HIERARCHICAL CLASSIFICATION FUSION FRAMEWORK

Asmar A. Khan, Costas Xydeas, Hassan Ahmed

School of Computing and Communications
Infolab21, Lancaster University, LA1 4WA
United Kingdom

ABSTRACT

This paper presents a novel hierarchical Classification Fusion (CF) framework which operates on Abstract and Measurement levels simultaneously and thus exploits information patterns resulting from the output labels and posterior beliefs of individual classifiers. Furthermore the proposed classification fusion methodology allows for the decomposition of the input data, which is used to design individual classifiers, into subsets. This in turn permits individual classifiers to be re-designed per subset and in a manner that increases overall system classification performance. Experimental results are presented which demonstrate the potential of the proposed methodology in the case of multi-modal, multi-feature binary data classification problems. In addition the proposed CF design framework can be applied to multi class problems and is independent of the type of classifiers employed in the system.

Index Terms— Ensemble Methods, Classification Fusion

1. INTRODUCTION

Most of real life Data Classification (DC) applications involve multidimensional input feature spaces which in turn are characterized by high degrees of variability. In such difficult cases and while well-trained single classifiers attempt to learn and thus conform to input data statistical characteristics, their performance is often less than desirable. In response to the above situation, Ensemble Methods [1] have been developed and frequently used in machine learning applications where they have clearly out-performed conventional DC techniques based on single classifiers [2]. These single classifiers, named as "Base" Classifiers (BC), are normally combined by a meta-learner, which in turn could be itself a classifier operating on the outputs of Base Classifiers. In general, EM type of classification systems can be grouped in two major categories. The first type enhances DC system performance by applying algebraic or fuzzy formulations [3, 4]. The second type relies on often iterative training procedures which are designed to further exploit input data distribution characteristics. Thus, Ensemble Methods operate on more than one Base Classifiers (BC) and aim at improving overall system performance by either "Fusing" results from individual classifiers or improving DC performance by using "iterative" learning of input data characteristics. Furthermore and as the name suggests, Classification Fusion (CF) attempts to "combine" Base Classifier output information and by doing so outperform individual classifiers. In principle, CF recognizes that even a relatively weak classifier might outperform an overall better classifier for some particular input instances and can therefore enhance overall system performance. Naturally, the way information fusion is performed is of key importance in CF. In general, and within the context of CF data classification, information fusion can be performed in different domains i.e. at:

1. Abstract Level (AL) where each BC output is represented by a classification label and where a label fusion scheme follows e.g. majority rule, AND/OR type of logic. For example, Kittler in [3] explored in detail CF with fixed label combination rules of the majority, mean and median type. Logic based rules can also be found in [5].

2. Rank Level (RL) where each Base Classifier provides a ranked list of labels with the top element in the list being the 1st choice. Given these prioritized lists, Label fusion techniques like the Borda Count [6] are then applied to produce a final output.

3. Measurement Level (ML) where each classifier provides a measure of confidence associated with each class. Examples of such CF systems are the Bayesian [3] and Posterior Probability Estimation schemes [4]. Recently a classifier that is developed via a non-Bayesian framework has been reported [7].

Other important types of Ensemble Methods deal with the process of learning the statistical characteristics of the input data and iteratively train relatively simple classifiers over time. Notable examples are Bagging Predictors [8], Boosting [9, 10] and a number of their variations. Boosting techniques are particularly successful in maximizing classification performance and while using iterative learning, employ different votes/weights in exclusive subsets of input data. Finally the recent modeling of input data work in [11] is noted as an in-

teresting way of building complex classifiers by combining simple hypotheses.

2. PROPOSED AL/ML-CF FRAMEWORK

A number of classification fusion (CF) techniques operate on Abstract Level (AL-CF), by fusing class labels [3], or on Measurement Level (ML-CF) where posterior probabilities of individual classifiers are combined [4]. The proposed AL/ML-CF methodology represents an amalgamation of these two levels in general, where input data labeling, obtained through conventional classification, is improved by an additional meta-learner fusion process that operates on i) classification labels and ii) the posterior probability outputs of these classifiers. Furthermore this methodology allows for the decomposition of the input data into subsets, which are then used to train corresponding individual classifiers. This effectively permits individual classifiers to be re-trained per subset and in a manner that increases overall system classification performance.

Thus in this paper a novel CF design framework is presented and several AL/ML-CF schemes are proposed and shown to improve upon the classification performance of Baseline (B/CF) systems and more importantly state of the art Ensemble methodologies. B1/CF, a first baseline scheme that is considered here, is based on Meta-Learning, where output posterior class probabilities from n conventional Base Classifiers are fused using a Neural Network (NN) as the Fusion element. Figure 1 shows for simplicity an example of a binary B1/CF classification system with $n = 3$. Here input instances x_i where $i = \{1, ..R\}$ need to be classified by the system in one of two classes, say w_j where $j \in \{0, 1\}$. Thus given x_i and also assuming for example that x_i belongs to class w_j, each of the three Base Classifiers C_1, C_2, and C_3 provides corresponding posterior probabilities $P_k(w_0|x_i)$ and $P_k(w_1|x_i)$, $k = 1, 2, 3$. These are the inputs to the following NN whose task is to maximize the probability $P_k(w_j|x_i)$. The DC performance of this framework relates to that of the method of Stacking [12] which effectively selects the best i.e. maximum posterior probability classifier.

The development of the first of the proposed systems, i.e. AL/ML-CF1, has been motivated by the decomposition of the input data into N subsets in general and according to the classification label patterns obtained from the n conventional classifiers, in particular. Thus for n classifiers and m class input data, there are $N = m^n$ mutually exclusive subsets of the input data set $G = S\{x\}$, with each subset $G_p = S_p\{x\}$, $p = 1, 2, , N$ represented by an n long m-digit pattern like $[w_j^1 ... w_j^n]$ and $w_j \in \{0, 1, ..m\}$. Thus each subset $S_p\{x\}$ of input data is consistently characterized by a specific m-digit classification pattern. It is this classification behavior observed across BCs that is common per subset and it is therefore used to decompose the input data space into smaller sets. The general idea of decomposing input data into groups first

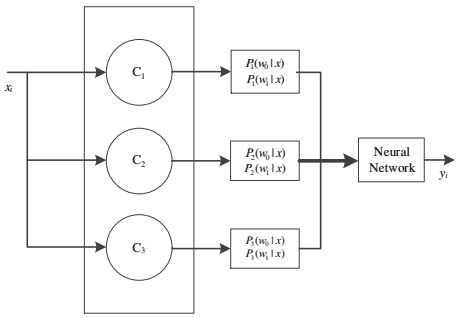

Fig. 1. Block diagram of B1/CF using $n = 3$ conventional classifiers, C_1, C_2, and C_3 , followed by NN type of Meta-Learning fusion. Input data x belongs to two classes i or j. Overall system output y indicates class i or j

appeared in the Behavior Knowledge Space (BKS) models scheme of [13]. However, AL/ML-CF operates in a much more detailed treatment of group populations and also in an "iterative" training way (as it will be showed later in the section), rather than just making a one-off decision. Thus in the case of $n = 3$ classifiers and $m = 2$ classes, the proposed data grouping enables AL/ML-CF to apply $2^3 = 8$ different meta-learners (NNs) on input data samples belonging to 8 different subsets, see top part of figure 2.

Furthermore this input data decomposition approach can be extended by applying again the n classifiers, this time on each of the N previously defined G_p subsets. The resulting AL/ML-CF2 system is now constructed with further $2^3 \times n$ classifiers and $2^3 \times 2^3$ NNs attached to corresponding input data subsets.

Note that these additional 2^{nd} stage BCs are trained on different sample spaces and therefore have different characteristics, as compared to the initial BCs. Once the 2^{nd} stage BCs are trained they are stored and later applied on testing data samples coming from a particular G_p subset, as specified by the output pattern of the initial base classifiers. Thus during the testing phase, initial BCs provide an n long m-digit pattern that acts as the address of the appropriate 2^{nd} stage set of BCs. These are then applied to classify again the same input data sample and their outputs are "fused" using a corresponding NN, see bottom part of figure 2. Now, System AL/ML-CF2 can be extended, very much in the same way as described above, to yield AL/ML-CF3. In this case the system's hierarchical structure includes a 3^{rd} stage with further $2^3 \times 2^3 \times n$ classifiers and $2^3 \times 2^3 \times 2^3$ data subsets and corresponding NNs. Obviously system design can be generalized for k stages to yield AL/ML-CFk. The adopted hierarchical data decomposition process which enables the application of even more specialized classifiers, has a major limitation. That is, with k increasing, system complexity increases exponentially.

Therefore, the main question to be asked at this point relates to how an appropriate trade-off between complexity and

classification performance can be found. Now, system complexity relates to the exponential growth of subsets, which is allowed across stages in the above defined "Uniform Tree", or Full type of Expansion of input data subsets.

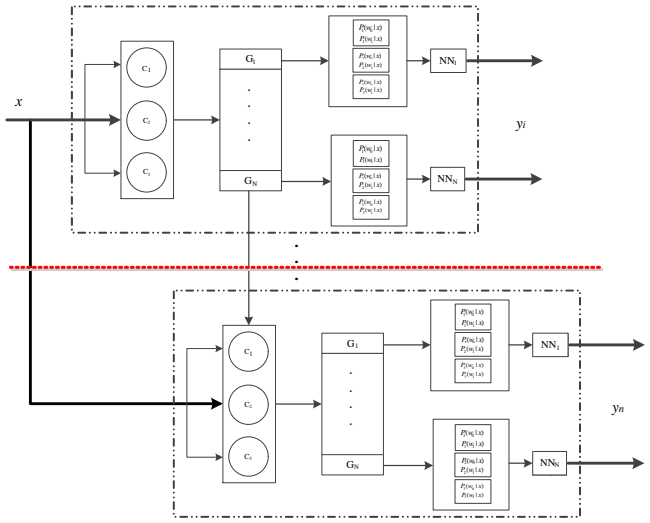

Fig. 2. Top part of diagram depicts AL/ML-CF1. Whole diagram without the NNs of top part represents AL/ML-CF2.

2.1. Hierarchical Non-Uniform Tree

This implies that a "Non-Uniform tree" or a Partial Expansion in the number of input data subsets can potentially limit system complexity. Furthermore, if this is linked to system DC performance, appropriate complexity versus performance trade-offs can be determined and system performance is maximized. Thus a system design methodology AL/ML-CFk* has been developed and applied, which incorporates DC performance based decisions to achieve partial expansion into subsets. In particular, during training a "complete" hierarchical structure is produced that covers full expansion up to a given number of levels, say k, and also includes the design of Base Classifiers for each data subset defined in the structure. Testing produces classification results for every subset in this general structure. These are collected and subsequently used to determine an appropriate partially expanded tree of input data decomposition. Currently the system is designed so that a specific "parent" subset is further divided only to those "children" subsets (i.e. expanded) which have better classification performance than the parent. Of course, one could consider here a "*how much better classification*" type of system parameter, which will effectively control the sparsity of the tree. The more demanding is the classification gain required to be achieved by children subsets, the more sparse is the tree and therefore the less complex will be the resulting AL/ML-CFk system.

Table 1. Classification Rate of existing CF schemes with the proposed AL/ML-CF schemes on two different datasets

Classification Fusion	MAGIC %	MiniNooBE %
Majority	73.92	89.83
Sum	73.0	89.76
Product	63.72	85.38
B1/CF	81.23	90.28
Bayesian	74.12	90.54
BKS	77.54	89.72
AdaBoost	81.55	92.66
AL/ML-CF1	81.49	91.07
AL/ML-CF2*	82.89	94.36

3. EXPERIMENTAL RESULTS

Systems were computer simulated and their performance was tested using two different large input datasets. One of the datasets is MiniBooNE particle identification [14] and the other is data simulating registration of gamma and hadron particles in Cherenkov imaging gamma ray telescope MAGIC [15]. The MiniBooNE dataset is composed of $130,065$ samples distinguishing electron neutrinos (signal) from muon neutrinos (background). Each example consists of 50 features. There are a total of $19,020$ samples in the MAGIC gamma telescope dataset representing two classes of gamma (signals) and hadron (backgrounds) with each sample having 10 attributes. Furthermore system performance was determined using a 10-fold cross validation process.

The proposed AL/ML-CF methodology was tested using the following three conventional Base Classifiers, i) Linear Discriminant Analysis, ii) Nave-Bayesian and iii) Gaussian Mixture Model. Figure 3 presents the % improvement of AL/ML-CF2* over AL/ML-CF1 and several other existing techniques. These include B1/CF systems, majority rule fusion, sum and product rule based fusion [16]. In addition three more advanced CF schemes i.e. Bayesian [4], Behavior Knowledge Space (BKS) [13] and Adaboost [9] are also included. This figure clearly shows that proposed AL/ML-CF2* system outperforms all other conventional fusion methodologies including the popular, iteratively learning based, Adaboost scheme. Table 1 shows the % classification rates performance of all schemes. Note that in these experiments AL/ML-CF2* operated with as much as 90% reduction at stage 2 expansion.

Furthermore AL/ML-CF2* is also compared with BKS and AdaBoost in terms of Receiver Operating Characteristics (ROC) curves see figure 4. Again ROC results are indicative of the advantage offered by the proposed AL/ML-CF framework, as compared to two conventional but powerful CF techniques.

4. DISCUSSION

The proposed CF methodology improves classification performance as compared to existing and commonly used classification fusion based DC schemes. Furthermore the AL/ML-CF framework is not limited to specific types of Base Classifiers or Meta-Learners (Fusion). It models and selectively exploits the information present in input data, by operating on data subsets, which are determined by the classification behavior of base Classifiers. To this extend AL/ML-CF can be thought as belonging to the same, general data modeling CF category as Bayesian Modeling Averaging [11], where input statistical characteristics are estimated by a non-linear, non-parametric fusion combiner. Note that the proposed framework is generic, modular and is based on independent classification parts. These can be replaced by any type of classifier. Also note that NNs operating on BC outputs can be replaced by other meta-learner, like Bayesian, SVM or AdaBoost. However experimentation has also shown that, within the proposed framework, Neural Network meta-learners outperformed both naive-Bayesian and SVM.

Finally computer simulation based experimental results, obtained with CF systems operating on two different types of input databases, show that AL/ML-CF2* can deliver improved classification performance of the order of $5-30\%$, when compared to conventional baseline Fusion classifiers and up to 2 to 4% when compared to the AdaBoost and the Bayesian combiner(fusion) systems. As a final point of discussion we raise the novelty of AL/ML-CF, which stems from the notion that input statistical characteristics can be modeled more accurately through the behavior of Base Classifiers.

5. CONCLUSION

In this paper a novel DC fusion framework (AL/ML-CF) is presented, which operates on abstract and measurement levels simultaneously and exploits the information resulting from classification labels and posterior beliefs. Experimental results refer to proposed DC hierarchical structures populated with multi-feature classifiers and meta-learners in a manner that improves binary classification performance. Two popular input data bases have been employed in these experiments i.e. MiniBooNE particle identification and MAGIC. This design framework can easily be applied to multi-class problems. Furthermore, AL/ML-CF is generic and is not restricted to a specific type of classifier at input and fusion levels. Computer simulation results show that classification improvements of the order of 2 to 30 %, are achieved using the above input datasets and when compared to several other well-known CF systems.

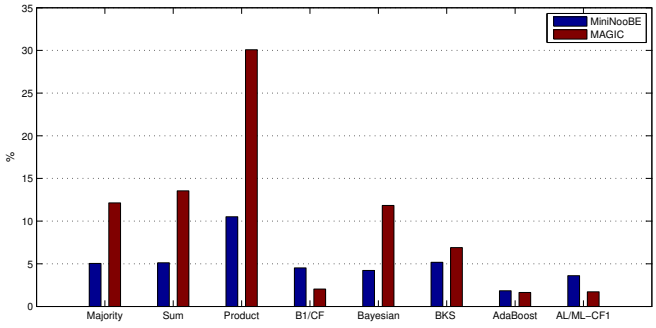

Fig. 3. %age improvement of AL/ML-CF2* with respect to other proposed and existing schemes

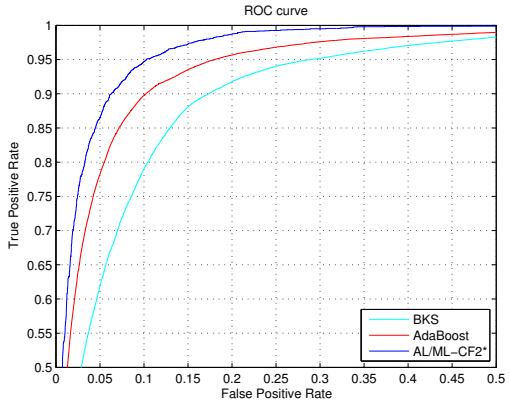

Fig. 4. ROC curve of AL/ML-CF2* compared with BKS and AdaBoost

6. REFERENCES

[1] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[2] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[3] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[4] L.I. Kuncheva, "A theoretical study on six classifier fusion strategies," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 281–286, 2002.

[5] Z. Stejić, Y. Takama, and K. Hirota, "Mathematical aggregation operators in image retrieval: effect on retrieval performance and role in relevance feedback," *Signal processing*, vol. 85, no. 2, pp. 297–324, 2005.

[6] M. Van and S. Erp, "L.: Variants of the borda count method for combining ranked classifier hypotheses," in *in the Seventh International Workshop on Frontiers in Handwriting Recognition*, 2000, pp. 443–452.

[7] O.R. Terrades, E. Valveny, and S. Tabbone, "Optimal classifier fusion in a non-bayesian probabilistic framework," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1630–1644, 2009.

[8] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[9] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*. Springer, 1995, pp. 23–37.

[10] P. Viola and M.J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[11] R. Sicard, T. Artières, and E. Petit, "Learning iteratively a classifier with the bayesian model averaging principle," *Pattern Recognition*, vol. 41, no. 3, pp. 930–938, 2008.

[12] JM Gibbons, GM Cox, ATA Wood, J. Craigon, SJ Ramsden, D. Tarsitano, and NMJ Crout, "Applying bayesian model averaging to mechanistic models: An example and comparison of methods," *Environmental Modelling & Software*, vol. 23, no. 8, pp. 973–985, 2008.

[13] Y.S. Huang and C.Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 1, pp. 90–94, 1995.

[14] AA Aguilar-Arevalo, AO Bazarko, SJ Brice, BC Brown, L. Bugel, J. Cao, L. Coney, JM Conrad, DC Cox, A. Curioni, et al., "Measurement of muon neutrino quasielastic scattering on carbon," *Physical review letters*, vol. 100, no. 3, pp. 32301, 2008.

[15] J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, C. Baixeras, JA Barrio, H. Bartko, D. Bastieri, et al., "Implementation of the random forest method for the imaging atmospheric cherenkov telescope magic," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 588, no. 3, pp. 424–432, 2008.

[16] L. Xu, A. Krzyzak, and C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 3, pp. 418–435, 1992.