Dynamic Ensemble Selection Methods for Heterogeneous Data Mining

Chris Ballard and Wenjia Wang*

***Abstract*— Big data is often collected from multiple sources with possibly different features, representations and granularity and hence is defined as heterogeneous data. Such multiple datasets need to be fused together in some ways for further analysis. Data fusion at feature level requires domain knowledge and can be time-consuming and ineffective, but it could be avoided if decision-level fusion is applied properly. Ensemble methods appear to be an appropriate paradigm to do just that as each subset of heterogeneous data sources can be separately used to induce models independently and their decisions are then aggregated by a decision fusion function in an ensemble. This study investigates how heterogeneous data can be used to generate more diverse classifiers to build more accurate ensembles. A Dynamic Ensemble Selection Optimisation (DESO) framework is proposed, using the local feature space of heterogeneous data to increase diversity among classifiers and Simulated Annealing for optimisation. An implementation example of DESO - BaggingDES is provided with Bagging as a base platform of DESO, to test its performance and also explore the relationship between diversity and accuracy. Experiments are carried out with some heterogeneous datasets derived from real-world benchmark datasets. The statistical analyses of the results show that BaggingDES performed significantly better than the baseline method - decision tree, and reasonably better than the classic Bagging.**

I. Introduction

In the era of big data, large volumes of data are often being generated from many different sources, such as user interactions with various social media, sensor arrays and mobile devices. Each data source could potentially contain some additional and/or possibly complementary information associated with the intended target and hence may be useful in analysis and modelling.

However, it can be very challenging [1] to use such *heterogeneous* data sources as they may have different formats of representation and/or levels of granularity and hence they have to be integrated and/or fused in some ways before being used for analyses. There are roughly two broad ways for data integration/fusion: *feature-level data fusion* and *decision-level data fusion*. *Feature-level data fusion* can be achieved by merging/combining and/or concatenating features together from all the data sources to produce a "single flat" dataset. Due to the nature of heterogeneous data sources, low level data fusion [2] based on existing and/or new features can be tricky and time consuming as it often requires domain knowledge, [3] and may also result in either some unnecessary redundancy or loss of the useful information. On the other hand, *decision-level data fusion* is about combining the decisions that are learned from various data sources separately to produce a final decision. In this sense, machine learning ensembles appear naturally to be an appropriate approach to solve this problem.

Chris Ballard is with Tribal, Norwich, UK.
Wenjia Wang is with the School of Computing Sciences, University of East Anglia(UEA), Norwich, UK.
*Corresponding author: wenjia.wang@uea.ac.uk

An ensemble in this context is a machine learning system that automatically induces multiple models (e.g. classifiers) from data for a given problem with one or some base learning algorithms, and then combines their outputs, i.e. decisions, with a decision fusion function to generate a final solution. A collection of N models is more likely to produce a more accurate and reliable decision than an individual model operating alone, provided that the classifiers in the ensemble are diverse enough from each other to avoid making the same errors on testing instances.

Ensembles are commonly built with classifiers that are trained from a single data source or dataset. However, as the classifiers that are induced from data are highly data-dependent, even though different parameters or conditions may be used for training, they will be less diverse. As a result, an ensemble of such classifiers usually produces little or even no gain in predictive performance.

On the other hand, with big data that is generated from heterogeneous data sources, it is possible that different data sources capture different perspectives of the problem and hence more likely to generate more diverse classifiers when each data source is used in training separately.

Thus, building ensembles from heterogeneous datasets offers two possible benefits: (1) it uses different feature sets directly in training, which overcomes the challenges of integrating data on feature-level; and (2) the classifiers generated in such a manner are likely more diverse and therefore their ensemble can be more accurate and reliable.

Selecting more diverse and accurate classifiers for building ensemble is important[4] and usually called ensemble selection. In *Static ensemble selection* a single best set of classifiers is selected to build an ensemble. In *dynamic ensemble selection*, diversity at a local level is dynamically used to find the best combination of classifiers to form a variable ensemble classify each individual sample.

This papers explores the application of diversity measures for dynamic ensemble selection of classifiers trained against multiple heterogeneous data sources. It will also assess the effectiveness of some diversity measures in active participation of ensemble generation and classifier combination at a local level.

II. Related Work

A. Ensemble selection strategies

There are three main ways that classifiers can be selected in order to optimise the overall ensemble accuracy [5].

1) Dynamic Classifier Selection: with classifier selection, a single best classifier is selected by evaluating their competence in the local region of an instance $\mathbf{x}$ after the classifiers have been trained.

2) Static Ensemble Selection(SES): starts from generating a pool of classifiers and selects a fixed (static) single "best" set of classifiers to form an ensemble.

3) Dynamic Ensemble Selection(DES): builds on the classifier selection approach. Rather than selecting a single best classifier, a set of classifiers is chosen for each sample. A key idea of DES is based on an assumption that different classifiers will perform better in different regions of the feature space. By using an appropriate evaluation metric within a region local to the instance being classified, we can determine the set of classifiers which perform optimally.

B. Dynamic Ensemble Selection

1) Finding the optimal ensemble: The requirements for finding the optimal ensemble to use to classify a single instance are the same as that for static ensemble selection. However, performance is an even greater consideration as the search will need to be repeated for each classified instance. Typically a combinatorial optimisation algorithm is used.

2) Local k-NN Ensemble Selection: Ko et al [5] suggest two methods for DES. *KNORA-ELIMINATE* considers the nearest K neighbours to $\mathbf{x}$ and uses classifiers that give a correct prediction for the K neighbouring points. *KNORA-UNION* only considers classifiers that correctly classify any of the N neighbouring instances. The algorithms were tested on three ensembles generated using random subspace, bagging and boosting. The results showed marginal but not uniform improvements in accuracy when tested against small data sets, but good improvement when tested on a relatively large data set. The issue here is that the results from just one dataset could not be generalised.

3) GMDH Adaptive Ensemble Selection: Xiao and He [6] propose an algorithm called *GDES* which uses Group Method of Data Handling (GMDH) to find the optimal ensemble to classify each test instance. GMDH uses a feed-forward neural network structure to combine candidate models. Bagging was used to generate a pool of classifiers. For each instance to be classified, the K-nearest neighbours are located in the training set, which are then classified by each classifier in the pool. GMDH is used to find the "optimum" ensemble. Their results showed that GDES obtained marginally better results than KNORA-ELIMINATE.

4) Dynamic ensemble selection using local competence and diversity: Soares et al [7] propose two approaches to dynamic ensemble selection which are conceptually similar to KNORA. However, their method takes into account measures of local competence *and* diversity to select classifiers. By comparing their approach to static classifier selection using accuracy and diversity, and dynamic classifier selection using local accuracy, they found that the clustering approach gave the best performance, but only statistically significant on one out of the two datasets used. For these reasons, the results are hardly convincing.

5) Dynamic ensemble selection using a randomised reference classifier: Woloszynski and Kurzynski [8] considers classifier competence as part of a probabilistic model. In their experiments, this approach led to statistically significant improvement over majority vote and KNORA-ELIMIATE. This approach was extended to take into account both local competence and diversity [9]. The competence of classifier and pairwise diversity are calculated for the pool of classifiers. However, this work is limited by the fact that it only employed a pairwised diversity measure, which is known to be ineffective, as a criterion for selecting models, although it is combined with a competence measure. It may be interesting to see if non-pairwise diversity measures can do better when used in selecting models for building an ensemble.

III. Proposed Dynamic Ensemble Selection(DES) Methods

A. A framework of DES with optimisation (DESO)

To overcome the issues described above, we propose a generic DES framework with an optimisation function (DESO), by taking the advantages of their algorithms, i.e. the k-means version of algorithm in [7] and simulated annealing in [9] for finding the best combination of classifiers in each cluster sub-ensemble, albeit for a different DES approach. More importantly, it should be noted that our DESO method employs some non-pairwise diversity measures, specifically the ones that appeared to be more effective among the existing definitions[10], such as, Coincident Failure Diversity (CFD) and Minority Failure Diversity (MFD), which will be described in a bit more detail in the next section, in attempt to overcome the issues associated with pairwise diversity measures. The proposed framework of DES with optimisation (DESO) firstly generates a pool of N classifiers using a given base learning algorithm. During training, k-means is used to cluster the validation set to form K clusters, $c_1, c_2, \ldots, c_K$. Simulated Annealing is then run for each cluster c_k, to find the combination of classifiers E_{c_k} which should have a maximum diversity D.

When classifying a test instance each instance is assigned to its nearest cluster c_k. The sub-ensemble for that cluster is then used to classify the instance. Finally, the decisions from the sub-ensemble classifiers are combined using a decision aggregation method.

The full algorithm is shown in Figure 1.

The key idea of DESO is that different classifiers should perform better in local regions of the feature space, trained from heterogeneous data sources independently. It also levers the idea that the best ensembles should be composed of the most diverse classifiers, which are optimised by simulated annealing with an energy function based on diversity.

Input parameters

- K clusters defining local regions in feature space
- N trained classifiers $H = \{h_1, h_2, \ldots, h_N\}$
- T_{c_k} classifiers in each sub-ensemble E_{c_k} in cluster c_k

Ensemble training

1) Cluster the validation set into K groups using k-means to generate clusters at points $C = \{c_1, c_2, \ldots, c_k\}$.
2) For each cluster c_k:
   a) Find the ensemble E_{c_k} as a solution to the following optimisation problem:

$$D(E_{c_k}|\mathbf{x}_{c_k}) = \max_{E_{c_k} \in H, T_{c_k}} D(E_{c_k}|x_{c_k})$$

   where D is a non-pairwise measure of diversity and $\mathbf{x}_{c_k}$ are instances belonging to cluster c_k in the validation set.

Classification

1) For each test pattern $\mathbf{x}$:
   a) Assign $\mathbf{x}$ to cluster c_k which has the nearest centroid based on the Euclidean distance
   b) Classify test pattern using the ensemble classifiers in E_{c_k}
   c) Use a decision combination method to combine the classifier decisions and generate an overall decision for $\mathbf{x}$.

Fig. 1. DESO algorithm with k-means clustering and combinatorial optimisation.

B. Heterogeneous DES method: BaggingDES

In principle, any ensemble construction algorithm can be used to implement the DESO. In this study, a specific implementation of the DESO framework is achieved with Bagging as its vehicle, hence named as BaggingDES. This was chosen because Bagging works with partitioned data subsets and hence fits the basic mechanism of the DESO and heterogeneous feature sets naturally.

In BaggingDES, a sub-ensemble is generated for each feature set using Bagging. This generates a pool of N classifiers in each feature set. DESO is then applied to each feature sub-ensemble to find the sub-ensemble in each cluster with optimal diversity. The algorithm is given in Figure 2.

C. Conditions affecting DESO performance

The performance of BaggingDES can be affected by the following parameters, which are used to configure DESO.

1) Size of base classifier pool: DESO selects classifiers from a pool generated using Bagging. The number of classifiers in the pool may affect the performance of the DESO ensembles selected for each cluster. As bagging iterations are increased, the differentiation between the classifiers may reduce. Consequently with a large number of classifiers, DESO may not be able to find an optimal solution. But as this is rather common issue in ensemble methods and has been investigated quite intensively in the previous research, hence not explored particularly in this study.

Input parameters

- L feature sets $\mathbf{FS}$, where $FS_l \in FS_1, FS_2, \ldots, FS_L$ split into train, test and validation partitions.
- Number of candidate classifiers to generate N

Classifier generation

1) For each FS_l in $\mathbf{FS}$:
   a) Generate N classifiers $\mathbf{h}^{FS_l} = \{h_1^{FS_l}, h_2^{FS_l}, \ldots, h_N^{FS_l}\}$ using Bagging.
   b) Apply DES to the validation set to generate C^{FS_l} clusters and identify the sub-ensemble $E_{c_k^{FS_l}}$ in each cluster with optimal diversity (see Figure 1).

Classification

1) For each Feature Set FS_l:
   a) for each test pattern $\mathbf{x}^{FS_l}$ in FS_l:
      i) Classify $\mathbf{x}^{FS_l}$ using the cluster sub-ensemble $E_{c_k^{FS_l}}$ with the nearest centroid based on Euclidean distance.
2) Use a decision combination method to combine the feature set decisions and generate an overall decision for FS.

Fig. 2. BaggingDES - Dynamic Ensemble Selection with bagged sub-ensembles

2) k-means clustering: The purpose of clustering is to group instances that are similar. Therefore the parameters of k-means may have an effect on performance. The number of clusters which are generated could also have an impact. Since the validation set usually has a small proportion of instances, it is important to balance its size with the number of clusters. A large number of clusters will reduce the number of instances per cluster. This will make it difficult for DESO to find any differentiation between the classifiers. This issue was investigated in this study but the results are not presented in this paper primarily due to the page's limit.

3) Size of DESO ensemble: Simulated annealing will generate a sub-ensemble of fixed size for each cluster. The number of classifiers selected for each one will have an impact on the ensemble's performance. With a small number of classifiers, there may not be sufficient diversity between them. Conversely, with a larger number, the performance of DESO will reduce as the ensemble size approaches the size of the base classifier pool. This is a specific issue in DESO and hence investigated in detail in this study and results are presented in Section V-E.

In addition, as simulated annealing optimises the value of an energy function, it is important to build an energy function that can well represent the truly useful performance of DESO ensemble. This could be any function such as a diversity or accuracy measure. The non-pairwise diversity measures are used in this study and will be compared to a simple measure of classification error. In the implementation of simulated annealing, for simplicity, optimisation will run for a specified number of iterations. With a small number of iterations, the optimal solution may not be found.

The chosen values for these parameters will be given in detail in the experiment Section.

IV. Measuring Ensemble Diversity

The diversity of a set of classifiers in an ensemble can be assessed by using diversity measures. Different measures are available, and all of them broadly attempt to measure the degree of differentiation between the base classifiers.

Previous research

The type of diversity in an ensemble can greatly affect the outcome [11]. To assess the impact of diversity on the accuracy of an ensemble, quantitative measurements of diversity need to be used. These measures attempt to quantify the difference between the errors that the classifiers make. There are two types of diversity measures: pairwise and non-pairwise[10].

Pairwise diversity measures compare the difference between the predictions of pairs of classifiers. Where an ensemble consists of more than two classifiers, the statistic is generated for all possible classifier pairs and an average value obtained. Non-pairwise measures are generally more complex and are designed to measure the difference between predictions of sets of multiple classifiers[11], [12], [13].

A. Non-pairwise diversity measures

In this study, two non-pairwise diversity definitions: Coincident Failure Diversity (CFD)[14] and Minority-Failure Diversity (MFD)[11] are used, instead of pairwise diversity measures used in the other studies reviewed in the related work.

CFD measures the probability that n classifiers in an ensemble fail on randomly selected test data[14]. It is considered as one of few relatively effective non-pairwise diversity measures.

MFD is a unique non-pairwise diversity measure as it is related to the decision fusion function of an ensemble. It measures the probability that minority of classifiers in an ensemble fail, or majority of classifiers succeed, and therefore is considered as more appropriate , when the majority voting is used as the fusion function[11].

B. Actively using diversity

Diversity measures can be used in two ways. They can be used as a passive "downstream" measure to assess the diversity of a classifier system once it has been trained. However, they can also be *actively* used as part of the ensemble selection processes as described in the earlier sections.

This study investigates the application of the above mentioned three diversity measures for dynamic ensemble selection of classifier trained with heterogeneous data sources and evaluate their effectiveness in BaggingDES.

V. Experiments

A. Objectives of the experiments

With the proposed DESO and BaggingDES algorithms, we designed our experiments with the following objectives and set-ups.

1) Benchmark test: Experiments which compare the performance of the proposed BaggingDES algorithm against two baseline methods - single classifier Decision Trees and Bagging.
2) DESO Conditions - experiments investigating the performance of BaggingDES under different conditions, as described in the earlier section.
3) Relationship between diversity and accuracy - experiments designed to identify whether a link exists between diversity and accuracy when using DESO.

B. Datasets

As no benchmark heterogeneous dataset was found publicly available, we decided to construct some based on appropriate real-world benchmark datasets available from the UCI Machine Learning Repository [15]. The datasets were chosen based on the number of class labels and features. In order to simulate multiple heterogeneous data sources, each dataset was split into several separate feature sets based on the "natural" (such as phenotypic characteristics and/or representations) grouping of the features.

The details of the "semi-constructed" data sets used for the experiments are shown in Table I. The number of feature sets used for the multiple source experiments is shown for each dataset.

TABLE I

Datasets used in the experiments. Note: F-sets: Number of heterogeneous feature sets.

Dataset	F-Sets	Instances	Features	Classes
Arrhythmia	7	452	279	2
Biodegradation (QSAR)	4	1,055	41	2
Heart Disease	3	303	14	2
Hill Valley	4	606	100	2
Ionosphere	2	351	34	2
LSVT	4	126	309	2
Phoneme	10	5,404	5	2
SPECTF	4	267	45	2
WDBC	3	569	30	2

C. Experimental procedure and conditions

Five-fold cross validation is employed in the experiments. In all the experiments, DESO parameters are chosen (e.g. base classifiers N =100, annealing steps=150, K=3) and kept constant, except the parameter that is being tested.

D. Experimental results and comparison

The mean accuracy percentage and standard deviation of the experiments for BaggingDES and the two baseline methods- Decision tree and Bagging are shown in Table II. It can be seen that the BaggingDES ensembles outperform the classic Bagging ensembles and also single classifiers Decision trees on all the datasets, except two particular datasets *Hill Valley* and *Phoneme*, on which the single classifiers decision trees performed slightly better.

Some statistical analyses (non-parametric Friedman and Nemenyi post-hoc test, with $\alpha = 0.05$) were carried out to test whether the differences between these methods in terms

TABLE II

MEAN ACCURACY (%) AND STANDARD DEVIATION OF BAGGINGDES COMPARED TO TWO BASELINE METHODS: SINGLE CLASSIFIER DECISION TREES AND CLASSIC BAGGING. (NOTE: THE NUMBER IN BOLD FONT INDICATES THE BEST RESULT.)

——Method:	Decision Tree		Bagging		BaggingDES	
Dataset	Acc	SD	Acc	SD	Acc	SD
Arrhythmia	71.90	4.62	73.33	4.91	**76.19**	2.61
Bio Deg	82.94	2.16	85.50	3.12	**85.78**	1.16
Heart Disease	71.38	4.52	78.88	4.07	**79.22**	3.01
Hill Valley	**53.31**	2.59	51.98	3.94	50.82	4.03
Ionosphere	89.18	1.90	90.30	3.89	**91.18**	2.89
LSVT	72.92	10.07	76.22	2.21	**79.48**	12.17
Phoneme	**87.21**	0.59	85.07	0.58	84.92	0.98
spectf	64.77	6.81	70.78	3.89	**73.38**	8.47
WDBC	91.03	2.12	94.90	1.97	**95.60**	1.37

of their ranking positions on all the datasets are critical or not and the results are illustrated by Figure 3. This shows that BaggingDES (BagDES) is ranked as the best with an average rank value of 1.44 and is significantly better than Decision Tree with a Critical Distance (CD) = 1.10, but not significantly better than Bagging, which is in turn not significantly better than Decision Tree. It should be noted, however, that these results and interpretations should be taken with caution as the testing sample size is fairly small. Nevertheless, it is reasonable to state that this Dynamic Ensemble Selection method has shown very encouraging potential in dealing with heterogeneous data sources.

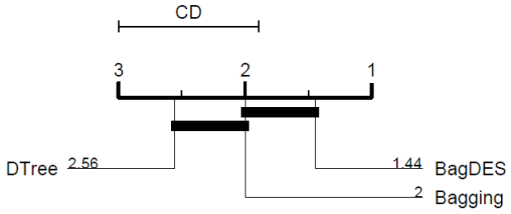

Fig. 3. Critical differences between the ranks of the results obtained from three methods: Decision Trees(DTree), Bagging and BaggingDES(BagDES) on all the datasets. A thick bar groups the methods that are not statistically significantly different at CD = 1.10.

E. Diversity, Error and Size of DESO Ensembles

As described earlier in Section IV, we chose two non-pairwise diversity measures to investigate their influence on the performance of DESO ensembles under some conditions, such as the number of DESO clusters and the ensemble size.

1) Coincident Failure Diversity(CFD): The fitting curves in Figure 4 show mean train and test error and mean CFD as a function of DESO ensemble size by dataset.

These plots show that CFD has a relatively high degree of correlation with the training error and a low correlation with the test error. That means that the relationship between the CFD and the test performance of DESO ensembles is not clear when optimising DESO ensemble.

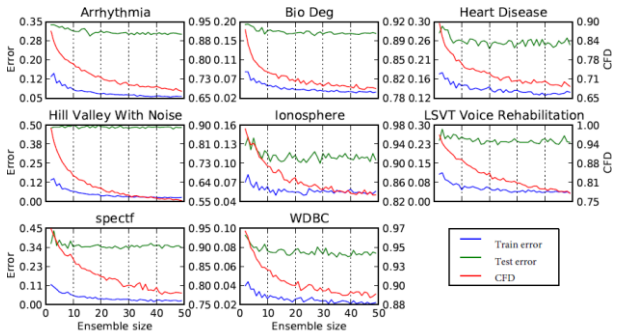

Fig. 4. Mean train & test error and CFD as DESO ensemble size varies.

2) Minority Failure Diversity(MFD): The fitting curves in figure 5 show the mean train and test error, and mean MFD as a function of DESO ensemble size on each dataset. It can be seen that these curves are different from the ones CFD and closer to or almost in line with the test errors after the certain sizes. That means that MFD is a better predictor of the testing accuracy of DESO.

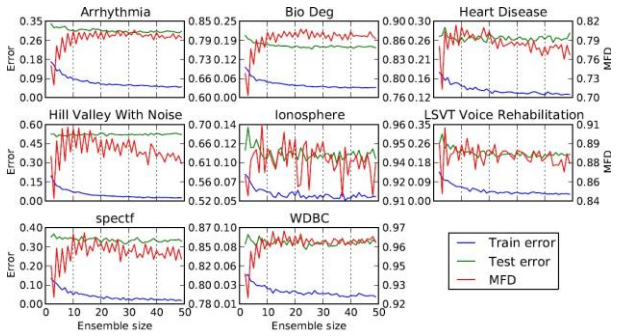

Fig. 5. Mean train & test error and MFD as DESO ensemble size varies.

In summary, these results show quite clearly that diversity measure CFD tends to behave just like the training errors, which are obviously different from the testing error on all the datasets used in this study and hence it is difficult to use it to predict how good the ensembles may perform on the test data. On the other hand, the diversity measure MFD appears to have closer approximations to the test errors, when the number of the classifiers in ensembles reaches a certain threshold, e.g. around 5 in this study, regardless of the differences in the data sets and the number of classifiers used in ensembles. Another common characteristic pattern, showed by all the curves, is that the diversity and test errors tend to be levelled off after the size of the ensembles reaches around 10, which indicates it is not necessary to generate a larger number of models to build an ensemble as they do not contribute much to the improvement of the performance of DESO ensembles, whilst consuming more time and memory space in learning.

VI. CONCLUSION

This paper presents a generic dynamic ensemble selection and optimisation (DESO) framework and an example of its

implementation - BaggingDES, taking a common ensemble method - Bagging, as a testing vehicle, for mining heterogeneous data sources.

This work is significant simply because big data is often collected and/or aggregated from multiple sources in different representations and granularity over different temporal and spatial periods, and these heterogeneous data sources are more likely to capture and represent local, diverse and/or complementary information of a given problem, it is then logical to utilise these heterogeneous data sources "independently" to induce models that could be more diverse from each other, compared with the models trained from a single homogeneous data source.

So, the learning process of the proposed DESO starts by generating models or classifiers with different sources of the data - represented with different regional feature spaces, and these models are clustered by using k-means into groups or sub-ensembles and optimised with simulated annealing algorithm based on the diversity among the models. When classifying, a new data instance is presented to the nearest sub-ensembles and their decisions are then aggregated with a consensus function, such as voting, ot produce a final classification decision.

The BaggingDES is tested on the eight semi-simulated heterogeneous datasets with various numbers of feature sets - generated from 7 real-world benchmark datasets, and the experimental results show that the performance of BaggingDES is significantly better than the reference baseline method - decision tree, and much better the classic Bagging method, though not significantly.

The experimental results of examining the relationships between the diversity, error and size of DESO ensembles show that there are some correlations between the diversity and error but it is not stronger enough to be used as a reliable predictor of the performance of ensembles on test data.

Another finding is that with heterogeneous data, it is not necessary to build a DESO-ensemble with large number of the models that are induced from the same course as they make no or negligible contribution to the performance improvement of DESO-ensembles.

With these encouraging results, further work should include running more experiments on more big datasets, which consist ideally of true heterogeneous data sources to verify the findings of this study and also investigating in more depth on the relationship between diversity and accuracy and reliability of DESO ensembles.

REFERENCES

[1] S. Holmes, "Seminar slides: Heterogeneous data challenge: Combining complex data." [Online]. Available: http://web.stanford.edu/group/mmds/slides2010/Holmes.pdf

[2] A. Mojahed, J. Bettencourt-Silva, W. Wang, and B. de la Iglesia, "Applying clustering analysis to heterogeneous data using similarity matrix fusion," in *Machine Learning and Data Mining in Pattern Recognition*, 2015, pp. 251–265.

[3] S. Yu, B. de Morr, and Y. Moreau, "L2-norm multiple kernel learning and its application to biomedical data fusion." *BMC Bioinformatics*, vol. 11, p. 153, 2010.

[4] G. Richards and W. Wang, "What influences the accuracy of decision tree ensembles?" *Journal of Intelligent Information Systems*, vol. 39, no. 3, pp. 627–650, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10844-012-0206-7

[5] A. H. Ko, R. Sabourin, and A. S. Britto Jr, "From dynamic classifier selection to dynamic ensemble selection," *Pattern Recognition*, vol. 41, no. 5, pp. 1718–1731, 2008.

[6] J. Xiao and C. He, "Dynamic classifier ensemble selection based on gmdh," in *Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on*, vol. 1. IEEE, 2009, pp. 731–734.

[7] R. G. Soares, A. Santana, A. M. Canuto, and M. C. P. de Souto, "Using accuracy and diversity to select classifiers to build ensembles," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1310–1316.

[8] T. Woloszynski and M. Kurzynski, "A measure of competence based on randomized reference classifier for dynamic ensemble selection," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 4194–4197.

[9] R. Lysiak, M. Kurzynski, and T. Woloszynski, "Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers," *Neurocomputing*, vol. 126, pp. 29–35, 2014.

[10] S. Bian and W. Wang, "On diversity and accuracy of homogeneous and heterogeneous ensembles," *International Journal of Hybrid Intelligent Systems*, vol. 4, pp. 103–128, 2007.

[11] W. Wang, "Some fundamental issues in ensemble methods," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008, pp. 2243–2250.

[12] P. Adhvaryn and M. Panchal, "A review on diverse ensemble methods for classification," *IOSR Journal of Computer Engineering*, vol. 1, no. 4, pp. 27–32, 2012.

[13] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181–207, 2003.

[14] D. Partridge and W. Krzanowski, "Software diversity: practical statistics for its measurement and exploitation," *Information and software technology*, vol. 39, no. 10, pp. 707–717, 1997.

[15] K. Bache and M. Lichman, "UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2013)," 2013.