Generalised Decision Level Ensemble Method for Classifying Multi-media Data

Saleh Alyahyan[1,2] and Wenjia Wang[1]

[1] University of East Anglia, Norwich, UK
[2] Shaqra University, Shaqra, Saudi Arabia
{S.Alyahyan,Wenjia.Wang}@uea.ac.uk

Abstract. In recent decades, multimedia data have been commonly generated and used in various domains, such as in healthcare and social media due to their ability of capturing rich information. But as they are unstructured and separated, how to fuse and integrate multimedia datasets and then learn from them effectively have been a main challenge to machine learning. We present a novel generalised decision level ensemble method (GDLEM) that combines the multimedia datasets at decision level. After extracting features from each of multimedia datasets separately, the method trains models independently on each media dataset and then employs a generalised selection function to choose the appropriate models to construct a heterogeneous ensemble. The selection function is defined as a weighted combination of two criteria: the accuracy of individual models and the diversity among the models. The framework is tested on multimedia data and compared with other heterogeneous ensembles. The results show that the GDLEM is more flexible and effective.

Keywords: Multimedia· Classification · Heterogeneous Ensemble · Diversity · Model Selection.

1 Introduction

Multimedia data usually consist of multiple media types of data, such as text, images, graphics, audio and video etc.[1][2]. They can be generated in many fields and internet. For example, in healthcare, a patient with a potential complex disease may be examined through a series of tests to investigate their underlying problems with blood test, screening (X-rays, CT, MRI), Electroencephalography(EEG), or Electrocardiogram(ECG), Endoscopy(video), etc. These tests generate data represented in multimedia, as illustrated by Fig. 1, and they need to be considered together by doctors in order to make an accurate diagnosis and carry out appropriate treatments.

Then, as more and more multimedia data have been rapidly generated and accumulated in various applications, analysing, discovering useful knowledge and learning from them have become a challenge in machine learning and data mining fields. Although many different algorithms have been developed for dealing with multimedia data with different strategies, how best to combine many sets of
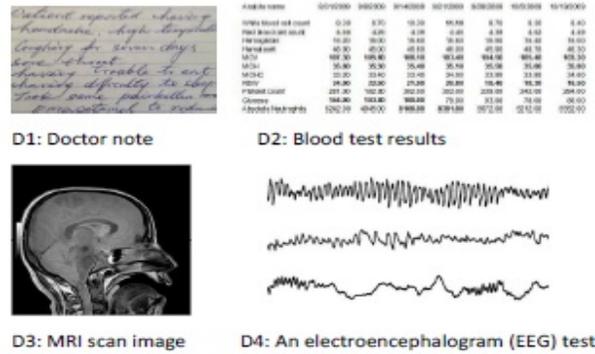
Fig. 1. An example of multimedia medical data.

different media data for effective machine learning is still a research issue [3][4]. It should be stressed that there seems a misrepresentation of multimedia data in some published researches which were presented as the work on multimedia whilst actually used only one media of data, mostly just imagery or video, instead of several media datasets together[5][6].

So for clarity, we define Multi-Media Data (MMD) as a collection of several sub datasets that are represented by at least two or more different media formats: numbers, text, image, video, graphics, animation, audio, time series data, etc.

Multimedia data are usually unstructured so they need to be transformed into a structured representation through feature extraction. Then their features are aggregated to form a big flat single dataset for further analysis. This is arguably the most common approach for integrating multimedia data Then the analysis can be done just like any other data. In this way, one obvious possible problem is that the integrated dataset may have a very high dimensionality, i.e. too many features that can often overwhelm machine learning and data mining algorithms to produce good results[7].

In this research, we apply another strategy, contrary to above approach of integrating multimedia datasets into one big dataset. Firstly we use each dataset to generate some models, and then combine these models' decisions to produce the final solution. This is called the decision-level fusion. Moreover, with this approach, it provides us a natural platform to build heterogeneous ensembles for classifying multimedia data.

A heterogeneous ensemble for classification combines multiple classifiers that are created by using different algorithms on different or same datasets, with an aim of making the classifiers more diverse and hence possibly increasing accuracy [8][9][10].

Following our previous studies[11][12], this study presents a new generalised function for selecting classifiers to build a heterogeneous ensemble based on two criteria - accuracy and diversity, combined with variable weights.

The rest of the paper is organized as follows. Section 2 briefly reviews some related previous studies. Section 3 describes our proposed methods in detail, including the tools and programs used in the research. Section 4 provides details of the experiment conducted and our results. Section 5 gives conclusions and suggestions for the further work.

2   Related Work

As mentioned, although there are many publications on multimedia domains, they mostly only dealt with only a single type of media data. For instance, on a major international conference on multimedia big data in 2017, only about 15% of the published papers used more than one type of media data. So, there are actually not many published researches truly working on multimedia data, nor many multimedia data sets are publicly available.

There are, however, some studies that have used multimedia datasets for machine learning. For instance, Aalaa et al. [7][13] generated five small heterogeneous datasets containing a mixture of both structured and unstructured datasets and then applied clustering ensemble methods to them. Their experiments showed that the results using all available types of media outperform the results using the best individual types of media.

Yamanishi and Jenjia [14] conducted a study of the distributed learning system for Bayesian learning strategies. In their system each instance was observed by different classifiers which were called agents. They aggregated the outputs from the agents to give significantly better results. They demonstrated that distributed learning systems work approximately (or sometimes exactly) as well as the non-distributed Bayesian learning strategy. Thus, by employing their method, they were able to achieve a significant speeding-up of learning.

Onan [15] applied ensemble classification methods to text datasets. In his experiment the data sets ware represented by 5 different formats. Five types of classifiers were used: Naive Bayes, Support Vector Machine, K-Nearest Neighbour, Logistic Regression and Random forest. He compared individual classifiers and their homogeneous ensemble using Bagging and Boosting. The results showed that ensembles out-performed individuals.

Ballard and Wang[16] developed a dynamic ensemble selection methods for heterogeneous data mining. Although their datasets are not multimedia, their basic idea of combining multiple datasets at decision level inspired this work.

Some recent studies published in a major international conference on Multimedia Big Data (2017) used multimedia data. For instances, Amato et al.[17] used a multimedia dataset collected from the social media networks to develop a recommendation system. The data they used include text messages, tags, photos and possible interactions between the users. Their results show that using these multimedia data in their filtering and ranking based method helped to achieve a promising accuracy of recommendation, even though they did not use any machine learn methods. Liu et al.[18] used both image and audio data to

detect drones. They used the feature level integration approach to generate a big dataset for training SVM models for classifying drones.

In summary, previous studies have used decision-level combination methods and different machine learning approaches to analyse so-called heterogeneous datasets, whilst in fact their datasets mostly come from different data sources of the same type. Thus, these studies were limited by their single medial of data and how their methods may perform on multimedia datasets is unknown. Also, much of this previous work is limited to applying some issues which affect the ensemble out comes, including diversity and accuracy for candidate models. In other words, model selection criteria have not been applied.

3   The Generalised Decision Level Ensemble Method(GDLEM)

3.1   The Decision Level Ensemble Method Framework

Our decision-level ensemble method(DLEM), as shown in Fig. 2, consists of four modules namely: (1) the multimedia data representation and feature extraction, (2) the modelling, (3) the model selection and, (4) the combination. In the first
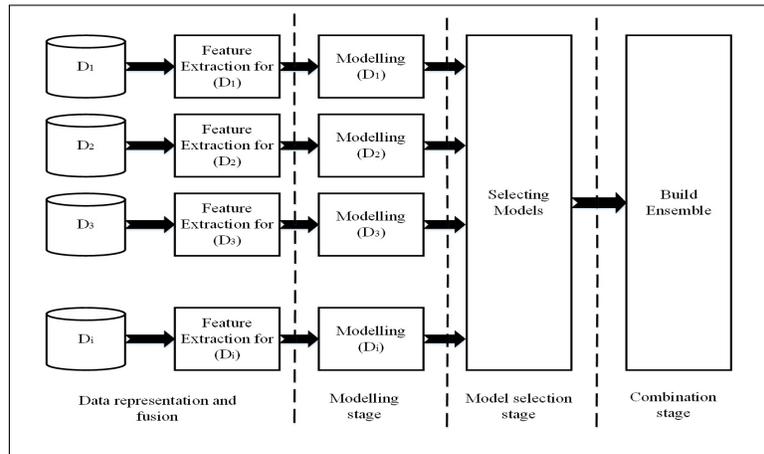

Fig. 2. The general framework for DLEM

stage, the DLEM extracts features from each subset of media data to create D_i's $(1 < i < n)$ such that each D_i represents the unique type of media features, i, for each instance.

In the second stage, the DLEM employs some heterogeneous machine learning algorithms to generate individual models for each dataset D_i. The total number of the generated individual models for the MMD, D, is determined by $m * n$.

This modelling stage produces a pool of models, PM, with members PM_{ij} representing the individual model fitted using D_i with the base classifier method $B_j, 1 < j < m$.

The third stage selects models from the model pool PM using accuracy and diversity as selection criteria, either individually or jointly in some predefined rules. Using these criteria, three different rules: $R0$, $R1$, $R2$, were derived. After some intensive experiments in our earlier studies, we then devised a new rule that uses a function to combine accuracy and diversity in a more generalised manner to select the models.

Rules R0, R1 and R2 are described in our earlier work [11][12]. We give a brief summary of them here for convenience as they are the bases of the new rule.

R0: This rule only uses *accuracy* as the criterion for model selection. The DLEM firstly computes the accuracy, $(Acc(m_i))$, for each of the n models in the model pool PM and sort them in a descending order based their accuracy. Then the DLEM selects the N most accurate models from the PM, i.e., $m_i = max\{Acc(m_j), m_j \in PM\}\, i = 1...N$, and add them to the ensemble, ϕ.

R1: It uses both *accuracy* and *diversity* as criteria separately to select models at different stages. The DLEM first removes the most accurate model (MAM) from PM; using $m_1 = max\{Acc(m_j), m_j \in PM\}$ and add it to the ensemble, ϕ. Then the pairwise diversities between MAM and remaining models in PM, are calculated by the Double Fault (DF) method[19] and the models in the PM are sorted in a decreasing order based on the magnitude of the DF's. The (N-1) most diverse models from the sorted PM are selected (Equation 1) and added to the ensemble, ϕ. Therefore ϕ now contains MAM and the (N-1) most diverse models from PM.

$$m_i = max\{DF(m_1, m_j), m_j \in PM\}\, i = 2...N \qquad (1)$$

R2: This rule uses both *accuracy* and two types of *diversity* measures, namely the *DF* method and the *Coincident Failure Diversity* (CFD) method [20]. Firstly, the MAM is selected and removed from the PM and added to Φ. Then the most diverse model (MDM) is determined from the PM using $MDM = max\{DF(m_1, m_j), m_j \in PM\}$ and added to Φ which now contains both the MAM and MDM models. All possible combinations between Φ and each of the remaining (N-2) members of the PM are generated to create J number of ensembles, ϕ_i, where $1 \le i \le J$ and J is given by $J = \binom{|PM|}{N-2}$. For each ϕ_i, its diversity is computed with CFD. The ensemble with the maximum CFD diversity, Φ_{md}, is then selected as the final ensemble.

$$\Phi_{md} = max\{CFD(\Phi \Leftarrow m_j), m_j \in PM\} \qquad (2)$$

The relevant literature shows that most studies, including some of our own rules. e.g. R0, only used one measure, either accuracy or diversity as the model selecting criterion, but our results as well as others demonstrated that the ensembles built in such ways are not really utilizing the balanced strengths of individual models but only focused on one aspect over others. Although our rules R1 and

R2 use two measures but they use them separately in a sequential manner. So in this study, we proposed a new rule that combines both accuracy and diversity with a weighted function and the combined score is used for selecting candidate models.

R3: This new rule uses a combination of accuracy (Acc) and diversity(Div) as a generalised criterion for selecting models to build an ensemble. The combined measure is defined below.

$$\gamma_i = \alpha(Acc)_i + \beta(Div)_i .\tag{3}$$

Where α and β are the weights for accuracy, Acc, and diversity ,Div, of model m_i ($1 \leq$ i$\leq n - 1$) in the PM. The diversity measure Div in this rule is flexible can be a pairwise on non-pairwise diversity as long as it is considered as appropriate. In this study, we use the CFD.

After taking the best model out from the model pool PM, the combined score, γ_i, is calculated for the remaining $n - 2$ models in PM. The model with $max(\gamma_i)$ is selected from PM and added to Φ.

R3 is considered as a generalised rule because all other three rules R0, R1 and R2 are just its special cases with specific values for the weights and the diversity measure. When set $\alpha = 1$ and $\beta = 0$, R3 becomes R0. If we use the DF as the diversity measure and set $\alpha = 0$ and $\beta = 1$, then R3 becomes R1. if we use non-pairwise diversity measure such as the CFD and set $\alpha = 0$ and $\beta = 1$, R3 becomes R2.

Based on this new rule, a corresponding algorithm for building a decision level ensemble is derived and named as generalised Decision-level Ensemble Method (GDLEM) as it is flexible with R3 to apply various rules for selecting models by manipulating the weights or changing the measures used in the relationship γ in equation 3. The GDELM as follows. The first step is the same as that of the other three rules, i.e. choosing the MAM from PM as the first member of Φ. The key difference starts from the second step where the selection of candidate models uses the newly defined γ_i. This second step is repeated until N models with $max(\gamma_i)$ completely fills Φ.

3.2   Implementation of the GDLEM

The experiment was carried out on a normal PC, with an I7 processor and 16 GB RAM. As the GDLEM is flexible for selecting candidate classifiers, we have selected 10 efferent base classifiers that are provided in the WEKA library [21]. These base classifiers are: trees(*J48, RandomTree, REP-Tree*), bayes(*NaiveBayes, BayesNet*), function(*SMO*), rules(*JRip, PART*) and Lazy(*IBk, LWL*).

4 Experiment Design and Results

4.1 Dataset

We conducted a series of experiments using a benchmark dataset − 8 Scene Categories Dataset[22], which contains two subsets in different media: 2688 images and their annotations represented by XML files. The images are categorized into eight classes in according to their scenes and objects captured by the images. Each XML file contained a number of tags that describe an image. The annotations were dealt with as text and 782 textual features were extracted out from the texts to form a data subset D_t. For the imagery data, 567 features were extracted out from the images using Histograms of Oriented Gradients (HOG) [23] to form another data subset, i.e. imagery data D_g.

Ten base classifiers were used to learn from the textual and the imagery features subsets, which gave twenty heterogeneous models in total. This gave the GDLEM the opportunity to have more variety of models.

4.2 Experiment Design and Results

We carried out a series of experiments to investigate the performance of the GDLEM, using three selection rules separately, on the multimedia data. The investigated issues included (1) the performance measures and classifier selection criteria represented by the rules: R0, R1, R2 and R3, and (2) the ensemble size. A total of 135 experiments were conducted. This involved running all possible combination of these parameters. Each experiment was repeated five times with different samplings of the datasets.

In parallel, we conducted the experiments to investigate the influence of CFDs on the accuracy of all the ensembles built with the first three rules although the CFD is not used by R0 and R1. With R3, through varying the values of the weights α and β from 0 to 1 with an increment of 0.1, such that $\alpha + \beta = 1$, and using the above experiment settings, 850 experiments were carried out in total.

Some summarised results are shown in Fig.3. They clearly shows that the GDLEMs built with the three rules are generally superior to individual classifiers, because that the mean accuracies (shown in red lines on the figures) of the GDLEMs are approximately 10% higher than the mean accuracies (illustrated by blue lines) of the individual classifiers in the GDLEMs. In addition, it further demonstrated that our ensemble results have a higher level of accuracy overall than the best individual models, the MAMs. Hence, our GDLEM had the best reliability overall because the reliability of an MAM was not consistent over a succession of experiments. On the hand, the ensembles built with our method, the GDLEM, are more consistent and reliable as well as more accurate.

Fig.4 compares the results of GDLEMs built with the three rules and variable sizes from 3, 5, 7 to 19 on the test data. This shows the weakness of R1. Our previous studies indicated that there were accuracy issues with this rule. However, these became much more apparent in the current work when the high number of models were used. The increase in model numbers highlighted very
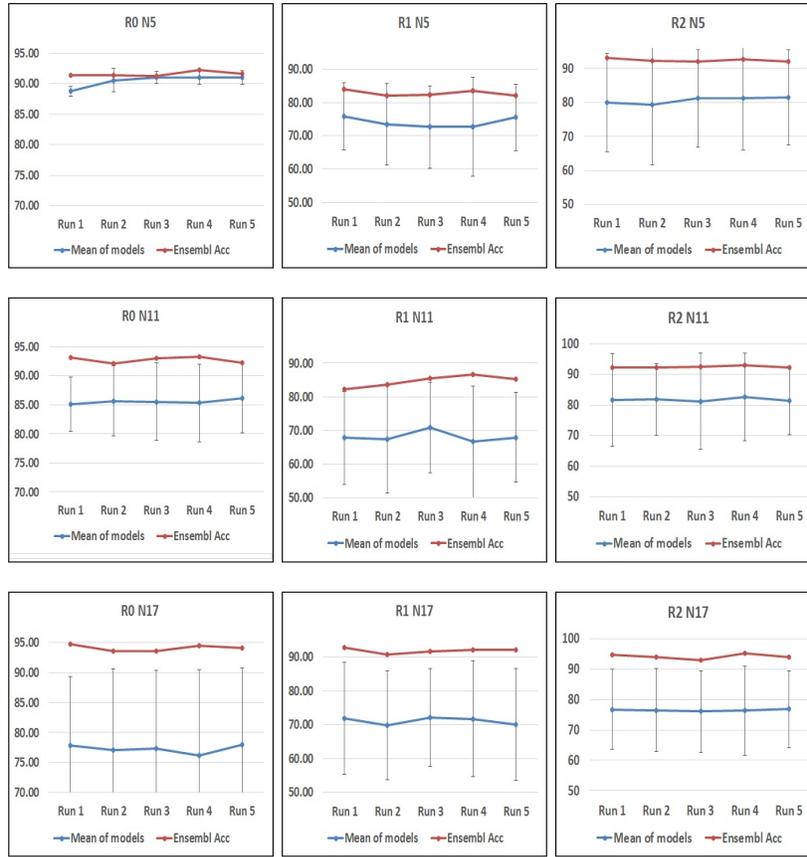
Fig. 3. Summaries of GDLEM results for rules R0, R1 and R2 and the size of ensemble 5, 11 and 17 are shown in each sub-graph. Tow lines (read and blue) are the accuracy of GDLEM and the mean accuracy for models that are chosen for the GDLEM respectively. The stranded deviation is shown by whiskers over 5 runs.

clearly the disadvantages of R1. As can be seen, its accuracy levels varied inconsistently: started low and wend further lower; it only improved when N = 11. On all the way up to N = 19, it is still worse than two other rules.

R0 performed reasonably well because it combines all the models in the PM, which have the best accuracies. R2 is same as R0 when N=3, but improved while R0 went down when the size increased, although they are similar after N = 11. But R2 is more favourable as it performed better when the size of ensembles is smaller, which means it is more efficient.

Fig.5 shows the average values of the CFD in the ensembles built with R0, R1 and R2, although the CFD is not used in R0 and R1. The purpose is to see if the CFD can be used to explain why some ensembles are better than others. These results show that in R0 the CFD is increasing to give the best results at
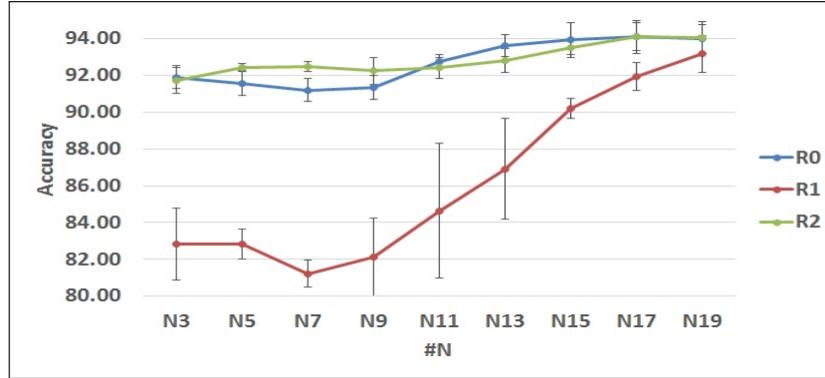
Fig. 4. Comparing the results produced by all three rules in nine different sizes of the GDLEM.

N11. When we link this result with the accuracy level for R0 shown in Fig. 4, we can see that the best ensemble results were gained when we combined models that have best accuracy and CFD when N = 11 and 19.
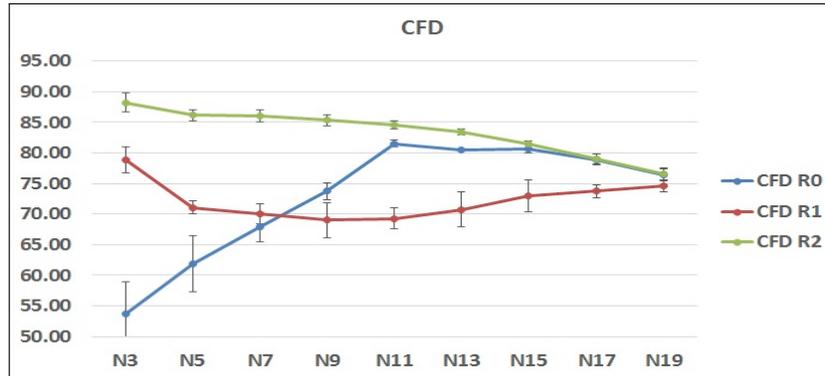

Fig. 5. Comparing the CFDs for all three rules in nine different sizes of the ensembles.

As R3 is a generalised and flexible rule, it gives the chance to do more investigation into the influences of the CFD in the ensemble. Fig.6 shows improvement in accuracy for model selection for some ensembles of size 3 to 9, given α is set between 0.5 and 0.6. For ensembles of size greater than 9, varying α and β does not have much impact on the accuracy level for the ensemble (See Fig.7) and that is because the size of the model pool is too small. When the size of the ensemble reaches and succeeds the 50% of the model pool, there is not much space for selecting models and hence the ensembles could be more or less the same regardless of whatsoever models are chosen.
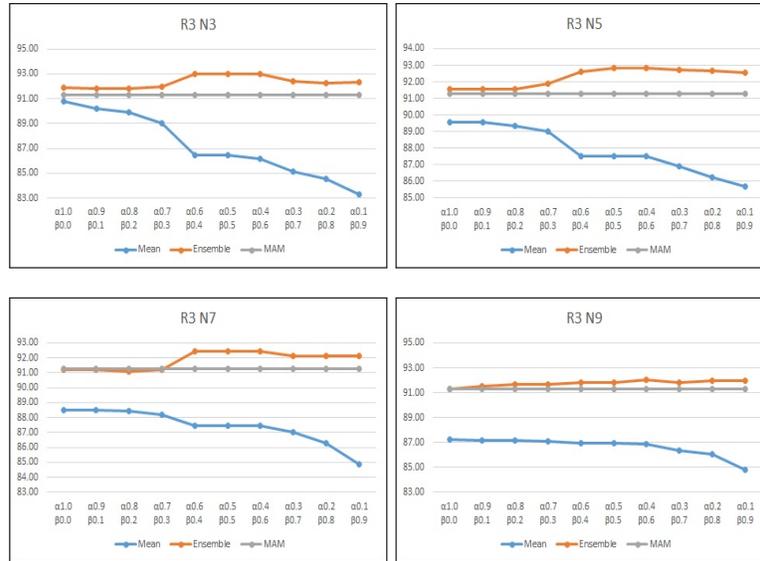
Fig. 6. Sample of GDLEM results for the generalised rule R3 with ensemble size 3, 5, 7 and 9 are shown in each sub-graph. three lines: read for the accuracy of GDLEM, blue for the mean accuracy for models that are chosen for the GDLEM and grey for MAM.

The best results are produced by R3 when the weight of the accuracy α is equal to 0.4 as it is shown in the critical difference digram in Fig.8 and the weight for diversity is 0.6. And this means, when more weight, about 20% , is put on the diversity than on the accuracy and CFD, the ensembles with less accurate but more diverse modules achieved the best results. Moreover, the digram shows that the range between 0.4 and 0.6 for α performs better than others.

Thus, it can be seen that the generalised selection rule R3 - a combination of accuracy and CFD measures, gives chances to the GDLEM to select the models that can help improve the accuracy of heterogeneous ensembles. The systematic empirical investigations found the best ensembles are produced when the weights for accuracy and diversity are split at 0.4 to 0.6 respectively. That effect clearer when a large pool of models and we select number of models below the half. In summary, the ensembles built with model selection criteria that use a combination of CFD and DF and accuracy measures, gives good results. They are superior to those results obtained using either pair-wise diversity (R1) or just accuracy (R0).

4.3   Critical comparison with other ensembles

The results of GDLEM were compared with the feature-level ensemble method(FLEM) and various heterogeneous ensembles based on the single media
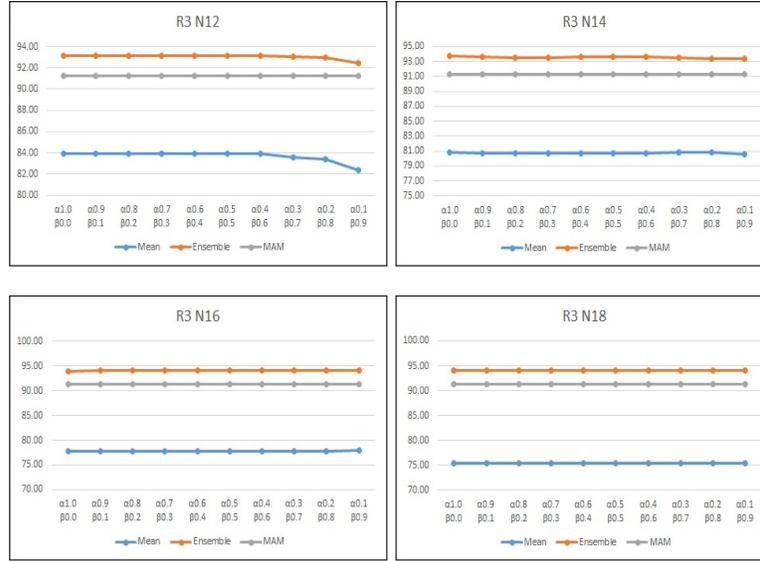
Fig. 7. The summary of the GDLEM results of ensembles with 12, 14, 16 and 18 models selected by the generalised rule R3. Red line represents the accuracy; blue = the mean accuracy of the models in the ensembles, and grey = MAM in the ensembles.

data, text (HEST) and image data (HESG). The full comparative results between the FLEM and the HESG were published in [12] and the full results for the HEST were published in [11]. Figure. 9 shows the critical difference diagram for the GDLEM, DLEM, FLEM, HEST and HESG, with all rules R0, R1, R2 and R3. The GDLEM-R3 is the best on average and a credible explanation is this R3 with appropriate weights can produce the optimal combination between the accuracy and CFD to improve classification accuracy.

5   Conclusion and Future Work

In this study, we developed a generalised heterogeneous ensemble method to classify multi-media datasets at the decision level (GDLEM) aiming to achieve the best and most reliable accuracy results. Our GDLEM consists of four stages: extracting features from multi-media subsets, modelling the subsets datasets, selecting models with different rules based on various criteria, and building heterogeneous ensembles. The new model selection rule, R3, has been demonstrated to have a capability to select the individual models that are less accurate but more diverse. Hence, in some points, e.g. accuracy weight from 40% to 60%, it achieved the best level of ensemble accuracy beating those obtained by other ensembles, including DLEM, FLEM, HEST and HESG in the same dataset. Other obvious observation from this study is that heterogeneous ensemble gives
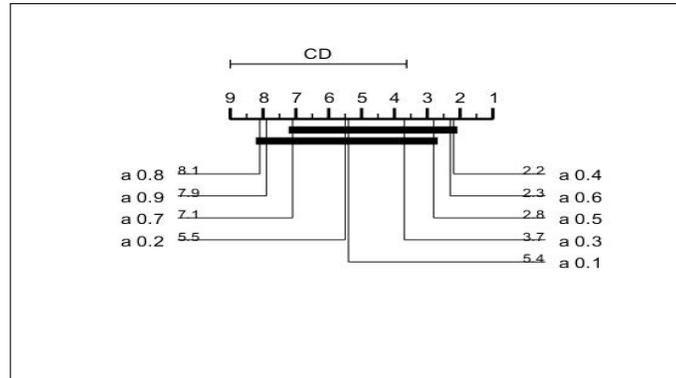
Fig. 8. A diagram shows critical differences of the average results of ensembles with different sizes from 3 to 19, when the accuracy weight α varied from 0.1 to 1.0 with a step size of 0.1.

better results when we combine accuracy and diversity measurements for model selection.

Suggestions for future work to improve our approach include (1) increasing model pool size so that there are more choices for model selection, (2) creating other complex selection rules by adding more measures in the relationship on the generalised R3, (3) of course, the approach should be tested on more multimedia datasets. Finally it should be extended to other different problems such as time serious classification.
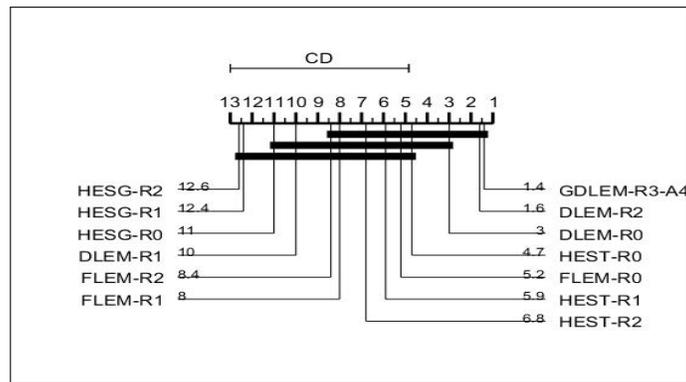
Fig. 9. Critical difference diagram for the ensembles built with GDELM, DLEM, Feature-Level Ensemble Method(FLEM), Hybrid Ensembles Built with Textual Data(HEST) and with Imagery Data(HESG) for all rules R0, R1, R2 and R3. It shows that the GDLEM with R3 is the best.

References

1. S. More and D. K. Mishra, "Multimedia data mining: A survey," *International Journal of Science, Spirituality, Business And Technology*, vol. 1, no. 1, pp. 49–55, 2012.
2. S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. Iyengar, "Multimedia big data analytics: A survey," *ACM Computing Surveys*, vol. 51, pp. 1–34, 01 2018.
3. W. Zhu, P. Cui, Z. Wang, and G. Hua, "Multimedia big data computing," *IEEE multimedia*, vol. 22, no. 3, pp. 96–c3, 2015.
4. P. Wlodarczak, J. Soar, and M. Ally, "Multimedia data mining using deep learning," in *Digital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on*, pp. 190–196, IEEE, 2015.
5. M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.
6. Z. Kuang, Z. Li, T. Zhao, and J. Fan, "Deep multi-task learning for large-scale image classification," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pp. 310–317, April 2017.
7. A. Mojahed, J. H. Bettencourt-Silva, W. Wang, and B. de la Iglesia, "Applying clustering analysis to heterogeneous data using similarity matrix fusion (smf)," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 251–265, Springer, 2015.
8. 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*, pp. 2243–2250, IEEE, 2008.
9. B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017.
10. M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3–17, 2014.
11. S. Alyahyan, M. Farrash, and W. Wang, "Heterogeneous ensemble for imaginary scene classification," in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pp. 197–204, 2016.
12. S. Alyahyan and W. Wang, "Feature level ensemble method for classifying multimedia data," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 235–249, Springer, 2017.
13. A. Mojahed and B. de la Iglesia, "An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach," *Knowledge and Information Systems*, vol. 50, no. 1, pp. 27–52, 2017.
14. K. Yamanishi, "Distributed cooperative bayesian learning strategies," *Information and Computation*, vol. 150, no. 1, pp. 22–56, 1999.
15. A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *Journal of Information Science*, p. 0165551516677911, 2016.
16. C. Ballard and W. Wang, "Dynamic ensemble selection methods for heterogeneous data mining," in *12th World Congress on Intelligent Control and Automation (WCICA)*, pp. 1021–1026, June 2016.
17. F. Amato, V. Moscato, A. Picariello, and G. Sperlì, "Recommendation in social media networks," in *Third IEEE International Conference on Multimedia Big Data, Laguna Hills, CA, USA, April 19-21, 2017*, pp. 213–216, IEEE Computer Society, 2017.

18. H. Liu, Z. Wei, Y. Chen, J. Pan, L. Lin, and Y. Ren, "Drone detection based on an audio-assisted camera array," in *Third IEEE International Conference on Multimedia Big Data, Laguna Hills, CA, USA, April 19-21, 2017*, pp. 402–406, IEEE Computer Society, 2017.

19. G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9, pp. 699–707, 2001.

20. 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.

21. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

22. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

23. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.