Wireless Networks
https://doi.org/10.1007/s11276-018-01906-3

Decision level ensemble method for classifying multi-media data

Saleh Alyahyan[1,2] · Wenjia Wang[1]

Abstract
In the digital era, the data, for a given analytical task, can be collected in different formats, such as text, images and audio etc. The data with multiple formats are called multimedia data. Integrating and fusing multimedia datasets has become a challenging task in machine learning and data mining. In this paper, we present heterogeneous ensemble method that combines multi-media datasets at the decision level. Our method consists of several components, including extracting the features from multimedia datasets that are not represented by features, modelling independently on each of multimedia datasets, selecting models based on their accuracy and diversity and building the ensemble at the decision level. Hence our method is called decision level ensemble method (DLEM). The method is tested on multimedia data and compared with other heterogeneous ensemble based methods. The results show that the DLEM outperformed these methods significantly.

Keywords Multi-media data · Classification · Ensemble · Decision level fusion · Diversity · Models selection

1 Introduction

In recent decades, multimedia has been increasingly generated and used in various fields and applications such as, in healthcare, where numerical data (test results), images (X-rays, CT, MRI scans), time series (EEG or ECG), video (endoscopy), audio (recorded doctor's voice), and textual data (test reports, doctor's notes etc.), as illustrated by Fig. 1, are often generated when trying to make a diagnosis for a complex disease. It is more popular on social media where text, emoticons, images, video and audio talks, are also often uploaded and displayed to help enhance the meaning and understanding of a conversation or a concept.

Then, integrating and/or fusing these multimedia datasets together to obtain as much useful information as possible to improve machine learning, has become a challenging task [17, 20].

✉ Wenjia Wang
Wenjia.Wang@uea.ac.uk

Saleh Alyahyan
S.Alyahyan@uea.ac.uk

[1] School of Computing Sciences, University of East Anglia, Norwich, UK

[2] College of Computer Science, Shaqra University, Shaqra, Saudi Arabia

However, it should be noted that multimedia data has been inappropriately interpreted in some studies and published literature where in fact they used just one single media data, usually just image data, as multimedia data [5].

In this research, we define Multi-Media Data (MMD) as a collection of several datasets that are represented by at least two or more different media formats: numerics, text, image, video, graphics, audio, time series data etc.

In order to analyse multimedia data, a very common method is to combine all the sub-datasets into a big flat single dataset, which is done by integrating all the features extracted from multimedia datasets. Then the analysis can be done just like any other data. In this way, one obvious possible problem is that the integrated dataset may be too big with a very high dimensionality, i.e. too many features that may overwhelm a machine learning and data mining algorithm to produce good results [10].

In this research, we apply another approach, that is, instead of fusing all datasets into a big dataset, we firstly use each dataset to generate a model or 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 with 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 the aim of

Published online: 11 December 2018

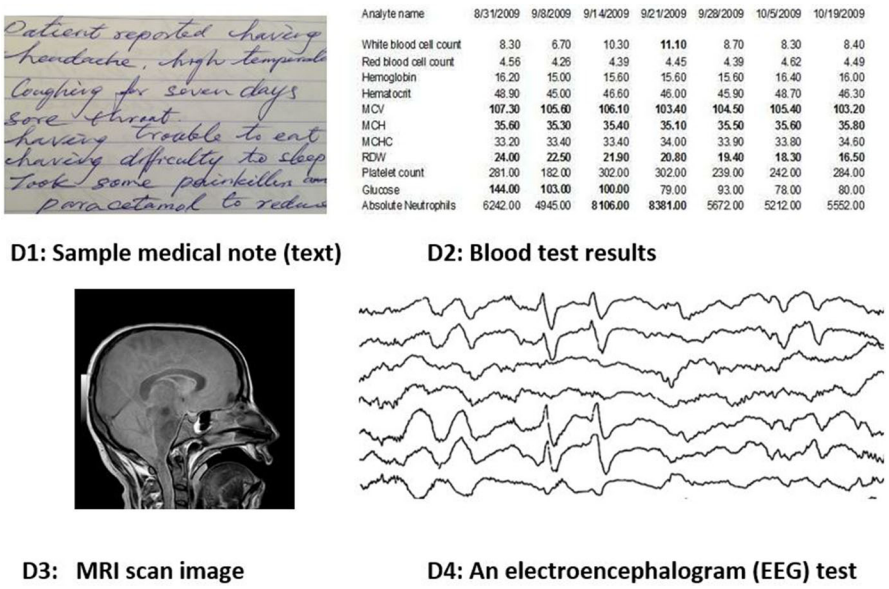


Fig. 1 An example of multimedia medical data

making the classifiers more diverse and hence possibly increasing accuracy [8, 15, 18]. This work is a continuation of our previous research [1, 2]. In this study we will investigate the problems of classifying multimedia data by combing them at decision level with heterogeneous ensembles.

The remainder of the paper is structured into four parts: Sect. 2 provides a brief review of some earlier studies related to the current work. Section 3 gives a detailed description of the methods used. This covers the programs and the tools. Section 4 gives points of interest of the investigation directed and our outcomes. Section 5 sets out our conclusions and suggests further work which could be undertaken.

2 Related work

There are several studies that have applied machine learning methods to multimedia datasets. Mojahed et al. [10, 11] applied clustering ensemble methods to multimedia datasets. They generated five different heterogeneous datasets containing a mixture of both structured and unstructured datasets. Their experiments showed that the clustering results using all available types of media outperform the clustering results using the best individual types of media.

Bagnall et al applied ensemble methods to time series data analysis [3, 4]. In their work, massive time series datasets are transformed into four different representations, which are equivalent to multimedia datasets, and then each subset is used to train seven different base classifiers including Random Forest, Naive Bayes, Decision Tree and Support Vector Machine. Some of these classifiers are then used to build ensembles. They demonstrated that they could achieve significantly improved accuracy results on more than 75 datasets. Do et al. [6] conducted experiments in the same area using series Nearest Neighbours classification. Their methods out-performed other methods, including Random-Forest and Support Vector Machine.

Yamanishi [19] 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. In their study, 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 [13] applied ensemble classification methods to text datasets. In his experiment the data sets ware represented by 5 different representations. The classifiers which were used were five Naive Bayse, Support Vector Machine, K-Nearest Neighbour, Logistic Regression and Random forest. In his experiments, Onan compared individual classifiers and their homogeneous ensemble using Bagging and Boosting. The results show ensembles out-perform individuals.

To sum up, as can be seen clearly, although decision level combination idea has been used in some earlier studies when building various types of ensembles, their studies are limited in several aspects. The most critical one is that they did not explore the idea deeply on, for example, how to select appropriate models from each subset of models to build more effective ensembles. Then, the datasets used in these studies are predominantly derived

from the same types of data and different sources and those datasets are not of truly multimedia. Therefore, it is not possible to know how their methods would work with multimedia datasets. These issues will be addressed in our study. We will present a framework that deals with true multimedia data and employs several rules for selecting modules based on different criteria, accuracy and diversity, used independently or sequentially.

3 Decision level ensemble method (DLEM)

3.1 The decision level ensemble method framework

Our Proposed 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 stage of 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.

Let B_j $(1 < j < m)$ be the number of base classifiers. The modelling stage under the DLEM generates individual models for each D_i, such that the total number of generated individual models for the MMD is given by $m * n$. A pool of models, PM, with members PM_{ij} representing the individual model fitted using D_i under the base classifier method, B_j, is created.

The third stage selects models from the model pool PM using accuracy and diversity as selection criteria, either individually or jointly. Using these criteria, three different rules, named $R0$, $R1$ and $R2$, are derived as follows.

Figure 3 illustrates the three rules, $R0$, $R1$, $R2$, devised for model selection using various criteria.

R0 This rule only uses *accuracy* as a criterion for model selection. The DLEM firstly computes the accuracy, $(Acc(m_i))$, for each of the n models in the PM and sort them in descending order based on the magnitude of each model's $(Acc(m_i))$. Then the DLEM selects the N most accurate models from PM, i.e., $m_i = max\{Acc(m_j), m_j \in PM\} i = 1...N$, and add them to the ensemble, φ, as shown in Fig 3(a).

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 [7] 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 (Eq. 1) and added to the ensemble, φ. Therefore φ now contains MAM and the $(N - 1)$ most diverse models from PM.

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

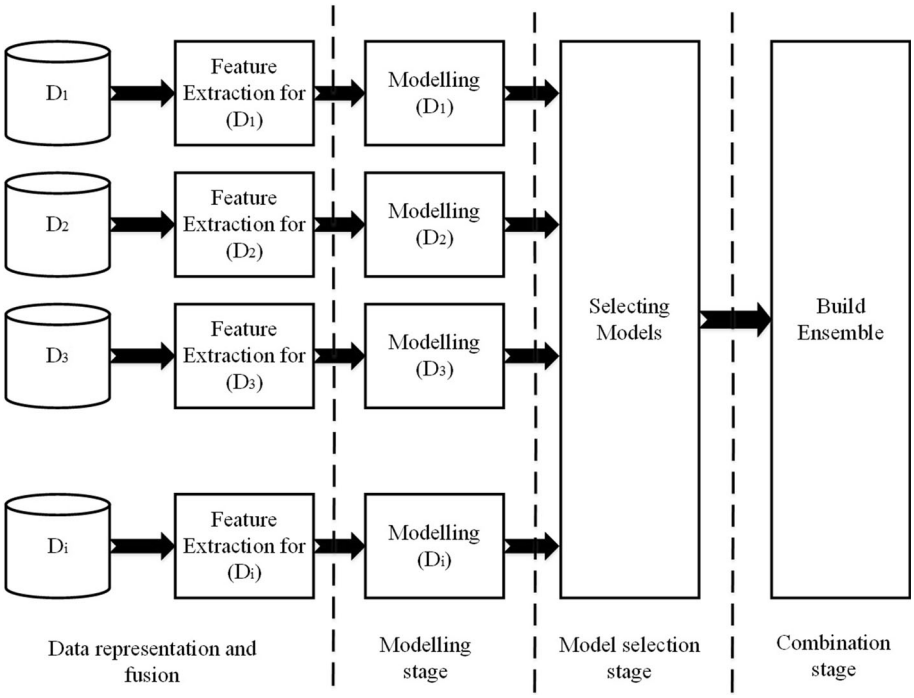


Fig. 2 The general framework for DLEM

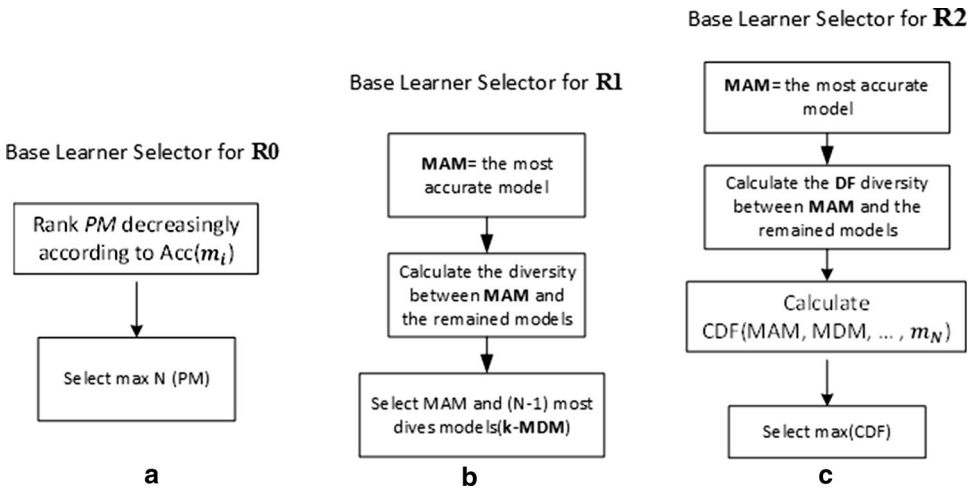


Fig. 3 Main steps for R0, R1 and R2 in HES [1]

R2 This rule uses both *accuracy* and two types of *diversity* measures, namely the *DF* method and the *Coincident Failure Diversity* (CFD) method [14]. 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 $1 \leq i \leq J$ and J is given by $J = \binom{|PM|}{N-2}$. For each ϕ_i, the diversity index of type CFD are computed. The ensemble with the maximum CFD diversity index is then selected as the final ensemble, Φ using $\Phi = max\{CFD(\Phi \Leftarrow m_j), m_j \in PM\}$.

The fourth and final stage combines the decisions from each of the models in the ensemble, using a fusion function. In this experiment, s simple majority method [9] is applied to obtain the final decision of the ensemble.

3.2 Implementation of the DLEM

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

4 Experiment design and results

4.1 Dataset

Our experiment was conducted using a benchmark dataset, which called 8—Scene Categories dataset [12]. It consists of 2688 instances and introduced by two parts whose media are not the same. In the first part, XML files were used to represent image's notation. Within each of the XML files there are tags which treated as a text sub-dataset D_t. In the second part, the actual images were represented to produce the images sub-dataset D_g. The numbers of features obtained were 782 features from D_t and 567 features from D_g. The dataset was categorised by eight classes.

Ten base classifier were learned from the textual and the imagery features subsets, which gave twenty heterogeneous models in total. This gave the DLEM 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 DLEM, using three selection rules separately, on the multimedia data. The investigated issues included (1) the performance measures and classifier selection criteria. These were represented by the three rules: R0, R1 and R2, and (2) the variation of the ensemble size from 3, 5, 7 to 19. A total of 135 experiments were conducted. This involved running all possible combination of these parameters. Repetition was undertaken for five different runs.

In parallel, we conducted the a series of experiments aiming to investigate the influence of the CFD on both R0 and R1. The main reason is that R2 produced the best performance, and it is using the CFD measure to select the ensemble models.

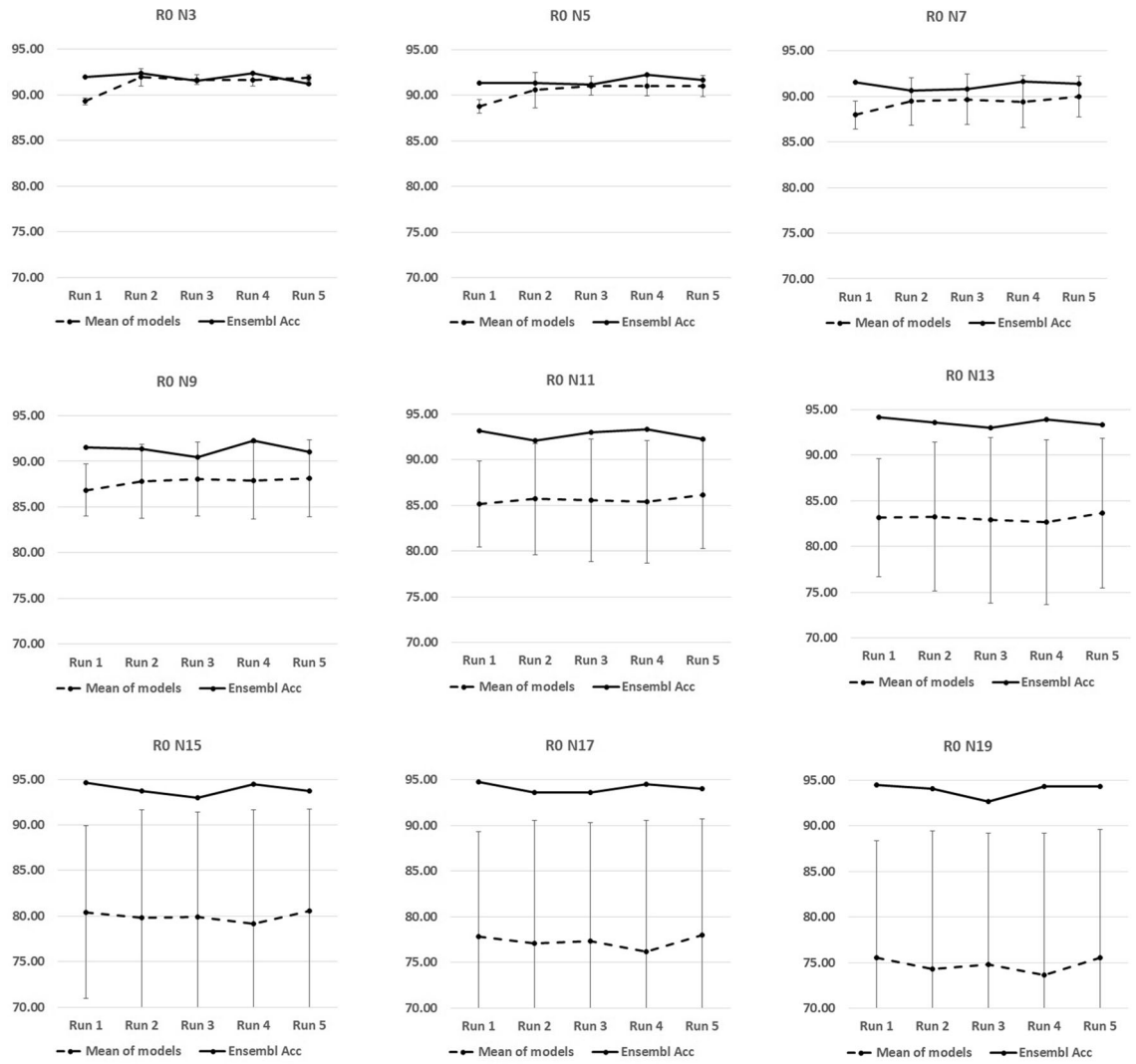


Fig. 4 Shows the DLEM results obtained using the rule R0, and ensemble sizes 3, 5, 19 are presented in the sub-graphs. The red lines represent DLEMs accuracy and the blue lines show the mean accuracy for models selected for DLEM. The vertical lines show standard deviation over 5 runs

Figures 4, 5 and 6 present some of the results (means and standard deviations). These figures show clearly that the DLEMs built with the three rules are, on average, generally superior to individual classifiers. The reason for this is that the mean accuracies (shown in red lines on the figures) of the DLEMs are approximately 10% higher than the mean accuracies (illustrated by blue lines) of the individual classifiers in the DLEMs. Also, it had already been demonstrated, in our earlier work, that our ensemble results have a higher level of accuracy overall than the best individual model, the MAM. Hence, our DLEM had the best reliability overall because the reliability of a MAM was not consistent over a succession of runs. In one run, the current MAM could perform best, but in others it might perform much worse. It is the consistency of its accuracy level which gives our method, the DLEM, its advantage.

Figure 7 compares the results of DLEMs 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 had indicated that there were accuracy issues with this rule. However, these became much more apparent in the current piece of work due to the high number of models which were tried. The increase in model numbers highlighted very clearly the disadvantages of R1. R0 performed reasonably well because it combined all the models in PM, which have the best accuracies. However, in comparison with R2, its results are not as steady and consistent. As can be seen, its accuracy level only improved at N9. It therefore does not have the required level of consistency.

Figure 8 shows the average results of CFD in the ensembles built with R0, R1 and R3, although the CFD is not used in R0 and R1. The purpose is to see if the CFD can

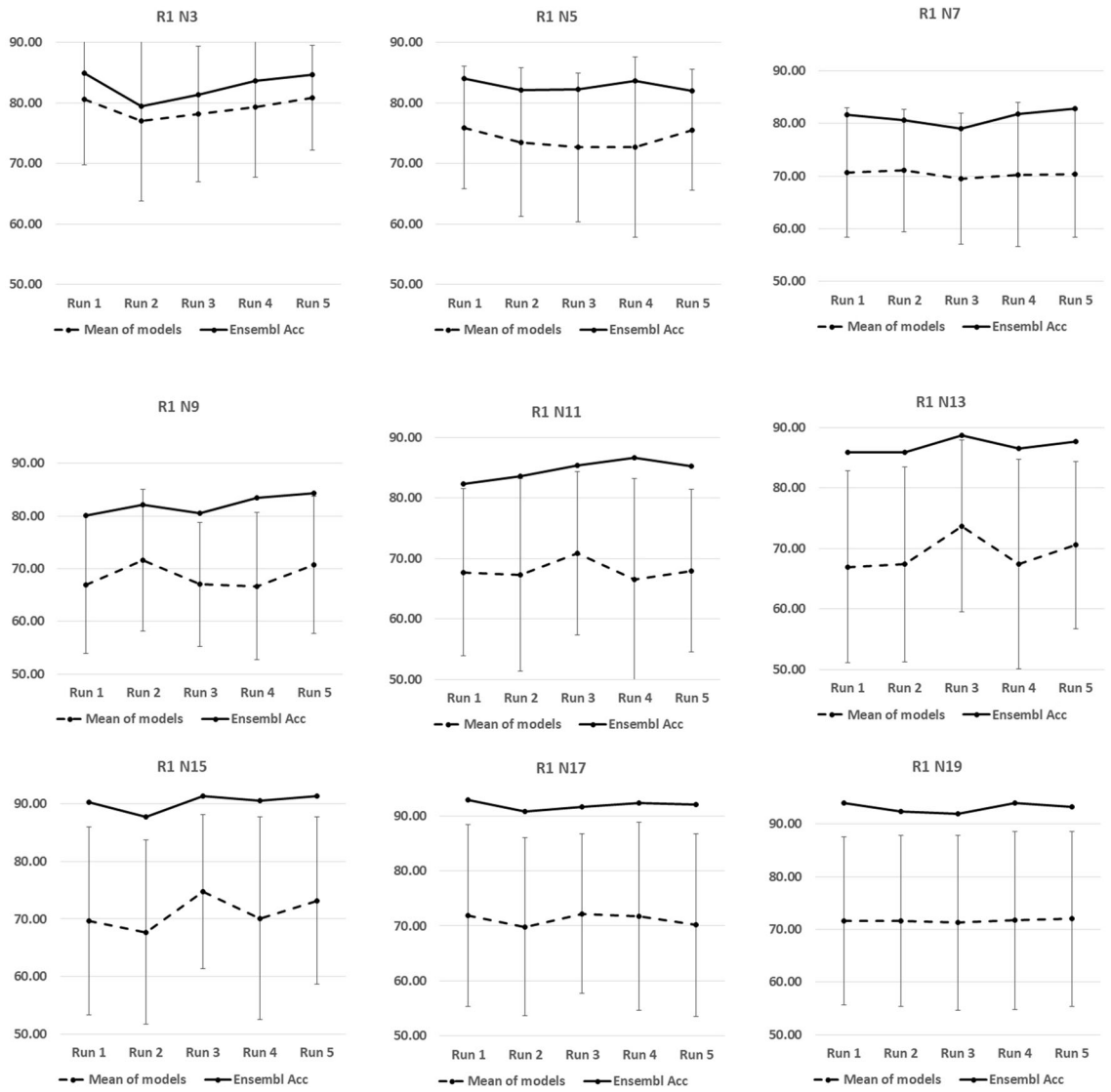


Fig. 5 Shows the DLEM results obtained using the rule R1, and ensemble sizes 3, 5, 19 are presented in the sub-graphs. The red lines represent DLEMs accuracy and the blue lines show the mean accuracy for models selected for DLEM. The vertical lines show standard deviation over 5 runs

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 N11. When we link this result with the accuracy level for R0 shown in Fig. 3, we can see that the best ensemble results were gained when we combined models that have best accuracy and CFD, which are N11 to N19.

Thus, it can be concluded that the ensemble with model selection criteria using a combination of CFD and DF and accuracy measures (R2), gives the best results. These are superior to those results obtained using either pair-wise diversity (R1) or just accuracy (R0). It is also important to take into account both diversity and accuracy when building large ensembles for classification. Furthermore, there is nothing to be gained from using R1 in any future work and we will therefore not use it again.

4.3 Comparison

The results were compared with the feature-level ensemble method (FLEM) and various heterogeneous ensembles based on the single media data, text (HEST) and image data (HESG). The full comparative results between the FLEM and the HESG were published in [2] and the full results for the HEST were published in [1]. Figure 9 shows the critical difference diagram for the DLEM, the FLEM, the HEST and the HESG, for all three rules R0, R1 and R2. The DLEM-R2 is the best on average for all five runs. An immediate question was raised as to why it is the best. That

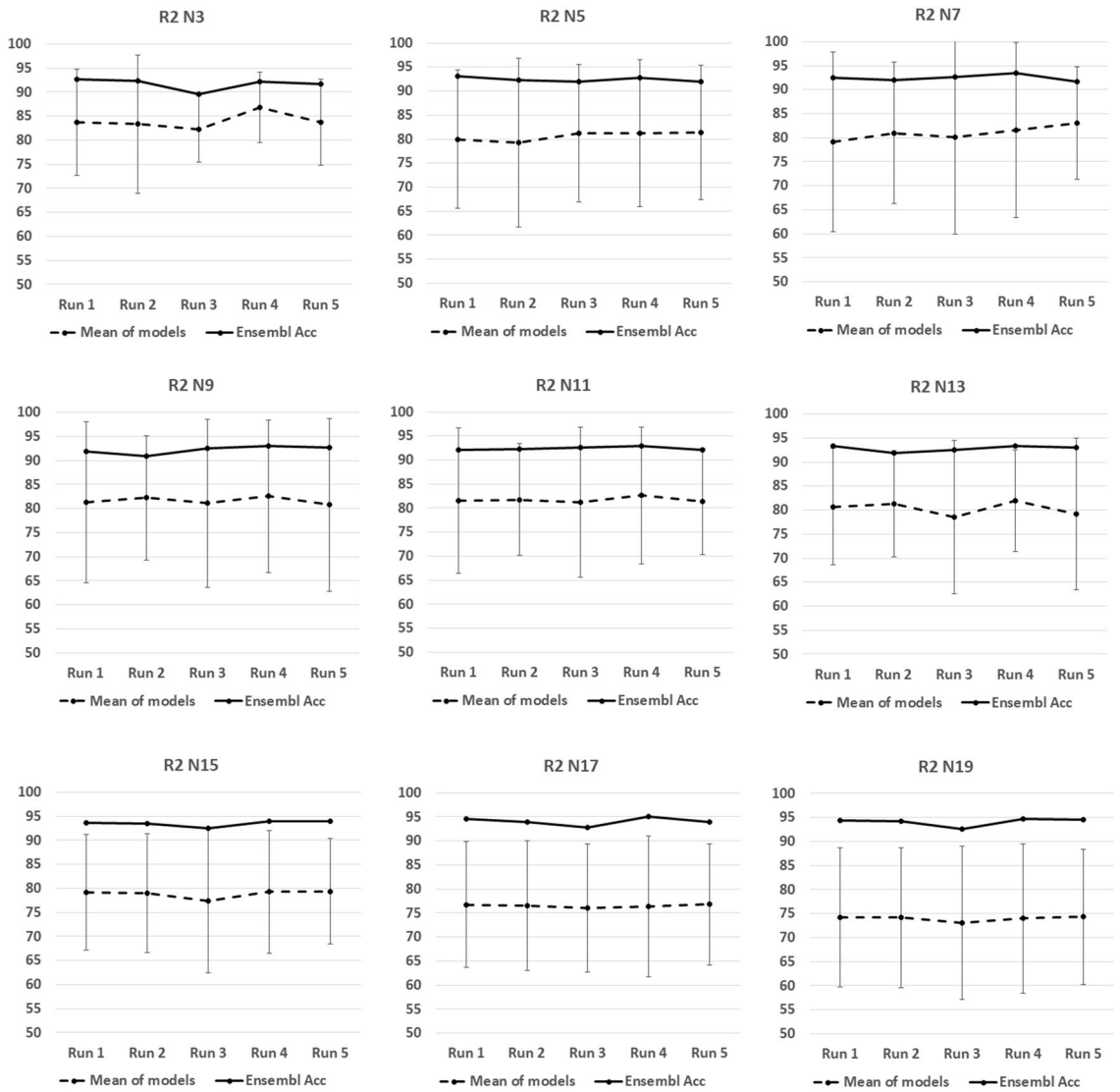


Fig. 6 Shows the DLEM results obtained using the rule R2, and ensemble sizes 3, 5, 19 are presented in the sub-graphs. The red lines represent DLEMs accuracy and the blue lines show the mean accuracy for models selected for DLEM. The vertical lines show standard deviation over 5 runs

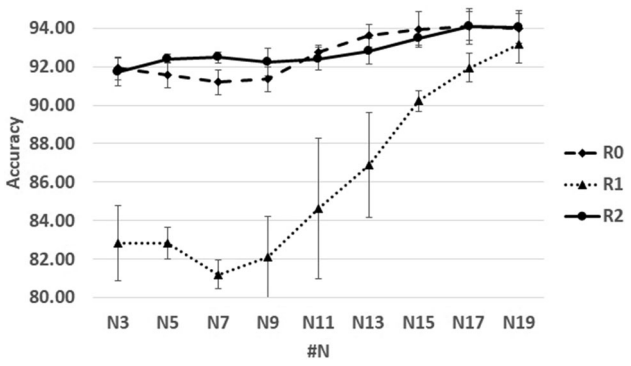


Fig. 7 Comparing all three rules in nine different sizes of the DLEM

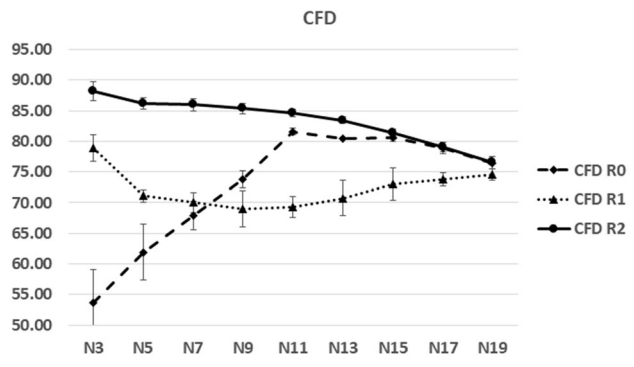


Fig. 8 Comparing CFD for all three rules in nine different sizes of the DLEM

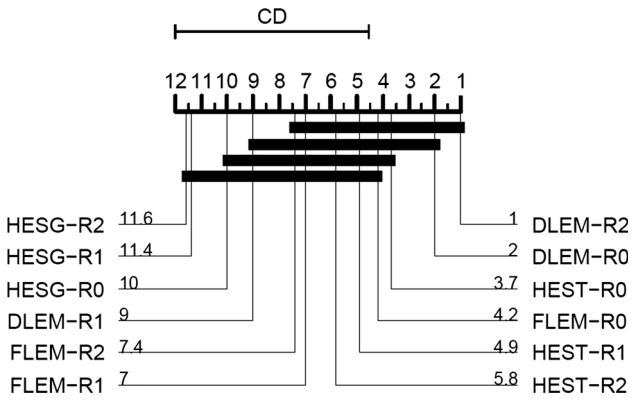


Fig. 9 Critical difference diagram for DLEM, FLEM, HEST and HESG for all three ruls R0, R1 and R2

led us to investigate further, to examine the effect that the CFD has on the ensemble.

5 Conclusion and future work

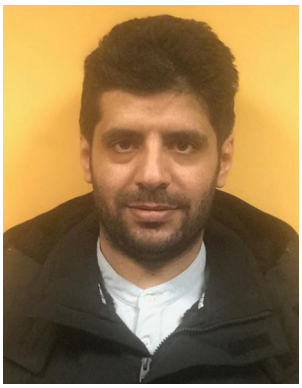
In this study, we developed a heterogeneous ensemble method to classify multi-media datasets at the decision level (DLEM) aiming to achieve the best and most reliable accuracy results. Our DLEM consists of four stages: extracting features from multi-media subsets, modelling the subsets datasets, selecting models with different rules based on various criteria,and building heterogeneous ensembles. There are some observable outcomes from our results. Firstly, the ensemble results obtained by the DLEM are better than all the results that we have already obtained by FLME, HEST and HESG in the same dataset. Secondly, the best ensemble results for classifying multi-media datasets could be obtained by combining models with higher accuracies and CFDs. Thirdly, R1 is not useful for classifying multi-media datasets especially when the number of modes are high.

Suggestions for future work to improve our approach include, firstly, creating other different rules for the model's selection by giving more attention to the CFD. Secondly, applying this approach on different classification problems like time serious classification. Thirdly, more experiments will be conducted by using more multi-media datasets.

References

1. Alyahyan, S., Farrash, M., & Wang, W. (2016). Heterogeneous ensemble for imaginary scene classification. In *Proceedings of the 8th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K 2016): KDIR, Porto—Portugal*, November 9–11, 2016. (Vol. 1, pp. 197–204). https://doi.org/10.5220/0006037101970204.
2. Alyahyan, S., & Wang, W. (2017). Feature level ensemble method for classifying multi-media data. In: *International conference on innovative techniques and applications of artificial intelligence* (pp. 235–249). Springer.
3. Bagnall, A., Davis, L., Hills, J., & Lines, J. (2012). Transformation based ensembles for time series classification. In *Proceedings of the 2012 SIAM international conference on data mining* (pp. 307–318). SIAM.
4. Bagnall, A., Lines, J., Hills, J., & Bostrom, A. (2015). Time-series classification with cote: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *27*(9), 2522–2535.
5. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209.
6. Do, C. T., Douzal-Chouakria, A., Marié, S., Rombaut, M., & Varasteh, S. (2017). Multi-modal and multi-scale temporal metric learning for a robust time series nearest neighbors classification. *Information Sciences*, *418*, 272–285.
7. Giacinto, G., & Roli, F. (2001). Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, *19*(9), 699–707.
8. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, *37*, 132–156.
9. Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. London: Wiley.
10. Mojahed, A., Bettencourt-Silva, J. H., Wang, W., & de la Iglesia, B. (2015). 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.
11. Mojahed, A., & de la Iglesia, B. (2017). An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach. *Knowledge and Information Systems*, *50*(1), 27–52.
12. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.
13. Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28–47. https://doi.org/10.1177/0165551516677911.
14. Partridge, D., & Krzanowski, W. (1997). Software diversity: Practical statistics for its measurement and exploitation. *Information and Software Technology*, *39*(10), 707–717.
15. Wang, W. (2008). Some fundamental issues in ensemble methods. In *IEEE international joint conference on neural networks, 2008. IJCNN 2008. IEEE world congress on computational intelligence* (pp. 2243–2250). IEEE.
16. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.
17. Wlodarczak, P., Soar, J., & Ally, M. (2015). Multimedia data mining using deep learning (pp. 190–196).
18. Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, *16*, 3–17.

19. Yamanishi, K. (1999). Distributed cooperative bayesian learning strategies. *Information and Computation*, *150*(1), 22–56.
20. Zhu, W., Cui, P., Wang, Z., & Hua, G. (2015). Multimedia big data computing. *IEEE Multimedia*, *22*(3), 96-c3.



Saleh Alyahyan received his B.S (2006) from King Faisal University, Saudi Arabia and MIT (2011) from the University of Newcastle, Australia. He is expected to finish his PhD in Data Mining and Machine Learning from the University of East Anglia by the end of 2019. He joined Shagra University scenes 2008 and they provide him with Master and PhD scholarships.

Wenjia Wang received his BEng (1982) and MEng (1985) degrees from NEU (North-Eastern University, China) in Automatic Control Engineering, and PhD degree in Advanced Computing in 1996 from the University of Manchester Institute of Science and Technology (UMIST), UK. He joined the School of Computing Sciences, University of East Anglia, as a senior lecturer in 2002. He develops research in the areas of data mining/knowledge discovery, ensemble approach and artificial intelligence. He has been the Director of MSc in Computer Science since 2003.