

Graph-Based Cross-validated Committees Ensembles

Nils Ever Murrugarra Llerena
Computer Science Department
University of Pittsburgh
Pittsburgh, PA, USA
nineil@cs.pitt.edu

Lilian Berton, Alneu de Andrade Lopes
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
13560-970 São Carlos, SP, Brazil
{lberton, alneu}@icmc.usp.br

Abstract—Ensemble techniques combine several individual classifiers to obtain a composite classifier that outperforms each of them alone. Despite of these techniques have been successfully applied to many domains, their applications on networked data still need investigation. There are not many known strategies for sampling with replacement from interconnected relational data. To contribute in this direction, we propose a cross-validated committee ensemble procedure applied to graph-based classifiers. The proposed ensemble either maintains or significantly improves the accuracy of the tested relational graph-based classifiers. We also investigate the role played by diversity among the several individual classifiers, i.e., how much they agree in their predictions, to explain the technique success or failure.

Keywords—ensembles; cross-validated committees; graph-based learning;

I. INTRODUCTION

Ensembles algorithms encompass important techniques to improve model accuracy in classification tasks by reducing variance and improving the stability of predictions. The main idea behind an ensemble is to combine individual classifiers to obtain a classifier that outperforms each of them separately used. Studies on this technique have led to an active area of research in machine learning successfully applied in a variety of applications ranging from bioinformatics [19] to geography [2].

Currently, data represented by graphs (networked data) such as social networks, web pages, among others, are very common and fast growing. Hence, researches on techniques to improve graph-based relational classifiers are certainly relevant. However, there are not many known strategies for sampling with replacement from interconnected relational data represented as a graph with dependencies among related instances. Furthermore, as Jensen and Neville [9] evidence some characteristics of relational data, like concentrated linkage and relational autocorrelation, reduce the effective sample size of relational datasets.

In a previous work we proposed a graph-based bagging technique in which we investigate a bagging procedure applied to networked data [14]. Here, we extend such investigation considering another ensemble strategy for graph-based classifier. The main difference between this work and [14] is the way of considering the sampling phase. In this

work, we present a new technique referred to as Graph-Based Cross-Validated Committees. The proposed algorithm either maintains or significantly improves the performance of the base classifiers.

The remainder of this paper is organized as follows. Section 2 discusses background and related work on ensemble techniques. Section 3 describes the proposed technique. Section 4 describes the experimental evaluation and results achieved. Section 5 presents a discussion of the results. Finally, Section 6 presents the conclusions.

II. BACKGROUND AND RELATED WORK

Ensemble techniques were initially developed for independent and identically distributed (i.i.d) tabular data, but some recent work has extended them to relational data. Preisach and Schmidt-Thieme [17] develop voting and stacking methods to combine relational data with multiple relations. Eldardiry and Neville [6] develop a method for sampling relational data that can improving bagging in relational domains. After that Eldardiry and Neville [7] propose a relational ensemble framework that reduces learning and inference variance. Natajara et. al [15] propose an algorithm based on functional-gradient boosting that learns the structure and parameters of the Relational Dependency Network models simultaneously.

Eldardiry and Neville [6] propose a relational sampling technique, in which the key idea is to sample subgraphs with replacement from the original data, preserving local link and attribute structures within subgraphs. They show that their method produces more accurate variance estimates than naive i.i.d. resampling for a ranking task. They also compare the resampling approach to i.i.d. resampling using ensemble methods for classification of relational data. The authors state that bagging with their approach results in significant improvement in accuracy for both synthetic and real-world relational data.

The same authors propose other technique [7] to propagate inference information across simultaneous collective inference processes running on the base models of the ensemble to reduce *inference variance*. Then the algorithm combines the final model predictions to reduce *learning variance*. The authors consider the problem of relational learning and

collective classification of a single set of connected objects in multiple link graphs. This method uses the complete set of nodes in the training network for learning each model.

Both work are related to our, but there are some differences. In our work the sampling process resembles the ones for i.i.d. data, nevertheless (i) the training and test data are kept together, as in transductive inference; and (ii) the edges selection process considers one different strategy leading to a different way of wiring the network for the learning phase. We use one cross-validated committees approach. At the end we investigate the role played by diversity among the several individual classifiers. In what follows, we present ensemble techniques applied to traditional classifiers (for tabular data) emphasizing the approaches for sampling. We follow the category proposed in [18] which classifies the ensemble methods in *dependent* and *independent* ensembles.

Dependent methods present an interaction between classifiers in the ensemble. Therefore, it is possible to harness the knowledge generated in previous iterations to guide learning in the subsequent ones. This process, illustrated in Figure 1, considering the dotted lines, consists of training and classification phases. The training phase takes the following steps: (i) selection of a set of examples to create the training set E_{tr_i} , (ii) induction of the model h_i by applying the base classifier L to the set E_{tr_i} , (iii) selection of training data in the set $E_{tr(i+1)}$ considering the model h_i and the initial training set E_{tr} , (iv) storage induced models (h_1, h_2, \dots, h_T) . Subsequently, in the classification step, all the induced models are combined in a final classifier as in (*Boosting*) [8]. Another possibility is to use only the last classifier produced as in *Incremental Batch Learning* [18].

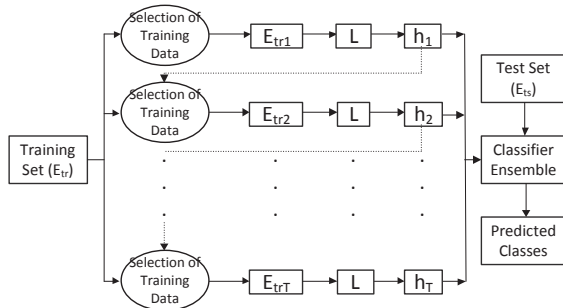


Figure 1. Dependent (considering the dotted lines) and Independent methods (disregarding the dotted lines).

In the independent methods, the original data set is divided into several subsets from which different classifiers are induced. These subsets can be disjoint or overlapping. This behavior is illustrated also in Figure 1, now disregarding the dotted lines. Here the step of creating the training set $E_{tr(i+1)}$ is generated only considering the initial training set E_{tr} , without considering the model h_i . One technique

representative of this approach is the *Cross Validated Committees* presented next.

Cross validated committees [16] constructs training sets excluding some subsets of training examples. For example, the training set can be divided into T disjoint subsets. Then it constructs T different training sets excluding only one of these T subsets in each iteration. For each training set a learning algorithm L is applied to induce a classification model h_i . Finally, to classify new examples the classification models (h_1, h_2, \dots, h_T) are combined.

III. GRAPH-BASED ENSEMBLES

In this section, we present the proposed ensemble of graph-based classifiers to be applied in the context of networked data. Firstly we present the graph-based classifiers selected as base classifiers for the proposed ensembles. Moreover, we observe that for networked data some peculiarities for generating training and test data must be considered and are explained next.

A. Networked data

Networked data is data represented in a graph, $G = (V, E)$, where V is a set of vertices representing objects and E a set of edges representing relation among vertices. For this data, the subset of vertices with known labels V^K (training set) can be created by selecting a class-stratified random sample from V . Hence, the test set (V^U) is defined as $V - V^K$. Although, it could be desirable to keep the test data disjoint as in traditional machine learning setting, this is not applicable for datasets in graph-based learning scenarios [12]. An example of this representation is showed in Figure 2, where training set, V^K (gray and white nodes), and test set, V^U (with “?” mark), are kept together in the graphs for the cross-validation process.

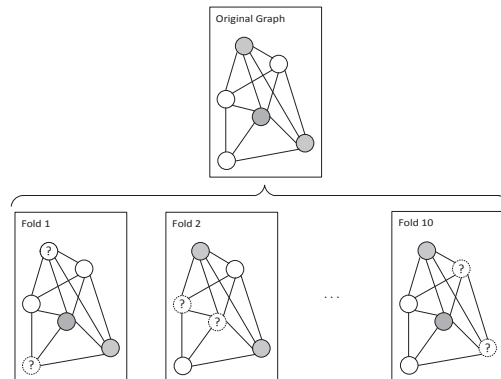


Figure 2. A 10-fold stratified cross-validation process in graph data.

B. Graph-Based Relational Classifier

In general, graph-based classifiers consider three components: (1) a relational model to classify unlabeled vertices considering links to neighbors, (2) a collective inference procedure, to spread labels to unlabeled interconnected vertices, and (3) a non relational model which uses local information, when available, for the classification. The non relational model generates prior probabilities for the relational learning and collective inference processes.

We evaluate the ensemble versions of the following well known relational classifiers: (1) Weighted-Vote Relational Neighbor classifier (*wvrn*) [12] which estimates class of a vertex considering the weights of the edges in its neighborhood; (2) Probabilistic Relational Neighbor classifier (*prn*) [11], this classifier is a probabilistic version of the Weighted-Vote Relational Neighbor classifier; (3) Class-Distribution Relational Neighbor classifier (*cdrn*) [12] which considers two vectors: (a) a Class Vector $CV(v_i)$ to be the vector of summed linkage weights to the various (known) classes and (b) a class c Reference Vector $RV(c)$ to be the normalization of the class vectors CV for known nodes of class c . Finally, to estimate the predicted class, it considers a similarity distance between these vectors; (4) Network-only Bayes classifier (*nbc*) [3], [12] which considers an adapted Bayes classifier based on neighborhood; (5) Network-only Link-Based classification (*nlb*) [10], [12] which creates a feature vector for each node considering the labels of neighboring nodes. Then, in the classification phase it uses a logistic regression. It considers different methods to create the feature vector: existence (*nlb - b*), mode (*nlb - m*), value counts (*nlb - c*) and normalization of *nlb - c* (*nlb - d*).

In the following section, the proposed algorithm: Graph-Based Cross-Validated Committees is detailed.

C. Graph-Based Cross-Validated Committees (GBCVC)

The algorithm *GBCVC* has two stages: the training and classification stages. In the training phase, summarized in Algorithm 1, firstly, the training set of vertices V_{tr} is divided into T sets (cv_1, cv_2, \dots, cv_T), i.e., the set of vertices V_{tr} is partitioned in T disjoint sets of equal size. This partitioning is stratified considering the classes of vertices. Secondly, in each iteration of the algorithm, one of these sets, cv_t , is not considered in the generation of the new training set V_{tr_t} . In the graph sampling, shown in Figure 3, a subgraph G_{tr_t} is selected from G_{tr} taking into consideration the vertices V_{tr_t} . Then the model h_t is induced considering the classifier L and the graph G_{tr_t} . Finally, the algorithm stores the model h_t and returns a set of models h_1, h_2, \dots, h_T .

In the classification stage, the final model predicts the class of a test example (x_i) considering Equation 1. Where, M_j denotes the classifier j and $P_{M_j}(y_i = c_k | N_i)$ denotes the probability of y_i be the class c_k given the neighborhood N_i of x_i .

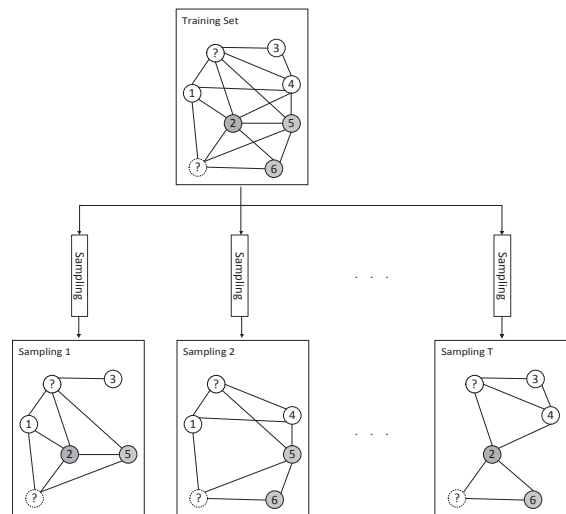


Figure 3. Sampling in the Algorithm *GBCVC*.

Input:

- $G_{tr} = (V_{tr}, A_{tr})$: training set graph.
- L : relational base classifier.
- T : number of iterations.

Output:

- h_1, h_2, \dots, h_T : T generated models.

Algorithm:

(cv_1, cv_2, \dots, cv_T) \leftarrow Divide the nodes V_{tr} in T disjoint sets in a stratified way;

for $t \leftarrow 1$ **until** T **do**

$V_{tr_t} \leftarrow$ Remove the nodes cv_i from the set V_{tr} ;

$G_{tr_t} \leftarrow$ createGraph(G_{tr}, V_{tr_t});

$h_t \leftarrow$ Apply the learning algorithm (L) to the graph (G_{tr_t});

 Store the resulting model h_t ;

end

return h_1, h_2, \dots, h_T

Algorithm 1: Classifier *GBCVC* - Learning Phase.

$$P(y_i = c_k | N_i) = \underset{c_k \in \text{dom}(\text{classes})}{\text{argmax}} \left(\sum_j P_{M_j}(y_i = c_k | N_i) \right) \quad (1)$$

IV. EXPERIMENTAL EVALUATION

The evaluations were carried out on data originally represented as graph from the Mark Newman's site¹ and Netkit platform². The elements in the networks ranged from 105 to 2189 vertices, 850 to 511 edges, and 2 to 12 classes and they are also presented in Table I.

In these evaluations, due to the number of datasets and algorithms it was used boxplots graphics to summarize the results obtained. It was applied the method proposed in [4] to

¹<http://www-personal.umich.edu/~mejn/netdata/>

²<http://netkit-srl.sourceforge.net/>

determine statistical differences between the algorithms with confidence level of $\alpha = 0.05$ and also, it was employed the 10-fold stratified cross-validation process, considering the scheme for networked data, described in Section III-A.

The following graph-based classifiers were evaluated: Weighted-Vote Relational Neighbor classifier (*wvrn*), Probabilistic Relational Neighbor classifier (*prn*), Class-Distribution Relational Neighbor classifier (*cdrn*), Network-only Bayes classifier (*nbc*) and the Network-only Link-Based classification (*nlb*) with four variations: *nlb - m*, *nlb - d*, *nlb - c* and *nlb - b*. Then, each of these classifiers have been employed the technique of Cross-Validated Committees. Finally, we compared the versions of the individual classifiers and the ensembles using the statistical test of [20] to verify if the ensemble improved the base classifier.

Table I
DOMAINS SPECIFICATIONS.

Domain	# Vertices	# Edges	# Classes
adjnoun (adj)	112	850	2
football (foot)	115	1232	12
imdb-all (iall)	1441	51481	2
imdb-prodco (iprod)	1441	20317	2
industry-pr (ipr)	2189	13062	12
industry-yh (iyh)	1798	14165	12
polblogs (pblogs)	1490	19078	2
polbooks (pbooks)	105	882	3
WebKB-cornell-cocite (wcc)	351	26832	6
WebKB-cornell-link1 (wcl)	351	1393	6
WebKB-texas-cocite (wtc)	338	32988	6
WebKB-texas-link1 (wtl)	338	1002	6
WebKB-washington-cocite (wwac)	434	30463	6
WebKB-washington-link1 (wwal)	434	1941	6
WebKB-wisconsin-cocite (wwic)	354	33250	6
WebKB-wisconsin-link1 (wwil)	354	1155	6

Next we describe the results of comparisons between the base classifiers and their Cross-Validated versions. Table II summarizes the comparisons, where the symbol “ Δ ” represent a significant difference (the ensemble improved the base classifier) and “—” represents no significant difference between the pair of classifiers. Also, Figure 4 presents boxplots graphics of the comparisons among the algorithms.

In the following analysis, an ensemble technique X with a base classifier Y is referred to as classifier XY . Hence, the classifier *GBCVCprn* showed a median greater than the median of *prn*. On the other hand, the classifiers *cdrn*, *nlb-m* and *nlb-b* had similar medians with their cross-validated committees version, but the lowest limits of their ensemble version were higher. In addition, the other classifiers showed no significant difference between their versions. Finally, the Wilcoxon test showed that the ensemble significantly improved the *prn* and *cdrn* classifiers (Table II).

V. DISCUSSION

To analyze the behavior of the proposed ensemble method, it was used k -error diagrams [13]. A k -error diagram consists of a measure kappa (k) [1] to represent the degree of agreement between classifiers, and a measure of *error*, to represent the average error rates of these classifiers. These

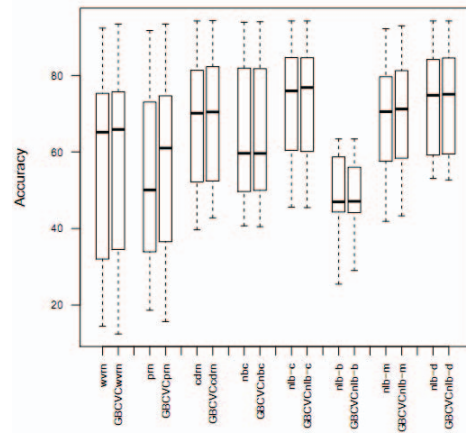


Figure 4. Comparative Results between graph-based classifiers (*wvrn*, *prn*, *cdn*, *nbc*, *nlb-m*, *nlb-d*, *nlb-c*, *nlb-b*) and the ensemble *GBCVC* in 16 data sets.

diagrams help to visualize the relation between error and diversity of each classifier of the ensemble [5].

For each pair of classifiers, the error is represented as the average of their error rates on test data, and the diversity is represented by a measure of agreement (*kappa*) between these two classifiers. This measure presents some cases: (a) $kappa = 0$, the agreement between the two classifiers equals that expected by chance; (b) $kappa = 1$, the two classifiers agree in every example; (c) when $kappa < 0$, exists a disagreement between the classifiers and (d) $kappa$ in the interval $[0, 1]$ indicates a degree of agreement between the classifiers.

In the experiments carried out for each fold in the ensemble were considered 10 classifiers. These 10 classifiers were combined 2 by 2 leading to $C_2^{10} = 45$ points for each fold in the k -error diagram. Finally, since there was 10 folds in total, it was generated $45 * 10 = 450$ points in the k -error diagram for each data set.

To analyze the proposed algorithm, we selected settings (datasets and classifiers) where this technique substantially improves, decreases or maintains the accuracy of the base classifier. These cases were chosen to try to understand when the ensemble succeeds or fails.

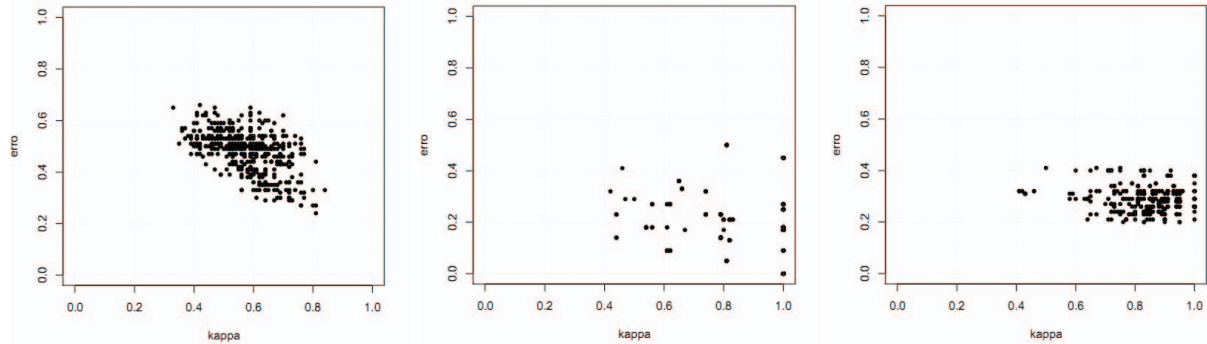
In the first case study, illustrated in Figure 5(a), the ensemble substantially improves the accuracy in the dataset *WebKB-texas-cocite* with the base classifier *prn*. In this case, the error rate of the base classifier is $ecb = 53.17\%$ and 75.11% of the individual classifiers of the ensemble present a lower error rate than *ecb*. Additionally, most of these classifiers (62.66%) have a *kappa* in the interval $[0.5; 0.7]$. These results indicate that smaller error rates than the base classifier error rate and a reasonably diversity improve accuracy of the ensemble.

In the second case study, showed in Figure 5(b), the ensemble keep the accuracy in the dataset *adjnoun* with

Table II

COMPARATIVE RESULTS OF THE ACCURACIES OF THE ALGORITHMS WVRN, PRN, CDRN, NBC, NLB-M, NLB-D, NLB-C AND NLB-B AND THEIR *GBCVC* VERSIONS (*GBCVC*WVRN, *GBCVC*PRN, *GBCVC*CDRN, *GBCVC*NBC, *GBCVC*NLB-M, *GBCVC*NLB-D, *GBCVC*NLB-C AND *GBCVC*NLB-B). EACH RESULT IS FOLLOWED BY ITS WILCOXON TEST.

	wvrn	GBCVCwvrn	prn	GBCVCprn	cdrn	GBCVCcdrn	nbc	GBCVCnbc	nlb-m	GBCVCnlb-m	nlb-d	GBCVCnlb-d	nlb-c	GBCVCnlb-c	nlb-b	GBCVCnlb-b
adj	20.38	17.65	26.59	28.48	80.53	82.42	82.20	82.20	81.44	82.42	85.98	86.82	88.71	87.80	63.41	63.41
foot	92.12	91.21	91.29	89.47	86.82	87.65	86.82	86.82	91.21	89.47	89.55	89.55	87.80	89.47	25.45	29.02
iall	73.63	73.35	70.79	71.34	77.59	77.86	61.90	61.84	74.53	76.75	79.53	79.60	84.60	84.53	56.77	56.77
iproduct	76.06	77.45	75.30	78.00	82.17	82.17	81.61	81.33	77.87	80.08	82.38	82.31	79.53	79.60	62.53	62.53
ipr	54.27	54.59	51.58	52.72	54.73	54.55	40.70	40.48	53.82	53.86	53.08	53.31	45.59	45.46	41.62	41.98
iyh	64.18	64.91	59.12	60.12	64.46	64.69	48.67	48.39	64.85	64.63	62.74	62.68	58.68	59.68	38.43	39.60
pblogs	92.35	93.42	91.74	93.42	94.23	94.30	93.89	94.03	92.21	92.95	94.23	94.23	94.23	94.23	60.67	62.42
pbooks	88.73	88.73	85.82	86.82	85.82	85.82	86.73	86.73	86.82	85.82	86.73	87.73	84.73	84.73	62.82	55.27
wcc	61.25	61.53	38.75	45.03	58.43	58.43	41.88	41.60	62.68	63.54	66.68	66.12	67.24	66.95	46.15	46.44
wel	23.92	28.21	29.06	29.35	39.65	42.80	43.90	44.47	41.88	43.31	54.72	56.15	54.13	54.98	43.91	43.63
wtc	73.73	73.43	46.83	61.90	72.24	71.93	55.31	55.32	72.82	74.00	78.42	78.72	79.32	80.51	45.59	45.59
wtl	14.49	12.43	18.63	15.67	44.71	46.19	50.64	51.52	50.02	50.94	53.29	52.70	60.42	60.43	45.00	46.18
wwac	66.15	66.84	49.31	66.14	69.34	69.12	57.39	57.39	68.22	68.44	71.21	71.44	73.95	75.11	44.69	44.69
wwal	40.11	40.80	43.31	43.77	49.09	49.08	64.73	65.18	61.28	62.90	69.80	69.34	68.18	67.50	52.33	51.87
wwic	74.59	74.03	50.85	64.42	70.91	71.77	63.57	63.28	73.47	74.03	79.11	79.67	77.99	78.56	47.75	47.75
wwil	22.04	19.52	22.04	20.63	49.66	50.23	54.79	55.07	52.53	50.56	55.66	56.22	60.42	59.85	48.86	48.86
Wilcoxon	-	-	▽	△	▽	△	-	-	-	-	-	-	-	-	-	-



(a) *WebKB - texas - cocite* using *prn* (substantial gain)

(b) *adjnoun* using *nbc* (draw)

(c) *WebKB - texas - cocite* using *cdrn* (small reduction)

Figure 5. *GBCVC* case studies.

the base classifier *nbc*. In this case, the error rate of the base classifier is $ecb = 17.80\%$ and 55.77% of the rates of individual classifiers of the ensemble are higher than ecb . Also, 85.77% of these classifiers have a $kappa$ in the interval $[0.7, 1.0]$. These data indicate that error rates of individual classifiers greater than error rate of the base classifier with less diversity in the classifiers does not improve accuracy of the ensemble (in this case the accuracy was maintained).

In the third case study, showed in Figure 5(c), the ensemble slightly decreases the accuracy in the dataset *WebKB - texas - cocite* with the base classifier *cdrn*. In this case, the error rate of the base classifier is $ecb = 27.76\%$ and the rates of individual classifiers of the ensemble are higher than ecb in 64.44% of the cases. Also, it was calculated that 79.11% of these classifiers have a $kappa$ in the interval $[0.8, 1]$. These data indicate that error rates of individual classifiers greater than error rate of the base classifier and a less diversity does not improve accuracy of the ensemble (in this case the accuracy was slightly decreased). Note that in this case, the percentage of the individual classifiers with higher error rate increased compared with the previous case

and displays a lower degree of diversity (high-kappa value).

VI. CONCLUSION

Most of the ensemble algorithms is used with structured data in an attribute-value representation for induction of computational models. This representation is limited to describe individual characteristics of objects without considering relationships between them. However, there are domains where data are naturally represented in a relational format, as in co-authorship networks, telephone calls networks, among others.

To address the problem of combining classifiers for relational data based on graphs, in this article we proposed one ensemble algorithm for graph-based relational classifiers. Two aspects of this research should be highlighted. The first was in the adaptation of the sampling step in the Cross-Validated Committees classifiers for networked data. The second was the analysis on the role played by diversity between classifiers in graph-based ensemble algorithms.

Regarding the Graph-Based Cross-Validated Committees classifiers, they performed better than their base classifiers.

Classifiers *prn* and *cdrn* present better performance with significant statistical difference. Furthermore, regarding the diversity analysis based on *k*-error diagrams showed the importance of diversity in improving the base classifier accuracy.

ACKNOWLEDGMENT

This research was supported by the Brazilian agencies CNPq and FAPESP.

REFERENCES

- [1] Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition.
- [2] Bruzzone, L., Cossu, R., and Vernazza, G. (2004). Detection of land-cover transitions by combining multivariate classifiers. *Pattern Recogn. Lett.*, 25:1491–1500.
- [3] Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *SIGMOD Rec.*, 27(2):307-318.
- [4] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.
- [5] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.*, 40:139–157.
- [6] Eldardiry, H., and Neville, J. (2008). A resampling technique for relational data graphs. In *Proc. of SNA-SIGKDD'08*.
- [7] Eldardiry, H., and Neville, J. (2011). Across-Model Collective Ensemble Classification. In *Proc. of AAAI'11*.
- [8] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 148–156.
- [9] Jensen, D., and Neville, J. (2002). Linkage and autocorrelation cause feature selection bias in relational learning. In *Proc. of ICML'02*.
- [10] Lu, Q. and Getoor, L. (2003). Link-based classification. In Fawcett, T., Mishra, N., Fawcett, T., e Mishra, N., editors, *ICML*, pages 496–503. AAAI Press.
- [11] Macskassy, S. A. and Provost, F. (2003). A simple relational classifier. In *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, 64–76.
- [12] Macskassy, S. A. and Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983.
- [13] Margineantu, D. D. and Dietterich, T. G. (1997). Pruning adaptive boosting. In *Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97*, 211–218, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [14] Murrugarra-Llerena, N. and de Andrade Lopes, A. (2011). A graph-based bagging. In *Proceedings of the CoLISD: Collective Learning and Inference on Structured Data*, 25–36, Atenas, Grecia. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.
- [15] Natarajan, S., Khot, T., Kersting, K., Gutmann, B., and Shavlik, J. (2012). Gradient-based Boosting for Statistical Relational Learning: The Relational Dependency Network Case. *Journal Machine Learning*, vol. 86 (1).
- [16] Parmanto, B., Munro, P. W., and Doyle, H. R. (1995). Improving committee diagnosis with resampling techniques. In *NIPS*, 882–888.
- [17] Preisach, C., and Schmidt-Thieme, L. (2006). Relational ensemble classification. In *Proc. of ICMD'06*.
- [18] Rokach, L. (2010). *Pattern classification using ensemble methods*. Series in machine perception and artificial intelligence. World Scientific.
- [19] Tan, A. C. and Gilbert, D. (2003). Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics*, 14:206–217.
- [20] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.