A Graph-based Bagging

Nils Murrugarra-Llerena and Alneu de Andrade Lopes

Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo – USP
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP – Brasil
{nineil,alneu}@icmc.usp.br

Abstract. The ensemble technique weights several individual classifiers and combines them to obtain a composite classifier that outperforms each of them alone. Despite of this technique has been successfully applied in many areas, in the literature review we did not find its application on networked data. To tackle with this lack, we propose a bagging procedure applied to graph-based classifiers. Our contribution was to develop a bagging procedure for networked data which either maintains or significantly improves the performance of the tested relational classifiers. Additionally, we investigate the role played by diversity among the several individual classifiers to explain when the technique can be successfully applied.

Keywords: ensembles, bagging, graph-based classifiers

1 Introduction

The main idea behind ensemble technique is to weight several individual classifiers and combine them to obtain a classifier that outperforms each of them alone. This methodology imitates our reasoning when we consider several opinions and combine them in some way to reach a final decision [23]. Studies on this technique have led to an active area of research in machine learning with successful applications in many fields, such as finance [11], bioinformatics [24], cheminformatics [20], medicine [18], image retrieval [12], manufacturing [17] and geography [4].

A branch of these studies is related to the Bagging technique [3]. In this approach, several classifiers are trained on samples of instances taken with replacement from the training set. These samples have the same size of the training set. Hence, in a given sampling some instances may not be included and others may appear more than once. To classify a new instance, each classifier returns its prediction and the composite classifier combines them. The bagging technique produces as result a combined model which often outperforms a single classifier trained with the original training set.

The main finding in these researches is that ensemble is generally more accurate than the individual base classifier that originated it [8]. Additionally,

the accuracy and diversity of individual classifiers impact on the accuracy of the combined model [10]. However, to the extent of our knowledge, no previous study has evaluated ensemble techniques applied in the context of networked data. Since this type of data is very common, research on techniques to enhance relational classifiers are certainly pertinent.

In order to tackle the previously mentioned lack we propose a bagging procedure applied to graph-based classifiers. The main difference of the proposed Graph-based Bagging with the propositional bagging procedure is in the sampling step. In the former, each classifier considers a sample of nodes taken with replacement from the training set. Our contribution is to develop a bagging procedure for networked data when a relational classification strategy is applied. Additionally, we investigate possible differences in the learning setting to explain when the technique is successfully applied.

Preliminary results on sixteen relational data sets comparing our proposal with other well-known eight graph-based classifiers show that the proposed technique either significantly improves or maintains the accuracy of the classifiers. In the evaluation, Demšar's method [7] and Wilcoxon signed-rank test [26] were used to compare the classifiers with their bagging versions.

The reminder of the paper is organized as follows. Next section describes related work on bagging approaches. Section 3 describes graph-based classifiers and the proposal algorithm (Graph-based Bagging). Results are presented in Section 4. Then, Section 5 presents discussion and, finally, the last section shows conclusions and future works.

2 Related Work

A variation of the bagging technique is Wagging (**W**eight **Agg**regation)[1], where each classifier is trained on the entire training set, but for each instance is assigned a weight. While the bagging technique considers discrete values for weights, the wagging technique considers continuous values.

Another strategy takes into consideration diversity-measures together with the bagging technique, as in the Bagging Using Diversity (BUD) algorithm [25]. BUD algorithm considers that the more diverse the classifiers are, the better the result achieved by the combined model. The algorithm starts generating a set of base classifiers from training instances. Next, a subset of classifiers is selected considering diversity-measures, such as disagreement measure, double fault measure, Kohavi Wolpert variance, inter-rate measure and generalized diversity. Finally, the subset selected is combined to classify new examples.

Brill and colleagues propose an Attribute Bagging (AB) strategy which combines random subsets of features [2]. Firstly, AB finds an appropriate size for

the feature subset dimensionality by random search. Then, it randomly selects subsets of features, creating projections of the training set. Next, the classifiers are trained on each projection. Finally, the classifiers are combined by voting. In a more recent study, [6] develops a Trimmed Bagging (TB) which aims to prune classifiers with highest error rates. Experiments showed that TB performs comparable to standard bagging in unstable base classifiers (as decision trees). On the other hand, when applied to stable base classifiers, like support vector machines, its performance is improved.

To the extent of our knowledge, the aforementioned techniques are applied only on propositional data, hence employing propositional base classifiers.

2.1 A Graph-based Bagging (GBB)

In this section, we present a Graph-Based Bagging (GBB) technique which aims to improve the accuracy of graph-based classifiers using an ensemble of graph-based classifiers.

Specifically, given a dataset $G = (V, E)$, where V represents a set of vertices and E a set of edges. The subset of vertices with known labels V^K (training set) is created by selecting a class-stratified random sample of V. Then, the test set (V^U) is defined as $V - V^K$. Although, it could be desirable to keep the test data disjoint as done in traditional machine learning setting, this is not applicable for datasets in graph-based learning [16]. An example of this representation is showed in Fig. 1. In figure, 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.

The technique consists of two main steps: the learning and classification phases. The learning phase, as presented in Fig. 2 and Algorithm 1, consists of the following steps: (i) Sampling n (number of nodes in the training set) nodes from the training set with replacement and store in C_t; (ii) Construction of a graph (G_t) using the training graph set and C_t. In the construction of this graph considering C_t, some nodes will appear more than once and others may not appear. The nodes that appear more than once replicate their connections with their neighbors and nodes that do not appear have their connections removed. Hence, different graphs is generated in the learning phase; (iii)Induction of the models h_t by applying a graph-based learning algorithm (L) to each graph G_t; Following that, (iv) Storage of the induced models $h_1, h_2, ..., h_T$.

In the classification phase, as summarized in Fig. 2, the final model predicts the class for a test instance (x) considering Equation 1, where M_k denotes the k-th classifier and $P_{M_k}(y = c_i|x)$ denotes the probability of y obtaining the value c_i given an instance x.

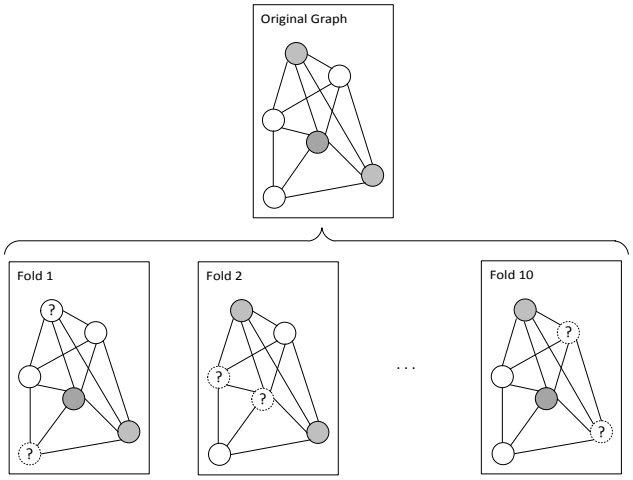

Fig. 1. A 10-fold for stratified cross-validation process in graph data.

$$class(x) = \underset{c_i \in dom(classes)}{\operatorname{argmax}} \left(\sum_k P_{M_k}(y = c_i|x) \right) \tag{1}$$

3 Results

In this section we present classification results on real data using the proposed Graph-Based Bagging algorithm. Here, due to the use of networked data, graph-based classifiers must be employed. In general, graph-based classifiers consider three components: (1) a relational model, to classify unlabeled nodes taking into consideration their neighborhoods; (2) a collective inference procedure, to infer on interconnected unlabeled nodes and (3) a non-relational model which uses, when available, local information to classify unlabeled nodes. Non-relational models generate priors that comprise the initial state for the relational learning and collective inference procedure.

We evaluate the bagging versions of the following well known relational classifiers: (1) Weighted-Vote Relational Neighbor classifier [15] which estimates the predicted class considering weights of the edges in its neighborhood; (2) Probabilistic Relational Neighbor classifier (*prn*) [14], this classifier is a probabilistic version of the Weighted-Vote Relational Neighbor classifier (*wvrn*);(3) Class-Distribution Neighbor classifier (*cdrn*) [15] which considers 2 vectors: (a) a node $v_i's$ 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 average of the class vectors for known nodes of class c. Finally, to estimate the predicted

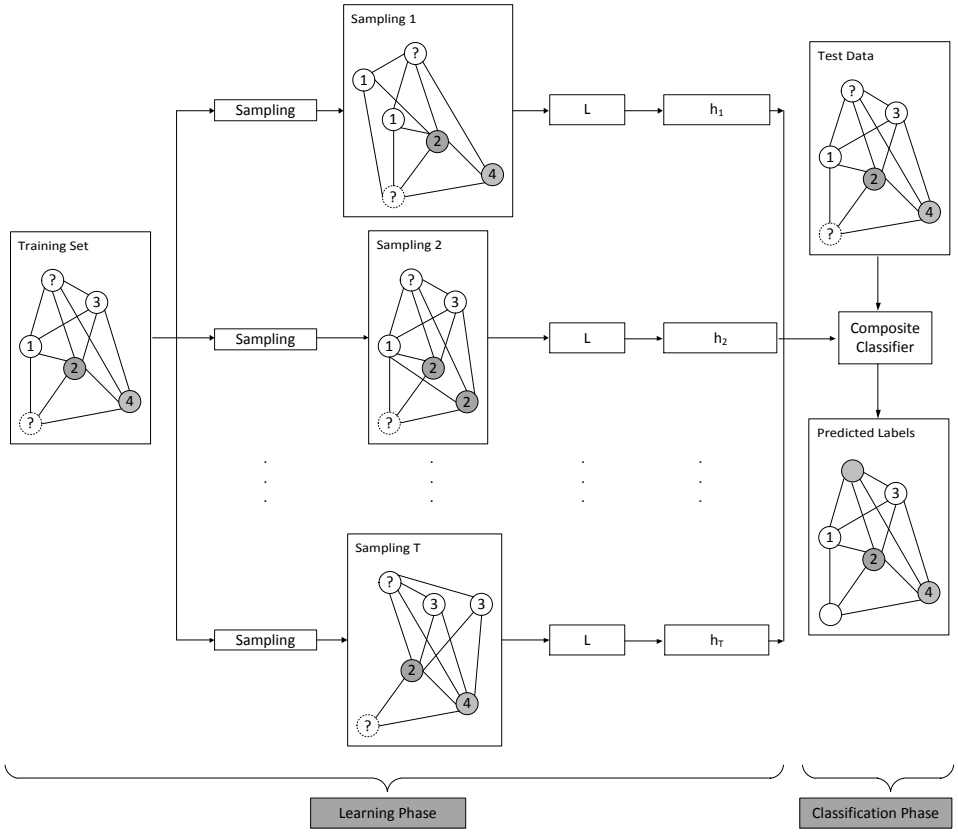

Fig. 2. Graph-based Bagging learning phase.

class, it considers a similarity distance between these vectors; (4) Network-Only Bayes classifier (*nobayes*)[5] [16] which considers an adapted Bayes classifier based on neighborhood to classify an instance; (5) Network-Only Link-Based classification (*nolb*) [13] [16] 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 ($nolb - b$), mode ($nolb - m$) and value counts ($nolb - c$).

The tests were carried out on sixteen data sets from the Netkit [21] and Mark Newman's site [1] presented in Table 1. The results were obtained through a 10-fold stratified cross-validation process. In order to analyze the results, we used the Demšar's method [7] to compare two classifiers over multiple data sets. This method called Wilcoxon signed-rank test [26] ranks the difference in per-

[1] http://www-personal.umich.edu/ mejn/netdata/

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

Output :

- $h_1, h_2, ..., h_T$ `//` T `generated models.`

Algorithm:
Let n be the number of nodes of V_{tr}.;
for $t \leftarrow 1$ *until* T **do**
  $C_t \leftarrow$ Sample n nodes with replacement from V_{tr}.;
  $G_t \leftarrow$ createGraph(G_{tr},C_t).;
  $h_t \leftarrow$ Apply the learning algorithm (L) to the graph (G_t).;
  Store the resulting model h_t.;
end
return $h_1, h_2, ..., h_T$

Algorithm 1: GBB learning phase

formance between classifiers for each data set, ignoring the signs, and compares the ranks for the positive and negative differences.

In the experiment, we compare the accuracy of each classifier with its bagging version using a confidence level (CL) of $alpha = 0.05$. Wilcoxon test showed statistical difference among the classifiers prn, $nolb{-}b$ and their bagging versions. Additionally, it did not show statistical difference with the others classifiers, as presented in Table 2 and 3, where $\triangle$ and $\bigtriangledown$ represent a significant difference and - represents no significant difference.

4 Discussion

To analyze the behavior of the proposed ensemble method, it was used k-error diagrams [19]. These diagrams help to visualize the relation between accuracy and diversity of each classifier of the ensemble [9]. For each pair of classifiers, the accuracy is represented as the average of their error rates (er) on test data, and the diversity is represented by a measure of agreement (mag) between these two classifiers. This measure presents some cases, (a) when $mag = 0$, the agreement of the two classifiers equals that expected by chance; (b) when $mag = 1$, the two classifiers agree on every example; (c) when $mag < 0$, exist a disagreement between the classifiers and (d) when mag in the interval $[0, 1]$, indicates a degree of agreement between the classifiers.

To analyze the Graph-based Bagging, we selected settings (datasets and classifiers) where this technique substantially improves, decreases or maintains the

Table 1. 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

accuracy of the base classifier.

In the first case study illustrated in Fig. 3(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 $esc = 53.17\%$ and the rates of individual classifiers of the ensemble are smaller than *esc*. Additionaly, most of these classifiers (66%) have a *mag* in the interval $[0.5, 0.7]$. These results indicate that a error rates smaller than the base classifier error rate and a reasonably diversity improve accuracy of the ensemble.

The second case study showed in Fig. 3(b), the ensemble decreases a bit the accuracy in the dataset $football$ with the base classifier *prn*. In this case, the error rate of the base classifier is $esc = 8.71\%$ and 68.89% of the rates of individual classifiers of the ensemble are higher than *esc*. Also, 90.22% of these classifiers have a *mag* in the interval $[0.8, 1]$. 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.

The third case study showed in Fig. 3(c), the ensemble substantially decreases the accuracy in the dataset $WebKB - texas - cocite$ with the base classifier $nolb - m$. In this case, the error rate of the base classifier is $esc = 27.18\%$ and the rates of individual classifiers of the ensemble are higher than *esc* in 69.78% of the cases. Also, it was calculated that 72.68% of these classifiers have a *mag* in the interval $[0.7, 0.9]$. These data indicate that error rates of individual classifiers greater than error rate of the base classifier and a less diversity decrease

Table 2. Comparative results among $wvrn, prn, cdrnandnobayes$ and their bagging versions ($bag-wvrn, bag-prn, bag-cdrn$ and $bag-nobayes$). Each result is followed by the Wilcoxon's Test.

Domain	wvrn	Bag-wvrn	prn	Bag-prn	cdrn	Bag-cdrn	nobayes	Bag-nobayes
adj	20.38	**23.86**	26.59	**33.86**	**80.53**	77.80	82.20	**84.02**
foot	**92.12**	90.38	**91.29**	89.47	**86.82**	**86.82**	**86.82**	85.98
iall	73.63	**73.77**	70.79	**71.00**	77.59	**78.07**	**61.90**	60.79
iprod	76.06	**77.72**	75.30	**76.20**	**82.17**	81.06	**81.61**	81.33
ipr	**54.27**	53.09	51.58	**52.35**	**54.73**	53.50	**40.70**	39.74
iyh	64.18	**64.79**	59.12	**60.73**	**64.46**	64.13	**48.67**	47.05
pblogs	**92.35**	**92.35**	91.74	**93.02**	**94.23**	94.03	93.89	**93.96**
pbooks	**88.73**	85.91	**85.82**	84.00	**85.82**	**85.82**	**86.73**	85.82
wcc	**61.25**	58.98	38.75	**49.33**	58.43	**58.44**	**41.88**	41.31
wcl	23.92	**31.34**	29.06	**30.49**	39.65	**49.02**	43.90	**45.88**
wtc	**73.73**	72.83	46.83	**68.07**	**72.24**	69.55	55.31	**55.61**
wtl	14.49	**17.48**	18.63	**18.66**	44.71	**49.21**	**50.64**	49.16
wwac	66.15	**66.84**	49.31	**66.82**	69.34	**69.36**	**57.39**	55.08
wwal	40.11	**41.95**	43.31	**44.48**	49.09	**54.12**	**64.73**	64.05
wwic	**74.59**	73.74	50.85	**66.68**	70.91	**72.90**	63.57	**64.43**
wwil	**22.04**	22.03	22.04	**22.32**	49.66	**49.94**	**54.79**	53.10
Wilcoxon	-	-	$\bigtriangledown$	$\bigtriangleup$	-	-	-	-

Table 3. Comparative results among $nolb-m, nolb-d, nolb-candnolb-b$ and their bagging versions ($bag-nolb-m, bag-nolb-d, bag-nolb-c$ and $bag-nolb-b$). Each result is followed by the Wilcoxon's Test.

Domain	nolb-m	Bag-nolb-m	nolb-d	Bag-nolb-d	nolb-c	Bag-nolb-c	nolb-b	Bag-nolb-b
adj	81.44	**82.35**	**85.98**	85.91	**88.71**	87.65	63.41	**65.23**
foot	**91.21**	87.65	**89.55**	87.80	**87.80**	**87.80**	25.45	**38.48**
iall	74.53	**78.00**	**79.53**	78.98	**84.60**	83.83	**56.77**	**56.77**
iprod	77.87	**80.01**	**82.38**	82.10	79.53	**79.60**	62.53	**63.98**
ipr	**53.82**	53.13	**53.08**	51.94	**45.59**	45.46	41.62	**42.62**
iyh	**64.85**	63.24	**62.74**	61.79	58.68	**60.12**	38.43	**40.99**
pblogs	92.21	**92.48**	**94.23**	94.09	**94.23**	94.09	60.67	**62.75**
pbooks	**86.82**	84.91	**86.73**	85.73	84.73	**85.73**	62.82	**62.91**
wcc	62.68	**63.26**	**66.68**	**66.68**	**67.24**	66.13	**46.15**	**46.15**
wcl	41.88	**45.87**	54.72	**57.27**	54.13	**55.27**	43.91	**48.47**
wtc	**72.82**	55.61	78.42	**79.01**	79.32	**81.38**	45.59	**46.77**
wtl	**50.02**	49.16	**53.29**	51.83	60.42	**60.71**	45.00	**47.99**
wwac	**68.22**	55.08	**71.21**	70.52	73.95	**74.17**	**44.69**	44.46
wwal	61.28	**64.05**	**69.80**	68.42	**68.18**	67.72	**52.33**	51.86
wwic	**73.47**	64.43	79.11	**79.65**	77.99	**81.37**	**47.75**	**47.75**
wwil	52.53	**53.10**	**55.66**	55.65	**60.42**	56.48	48.86	**49.42**
Wilcoxon	-	-	-	-	-	-	$\bigtriangledown$	$\bigtriangleup$

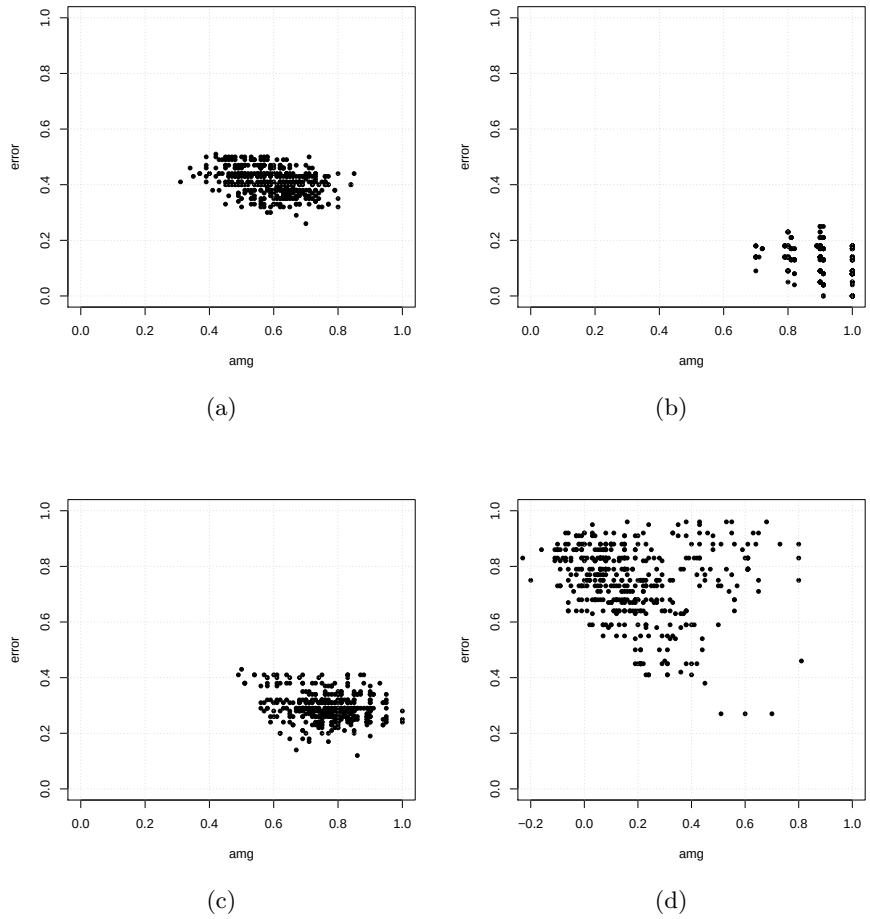

Fig. 3. Graph-based Bagging case studies. (a) $WebKB - texas - cocite$ using prn (substantial gain), (b) $football$ using prn (small reduction), (c) $WebKB-texas-cocite$ using $nolb-m$ (substantial reduction), and (d) $football$ using $nolb-b$ (substantial gain

substantially the accuracy of the ensemble.

Finally, in the fourth case study showed in Fig. 3(d), the ensemble substantially increases the accuracy in the dataset $football$ with the base classifier $nolb - b$. In this case, the error rate of the base classifier is $esc = 74.55\%$ and 55.77% of the rates of individual classifiers of the ensemble are higher than esc. Also, it was calculated that 60.22% of these classifiers have a mag in the interval $[0, 0.3]$. These data indicate that even with individual classifiers with error rates

greater than the error of the base classifier, the ensemble significatly improves the accuracy. This result shows the relevance of the role played by diversity in the accuracy of the ensemble.

5 Conclusions

This paper presents a graph-based bagging for networked data and relational classifiers. The experimental results indicate that the proposed technique significantly improved the classification accuracy of two of tested classifiers (*prn* and $nolb-b$) and kept the performance of the others. Furthermore, an analysis based on k-error diagrams showed the importance of diversity in improving the base classifier accuracy.

Future work includes adapting and testing the proposed technique with other ensembles such as cross-validated committees and/or wagging. The difference, in these cases, will be the way of considering the sampling phase.

6 Acknowledgments

This research was supported by the Brazilian agency CNPq.

References

1. Bauer, E. and Kohavi, R.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. In: Proceedings of Machine Learning Journal, pp. 105–139. Kluwer Academic Publishers, Hingham, MA, USA (1999).
2. Bryll, R., Osuna-Guttierez, R. and Quek, F.: Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets. In: Proceedings of Pattern Recognition Journal, pp. 1291-1302. (2003).
3. Breiman, L. : Bagging Predictors. In: Machine Learning, pp. 123–140. (1996).
4. Bruzzone, L., Cossu, R., Vernazza, G.: Detection of land-cover transitions by combining multidate classifiers. In: Pattern Recognition Letters 25, vol. 13, pp. 1491–1500. (2004).
5. Chakrabarti, S., Dom, B. and Indyk, P. : Enhanced hypertext categorization using hyperlinks. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pp. 307–318 . Seattle, Washington, United States (1998).
6. Croux, C., Joossens, K. and Lemmens, A.: Trimmed bagging. In: Proceedings of Computational Statistics and Data Analysis, pp. 362–368. Katholieke Universiteit Leuven (2007).
7. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. In: Journal of Machine Learning Research. vol. 7, pp. 1–30. JMLR.org. (2006).
8. Dietterich, T. G.: Ensemble Methods in Machine Learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems, pp. 1–15. Springer-Verlag, London, UK (2000).
9. Dietterich, T. G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. In: Machine Learning, pp. 139–157. Kluwer Academic Publishers, Hingham, MA, USA (2000).

10. Hansen, L. K. and Salamon, P.: Neural Network Ensembles. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 993–1001. IEEE Computer Society. Washington, DC, USA (1990).
11. Leigh, W. and Purvis, R. and Ragusa, J. M.: Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural networks, and genetic algorithm: a case study in romantic decision support. In: Decision Support Systems, vol. 32,pp. 361–184. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands (2002).
12. Lin, H., Kao, Y., Yang, F., Wang, P.: Content-based image retrieval trained by AdaBoost for mobile applications. In: International Journal of Pattern Recognition and Artificial Intelligence, vol. 20, pp. 525–541. (2006).
13. Lu, Q. and Getoor, L. : Link-based Classification. In Proceedings of the Twentieth International Conference (ICML 2003), pp. 496-503. AAAI Press, Washington, DC, USA (2003).
14. Macskassy, S. A. and Provost, F. : A simple relational classifier. In Proceedings of the Multi-Relational Data MiningWorkshop (MRDM) at the Ninth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 64–76 (2003).
15. Macskassy, S. A. and Provost, F. : Classification in Networked Data: A Toolkit and a Univariate Case Study. CeDER Working Paper CeDER-04-08. Stern School of Business, New York University(2004).
16. Macskassy, S. A. and Provost, F. : Classification in Networked Data: A Toolkit and a Univariate Case Study. In Proceedings of the Journal of Machine Learning Research, pp. 935–983 (2007).
17. Maimon, O., Rokach, L.: Ensemble of Decision Trees for Mining Manufacturing Data Sets. In: Machine Engineering, vol. 4, pp. 32–57. (2004).
18. Mangiameli, P., West, D. and Rampal, R.: Model selection for medical diagnosis decision support systems. In: Decision Support System, vol. 36, pp. 247–259. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands (2004).
19. Margineantu, D. D. and Dietterich, T. G.: Pruning Adaptive Boosting. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 211–218. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997).
20. Merkwirth, C., Mauser, H., Schulz-Gasch, T., Roche, O., Stahl, M., Lengauer, T.: Ensemble Methods for Classification in Cheminformatics. In: Journal of Chemical Information and Modeling, vol. 44, pp. 1971–1978. (2004).
21. Network Learning Toolkit for Statistical Relational Learning, `http://netkit-srl.sourceforge.net/`.
22. Raedt, L. D.: Logical and Relational Learning: From ILP to MRDM (Cognitive Technologies). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2008).
23. Polikar, R: Ensemble based systems in decision making. In: IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45. IEEE Press (2006).
24. Tan A. C., Gilbert D., Deville Y.: Multi-class Protein Fold Classification using a New Ensemble Machine Learning Approach. In: Genomic Informatics, vol. 14, pp. 206–217. Japan (2003).
25. Tang, E. K., Suganthan, P. N. and Yao, X.: An analysis of diversity measures. In: Proceedings of Machine Learning Journal, pp. 247–271. Kluwer Academic Publishers, Hingham, MA, USA (2006).
26. Wilcoxon, F. : Individual Comparisons by Ranking Methods. In: Biometrics Bulleti, pp.80–83. International Biometric Society (1945).