22%	40%
0.2500	52%	56%	0.5000	-	26%

Conclusions

- Biologically inspired method for semi-supervised classification
  - Cooperation and competition among particles
  - Natural way of preventing error propagation from mislabeled nodes
- Analysis of the results
  - Critical points in the performance curve as the mislabeled samples subset grows

Future Work

- Different types of networks
- Real-world data-sets
- Comparison with other methods

Acknowledgements

- This work was supported by the State of São Paulo Research Foundation (FAPESP) and the Brazilian National Council of Technological and Scientific Development (CNPq)

IEEE World Conference on Computational Intelligence – WCCI 2010
International Joint Conference on Neural Networks – IJCNN 2010

Semi-Supervised Learning from Imperfect Data through Particle Cooperation and Competition

Fabricio A. Breve[1] fabricio@icmc.usp.br
Liang Zhao[1] zhao@icmc.usp.br
Marcos G. Quiles[2] quiles@unifesp.br

[1] Department of Computer Science, Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), São Carlos, SP, Brazil

[2] Department of Science and Technology (DCT), Federal University of São Paulo (Unifesp), São José dos Campos, SP, Brazil