Relationship between Diversity and Correlation in Multi-Classifier Systems

Kuo-Wei Hsu and Jaideep Srivastava

University of Minnesota, Minneapolis, MN 55455, USA
{kuowei,srivasta}@cs.umn.edu

Abstract. Diversity plays an important role in the design of Multi-Classifier Systems, but its relationship to classification accuracy is still unclear from a theoretical perspective. As a step towards the solution of this probelm, we take a different route and explore the relationship between diversity and correlation. In this paper we provide a theoretical analysis and present a nonlinear function that relates diversity to correlation, which hence can be further related to accuracy. This paper contributes to connecting existing research in diversity and correlation, and also providing a proxy to the relationship between diversity and accuracy. Our experimental results reveal deeper insights into the role of diversity in Multi-Classifier Systems.

Keywords: Diversity, Correlation, Multi-Classifier System (MCS).

1 Introduction

The design of Multi-Classifier Systems (MCSs) is inspired by the group decision making process [13,14]. The motivation behind MCSs is that each classifier has its own strengths and weaknesses, and hence a group of classifiers could potentially leverage the wisdom of crowds. If each classifier in an MCS has expertise in classifying samples in some portions of a data space, the final output that is aggregated from all classifiers would become more reliable. More precisely, effective classifiers in an MCS are those that are accurate and independent. The former means that a classifier in an MCS is expected to provide performance at least better than random guessing, while the latter means that correlation between outputs of classifiers is expected to be small. This also implies that their outputs are expected to be diverse.

Diversity could be captured by disagreements between classifiers in an MCS and it plays a significant role in the success of MCSs [10]. However, the following research question becomes important for the design of MCSs: *Is there a relationship between diversity (between the member classifiers of an ensemble) and accuracy (of the ensemble)?* We address this research question by taking a different route and building a relationship between diversity and correlation, which could be related to accuracy.

Fig. 1 illustrates the focus of this paper. The relationship between diversity and accuracy is ambiguous in theory (e.g. *that elusive diversity* [9]). The relationship between correlation and the accuracy is relatively clear to researchers.

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 500–506, 2010.

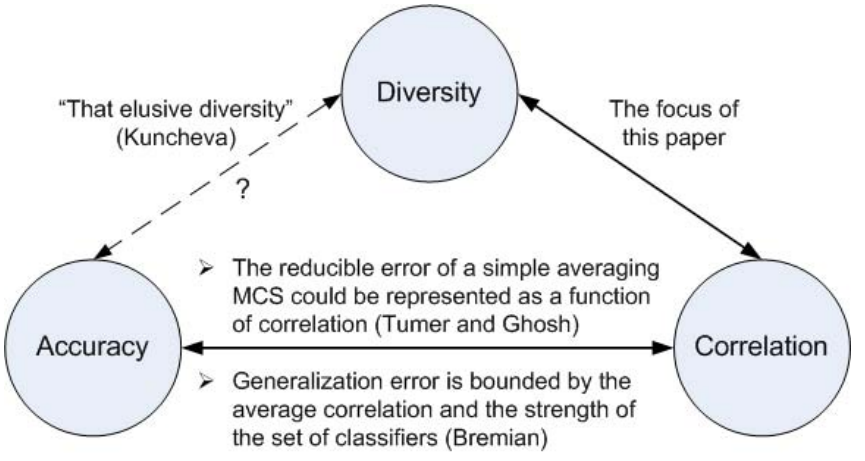

Fig. 1. Relationships among the accuracy, diversity, and correlation

Fig. 1 also gives two more examples of such relationships: Tumer and Ghosh build such a relationship for simple averaging ensemble [15], while Bremian relates correlation to performance of Random Forests [1].

This paper provides a proxy to the relationship between diversity and accuracy, while it has a potential to assist with a design guideline for MCSs.

The rest of this paper is structured as follows. Section 2 gives a theoretical analysis and Section 3 discusses experimental results. Section 4 is a brief review of some related work, while Section 5 gives conclusions and future work.

2 Theoretical Analysis of Diversity and Correlation

Diversity has been studied by many researchers [3,6], but its relationship to accuracy is not clear. One difficulty is that there exists an elegant bias-variance-covariance decomposition framework for regression tasks, but the framework does not directly apply to classification tasks [4]. Here we do not directly connect diversity to accuracy. Rather we build a relationship between diversity and correlation.

Notations. For a set of N instances and two classifiers, N_{11} and N_{00} denote the numbers of instances for which both classifiers are correct and incorrect, respectively; N_{10} and N_{01} denote the numbers of instances for which only the first and the second classifier is correct, respectively. The following definitions are with respect to outputs of classifiers i and j.

Definition 1. *Disagreement measure (Dis) representing diversity [10].*

$$Dis_{i,j} = \frac{N_{01} + N_{10}}{N_{11} + N_{10} + N_{01} + N_{00}} = \frac{N_{01} + N_{10}}{N}$$

Definition 2. *Q-statistic or Q [10,11].*

$$Q_{i,j} = \frac{N_{11} \cdot N_{00} - N_{01} \cdot N_{10}}{N_{11} \cdot N_{00} + N_{01} \cdot N_{10}}$$

Definition 3. *Correlation [10].*

$$\rho_{i,j} = \frac{N_{11} \cdot N_{00} - N_{01} \cdot N_{10}}{\sqrt{(N_{11} + N_{10}) \cdot (N_{01} + N_{00}) \cdot (N_{11} + N_{01}) \cdot (N_{10} + N_{00})}}$$

One could calculate system-wise values by averaging all pairs, so we ignore the subscripts i and j for concise representation. Using these definitions and the inequality of arithmetic-geometric-harmonic means, we obtain Corollary 1, as given below.

Corollary 1. *Relationship between disagreement measure and Q-statistic.*

$$Q \leq \frac{(1 - Dis)^2 \cdot N^2 - 4 \cdot Dis \cdot N}{(1 - Dis)^2 \cdot N^2 + 4 \cdot Dis \cdot N}$$

Corollary 1 helps us connect diversity to correlation, since a connection between Definition 2 and Definition 3 is that the absolute value of correlation will be bounded by the absolute value of $Q - statistic$. Next we define $f(x)$ based on Corollary 1.

$$f(x) = \frac{(1 - x)^2 \cdot N^2 - 4 \cdot x \cdot N}{(1 - x)^2 \cdot N^2 + 4 \cdot x \cdot N}, \text{ where } x = Dis$$

Since $x = Dis$ and hence $x \in (0, 1)$, we have $f(0) = 1$ and $f(1) = -1$. As the goal is to have zero correlation, we would like to know the interception of $f(x)$ and x-axis. We call the interception the *critical value* of x (x_c) or the *critical point* of Dis, and the following *critical value* is straightforward:

$$x_c = (1 + \frac{2}{N}) - 2 \cdot \sqrt{\frac{1}{N} - \frac{1}{N^2}}$$

Before this critical point, higher diversity reduces correlation. This supports the intuition that higher diversity between classifiers is usually associated with a better MCS. When diversity crosses the critical point, increasing diversity would increase correlation while highly correlated classifiers usually correspond to an inferior MCS.

3 Experiments and Discussion

For each trial for a data set, we randomly draw samples and accordingly train a decision tree (without pruning). Similarly, we generate a disjoint set of samples and use it as a test set for each trial for a data set. To control the variable in, we create a dummy classifier for each decision tree. We repeat this 100 times and create 100 pairs of classifiers in every experiment, using the corresponding test set to evaluate each pair of classifiers, calculating values of disagreements, $Q - statistic$, and correlation. Figures in Appendix illustrate the results. Our findings are summarized as below:

- The relationship between disagreement measure and $Q-statistic$ is not linear. Although curves of the theoretical upper bounds do not always match curves of the observed values, they do indicate trends of curves of the observed values.
- For some data sets, the theoretical upper bounds of the values of $Q-statistic$ are close to the observed values. For all data sets, they are close when diversity is lower and especially when N is smaller.
- There are exceptions that are larger than the theoretical upper bounds corresponding to them. They are exceptional cases where $Q - statistic$ is 1.
- As N increases, curves move to the right. The critical point is a function of N. This suggests that we need to increase diversity in order to obtain low (or even 0) correlation when the number of training samples increases.
- It is not always the case that we observe critical points in experiments. For those showing critical points, we observe that $Q - statistic$ and correlation move away from 0 as the diversity increases. This follows our analysis.

Now we take a couple of steps further and use our analysis result to explain some interesting phenomenon. [7] showed theoretically that heterogeneity (i.e. using different algorithms in an MCS) would improve diversity among member classifiers in an MCS. Furthermore, [8] showed empirically that one could obtain such an improvement more often in bagging setting than in boosting setting; in addition it empirically showed that AdaBoost with heterogeneous algorithms would work better when the data set is larger. Compared to bagging, AdaBoost often provides higher diversity. When we introduce heterogeneity into AdaBoost, diversity will probably be increased. As discussed earlier, increasing diversity has positive effect in the left region (between 0 and the critical point) of the graph of $f(x)$, but it has negative effect in the right region (between the critical point and 1) of the graph of $f(x)$. Moreover, the smaller the data set, the smaller the critical point, the smaller the left region. Therefore, using heterogeneous algorithms in AdaBoost on small data sets may actually have negative effect to the performance.

4 Related Work

The importance of reducing correlation between classifiers in an MCS has been recognized [2]. Tumer and Ghosh discuss a framework that quantifies the need to reduce correlation between classifiers in an MCS, and associate the number of training samples (i.e. the size of the training set) with the effect of correlation reduction [15]. Our analysis suggests that, for example, the critical point of Dis depends on N. Mane et al. prove that classifiers trained by using independent feature sets give more independent estimations and their combination gives more accurate estimations [12].

The term anti-correlation is confusing. In [13] McKay and Abbass describe it as a mechanism to promote diversity, but they do not explain why anti-correlation is equivalent to diversity promoting. Our analysis, however, explains this: When we promote diversity to a certain level (i.e. we have diversity in the neighborhood of the critical point), we decrease the upper bound of the absolute value of correlation and thus it is possible to observe very low or even negative correlation.

In [5] Chung et al. argue that, given the average the accuracy (or performance) of classifiers, there is a linear relationship between correlation and disagreement measure. Nevertheless, our analysis clearly shows that the relationship is not linear and our experimental results do not reveal the linear relationship as given in [5].

5 Conclusions and Future Work

In this paper we explored the relationship between diversity, represented by disagreement, and correlation between classifiers in MCSs, conducting a theoretical analysis and experiments for the relationship between diversity and correlation. As a result, we demonstrated a nonlinear function for the relationship, while the experimental results reveal some interesting insights. Therefore, this paper contributes to a better understanding of the role of diversity in MCSs.

Future work includes (1) investigating a tighter theoretical bound of Q-statistic, (2) integrating our analysis into those proposed by others in order to build a more elegant relationship between diversity and accuracy, and (3) using our analysis result to assist with classifier selection and/or combination algorithms for MCSs.

Acknowledgements. The research reported herein was supported by the National Aeronautics and Space Administration via award number NNX08AC36A, by the National Science Foundation via award number CNS-0931931, and a gift from Huawei Telecom. We gratefully acknowledge all our sponsors. The findings presented do not in any way represent, either directly or through implication, the policies of these organizations.

References

1. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
2. Brown, G., Wyatt, J., Tino, P.: Managing Diversity in Regression Ensembles. Journal of Machine Learning Research (JMLR) 6(September), 1621–1650 (2005)
3. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity Creation Methods: A Survey and Categorisation. Journal of Information Fusion 6(1), 5–20 (2005)
4. Brown, G.: Ensemble Learning. Encyclopedia of Machine Learning (2010)
5. Chung, Y., Hsu, D.F., Tang, C.Y.: On the Relationships Between Various Diversity Measures in Multiple Classifier Systems. In: International Symposium on Parallel Architectures, Algorithms, and Networks, pp. 184–190 (2008)
6. Ghosh, J.: Multiclassifier systems: Back to the future. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 1–15. Springer, Heidelberg (2002)
7. Hsu, K.-W., Srivastava, J.: Diversity in Combinations of Heterogeneous Classifiers. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 923–932. Springer, Heidelberg (2009)
8. Hsu, K.-W., Srivastava, J.: An Empirical Study of Applying Ensembles of Heterogeneous Classifiers on Imperfect Data. In: Workshop on Data Mining When Classes are Imbalanced and Errors Have Costs (2009)
9. Kuncheva, I.: That Elusive Diversity in Classifer Ensembles. In: Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), pp. 1126–1138 (2003)
10. Kuncheva, I., Whitaker, J.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51(2), 181–207 (2003)

11. Kuncheva, I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley Press, Chichester (2004)
12. Mane, S., Srivastava, J., Hwang, S.-Y.: Estimating missed actual positives using independent classifiers. In: International Conference on Knowledge Discovery and Data Mining (KDD), pp. 648–653 (2005)
13. McKay, R., Abbass, H.A.: Anti-correlation: A diversity promoting mechanism in ensemble learning. Australian Journal of Intelligence Information Processing Systems 7(3/4), 139–149 (2001)
14. Polikar, R.: Ensemble based systems in Decision making. IEEE Circuits and Systems Magazine 6(3), 21–45 (2006)
15. Tumer, K., Ghosh, J.: Error Correlation and Error Reduction in Ensemble Classifiers. Connection Science 8(3-4), 385–403 (1996)

Appendix A Experimental Results

In these figures, the x-axis is the value of disagreement measure (representing diversity) and y-axis corresponds to values of $Q - statistic$ or correlation ρ. A (blue) diamond and a (pink) square represent respectively an observed $Q - statistic$ and an observed correlation, while a (yellow) triangle gives an upper bound of the corresponding value of Q-statistic. We report results for 100 and 1000 training samples for each data set.

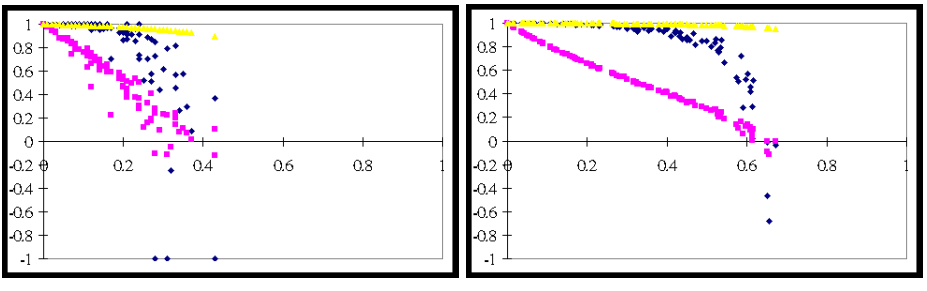

Fig. A1. Results for *Letter* with 100 (left) and 1000 (right) samples

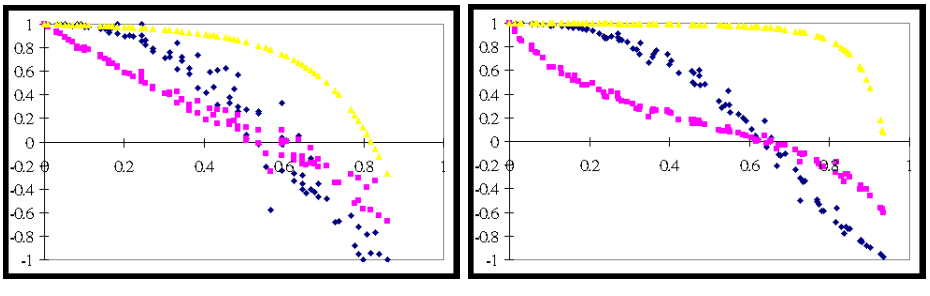

Fig. A2. Results for *Splice* with 100 (left) and 1000 (right) samples

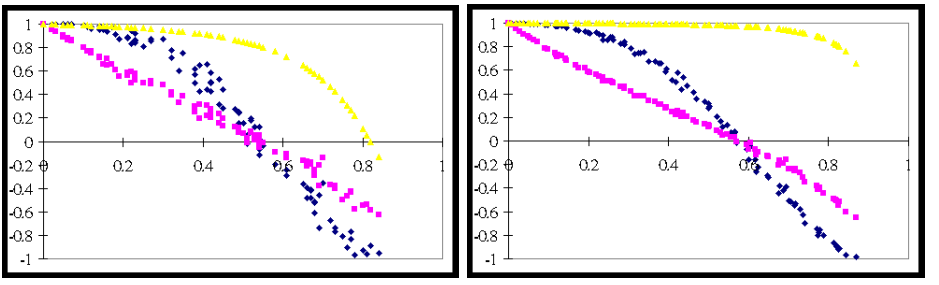

Fig. A3. Results for *Waveform-5000* with 100 (left) and 1000 (right) samples

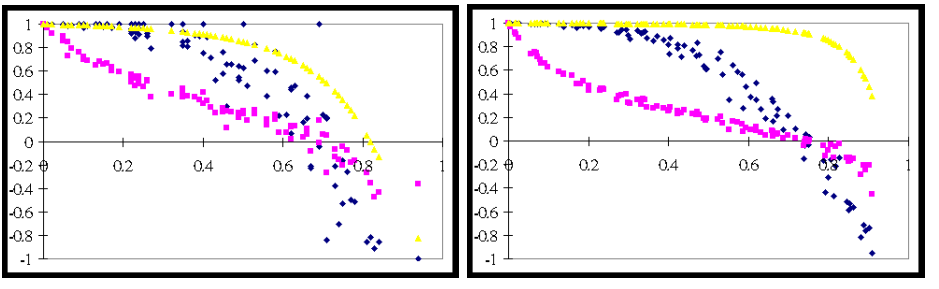

Fig. A4. Results for *Nursery* with 100 (left) and 1000 (right) samples

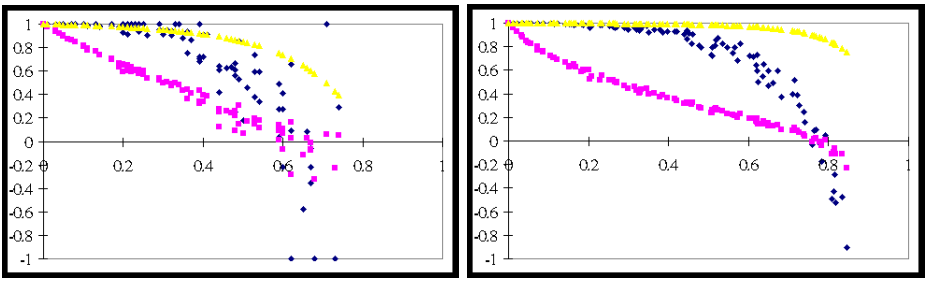

Fig. A5. Results for *Optdigits* with 100 (left) and 1000 (right) samples

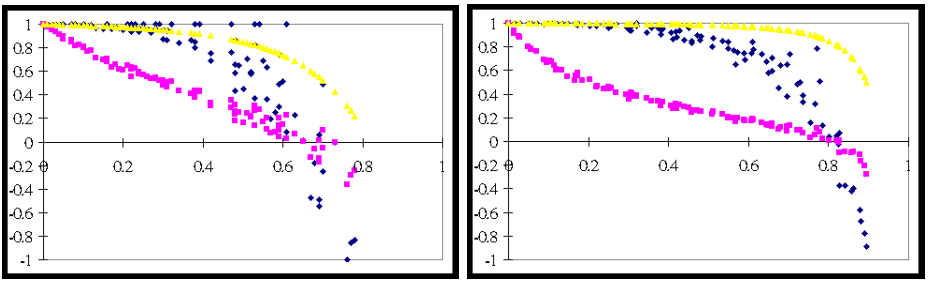

Fig. A6. Results for *Pendigits* with 100 (left) and 1000 (right) samples