Query Rules Study on Active Semi-Supervised Learning using Particle Competition and Cooperation

Fabricio Breve
Department of Statistics, Applied Mathematics and Computation (DEMAC)
Institute of Geosciences and Exact Sciences (IGCE), São Paulo State University (UNESP)
Rio Claro, São Paulo, Brazil
Email: fabricio@rc.unesp.br

Abstract

Semi-Supervised Learning and Active Learning are important machine learning classification techniques used mostly when labeled data are scarce. Semi-supervised learning techniques employ both labeled and unlabeled data in their training process, while active learning techniques interactively query a label source, like a human specialist, to get the labels of data points selected while the algorithm executes. Recently, a method combining those two approaches, based on particle competition and cooperation in networks, was proposed. It has two versions which differ on the rule used to query the information source. One is based on currently assigned label uncertainty and the other is based on nodes distance to closest labeled nodes. The best approach depends on the data set being analyzed. In this paper, a set of computer simulations is performed to better understand how the different query rules affect the classification accuracy of this method. First, a new query rule is introduced merging both previous rules into a single one. Later, the classification accuracy when the method is applied to some real-world data sets is compared with those achieved using the original rules. The new query rule lead to better classification accuracy than previous rules in most scenarios.

1. Introduction

Semi-Supervised Learning and Active Learning are important machine learning classification techniques used mostly on problems where unlabeled data is easily acquired, but the process of assigning them labels is expensive, time consuming, and/or requiring the intense work of human specialists [1]–[3]. Semi-supervised learning techniques employ both labeled and unlabeled data in their training process, overcoming a limitation of supervised and unsupervised learning methods which use only labeled or unlabeled data, respectively, in their training process. On the other hand, active learning techniques are able to interactively query a label source, like a human specialist, to get the labels of data points selected while the algorithm executes. They work on the assumption that fewer labeled items are needed if the algorithm is allowed to choose which of the data items will be labeled [4], [5].

Active learning methods are usually split in categories according to how they choose which data items will have their labels queried. One of the most used approaches is known as *uncertainty sampling*, in which the least confidence assigned labels are chosen to be queried [6]–[9]. Semi-supervised learning methods may also be split in some categories. The most active one is the graph-based methods category. It includes methods like Mincut [10], Local and Global Consistency [11], Particle Competition and Cooperation [12], label propagation techniques [13], [14], among others.

Recently, the semi-supervised learning graph-based method known as particle competition and cooperation [12] was extended to perform active learning as well [15]. In the extended method, a network is built from the data set using the distance among data items. A particle is then created for each labeled node. Particles walk in the network trying to dominate as many nodes as possible. Particles with the same label cooperate among themselves, while particles of different labels

This work was supported by the São Paulo State Research Foundation (FAPESP) and the Brazilian National Research Council (CNPq).

compete against each other. Labels are spread as particles move from node to node. The algorithm starts with as few as a single pre-labeled node per class. Then, each time the algorithm queries a node label, a new particle is created in the corresponding node. There is no retraining process, the algorithm dynamically chooses nodes to have their labels queried. Therefore, this method is usually much faster than those that require retraining after new queries. Particles usually dominate the unlabeled nodes which are closer to their corresponding node quickly, but the nodes on frontier regions and the nodes on dense regions without labeled nodes are often the stage of arduous disputes. Therefore, the algorithm query the labels of the nodes on those regions.

There are two versions of the algorithm. They differ on the rules used for choosing unlabeled nodes to be queried. The first version is a *querying by uncertainty* approach, where the algorithm always queries for the most dubious unlabeled network node, which is likely to be one of the stages of most disputes over time. The second version alternates between querying the most dubious unlabeled network node (like the first one) and querying the unlabeled network node which is more far away from any labeled node, according to distance tables dynamically built by the walking particles, and thus avoid classification mistakes in large regions. And which version achieves the higher classification accuracy? The answer depends on the data set in which the method is being applied.

In this paper, a set of experiments is developed to better understand how the different query rules affect classification accuracy of the active semi-supervised particle competition and cooperation method. First, a new query rule is introduced. It merges both rules from [15] into a single one, with a parameter to define weights to *uncertainty* and *distance to labeled nodes* criteria on the choice of the node to be queried. Later, the classification accuracy of the algorithm using the new query rule is compared with those achieved by the original rules on [15]. This study is important to better understand how the choice of query rule affects classification performance in each data set, so in the future the weighting parameter may be selected automatically, either before or during the algorithm execution.

The remaining of this paper is organized as follows. Section 2 presents an overview of the active semi-supervised learning particle competition and cooperation model. In Section 3, the new combined query rule is described. Section 4 shows some computer simulations. Discussion of the results is presented on Section 5. Finally, some conclusions are drawn on Section 6.

2. Active Semi-Supervised Learning Particle Competition and Cooperation

Overall, the active semi-supervised learning particle competition and cooperation method [12], [15] works as follows. First, the vector-based data set is converted to a non-weighted and undirected graph. Each data item becomes a graph node. Edges connecting the nodes are created according to the Euclidean distance between the nodes in the data feature space, no matter if they are labeled or not. Then, a particle is created for each labeled node. Particles with the same label belong to the same team and cooperate among themselves to dominate the unlabeled nodes in their neighborhood. On the other hand, particles with different labels compete against each other for the possession of the nodes. When the system runs, the particles walk in the graph, selecting the next node to visit according to a random-greedy rule. Each node has a set of domination levels, one level for each class of the problem. When a particle visits a node, it increases its class domination level on that node, at the same time that it decreases the domination level of the other classes. Each particle has a strength level, which lowers or raises according to the domination level of its class in the node being visited. Particles also have a distance table which they update dynamically as they walk on the graph. At the end of the iterative process, each data item is labeled after the class with the highest domination level on it. The original particle competition and cooperation algorithm is detailed described in [12].

The main novelty introduced by [15] is that every time the system reaches some stability levels, it queries one node label and creates a new particle corresponding to this new labeled node. The system runs until it stabilizes again, when it queries another label. This procedure is repeated until the target amount of labeled nodes is reached. The stability level is taken by monitoring the average maximum domination levels of the nodes $\langle v_i^{\omega_\ell} \rangle$, where $\ell = \arg\max_q v_i^{\omega_q}$ and $v_i^{\omega_q}$ is the domination of class/team q over node v_i. This measure usually increases fast when the system begins, but it quickly slows down and then it oscillates around the maximum point. When there is no more increase in the highest level achieved by $\langle v_i^{\omega_\ell} \rangle$, the system reached its highest stability level.

There are two version of the algorithm which differ in the rule used to choose the node which label will be queried. The first version, AL-PCC v1, selects the unlabeled node that the algorithm is most uncertain (least confidence) about which label it should have. The second version, AL-PCC v2, alternates between querying the most uncertain unlabeled network node and querying the unlabeled node which is more far away from any labeled node, according to the distances dynamically measured as particles walk.

To select the network node with most uncertain label at any given time, first the degree of uncertainty in each node is calculated using the following equation:

$$u_i(t) = \frac{v_i^{\ell**}(t)}{v_i^{\ell*}(t)}, \tag{1}$$

where $v_i^{\ell*}(t) = \arg\max_\ell v_i^\ell(t)$, $v_i^{\ell**}(t) = \arg\max_{\ell,\ell\neq v_i^{\ell*}(t)} v_i^\ell(t)$, and $v_i^{\omega_\ell}(t) \in [0 \quad 1]$ corresponds to the domination level from class ℓ over node v_i. $u_i \in [0 \quad 1]$, where $u_i = 0$ means completely confidence in the label assigned to a node, while $u_i = 1$ means total uncertainty in the label assigned. The node with the most uncertain label is then defined using:

$$q(t) = \arg\max_i u_i(t). \tag{2}$$

To select the node which is more far away from any labeled node, first the distance from each node to its closest labeled node is calculated as follows:

$$s_i(t) = \min_j \rho_j^{d_i}(t). \tag{3}$$

where $\rho_j^{d_i}(t) \in [0, \quad n-1]$ corresponds to the distance dynamically measured between the particle's corresponding node v_j and the node v_i. Then, the node which is more far away from any labeled node is defined using:

$$q(t) = \arg\max_i s_i(t). \tag{4}$$

AL-PCC v1 queries for a new label using (2), while AL-PCC v2 queries for a new label alternating between (2) and (4).

3. The New Query Rule

The new query rule is introduced to better understand how the query rules may affect classification accuracy. It combines both rules from (2) and (4) into a single one, including a parameter to define weights to the *assigned label uncertainty* and to the *distance to labeled nodes* criteria on the choice of the node to be queried.

The calculation of the combined query rule requires the previous calculation of uncertainty levels and distance levels using (1) and (3), respectively. Then, both vectors $u(t)$ and $u(t)$ must be normalized to the $[0 \quad 1]$ interval as follows:

$$u_i'(t) = \frac{u_i(t)}{\max_i u_i(t)}, \tag{5}$$

$$s_i'(t) = \frac{s_i(t)}{\max_i s_i(t)}. \tag{6}$$

Finally, the new combined measures are calculated as follows:

$$q(t) = \arg\max_i \beta u_i'(t) + (1-\beta)s_i'(t), \tag{7}$$

where $\beta \in [0 \quad 1]$ is the parameter to control the levels of *assigned label uncertainty* and *distance to labeled nodes* criteria to be used. When $\beta = 0$ only the *distance to labeled nodes* criterion is used. As β increases, the importance of the *assigned label uncertainty* in the combined query rule also increases while the *distance to labeled nodes* importance decreases. When $\beta = 1$ only the *assigned label uncertainty* criterion is used.

4. Computer Simulations

In the first set of computer simulations, the goal is to better understand the role of the query rule on the particle competition and cooperation model. Therefore, the particle competition and cooperation method (presented in Section 2), including the the new combined query rule (introduced in Section 3), is applied to some real-world data sets, presented on Table 1. The k parameter used during the graph construction step (each node connects to its k-nearest neighbors) is set to $k = 5$ in all experiments. This value usually produces optimal or near-optimal results in most scenarios, as observed in several empirical experiments. One could fine-tune it for each data set or even each individual configuration of the other parameters to achieve slightly higher classification accuracy.

Figure 1 shows the classification accuracy results when the proposed method is applied to 9 different data sets. In each simulation, the β parameter, which controls the combined query rule, is varied from 0 to 1, in 0.1 steps. The target amount of labeled nodes q

Table 1. Basic properties of the selected data sets

Data Set	Classes	Dimensions	Points	References
Iris	3	4	150	[16]
Wine	3	13	178	[16]
g241c	2	241	1500	[2]
Digit1	2	241	1500	[2]
USPS	2	241	1500	[2]
COIL	6	241	1500	[2]
COIL2	2	241	1500	[2]
BCI	2	117	400	[2]
Semeion Handwritten Digit	10	256	1593	[17], [18]

varies from 1% to 10% of the data set size, i.e., the algorithm starts with only one pre-labeled node per class, then it will query other nodes labels until the target amount is reached. Each point in the graphics is the average of 100 executions with different random selections of pre-labeled nodes. Notice that for Iris and Wine data sets (Figures 1a and 1b), simulations with only 1% and 2% labeled data set sizes are skipped, as only one labeled node per class would already pass the 2% mark, leaving no room for label queries.

In the second set of computer simulations, the classification accuracy results obtained from the first set are compared with those obtained using the original particle competition and cooperation method (PCC) [12] and those obtained using its active semi-supervised learning versions (AL-PCC v1 and AL-PCC v2), [15]. The following parameters were fixed for the PCC method: $p_{grd} = 0.5$ and $\Delta_v = 0.1$. For all the methods, $k = 5$ is fixed, as in the last set of experiments.

In each experiment, the PCC method is executed using 1% to 10% data items randomly chosen to be pre-labeled (since the algorithm start), as it requires. For AL-PCC v1 and AL-PCC v2, only one data item per class is randomly chosen to be pre-labeled. Then, each time the algorithm reaches stability, it queries the label of another node chosen according to its specific rules, like done in the first set of simulations. This is repeated until the defined amount of labeled items (1% to 10%) is reached. The results from the proposed method with the new combined query rule (CQR-AL-PCC) are taken from the first experiment, using both the best and the worst β choices.

Figure 2 shows the classification accuracy comparison when the methods are applied to the 9 different data sets. Once more, each point in the graphics from Figure 2 is the average of 100 executions with different pre-labeled nodes. In most scenarios, the method using the proposed combined query rule obtained better performance than the others.

5. Discussion

The first set of simulations confirmed that most data sets have some predilection for the query rule parameter β, even though the thresholds, the effective ranges of β, and the influence of a bad choice of β vary from one data set to another. The predilection for a particular choice of β seems to depend on data set properties, like data density, classes separation, etc. For instance, the emphasis on the *distance to labeled nodes* criterion may work better when classes have highly overlapped regions, many outliers, more than one cluster inside a single class, etc. In these scenarios, uncertainty measure may fail to detect large regions of the network completely dominated by the wrong team of particles, due to an outlier or the lack of correctly labeled nodes in that area. On the other hand, the *assigned label uncertainty* criterion may work better when classes are fairly well separated and there are not many outliers. In these scenarios, less particles may take care of a large region of the team territory, thus new particles created using the uncertainty criterion may help finding the classes boundaries.

The second set of simulations showed that the new combined query rule with the proper β parameter leads to the best classification accuracy in most scenarios, but a bad choice of β usually leads to major classification accuracy loss.

6. Conclusions

The computer simulations show how the different choices of query rules affect the classification accuracy of the active semi-supervised learning particle competition and cooperation method applied to different real-world data sets. The optimal choice of the newly introduced β parameter led to better classification accuracy in most scenarios. As future work, it is important to look for some possible correlation between information that can be extracted from the network a priori and the optimal β parameter, so it could be selected automatically.

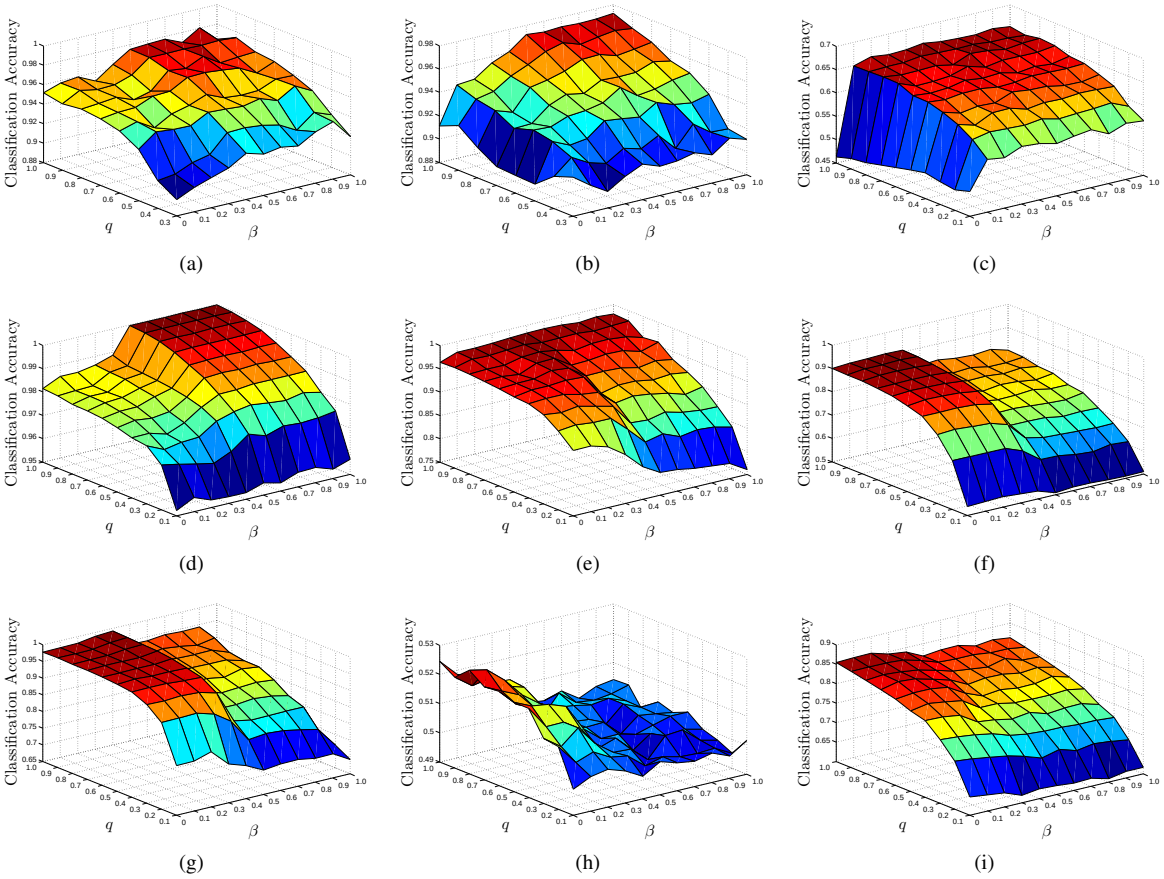

Figure 1. Classification accuracy when the proposed method is applied to different data sets with different β parameter values and labeled data set sizes (q). The data sets are: (a) Iris [16], (b) Wine [16], (c) g241c [2], (d) Digit1 [2], (e) USPS [2], (f) COIL [2], (g) COIL_2 [2], (h) BCI [2], and (i) Semeion Handwritten Digit [17], [18]

.

References

[1] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[2] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, 2006.

[3] S. Abney, *Semisupervised Learning for Computational Linguistics*. CRC Press, 2008.

[4] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[5] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," Swedish Institute of Computer Science, Box 1263, SE-164 29 Kista, Sweden, Tech. Rep. T2009:06, April 2009.

[6] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 3–12.

[7] D. Cohn, R. Ladner, and A. Waibel, "Improving generalization with active learning," in *Machine Learning*, 1994, pp. 201–221.

[8] G. Schohn and D. Cohn, "Less is more: Active learn-

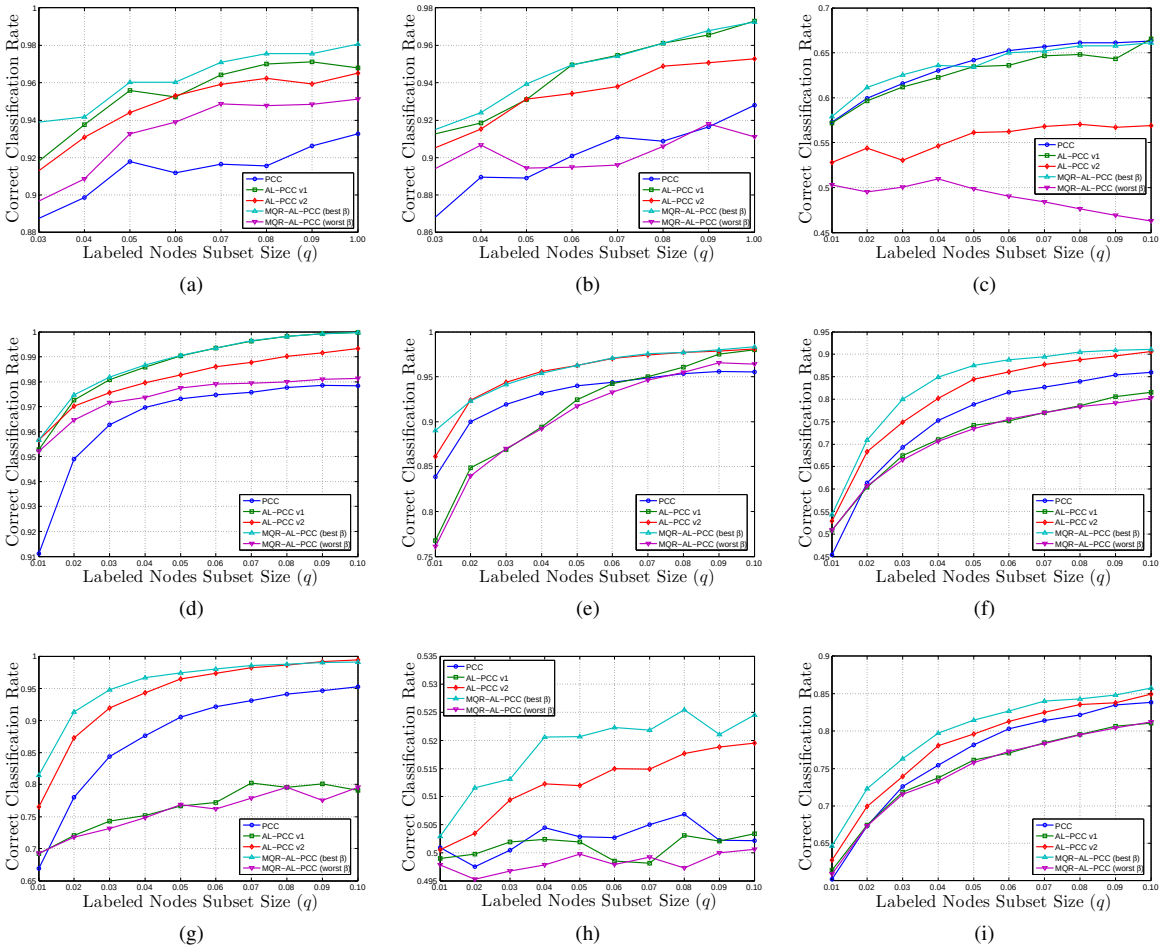

Figure 2. Comparison of the classification accuracy when all the methods are applied to different data sets with different labeled data set sizes (q). The data sets are: (a) Iris [16], (b) Wine [16], (c) g241c [2], (d) Digit1 [2], (e) USPS [2], (f) COIL [2], (g) COIL$_2$ [2], (h) BCI [2], and (i) Semeion Handwritten Digit [17], [18].

ing with support vector machines," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 839–846.

[9] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, ser. IDA '01. London, UK, UK: Springer-Verlag, 2001, pp. 309–318.

[10] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003, pp. 58–65.

[11] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, 2004, pp. 321–328.

[12] F. Breve, L. Zhao, M. Quiles, W. Pedrycz, and J. Liu, "Particle competition and cooperation in networks for semi-supervised learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 9, pp. 1686 –1698, sept. 2012.

[13] X. Zhu and Z. Ghahramani, "Learning from la-

beled and unlabeled data with label propagation," Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-CALD-02-107, 2002.

[14] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 55–67, Jan. 2008.

[15] F. Breve, "Active semi-supervised learning using particle competition and cooperation in networks," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Aug 2013, pp. 1–6.

[16] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[17] Semeion Research Center of Sciences of Communication, via Sersale 117, 00128 Rome, Italy.

[18] Tattile Via Gaetano Donizetti, 1-3-5,25030 Mairano (Brescia), Italy.