Proceedings of the 2015 INFORMS Workshop on Data Mining and Analytics
M. G. Baydogan, S. Huang, A. Oztekin, eds.

Diversity in Dynamic Class Prediction

Şenay Yaşar Sağlam
Department of Management Sciences
University of Iowa
Iowa City, IA
senay-yasarsaglam@uiowa.edu

W. Nick Street
Department of Management Sciences
University of Iowa
Iowa City, IA
nick-street@uiowa.edu

Abstract

Instead of using the same ensemble for all data instances, recent studies have focused on dynamic ensembles in which a new ensemble is chosen from a pool of classifiers for each new data instance. Classifiers agreement in the region where a new data instance resides in has been considered as a major factor in dynamic ensembles. We postulate that the classifiers chosen for a dynamic ensemble should behave similarly in the region in which the new instance resides, but differently outside of this area. In other words, we hypothesize that high local accuracy, combined with high diversity in other regions, is desirable. To verify the validity of this hypothesis we propose two approaches. The first approach focuses on finding the k-nearest data instances to the new instance, which then defines a neighborhood, and maximizes simultaneously local accuracy and distant diversity, based on data instances outside of the neighborhood. The second method makes use of an alternative definition of the neighborhood: all data instances are in the neighborhood. However, the importance of data instances for accuracy and diversity depends on the distance to the new instance. We demonstrate through several experiments that the distance-based diversity and accuracy outperform all benchmark methods.

Keywords: Classification, Dynamic Ensembles, Diversity, Local Accuracy.

Introduction

Recently, dynamic class prediction systems have been proposed in which each new data instance is treated differently when making a prediction. Since different instances are often associated with different classification difficulties, it is hypothesized that using different classifiers for the classification task rather than a single static ensemble of classifiers can improve performance. Hence, the system dynamically adjusts for each new data instance by choosing a classifier(s) from the existing pool of trained classifiers. Most of the studies in the field of dynamic class prediction have focused on finding the neighborhood of a new data instance and choosing the most competent classifier(s) to make the prediction. Unlike static ensembles, competency is generally defined in some form of the local accuracy of the classifier in the region. In addition, diversity among the classifiers in the ensemble is ignored as the system is not required to generalize. However, if we ignore the diversity of the classifiers in other regions, the result may be an ensemble with several similar classifiers, and should they predict the instance incorrectly, there is no way to correct that decision. Also, there is no benefit to having multiple similar classifiers in an ensemble, as they will provide no gain over a single classifier. In addition, relying solely on the competency in the region doesn't answer which ensemble to use in case there are several ensembles that make different predictions on the instance with same competency level.

We believe that the classifiers chosen for a dynamic ensemble should behave similarly in the region in which the instance resides, but differently outside of this area. In other words, we hypothesize that high *local* accuracy and confidence, combined with low *local* diversity but high diversity in other regions, may be desirable. Consequently, just like in static ensembles, diversity still plays a role in dynamic class prediction. In this study, we propose a method for making dynamic predictions by increasing local accuracy and nonlocal diversity.

First, we summarize the earlier studies regarding diversity in classifier ensembles. Then, we provide the general structure of our proposed system and explain the details of its implementation. We carry out several experiments to evaluate performance of our proposed measure. After presenting our results, the last section concludes the paper.

Literature Review

Two key factors affecting the static ensemble's performance are the accuracy of the constituent classifiers and the diversity among them, as no gain can be achieved by having the same classifier in the ensemble multiple times. Studies regarding static ensembles can be categorized into three: generating diverse classifiers by changing the training set [1], defining new diversity measures to explicitly maximize diversity to choose classifiers with high accuracy and high diversity to form the ensemble [7], and analyzing the relationship between diversity and accuracy [8]. These studies showed that diversity is beneficial only to a certain degree. In [2], they argued that methods implicitly integrating diversity should be used to form the ensembles.

Due to shortcomings of static classifier/ensemble selection methods, dynamic class prediction systems gained attention. Some of these studies have focused on finding k instances that are "closest" to the new data instance, and choosing the most accurate (competent) classifier or a set of classifiers (an ensemble) in the region is used to make the decision [5, 6]. With KNORA, [6] indicate that using the neighbors of a new instance for constructing an ensemble can prove useful. However, it has been shown that using neighbors of a new data instance not just for constructing the ensemble but also in fusion with that ensemble's decision further enhances the performance ([3]). Later [9] argued that classifiers' probability estimates are more informative compared to using the predictions alone to find the k-nearest neighbors and proposed a dynamic class prediction framework. In these studies, diversity was not considered as a factor affecting the performance as the system is not required to generalize. In other words, when focused on a single instance, the classifiers' performance over the different regions of the instance space was not relevant.

Proposed Method and Dynamic Diversity Concepts

We believe that the classifiers chosen for a dynamic ensemble should behave similarly in the region in which the new instance resides, but differently outside of this area. In other words, we hypothesize that high *local* accuracy, combined with high diversity in other regions, may be desirable, and that just like in static ensembles, diversity still plays a role in dynamic class prediction. In this section, we propose two measures to compute "Distant Diversity" and a method for making dynamic predictions by increasing local accuracy and distant diversity.

1. Problem Formulation: We formulate the problem of finding an ensemble with high *local* accuracy, combined with high diversity in other regions as an optimization problem. We attempt to simultaneously optimize two separate objectives: local accuracy (LAcc(E,t)), and distant diversity (DDiv(E,t)):

$$\max_E \{LAcc(E,t), DDiv(E,t)\} \text{ s.t. } |E| = n \tag{1}$$

Since our problem formulation involves two separate objectives, a method must be defined to combine them into a single objective function, f, for which a maximum can be sought. In this study, we consider linear combinations of the objectives. Since the relative scale and significance of the *local* accuracy and *distant* diversity is unknown in a dynamic setup, the weighting of each objective is dynamically varied to cover as much of the objective space as possible. Given the set of weights, $\mathcal{W}$:

$$f(E; w, t) = w\ LAcc(E,t) + (1-w)\ DDiv(E,t), w \in \mathcal{W}. \tag{2}$$

2. Unweighted Distant Diversity and Unweighted Local Accuracy: Given a set of classifiers $\mathcal{C} = \{C_a | a = 1, .., M\}$, a validation dataset $\mathcal{D}$, and a new data instance, t, we calculate the distance between t and v, $\forall v \in D$ and find the closest *k* instances. These instances form the neighborhood of instance t, defined as $\mathcal{D}_{in}(t)$. The data instances that are not in the neighborhood of t are defined as $\mathcal{D}_{out}(t) = \mathcal{D} \setminus \mathcal{D}_{in}(t)$. The ensemble diversity outside of neighborhood of instance t, denoted by $UwDDiv(E, t)$ is calculated on $\mathcal{D}_{out}(t)$ using a *Disagreement Measure*. The diversity of the ensemble is then the average of the pairwise diversities for each possible combination of classifiers in the ensemble:

$$UwDDiv_{C_aC_b} = \sum_{v\in\mathcal{D}_{out}} \frac{\delta(\ell_{v,a}\neq\ell_{v,b})}{|\mathcal{D}_{out}|} \tag{3}$$

$$UwDDiv(E,t) = \sum_{C_a\in E} \sum_{\substack{C_b\in E\\ C_b\neq C_a}} \frac{UwDDiv_{ab}(t)}{|E|\ (|E|\ -1)} \tag{4}$$

Unweighted local accuracy of an ensemble $UwLAcc(E, t)$, denoted by is formulated as follows:

$$UwLAcc_{C_a} = \sum_{v\in\mathcal{D}_{in}} \frac{\delta(\ell_{v,a}=\ell^{\star}_{v})}{|\mathcal{D}_{in}|} \tag{5}$$

$$UwLAcc(E,t) = \sum_{C_a\in E} \frac{UwLAcc_{C_a}}{|E|} \tag{6}$$

where $\ell_{v,a}$ is the assigned label for instance v by classifier C_a and $\ell^{\star}_{v}$ is the actual label of v.

3. Weighted Distant Diversity and Weighted Local Accuracy: We hypothesize that the ensemble composed of classifiers that have agreement on the data instances close to the new data and disagree on the data instances further away from the new instance should make more accurate prediction on this new instance. Therefore, when the local accuracy of a classifier is evaluated, correct predictions on the points that are close to the test instance should weigh more than the points further away. Similarly, when diversity between two classifiers is considered, the disagreement for the data instances further away from the new data instance should be more valuable. To avoid the effect of neighborhood size, we decide to weigh the accuracy of classifiers and diversity among them based on the distance between the new instance and all the other instances. The new classifier accuracy and disagreement measure definitions are as follows:

$$WLAcc_{C_a} = \left[\sum_{v\in\mathcal{D}} \delta(\ell_{v,a} = \ell^{\star}_{v})\frac{1}{d(t,v)}\right] / \sum_{v\in\mathcal{D}} \frac{1}{d(t,v)} \tag{7}$$

$$WDDiv_{ab}(t) = \left[\sum_{v\in\mathcal{D}} \delta(C_a(v) \neq C_b(v))d(t,v)\right] / \sum_{v\in\mathcal{D}} d(t,v) \tag{8}$$

$$WDDiv(E,t) = \left[\sum_{C_a\in E} \sum_{C_b\neq C_a\in E} WDDiv_{ab}(t)\right] /(|E|\ (|E|\ -1)) \tag{9}$$

where $d(t, v)$ is the distance between data instances t and v.

Heuristic Search Algorithm

A central challenge to solving the problem formulated in the previous section is the large number of possible ensembles that could be formed from a given pool of classifiers. To address this challenge, a heuristic local search algorithm is employed to find a high quality solution in a reasonable time. Based on this definition, the heuristic search algorithm operates as follows: it begins the search from the most locally accurate ensemble, s, as the purpose is to show the benefit of distant diversity alongside local accuracy, to have comparable results, and to evaluate the objective function for this solution to obtain f. For each iteration, the algorithm randomly chooses both a classifier $C_a \in s$ to remove and a classifier $C_b \notin s$ from the pool of classifiers to replace it, creating a new ensemble, s_{new}. Then, it evaluates the objective function for this new solution, f_{new}. If $f_{new} > f$, then the algorithm continues the search from s_{new}. It terminates when an iteration count, N_{stop}, has been reached. Even though this is a conservative approach and does not fully assess the benefit of diversity, we believe even in this situation we should benefit from diversity.

Distance Measure

For this study, we use probability-based template matching (PTM) [9], which maps each data instance into an alternate feature space constructed by using the probability estimates of each classifier in the pool as the values of the features.

Data Set	#Data Points	#Features	Class Ratio[a]	%Good Classifiers[b]	Diversity[c]
a1a	1605	119	0.33	100	0.126
australian	690	14	0.80	64.17	0.229
breast cancer	683	10	0.54	97.14	0.050
diabetes	768	8	0.54	99.74	0.222
german	1000	24	0.43	100	0.208
ionosphere	351	34	0.56	88.75	0.071
liver disorder	345	6	0.73	79.46	0.310
sonar	208	60	0.87	54.04	0.300
splice	1000	60	0.93	52.77	0.277
w1a	2477	300	0.03	100	0.004

[a] The variable *Class Ratio* represents the ratio of minority class to the majority class.

[b] The variable *%Good Classifiers* represents the ratio of classifiers with at least 50% accuracy to the total number of classifiers over 100 runs.

[c] The variable *Diversity* represents the diversity of the classifiers in the pool over 100 runs based disagreement measure.

Table 1: Summary of the data sets and classifiers

The probability estimates for the two-class problems are taken with respect to a particular class label, and the choice of label is arbitrary. The similarity between instances i and j, denoted by $PTM_{i,j}$, is calculated as the Euclidean distance between them in this alternate feature space:

$$PTM_{i,j} = \sqrt{\frac{\sum_{m=1}^{M}(\phi_{i,j,m})^2}{M}}, \text{ where } \phi_{i,j,m} = \widetilde{\Pr}_{i,m} - \widetilde{\Pr}_{j,m} \tag{10}$$

where $\widetilde{\Pr}_{i,n}$ represents the probability estimate in the alternative feature space for instance i returned by classifier C_m. Similar to the ED, the pair of instances (i, j) with the smallest $PTM_{i,j}$ is considered to be the most similar.

Experimental Setup and Results

In our experiments, we use 10 datasets with varying numbers of features and data instances retrieved from the LIBSVM website ([4]). A summary of the datasets and the classifiers generated for each is presented in Table 1. The programming code was written in MATLAB and LIBSVM ([4]) was used to construct RBF kernel SVM classifiers. The training parameters were chosen such that classifiers overfit the data. Each experiment was repeated 10 times, with each run registering a unique seed value for the random number generator. For each run, the datasets are randomly divided into three subsets such that 60% of the instances is used to train classifiers, 20% is used for validation, and the remaining 20% is used for testing. A pool of 1000 classifiers is then constructed for each data set using a combination of bootstrap instance sampling (as in bagging) and random subspace selection on the training set to generate highly diverse pool of classifiers. Finally, classifiers with an error rate above 50% are removed from the pool. The results specified in this section represents the mean performance over 10 runs.

Effect of Weights: In this experiment, we change the weight on local accuracy, w to analyze the effect of the chosen weights. Table 2 demonstrates the results of this experiment. The results indicate that the weight of the diversity $(1 - w)$ does not affect the performance, especially for $w > 0.5$. The results for the unweighted distant diversity presented in this table is for neighborhood size set to $k = 13$. We obtain similar results for $k = 1, 7, 19, 25$.

The results support our hypothesis that incorporating diversity in addition to accuracy in the dynamic setup results in higher performance. As we increase the weight of diversity in Equation 2, we do not see significant increase in the performance compared to the $w = 0.9$ case. One reason is that for the *breast cancer* and *w1a* datasets, the classifiers are similar and highly accurate. Therefore, the diversity among the classifiers is considerably low and constructing ensembles in this manner becomes meaningless. Another reason is that our search algorithm starts from the best locally accurate classifier. Even though we change the weights, we find similar ensembles at the end of optimization. In other words, we may not be fully exploring the search space.

Comparison Against Baseline Methods: We also perform experiments to evaluate the effect of distant diversity compared to global (Div) and local diversity (LDiv), while maximizing local accuracy (LAcc). Unlike unweighted distant

	Unweighted Distant Diversity							Weighted Distant Diversity						
Dataset / w	1	0.9	0.7	0.5	0.3	0.1	0	1	0.9	0.7	0.5	0.3	0.1	0
a1a	0.765	**0.805**	0.804	0.805	0.805	0.805	0.802	0.773	**0.810**	0.810	0.807	0.806	0.801	0.796
australian	0.845	**0.849**	0.849	0.849	0.846	0.840	0.846	0.849	**0.860**	0.861	0.853	0.849	0.841	0.834
breast cancer	**0.977**	0.973	0.973	0.972	0.973	0.976	0.971	**0.977**	0.975	0.975	0.972	0.976	0.974	0.974
diabetes	**0.765**	0.757	0.757	0.758	0.759	0.757	0.743	**0.770**	0.760	0.760	0.757	0.758	0.749	0.748
german	0.714	**0.729**	0.729	0.729	0.730	0.732	0.727	0.724	**0.730**	0.731	0.729	0.732	0.727	0.728
ionosphere	0.933	**0.937**	0.937	0.937	0.940	0.935	0.932	0.935	**0.938**	0.938	0.939	0.937	0.934	0.935
liver disorder	0.646	**0.696**	0.693	0.697	0.706	0.688	0.670	0.646	**0.700**	0.700	0.696	0.690	0.678	0.681
sonar	0.788	**0.846**	0.846	0.841	0.836	0.803	0.798	0.788	**0.858**	0.858	0.856	0.820	0.803	0.798
splice	0.566	**0.777**	0.775	0.774	0.764	0.746	0.689	0.566	**0.648**	0.653	0.610	0.594	0.591	0.591
w1a	0.971	**0.973**	0.973	0.973	0.973	0.973	0.972	0.973	**0.974**	0.974	0.974	0.974	0.974	0.973

Note: The column headers (w) denote the weight on local accuracy. The neighborhood size is set $k = 13$ for unweighted distant diversity.

Table 2: Performance of unweighted and weighted distant diversity for different weight values

Dataset	Div+LAcc	LDiv+LAcc	Div+Acc	UwDDiv+UwLAcc	WDDiv+WLAcc
a1a	0.805	0.805	0.772	0.805	**0.810**
australian	0.849	0.849	0.857	0.849	**0.860**
breast_cancer	0.973	0.973	0.972	0.973	**0.975**
diabetes	0.757	0.757	**0.765**	0.757	0.760
german	0.729	0.729	0.717	0.729	**0.730**
ionosphere	0.937	0.937	0.932	0.937	**0.938**
liver_disorder	0.696	0.696	0.690	0.696	**0.700**
sonar	0.846	0.846	0.808	0.846	**0.858**
splice	**0.777**	**0.777**	0.573	**0.777**	0.648
w1a	0.973	0.973	0.972	0.973	**0.974**

Table 3: Performance of proposed distant diversity concepts against other diversity concepts for $k = 13$ and $w = 0.9$

diversity, global diversity considers disagreements between classifiers on all data instances (both inside and outside of the neighborhood). Local diversity only considers the disagreements on the data instances in the neighborhood. In addition, we compare dynamic optimization with distant diversity and local accuracy against the static model with global accuracy (Acc) and global diversity.

When considering unweighted distant diversity, one may argue that computing diversity in this manner will not be able to capture its full benefit due the neighborhood size effect on it. Additionally, another argument can be made in favor of local diversity. Specifically, once the neighborhood of a new instance is defined, we should follow a similar approach to finding static ensemble by maximizing accuracy and diversity in this local region, since this chosen ensemble could be seen as the most competent ensemble in this given neighborhood. Table 3 displays the mean performance of each method over 10 runs for $k = 13$ and $w = 0.9$. Based on the results provided in Table 3, dynamically maximizing local accuracy and distant diversity is better than maximizing global diversity and accuracy for at least 8 datasets. We do not observe much difference among global, local, and unweighted distant diversity combined with local accuracy in terms of their contribution to the performance. This could also be an artifact of our search algorithm as we start the search from most locally accurate ensemble. The results clearly indicate that adjusting the weight of data instances while focusing on each new test case performs better than assuming every data instance has the same importance for the test case inside or outside the neighborhood.

We also compare the performance achieved by maximizing un/weighted distant diversity and un/weighted local accuracy against other benchmarks. Unlike our method, once the neighborhood is defined, KNN uses the labels of the neighbors to make the prediction. Hence, comparing our method against this method could inform us whether we should rely on the label of the data instances in the neighborhood. Similar to our method, KNORA methods find the competent classifiers in the neighborhood to make the predictions. In addition, having static methods as benchmarks could help us determine whether the trade-off between accuracy and computational cost is worth taking. Performance of these baselines is presented in Table 3. The results indicate that maximizing distant diversity along with weighted local accuracy dominates all of the benchmarks specified in this section.

Dataset	UwDDiv+	WDDiv+	KNN	KNORA_E	KNORA_U	Best	Best 25	Single SVM	Best Ens
	UwLAcc	WLAcc		_	_	Classifier			
a1a	0.805	**0.810**	0.798	0.732	0.732	0.756	0.778	0.751	0.764
australian	0.849	**0.860**	0.853	0.751	0.751	0.768	0.842	0.814	0.850
breast cancer	0.973	0.975	**0.976**	0.973	0.973	0.952	0.974	0.969	0.974
diabetes	0.757	0.760	0.750	0.632	0.632	0.654	0.651	0.738	**0.762**
german	0.729	**0.730**	0.729	0.668	0.668	0.697	0.700	0.696	0.712
ionosphere	0.937	**0.938**	0.934	0.917	0.917	0.905	0.929	0.926	0.935
liver disorder	0.696	0.700	0.687	0.538	0.538	0.620	0.628	**0.719**	0.664
sonar	0.846	**0.858**	0.820	0.699	0.699	0.714	0.755	0.719	0.779
splice	**0.777**	0.648	0.771	0.624	0.624	0.570	0.566	0.562	0.566
w1a	0.973	**0.974**	**0.974**	0.965	0.965	0.971	0.971	0.970	0.971

Table 4: Performance of our method with proposed distant diversity concepts against benchmarks

Conclusion

In this study, we have proposed two diversity concepts for dynamic ensemble selection. The first concept was distant diversity, which considered disagreements among the classifiers outside of the neighborhood. Based on our experimental results, we concluded that incorporating diversity with local accuracy improved the performance of dynamic ensembles regardless of the diversity concept (global, local, distant).

To avoid the issue of finding the optimal neighborhood size for each dataset, we proposed a second method that changes the definition of neighborhood. In this approach, all data instances were in the neighborhood. However, the importance of data instances for accuracy and diversity depended on the distance to the new instance. We demonstrated that weighted distant diversity and weighted local accuracy outperformed all benchmark methods.

References

[1] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[2] Gavin Brown and Ludmila I. Kuncheva. “good” and “bad” diversity in majority vote ensembles. In Neamat Gayar, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, volume 5997 of *Lecture Notes in Computer Science*, pages 124–133. Springer Berlin Heidelberg, 2010.

[3] PauloR. Cavalin, Robert Sabourin, and ChingY. Suen. Dynamic selection of ensembles of classifiers using contextual information. In Neamat Gayar, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, volume 5997 of *Lecture Notes in Computer Science*, pages 145–154. Springer Berlin Heidelberg, 2010.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] Luca Didaci, Giorgio Giacinto, Fabio Roli, and Gian Luca Marcialis. A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition*, 38(11):2188–2191, 2005.

[6] Albert HR Ko, Robert Sabourin, and Alceu Souza Britto Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008.

[7] Ron Kohavi, David H Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International*, pages 275–283, 1996.

[8] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[9] Şenay Yaşar Sağlam and W Nick Street. Dynamic class prediction with classifier based distance measure. In *Conferences in Research and Practice in Information Technology (CRPIT): Proceedings of The Twelfth Australasian Data Mining Conference*, volume 158 of *ICML-04*, 2014.