Estimating the Serial Combination's Performance from That of Individual Base Classifiers

Gian Luca Marcialis, Luca Didaci, and Fabio Roli

Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{marcialis,didaci,roli}@diee.unica.it
http:\\prag.diee.unica.it\en

Abstract. Although the large number of MCS topics, serial fusion of multiple classifiers has been poorly investigated so far. In this paper, we propose a model which, starting from the performance of individual classifiers and the traditional hypothesis of decision independence given the class, is able to estimate the performance, in terms of error rates, of the whole serial classification scheme. The model is tested on a large set of data sets and classifiers, and the importance of the basis hypothesis is evaluated under different scenarios, which can be in agreement or not with such hypothesis.

1 Introduction

Fusion of multiple classifiers can be performed in parallel or serially [1, 2]. Although the most of works relies on parallel fusion at several levels (feature-level, measurement-level, decision-level) [3, 4], very few papers deal with sequential, or serial, fusion of multiple classifiers [5–8, 16–18].

To the best of our knowledge, the only papers which try to analyze analytically serial fusion is the paper by Pudil et al. [5] and Trapeznikov et al. [18], where the problem is dealt by the point of view of the risk assessment. Other works analyze specifically a specific Unfortunately, Pudil's modelling does not allow to point out any specific pros and cons of serial fusion of multiple classifiers. In other cases, we have practical applications of the serial fusion with an experimental assessment of pros and cons [6, 7]. Theoretical and experimental approaches have been proposed for biometric applications [9–11]. In this paper, as in [9], we deal with the serial fusion of multiple classifiers from the point of view of the performance prediction: in other words, we do not try to find the optimal parameters allowing the minimization of a risk function as in [18], but tries to model the sequential fusion in order to have a prediction of the overall error rate, given the error rates of individual classifiers.

In this preliminary work, we derive the general expression of the error rate of a serial fusion of two classifiers for a two-classes classification problem, under the simple hypothesis of decisions independence among classifiers, and propose a

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 622–631, 2013.
© Springer-Verlag Berlin Heidelberg 2013

model which allows deriving further information on the performance achievable by the serial system with respect to the best individual classifier. The model is then tested on a battery of several UCI data sets, and on classifiers whose set up confirms or not the independence hypothesis, in order to see at which extent the proposed model is valid.

The rest of the paper is as follows. The model is described in Section 2. Section 3 report experiments. Conclusions are drawn in Section 4.

2 The Proposed Model

Fig. 1 describes the serial model used in this paper. Observation x_1, a statistical or structural representation of the pattern at hand, is input to the C_1 classifier, whose output is given in terms of d^1, where $d^1 \in \Omega$, being $\Omega = \{w_1, w_2\}$, the state of nature. On the basis of a reject option [12, 16, 18], we add to Ω a third class, namely, w_0. In this case, the final decision is left to the second classifier, named C_2. Thus, for sake of clarity, three decisions can be taken by the first classifier and we will indicate these with d_0^1, d_1^1, and d_2^1, corresponding to the three classes w_0, w_1, w_2 . In the case that C_1 decides for w_0, observation x_2, related to the same pattern, is taken as input by C_2, whose decision can be d_1^2 or d_2^2. In other words, no reject option is given for the second classifier.

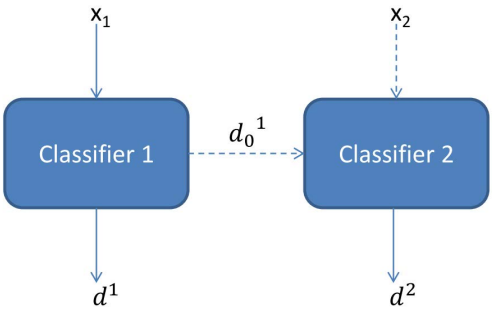

Fig. 1. The proposed model for serial fusion of multiple classifiers

Let d_i^{12} be the decision of the sequence $C_1 \rightarrow C_2$ for class w_i. The probability of error $P(d_i^{12}|w_j)$, $i \neq j$, is given as follows:

$$P(d_i^{12}|w_j) = P(d_i^1|w_j) + P(d_0^1 \bigcap d_i^2|w_j).$$

By hypothesizing the independence of d^1 and d^2, given w_j, we obtain:

$$P(d_i^{12}|w_j) = P(d_i^1|w_j) + P(d_0^1|w_j) \cdot P(d_i^2|w_j). \quad (1)$$

It is easy to see that, given a rejection region for C_1, the error rate is a linear function of the error rate of C_2, where errors of the first classifier represent the bias and the slope.

Aim of this paper is to verify this formula under different practical scenarios. In order to perform a complete investigation, we should perform experiments on all possible rejection regions.

However, in this paper, we will focus only on a particular rejection region, for sake of simplicity. The choice is not arbitrary, because this rejection region is acceptable in many cases.

We investigated the rejection region such that no classification errors are done by C_1. In other words:

$$P(d_i^1|w_j) = 0 \tag{2}$$

Actually, several applications, especially related to the information security, require that, on a chain of possible classifiers, no errors are made by the stages preceding the last one. For example, this occurs in personal identity verification through biometric systems, where rejecting an authorized user at first stage, as well as accepting an impostor at the intermediate stages of a serial system, is a serious drawback [9, 11].

Accordingly, the bias of the error rate indicated in Eq. 3 can be removed, thus obtaining:

$$P(d_i^{12}|w_j) = P(d_0^1|w_j) \cdot P(d_i^2|w_j). \tag{3}$$

This simple expressions leads us to an important, but general, property of serial fusion of multiple classifiers: if $P(d_i^2|w_j) < P(d_i^1|w_j)$, then the sequence $C_1 \rightarrow C_2$ allows to obtain a better performance than that of the best individual classifier, namely, C_2. This is evident from Eq. 3.

Thus, the *best* sequence is that which considers the best classifier at the second stage. It is worth to point out that this is not the "optimal" sequence in terms of error rate minimization, but only with respect to the best individual classifier. On the other hand, nothing can be said about the sequence $C_2 \rightarrow C_1$, under the hypothesis above.

In order to set the rejection region, we opt for a simple strategy suggested in [13], which is an extension of the so-called Chow's rule for rejection option [12]. In particular, this rule takes into account that in practical applications, individual classifiers do not provide the "real" posterior probability of each class but an estimation. In this case, Ref. [13] showed that this rule is able to derive the best trade-off between error rate and rejection rate with respect to the Chow's rule.

On the basis of the above observation, let's consider the output of the classifier C_k, indicated with p_i^k, as an estimation of the probability of w_i, given the pattern. Since we are considering only two-class classification problems, $p_2^k = 1 - p_1^k$.

According to the Chow's rule, we can set a rejection region as function of the interval $[\gamma_2^1, \gamma_1^1]$, and assigning decisions d_0^1, d_1^1, and d_2^1 as follows:

1. d_1^1 if $p_1^1 > \gamma_1^1$;
2. d_2^1 if $p_1^1 < \gamma_2^1 \rightarrow p_2^1 > 1 - \gamma_2^1$;
3. d_0^1 if $\gamma_2^1 \leqslant p_1^1 \leqslant \gamma_1^1$.

Worth noting, $\gamma_1^1 \in (0.5, 1]$ and $\gamma_2^1 \in [0, 0.5)$. In the case that $\gamma_1^1 = \gamma_2^1 = 0.5$, the second stage classifier is required only for the uncertainty value of posterior probability such that $p_1^1 = p_2^1 = 0.5$.

With regard to C_2:

1. d_1^2 if $p_1^2 > \gamma^2$;
2. d_2^2 if $p_1^2 \leqslant \gamma^2$;

Where $\gamma^2 \in [0, 1]$.

Thus, Eq. 3 becomes:

$$P(d_1^{12}|w_2) = P(\gamma_2^1 \leqslant p_1^1 \leqslant \gamma_1^1|w_2) \cdot P(p_1^2 > \gamma^2|w_2) \tag{4}$$
$$P(d_2^{12}|w_1) = P(\gamma_2^1 \leqslant p_1^1 \leqslant \gamma_1^1|w_1) \cdot P(p_1^2 \leqslant \gamma^2|w_1) \tag{5}$$

According to Eq. 2, Eqs. 4-5 can be simplified as:

$$P(d_1^{12}|w_2) = P(p_1^1 > \gamma_2^1|w_2) \cdot P(p_1^2 > \gamma^2|w_2) \tag{6}$$
$$P(d_2^{12}|w_1] = P(p_1^1 < \gamma_1^1|w_1) \cdot P(p_1^2 \leqslant \gamma^2|w_1) \tag{7}$$

This is the final scheme we followed in our experiments. This model allows to predict the performance of the serial system by starting from errors of the individual classifiers, under the assumption of decisions independence. In other words, a designer can draw the error rate curves in order to find the best sequence on the basis of simple evaluations derived from the classifiers taken individually. No complex estimation of joint probabilities is necessary, as well as further experiments on the overall system.

3 Experimental Results

Aim of this Section is to validate the model given by Eqs. 6-7, by assessing the prediction against the real error rate obtained by experiments.

3.1 Data Sets and Protocol

The experiments have been carried out on 32 data sets form the UCI repository [14] (see Table 1). The main assumption of our model is the conditional independence (given the class) between set of features used to represents the classification task at hand, as illustrated in Section 2. In order to create a setting that meets this assumption, we subdivide the original feature space into two views using the random-restart hill climbing method proposed in [15], which aims

at maximizing their independence. Basically, this method splits features set into two views in a random manner, then each feature is switched to the other view once a time, thus obtaining a group of new generated split. All these new splits are evaluated to check their independence, and the one that yields maximum independence is selected. In order to check the robustness of the proposed method when the independence assumption doesnt hold, we performed a second series of experiments in which datasets have been subdivided into two views in a random manner, each view being composed of half of the features, without forcing the independence between views.

In Table 1 we report the characteristics of the data sets used in the experiments. Columns #patterns indicate the number of patterns (total, class A and class B) whilst columns #features indicate the number of features for each view

The ample battery of data sets explored in this paper shows several and different experimental and practical conditions which may affect the design of multiple classifiers systems combined serially. First of all, we may see that some data sets exhibit a strongly unbalanced number of patterns for the involved classes (see for example, Audiology and Solar flare data sets). In other cases, the number of patterns is not balanced with respect to the size of feature set (see for example, Sonar and Breast Cancer data sets). Finally, we may see unbalanced performances among classifiers where the sizes of each feature set strongly differs. Therefore, the experimental results we report will show the reliability of our model as it may be expected in general.

Three different classifiers are adopted for exploring the generality of our observations: K-Nearest Neighbour, Naive Bayes, and Linear Log. We have arbitrarily chosen parameters for the classifiers (k=3 for the K-NN classifier and N=10 for the Nave Bayes classifier) because the aim of the paper is to calculate the final performance of the system from the performance of the individual classifiers; it is not essential to maximize the performance of individual classifiers.

Data sets were splitted training and test sets. The former is made of 70% of available patterns, whilst the latter of the remaining 30%.

3.2 Results

In Figs. 2 we report some ROC curves predicted on the test set by the model according to Eqs. 4-5, and the ones reported on the same data but during system's operations. The thresholds γ have been tuned on the training set in order to obtain zero error for the classifier at the first stage.

Four cases were chosen for showing the effectiveness of our model:

1. imbalanced distributions of patterns over classes and maximum independence among views;
2. imbalanced distributions of patterns over classes and no independence among views;
3. balanced distributions of patterns over classes and maximum independence among views;
4. balanced distributions of patterns over classes and no independence among views.

Table 1. Characteristics of UCI data sets adopted for experiments in terms of number of patterns for the two-class classification problem, and number of features per view (classifier)

Dataset	#patterns			#features - maxInd split		#features random split	
	tot	class A	class B	view 1	view 2	view 1	view 2
1-Audiology	200	48	152	31	24	28	27
2-Automobile	193	130	63	22	2	12	12
3-Congressional Voting Records	232	124	108	15	1	8	8
4-Contraceptive Method Choice	1473	629	844	2	7	5	4
5-Credit Approval	653	296	357	9	6	8	7
6-Dermatology	366	112	254	17	16	17	16
7-Ecoli	336	143	193	3	4	4	3
8-Flag	194	134	60	15	13	14	14
9-Glass identification	214	138	76	1	8	5	4
10-Heart statlog	270	150	120	4	9	7	6
11-Horse-colic	368	232	136	4	1	3	2
12-Ionosphere	351	225	126	6	27	17	16
13-kr-vs-kp	3196	1669	1527	13	23	18	18
14-Mushroom	8124	4208	3916	19	1	10	10
15-Pima indians diabetes	768	500	268	6	2	4	4
16-Sonar	208	97	111	34	26	30	30
17-Spambase	4601	2788	1813	21	36	29	28
18-Splice	3186	1532	1654	32	28	30	30
19-Tic-tac-toe	958	626	332	3	6	5	4
20-Breast Cancer-BCW	699	458	241	5	3	4	4
21-Breast Cancer-WDBC	569	212	357	11	19	15	15
22-Breast Cancer-WPBC1	194	46	148	11	22	17	16
23-Breast Cancer-WPBC2	198	47	151	13	19	16	16
24-Heart disease-Cleveland	303	164	139	5	6	6	5
25-Heart disease-Hungarian	294	188	106	1	3	2	2
26-Heart disease-LongBeachVA	200	144	56	2	2	2	2
27-Heart disease-Switzerland	123	75	48	1	2	2	1
28-Hepatits1	80	13	67	12	7	10	9
29-Hepatits2	155	32	123	1	3	2	2
30-Solar-flare-1	1389	1171	218	1	9	5	5
31-Solar-flare-2	1389	1321	68	8	2	5	5
32-Solar-flare-3	1389	1377	12	9	1	5	5

For reasons of clarity, Figs. 2 shows only two relevant examples of (almost) balanced and unbalanced datasets (respectively, Splice and Solar Flare 3), but the reported data does not deviate from the results obtained on other datasets.

Items 1-2 are shown by the ROC curves of Solar Flare 3 data set, whilst Splice data set has been chosen for items 3-4.

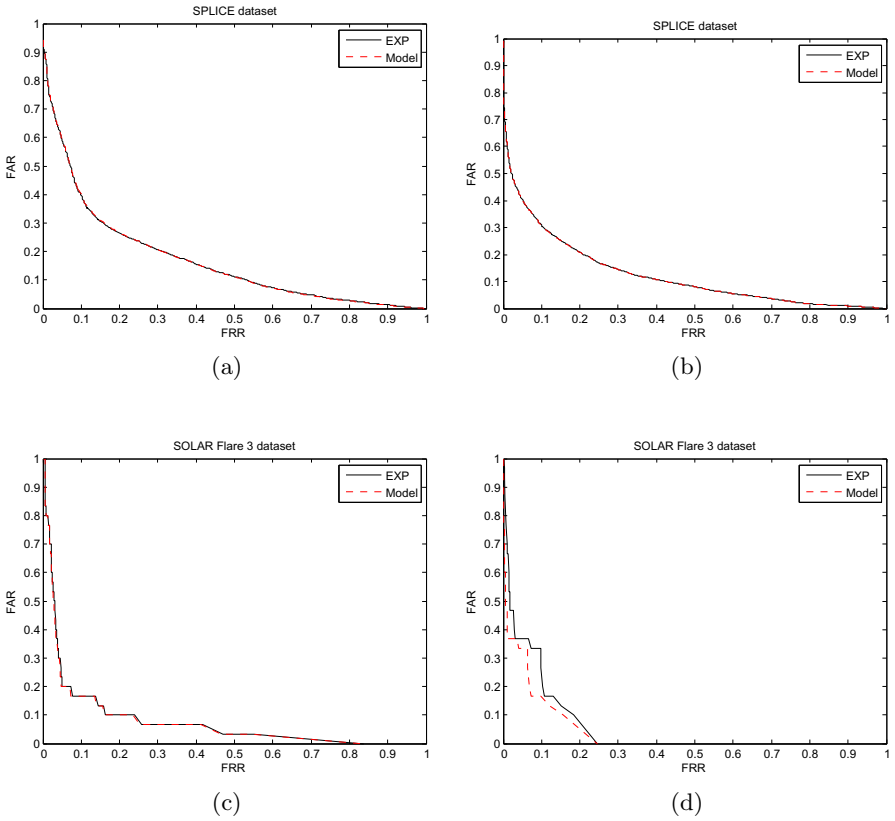

Fig. 2. ROC curves for serial system. Best classifier at second stage. (a) balanced data set, maximum independence split; (b) balanced data set, random split; (c) unbalanced data set, maximum independence split; (d) unbalanced data set, random split.

It is evident that ROC prediction in all cases is very near to the ROC computed by experiments. No appreciable difference is given when random selection of features for views generation is done.

In results reported in Figs. 2, the best classifier is always at the second stage, according to findings reported in Section 2. In the following tables, we explored the difference between ROCs predicted by our model and ROCs computed by experiments, even if the worst classifiers precedes the best one in the serial system.

Table 2 shows results in this last configuration. Reported values are the differences among AUCs for all datasets under consideration. In all cases, the very low values clearly show that our model correctly predict the final performance of the serial system. In fact, a difference of about 10^{-3} on average means that the error rate for one of two classes, in the worst case, is of about 10^{-1}, and on average, of $10^{-1.5}$, which corresponds to an error rate prediction of 3% on both classes.

This is true even when random split is performed for feature selection. In particular, we may see that difference between prediction on maximum independence and prediction on random split is negligible (last column of Table 2), even if, as it may be expected, the AUC difference is slightly higher in the random split case. This could mean that features on UCI data sets investigated are uncorrelated, thus even choosing them randomly does not impact on the model's predictions.

Table 3 confirms what reported in previous Table. For all datasets in 1, a very low AUC difference is reported.

Table 2. Difference of AUC values between the model's prediction and the experimental test. Results refer to the classifiers sequence where the worst classifier is at the first stage.

Classifier	**split**	**AUC difference** ($*10^{-3}$)			**diff**
		min	**max**	**mean** ($\pm stddev$)	
KNN- k=3	MAX IND	0	18.9	3.5($\pm$6.0)	0.7
	RANDOM	0	27.7	4.2($\pm$7.8)	
Naive Bayes - N=10	MAX IND	0	52.9	9.9($\pm$11.5)	2.1
	RANDOM	0	42.7	12.0($\pm$11.9)	
Linear Log	MAX IND	0	37.0	8.2($\pm$9.7)	2.7
	RANDOM	0	52.8	10.9($\pm$12.1)	

Table 3. Difference of AUC values between the model's prediction and the experimental test. Results refert to che classifiers sequence where the best classifier is at the first stage.

Classifier	**split**	**AUC difference** ($*10^{-3}$)			**diff**
		min	**max**	**mean** ($\pm stddev$)	
KNN- k=3	MAX IND	0	22.7	3.6($\pm$6.1)	1.5
	RANDOM	0	30.4	5.1($\pm$8.0)	
Naive Bayes - N=10	MAX IND	0	39.1	9.9($\pm$10.5)	7.5
	RANDOM	0	65.0	17.4($\pm$15.8)	
Linear Log	MAX IND	0	55.4	12.9($\pm$13.5)	1.1
	RANDOM	0	38.2	14.0($\pm$11.4)	

4 Conclusions

In this preliminary paper, we have shown the general expression of the error rate for a serial system of two classifiers in the case of two-classes classification problem. This expression allows predicting the performance of the system given the performance of the individual classifiers, under the assumption of decision independence.

The model has been validated by a large battery of UCI data sets, which have been set up in order to follow or not the assumption above. Results have shown

that the model is highly reliable, even in the case that independence hypothesis is not satisfied in practice.

Validation has been performed only on an operational point, namely, the point for which no errors are allowed for the first stage. Although this choice can be motivated for several practical applications, further works will rely on extending this investigation for any size of the rejection region. An ample set of experiments on realistic use-cases and scenarios will be also taken into account.

Finally, the case of more than two classifiers and the multiclass problem will be also considered in future works.

References

1. El Gayar, N., Kittler, J., Roli, F. (eds.): MCS 2010. LNCS, vol. 5997. Springer, Heidelberg (2010)
2. Sansone, C., Kittler, J., Roli, F. (eds.): MCS 2011. LNCS, vol. 6713. Springer, Heidelberg (2011)
3. Suen, C.Y., Lam, L.: Multiple Classifier Combination Methodologies for Different Output Levels. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 52–66. Springer, Heidelberg (2000)
4. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
5. Pudil, P., Novovicova, J., Blaha, S., Kittler, J.: Multistage Pattern Recognition with Reject Option. In: Proc. 11th IAPR-ICPR International Conference, vol. 2, pp. 92–95 (1992)
6. Last, M., Bunke, H., Kandel, A.: A feature-based serial approach to classifier combination. Pattern Analysis and Applications 5(4), 385–398 (2002)
7. Sansone, C., Vento, M.: Signature verification: increasing performance by a multistage system. Pattern Analysis and Applications 3, 169–181 (2000)
8. Pao, L., Trailovic, L.: The optimal order of processing sensor information in sequential multisensor fusion algorithms. IEEE Transactions on Automatic Control 45(8), 1532–1536 (2000)
9. Marcialis, G.L., Roli, F., Didaci, L.: Personal identity verification by serial fusion of fingerprint and face matchers. Pattern Recognition 42(11), 2807–2817 (2009)
10. Marcialis, G.L., Mastinu, P., Roli, F.: Serial fusion of multi-modal biometric systems. In: Proc. of IEEE International Workshop on Biometric Measurements and Systems for Security and Medical Applications, BioMS 2010 (2010), doi:10.1109/BIOMS.2010.5610438
11. Allano, L., Dorizzi, B., Garcia-Salicetti, S.: Tuning cost and performance in multimodal biometric systems: a novel and consistent view of fusion strategies based on the Sequential Probability Ratio Test (SPRT). Pattern Recognition Letters 31, 884–890 (2010)
12. Chow, C.K.: On optimum rejection error trade-off. IEEE Transactions on Information Theory 16(1), 41–46 (1970)
13. Fumera, G., Roli, F., Giacinto, G.: Reject Option with Multiple Thresholds. Pattern Recognition 33(12), 2099–2101 (2000)
14. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2010), `http://archive.ics.uci.edu/ml`

15. Du, J., Ling, C.X., Zhou, Z.-H.: When Does Co-Training Work in Real Data? IEEE Transactions on Knowledge and Data Engineering 23(35), 788–799 (2011)
16. Giusti, N., Masulli, F., Sperduti, A.: Theoretical and Experimental Analysis of a Two-Stage System for Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 893–904 (2002)
17. Senator, T.: Multi-stage classification. In: Proc. of IEEE 5th Int. Conf. on Data Mining, ICDM 2005 (2005)
18. Trapeznikov, K., Saligrama, V., Castanon, D.: Multi-stage classifier design. In: JMLR Asian Conference on Machine Learning (ACML 2012), pp. 1–16 (2012)