Global-Local Hybrid Ensemble Classifier for KDD 2004 Cup Particle Physics Dataset

Dustin Baumgartner and Gursel Serpen

***Abstract*—This paper presents a simulation-based performance evaluation of global-local hybrid ensemble (GLHE) in comparison to hundreds of classifiers which competed in the KDD 2004 Cup for the particle physics task. Simulations are performed on Weka machine learning software workbench. Performance metrics include prediction accuracy, area under the receiver operating characteristic (ROC) curve, cross entropy, and Q-score. Simulation results indicate that the difference between the performances of GLHE and the best submission for the KDD Cup 2004 particle physics task is relatively small for both prediction accuracy and area under the ROC curve. The performance of GLHE for prediction accuracy is only 3.2% worse than that of the best submission. On average over the four metrics, GLHE is ranked 56.5 out of 109. Noting that GLHE has not been fine-tuned or optimized with respect to any adjustable parameters or otherwise while the competition entries most likely were, the results are promising in support of the claims for the robustness of GLHE performance.**

***Index Terms*—Ensemble classifier, global-local learner, classifier diversity, KDD 2004 Cup, machine learning.**

I. Introduction

The success of ensembles of classifiers at improving the performance beyond single-classifier methods has led to increased research in the field [4]. Data manipulation ensembles, like bagging and boosting, have enjoyed widespread popularity and investigation in the past [17]. More recently however, empirical studies have shown that multiple learning algorithm (hybrid) ensembles are consistently better performing [11], [12], [14].

Comparative performance studies show that the "No Free Lunch Theorem" [23] applies – that is, although some classifiers consistently have robust performance, none is best for every problem or dataset [2], [3], [12], [13], [16], [18], [19]. Consequently, approaching a new classification problem requires the selection of an appropriate classifier, according to the performance needs and project constraints. Possible techniques for classifier selection mainly rely on three procedures as empirical evaluation, dataset characterization, and robust classifiers as below and organized from the most to the least complex process.

Empirical evaluation – A sample dataset is extracted from the problem. A variety of classifiers (e.g. decision trees, nearest neighbors, and ensembles) are run on this dataset and their performances are compared. Once the best classifier(s) is identified, parameter tweaking is performed to optimize performance for the problem. Empirical evaluation guarantees that the best performing classifier of those available is chosen, which is essential for domains that require high performance. One such field is the medical domain (e.g. predicting which patient is the best candidate for a kidney transplant). However, empirical evaluation is also a resource intensive process. Domain experts are needed to input knowledge of the problem, and machine learning specialists are needed to implement and configure parameters for each classifier. Furthermore, some classifiers have long run-times for training, and executing multiple iterations to obtain results suitable for comparison can be very time-consuming.

Dataset characterization – A sample dataset is extracted from the problem. Characteristics of the dataset – such as number of classes, number of instances, proportion of numerical or categorical features, skewness, and entropy – are calculated. Using these characteristics, a classifier is selected based on a repository (e.g. table) of exemplary dataset characteristics and best-performing classifiers for each dataset.

One interesting approach to this technique presents the task as a classification problem [1]. First, a set of classifiers are executed on 100 datasets. A new dataset is created with 100 instances, one for each of the original datasets. In this new dataset, an instance's features are the characteristics of a dataset and the class is the best performing classifier for the original dataset. Finally, a decision tree is used to create rules for predicting which classifier is best for new (previously unseen) dataset.

This technique does not guarantee the best performing classifier, although the study suggests that the chosen classifier will likely perform well [1]. A benefit of using dataset characteristics for new problems is that there is less effort needed than performing a complete empirical evaluation. Machine learning specialists are still required to create the initial repository dataset characteristics and classifier performance, but the remaining work of selecting a classifier for new problems can be performed solely by domain experts. This technique also takes much less time than empirical evaluations. There is a one-time calculation of characteristics before dataset characterization can be applied.

Robust classifiers – Select a classifier that is known to perform competitively for a spectrum of datasets (ideally from the same domain as the new problem). Ensembles are particularly good candidates, as they generally perform better than single-classifier methods. Selecting a robust classifier is similar to using dataset characterization since they both are merely an approximation of the optimal performing classifier. Alternatively, the two techniques differ since there are no

Manuscript received April 12, 2012; revised May 24, 2012.

D. Baumgartner was with the Electrical Engineering and Computer Science department, University of Toledo, Toledo, OH 43606 USA. He is now with Northrop Grumman Electronic Systems, Baltimore, MD USA (e-mail: dustin.baumgartner@gmail.com).

G. Serpen is with the Electrical Engineering and Computer Science department, University of Toledo, Toledo, OH 43606 USA (e-mail: gursel.serpen@utoledo.edu).

dataset characteristic calculations required for robust classifier selection.

There are benefits and detriments for each classifier selection technique. For example, should time be spent testing and modifying classifiers to achieve the optimal performance or should an existing robust classifier be used with the risk of performing poorly for the problem at hand. Generally, these are issues of concern for real-world users, whereas the researchers are responsible for creating better empirical tests, dataset characterizations, and robust classifiers.

In order to aid in the process of classifier selection, a design heuristic which aims to enhance robustness of performance for a hybrid ensemble was introduced in [4], [8]. This heuristic entails a mix of global and local learners in the base classifier composition of the ensemble. Accordingly an ensemble based on such a heuristic is termed the global-local hybrid ensemble (GLHE). In the following section, a brief overview of GLHE is provided and previous performance findings in the literature are reported. The remaining sections of the paper offer a new dimension to the performance assessment of GLHE with respect to classifiers that competed in the KDD 2004 Cup for the particle physics task [10].

II. Global-Local Hybrid Ensemble (GLHE)

The general architecture of the global-local hybrid ensemble is shown in Fig. 1 [4], [8]. The main feature of GLHE is the composition of the base classifiers. There are *n* base-classifiers from a global learner with different parameterizations. Likewise, there are *m* base-classifiers from a local learner with different parameterizations. The global and local learning algorithms provide the main source of diversity (through heterogeneity), while instantiation of multiple classifiers from each learner (through homogeneity) gives the ensemble an opportunity to benefit from the better performing learner numerous times rather than once. Any ensemble technique can be used to combine the base-classifier predictions.

GLHE exploits two sources of diversity in its base-classifiers, heterogeneous and homogeneous. Heterogeneity is achieved by using two types of learning algorithms – one global and one local. It is assumed that the intrinsic difference between global and local learning ensures high levels of diversity. Homogeneity uses multiple parameterizations of the same learning algorithm to allow both global and local learners to explore their respective region of the hypothesis space while also creating additional, albeit small, diversity among the base-classifiers. In previous studies, GLHE performance was compared to leading classifiers and found to be competitive and robust.

In [8], an analysis of the global-local hybrid ensemble design was performed and compared to traditional hybrid ensemble designs on 46 UCI Machine Learning repository data sets. The novel method showed benefits with respect to a variety of performance metrics, including prediction accuracy, CPU execution time, and diversity among the base classifiers. A comparison of GLHE to popular data manipulation ensembles, bagging and boosting, was conducted for the same data sets in [9]. This study showed equivalent performance of GLHE against three ensembles and better performance than another three ensemble configurations. Most importantly, [9] showed that GLHE is a reasonable alternative to the traditional practice of using bagging and boosting to support robust performance across multiple problem domains.

III. KDD Cup 2004 and Particle Physics Task

The Knowledge Discovery and Data Mining (KDD) Cup is an annual competition of data mining tasks, which is held in conjunction with the KDD Conference [10]. The 2004 competition focused on optimizing a variety of performance measures for large classification problems. We execute GLHE on one of the problems – particle physics task – and compare its performance to those that are reported from the competition results. An important factor to note is that we do not employ extra effort (e.g., parameter tweaking or understanding the data), and simply apply the basic or default GLHE design. The KDD Cup 2004 competitors, however, likely used many different techniques to optimize the performance of their classifiers for the problem. Thus, we are testing GLHE's performance against classifiers that are specialized and finely-tuned for the given problem.

The classification goal on the particle physics task from the KDD Cup 2004 is to differentiate between two types of particles that are generated in high energy collider experiments. It is a binary classification problem with 78 attributes. The training dataset has 50,000 instances while the testing dataset possesses 100,000 instances; however, we use only the training data because the actual values of class membership or labels of the testing data are not available. This methodology is still acceptable however under the assumption that the training and testing data correlated to a large degree. Therefore, the next best thing to do is to use 10-fold cross validation as the sampling technique to estimate performance metric values.

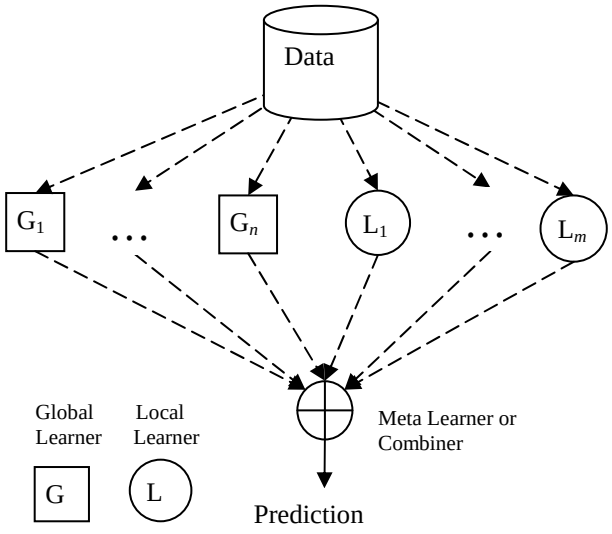

Fig. 1. Generic architecture of the GLHE classifier design.

Four performance metrics are considered [10].

- Prediction accuracy
- Area under the ROC curve
- Cross entropy – similar to squared error, such that it is a measurement of how close predicted values are to target values.

- Q-Score – a domain-specific metric devised by researchers at the Stanford Linear Accelerator (SLAC) to measure the performance of predictions made for certain kinds of particle physics problems. Smaller values indicate better performance.

These metrics are intended to support a variety of different perspectives on performance. Although one focus of the competition is to make researchers optimize their algorithms with respect to these different metrics, actual submission of results allows for a different set of predictions for each.

The KDD CUP 2004 website has two sets of results: the first contains only those submissions for competition, while the second has submissions before and after the competition ended. We chose the latter set to offer comparisons against, since it offers more classifiers for comparison. A total of 155 submissions were made, but only 109 of these reported on all four of the performance metrics.

IV. Simulation Study

GLHE requires global and local learner as well as homogeneous and heterogeneous diversities. The types of learners and diversities drive the composition of base classifiers. The heterogeneous diversity of the GLHE design comes from the use of two different types of base learning algorithms – one global and one local. For this study, the global learner is J48, a port of the C4.5 decision tree classifier in Weka [22], and the local learner is IBk, a nearest neighbor classifier. The categorization of these as global or local is consistent with the literature [15]. These learner choices constitute one of many options and is not intended to show any preference bias among all the options available for this decision making process. The homogenous diversity of GLHE is achieved by varying the parameter settings for each learner to create multiple classifiers. This requires three steps. First, the target number of classifiers from each learner is set. In this study, three classifiers are used from each learner. This sufficiently allows homogeneous tendencies to be incorporated while keeping the total number of base-classifiers at a reasonable level. Second, the base parameter design for each learner is established. For J48, the base parameters are Weka defaults. For IBk, the base parameters are also Weka defaults with the exception of distance weighting which has a value of "1/distance". Third, different parameter settings are selected for each learner to create the classifiers. For J48, the type of pruning is changed to one of standard pruning, reduced error pruning, and unpruned. For IBk, the number of nearest-neighbors is varied among 1, 5 and 10.

Simulations of GLHE with StackingC [20] as the combiner or meta learner were performed with Weka and the front-end utility for large experiments and evaluation [5], [6]. Calculation of performance metrics was done with the software utility supplied by the competition organizers, called PERF [10]. The performance of GLHE, its average rank relative to the competition submissions, the performance of the best submissions, and the total number of submissions for each metric are given in Table 1. The first observation is that the problem is relatively difficult to learn, since prediction accuracy of the best submission is just 70%. Secondly, the difference between GLHE's performance and that of the best submission is small for both predication accuracy and area under the ROC curve – even though GLHE is ranked 77 for prediction accuracy, its performance is only 3% worse as compared to that of the best submission. Considering each metric individually, GLHE appears to compete the best with regards to ROC, 56 out of 125, and the worst with regards to prediction accuracy, 77 out of 135. On average over the four metrics, GLHE is ranked 56.5 out of 109.

It is important to reiterate how the contest works in order to put the performance of GLHE into an appropriate perspective. First, the competitors most likely pursued and implemented some fine-tuning of classifiers and/or data preprocessing to achieve the best possible performance. Second, contestants could submit different sets of predictions for each performance metric, optimizing a classifier for each. These advantages are evident from the multiple submissions (sometimes over 30) by contestants. Furthermore, the winner of the contest used an approach with 639 predictors, including "data exploration, transformations, variable creation, modeling techniques, and customization of the predictions for four different evaluation criteria" [21]. Accordingly, the contest submissions may perform better for the particle physics task, but those same classifiers may not perform well across a spectrum of datasets given that no counter evidence is reported in the literature. On the other hand, global-local hybrid ensemble has been shown to exhibit a robust performance across a relatively larger spectrum of data sets or domains. Interpretation of the performance results of GLHE in conjunction with other and more comprehensive results for testing its performance as reported in [4], [7-9] provides a more complete assessment: GLHE performance benefits from the heuristic design rule that promotes inclusion of both global and local base learners.

TABLE I: Performance and Rank of GLHE, Best Submitted Performance, and the Total Submission Count for Each Metric from the KDD Cup 2004 Particle Physics Task.

Classifiers	Performance Measures			
	Accuracy	***ROC***	***Cross Entropy***	***Q Score***
GLHE	0.70228	0.79066	0.78848	0.26036
KDD'04 Best	0.73436	0.83101	0.70854	0.33396
GLHE (rank)	77	56	47	46
Total Submissions KDD'04	135	125	100	92

V. Conclusions

Comparative performance evaluation of global-local classification ensemble (GLHE) on the KDD Cup 2004 particle physics dataset against more than a hundred entries on the four performance metrics was accomplished. GLHE performed to surpass about half of the potentially highly-tuned entries in the competition although no optimization was pursued for any of its parameters, base learner composition or the ensemble meta learner. Simulation results suggest that GLHE demonstrated a

relatively robust performance on the KDD Cup 2004 particle physics dataset.

REFERENCES

[1] S. Ali and K. Smith, "On learning algorithm selection for classification," Applied Soft Computing , vol. 6, pp. 119-138, 2006.

[2] R. E. Banfield, L. O. Hall, K. W. Bowyer, D. Bhadoria, W. P. Kegelmeyer and S. Eschrich, "A comparison of ensemble creation techniques," Proc. of 5th International Conference on Multiple Classifier Systems, 2004.

[3] R. E. Banfield, L. O. Hall, K. W. Bowyer, D. Bhadoria, and W. P. Kegelmeyer, "A comparison of decision tree ensemble creation techniques," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 173-180, 2007.

[4] D. Baumgartner, Global-local hybrid classification ensembles: robust design with a reduced complexity. MSES Thesis. Electrical Engineering and Computer Science Department, University of Toledo, Toledo, OH, USA. 2009.

[5] D. Baumgartner and G. Serpen, "Large experiment and evaluation tool for WEKA classifiers," in Proceedings of the 5th International Conference on Data Mining, Las Vegas, pp. 340 – 346, 2009.

[6] D. Baumgartner and G. Serpen, "Fast preliminary evaluation of new machine learning algorithms for feasibility," in Proceedings of the 2nd International Conference on Machine Learning and Computing, Bangalore, pp. 113-115, 2010.

[7] D. Baumgartner and G. Serpen, "Comparative performance evaluation of global-local hybrid ensembles," International Journal of Hybrid Intelligent Systems, vol. 8(2), 2011.

[8] D. Baumgartner and G. Serpen, "A design heuristic for hybrid classification ensembles in machine learning," Intelligent Data Analysis, vol. 16, pp. 233-246, 2012.

[9] D. Baumgartner and G. Serpen, "Performance of global-local hybrid ensemble versus bagging and boosting ensembles," to appear in International Journal of Machine Learning and Cybernetics, 2012.

[10] R. Caruana and T. Joachims, (2004). Retrieved April 2012, from KDD Cup 2004: http://osmot.cs.cornell.edu/kddcup/

[11] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: An empirical analysis of supervised learning performance criteria," Proc. of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 69-78, 2004.

[12] R. Caruana, A. Niculescu-Mizil, G. Crew and A. Ksikes, "Ensemble selection from libraries of models," Proc. of 21st International Conference on Machine Learning, pp. 137-144, 2004.

[13] S. Dzeroski and B. Zenko, "Is combining classifiers with stacking better than selecting the best one?" Machine Learning, vol. 54, pp. 255-273, 2004.

[14] S. B. Kotsiantis and P. E. Pintelas, "A hybrid decision support tool - Using ensembles of classifiers," Proc. of International Conference on Enterprise Information Systems, pp. 448-456, 2004.

[15] T. M. Mitchell, Machine Learning. McGraw Hill. 1997.

[16] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," Journal of Artificial Intelligence Research, vol. 11, pp. 169-198, 1999.

[17] R. Polikar, Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6 (3), pp. 21-45, 2006.

[18] J. R. Quinlan, "Bagging, Boosting, and C4.5," Proc. of 13th National Conference on Artificial Intelligence, 1996, pp. 725-730.

[19] J. J. Rodriguez and L. I. Kuncheva, "Rotation forest: A new classifier ensemble method," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pp. 1619-1630, 2006.

[20] K. Seewald, "How to make stacking better and faster while also taking care of an unknown weakness," Proc. of 19th International Conference on Machine Learning, 2002.

[21] D. S. Vogel, E. Gottschalk and C. Wang, "Anti-matter detection: Particle Physics Model for KDD Cup 2004," SIGKDD Explorations, vol. 6 (2), pp. 109-112, 2004.

[22] H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (2 ed.). San Francisco: Morgan Kaufmann. 2005.

[23] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5(2), pp. 241-260, 1992.

Dustin Baumgartner is currently Senior Software Architect in the Automatic Target Recognition and Sensor Exploitation Technology Center at Northrop Grumman Electronic Systems in Baltimore, Maryland U.S.A. Dustin received his B.S. in Computer Science from Ohio Northern University in 2007 and his M.S. in Computer Science from the University of Toledo in 2009, and began work for Northrop Grumman in February 2010. Dustin has over 10 peer-reviewed and industry publications in the fields of machine learning, image processing for optical and infrared sensors, and target recognition/tracking.

Gursel Serpen received his PhD from the Electrical and Computer Engineering Department at the Old Dominion University, Norfolk, Virginia in December 1992. He worked as an applications engineer and then senior software engineer for Integrated Systems Inc. in Santa Clara, California between 1992 and 1993. He joined as a faculty member at the Computer Science and Engineering department of the University of Toledo in 1993. He is currently serving as a faculty member with the Electrical Engineering and Computer Science department at the University of Toledo. Dr. Serpen authored 70+ scholarly publications in the areas of neural networks, machine learning, computational intelligence, and artificial intelligence.