ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447423

SEQUENTIAL ACQUISITION OF FEATURES AND EXPERTS FOR DATUM–WISE CLASSIFICATION

Sachini Piyoni Ekanayake *Daphney–Stavroula Zois*

Department of Electrical and Computer Engineering
University at Albany, State University of New York, Albany, NY, USA
Emails: {sekanayake, dzois}@albany.edu

ABSTRACT

We present a sequential acquisition of features and experts framework for datum–wise classification. The goal is to accurately assign labels for each instance, minimizing the acquisition cost of features and experts. An expert uses domain knowledge to make decisions. Starting from a prior belief, features are sequentially acquired in a feature acquisition stage. When this stage terminates, the acquired subset of features is forwarded to an expert acquisition stage, where each expert provides their decision one at a time. At that time, contrary to prior work, the label assignment is reached based on the acquired experts' decisions thus far. We evaluate the framework's performance using six real–world datasets and compare it with existing methods. Experiments reveal that the proposed framework increases accuracy up to 56% compared to existing ensemble methods while acquiring 88% fewer features and, more importantly, 80% fewer experts on average.

Index Terms— supervised learning, instance–wise classification, feature selection, costly inputs, collective decision

1. INTRODUCTION

In many applications such as medical diagnosis, multiple experts (e.g., radiologist, primary care doctor) collaborate in addition to relying on feature observations (e.g., medical history, imaging scans). This is because, in critical situations, when arriving at a final diagnosis, collective decision ensures comprehensive and accurate medical decision–making [1, 2]. However, these experts and features are costly (e.g., expert decision cost, feature evaluation cost) [3, 4]. On the other hand, observations of individual patients may vary [5]; thus, person–wise decision–making is required. In this context, instance–wise feature and expert selection and classification have received considerable attention in machine learning [6–9].

Standard supervised classification (e.g., Support Vector Machines (SVM)) considers a batch–wise approach where all features are available and a single expert makes the decision. On the other hand, standard ensemble methods consider combining multiple experts' decisions (e.g., majority voting) [10]. However, they also rely on a batch–wise approach and assume that all features and experts are available during testing. Conversely, offline feature selection methods (e.g., L1–norm based (Lasso)) choose a subset of features in the training and use them for testing. On the contrary, dynamic instance–wise feature selection techniques [6, 8, 11] choose varying features for classifying data instances in testing relying on a single expert decision. Further, Dynamic Ensemble Selection (DES) techniques [12, 13] acquire all expert decisions during testing using all available features. If there is disagreement among experts, they execute instance–wise expert selection and final decision is reached based on selected subset of experts.

In contrast to prior work, in this paper, we propose an approach that does not require access to all features and expert decisions for classification during testing. Specifically, we propose a two–stage approach that sequentially acquires features and experts for instance–wise label assignment. Starting from an initial belief, features are sequentially acquired first. Then, the acquired subset of features and updated belief is used to drive expert acquisition. Finally, a label is assigned to the instance based on the acquired experts' decisions thus far. Performance is evaluated on six real–world datasets and compared to several existing approaches. We observe that the proposed approach leads to considerable performance improvement in accuracy, using less number of features and experts on average.

2. PROBLEM DESCRIPTION

Consider a *supervised classification* setting, where each data instance is associated with F number of features $X \triangleq [X_1, \ldots, X_F]^T$ and the value of X is denoted as $x \triangleq [x_1, \ldots, x_F]^T$. We define Y such that $Y = y$ denotes that the data instance has true label $y \in \{1, \ldots, N\}$, where N is the total number of labels. In standard supervised classification, the goal is to determine Y by using the feature values x where the predicted label is denoted as $\hat{y}$. Additionally, in this work, we use features to drive expert decisions, which are then used to finalize the label $\hat{y}$. In this context, we consider a set of

This material is based upon work supported by the National Science Foundation under Grants ECCS–1737443 & CNS–1942330.

experts as heterogeneous classifiers $C \triangleq \{C_1, \ldots, C_Z\}$ (e.g., $C \triangleq \{\text{SVM}\}$ for $Z = 1$), where Z is the number of experts. We define $\hat{X} \triangleq [\hat{X}_1, \ldots, \hat{X}_Z]^T$ as the set of decisions of each classifier using features $X_f \in X, f = 1, \ldots, F$, and the value of $\hat{X}$ is denoted as $\hat{x} \triangleq [\hat{x}_1, \ldots, \hat{x}_Z]^T$. Here $\hat{x}_z \in \{1, \ldots, N\}$ for $z \in \{1, \ldots, Z\}$. Note that accessing features and experts comes at a cost. Specifically, obtaining a decision of the classifier $C_z \in C$ involves a cost $c_z > 0, z = 1, \ldots, Z$. Also, features $X_f \in X, f = 1, \ldots, F$, have an associated cost, $e_f > 0$. In this setting, we propose a two–stage approach, as illustrated in Fig. 1, to assign a label for each data instance by first sequentially acquiring features (Feature Acquisition (FA) stage) and second sequentially acquiring experts' decisions (Expert Acquisition (EA) stage). To simplify the problem, we adopt the assumptions: (a) features and experts are ordered based on a certain performance measure, (b) features $X_f, f = 1, \ldots, F$, are conditionally independent given label Y, and (c) an expert decision is not affected by the other experts' decisions.

Optimization Problem. We define four random variables $S_{\text{FA}}, S_{\text{EA}}, D_{S_{\text{FA}}}$, and $D_{S_{\text{EA}}}$. We use $S_{\text{FA}} \in \{0, \ldots, F\}$ to denote the last feature acquired from the ordered feature set X at the end of the FA stage[1]. Similarly, we use $S_{\text{EA}} \in \{1, \ldots, Z\}$ to represent the last expert considered to obtain a decision. The decision that controls when the FA will terminate is denoted as $D_{S_{\text{FA}}} \in \{1, \ldots, N\}$. Furthermore, $D_{S_{\text{EA}}} \in \{1, \ldots, N\}$ indicates the final decision, i.e., the label assignment based on accumulated experts' decisions up to S_{EA}. We define two cost functions $J_{\text{FA}}(S_{\text{FA}}, D_{S_{\text{FA}}})$ and $J_{\text{EA}}(S_{\text{EA}}, D_{S_{\text{EA}}})$ for the FA and EA stages, respectively. First, we propose the following cost function for the FA stage:

$$J_{\text{FA}}(S_{\text{FA}}, D_{S_{\text{FA}}}) = \mathbb{E}\left[\sum_{f=1}^{S_{\text{FA}}} e_f\right] + \sum_{j=1}^{N}\sum_{i=1}^{N} M_{ij} P(D_{S_{\text{FA}}} = j, Y = i), \quad (1)$$

where $P(D_{S_{\text{FA}}} = j, Y = i)$ denotes the probability of assigning label j, while true label is i. The first expression in Eq. (1) represents the expected cost of acquiring S_{FA} features, while the second expression indicates the average cost of decision $D_{S_{\text{FA}}}$ in the FA stage. Here $M_{ij}, i, j \in \{1, \ldots, N\}$ is the cost of assigning label j to a data instance when the true label is i. Next, we propose the following cost function for the EA stage:

$$J_{\text{EA}}(S_{\text{EA}}, D_{S_{\text{EA}}}) = \mathbb{E}\left[\sum_{z=1}^{S_{\text{EA}}} c_z\right] + \sum_{j=1}^{N}\sum_{i=1}^{N} W_{ij} P(D_{S_{\text{EA}}} = j, Y = i), \quad (2)$$

where $P(D_{S_{\text{EA}}} = j, Y = i)$ denotes the probability of assigning label j, while true label is i. The first expression in Eq. (2) represents the expected cost of acquiring decisions from S_{EA} experts, while the second expression indicates the average cost of using the decisions of $D_{S_{\text{EA}}}$ in the EA stage. Here $W_{ij}, i, j \in \{1, \ldots, N\}$ is the cost of assigning label j to a data instance when the true label is i in the EA stage.

Sufficient Statistics. Let $\pi_f^i \triangleq P(Y = i | X_1 = x_1, \ldots, X_f = x_f)$, $f = 1, \ldots, F, i = 1, \ldots, N$, denote the posterior probability of the label of a data instance of interest being i when f features have been acquired so far in the FA stage. Using Bayes' rule, we update this posterior probability when a new feature is acquired, as follows:

$$\pi_f^i \triangleq \frac{P(X_f = x_f | Y = i)\pi_{f-1}^i}{\sum_{n=1}^{N} P(X_f = x_f | Y = n)\pi_{f-1}^n}. \quad (3)$$

Let $\boldsymbol{\pi}_f \triangleq [\pi_f^1, \ldots, \pi_f^N]^T, f = 1, \ldots, F$, represent the posterior probability vector, where $\boldsymbol{\pi}_0 \triangleq [\pi_0^1, \ldots, \pi_0^N]^T$ denotes the case where no features have been aquired. Here, $\pi_0^i = P(Y = i), i = 1, \ldots, N$, represents the prior probability of the true label being i. Next, in the EA stage, we define posterior $\phi_z^i \triangleq P_f(Y = i | \hat{X}_1 = \hat{x}_1, \ldots, \hat{X}_z = \hat{x}_z), z = 1, \ldots, Z, i = 1, \ldots, N$, as the probability of true label being i given z individual expert decisions. Let $\boldsymbol{\phi}_z \triangleq [\phi_z^1, \ldots, \phi_z^N]^T$, represents the posterior probability vector, where $\boldsymbol{\phi}_0 \triangleq [\phi_0^1, \ldots, \phi_0^N]^T$ denotes the case where no experts have been considered. Here, ϕ_0^i represents the prior probability of the data instance's true label being i before EA stage begins. We initialize ϕ_0^i to $\pi_{S_{\text{FA}}}^i$, which denotes the posterior probability determined at the end of the FA stage. Using Bayes' rule, we update this ϕ_z^i when a new expert decision is acquired as follows:

$$\phi_z^i \triangleq \frac{P(\hat{X}_z = \hat{x}_z | Y = i)\phi_{z-1}^i}{\sum_{n=1}^{N} P(\hat{X}_z = \hat{x}_z | Y = n)\phi_{z-1}^n}. \quad (4)$$

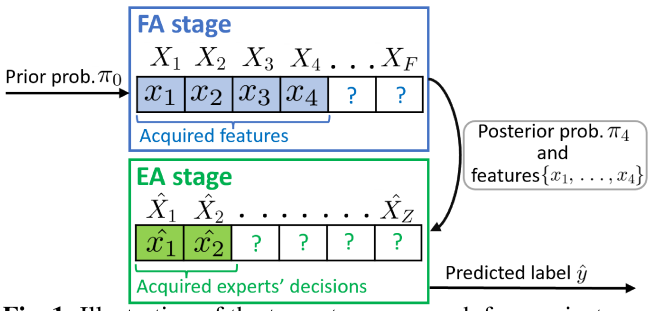

Fig. 1. Illustration of the two–stage approach for one instance.

3. TWO–STAGE SOLUTION

To determine the feature and expert acquisition strategies (i.e., S_{FA}^* and S_{EA}^*), we first minimize Eq. (1) and then Eq. (2).

FA Stage. First, $D_{S_{\text{FA}}}^*$ for a fixed S_{FA} is obtained. Eq. (1) is written in terms of the posterior probability and the indicator function $\mathbb{1}_A$[2] as [6]:

$$J_{\text{FA}}(S_{\text{FA}}, D_{S_{\text{FA}}}) = \mathbb{E}\left[\sum_{f=1}^{S_{\text{FA}}} e_f + \sum_{j=1}^{N}\sum_{i=1}^{N} M_{ij}\pi_{S_{\text{FA}}}^i \mathbb{1}_{D_{\{S_{\text{FA}}=j\}}}\right]. \quad (5)$$

Then $D_{S_{\text{FA}}}^*$ can be obtained by finding the lower bound of the second term of Eq. (5) as in prior work [6]. Therefore, $D_{S_{\text{FA}}}^*$ is given by:

$$D_{S_{\text{FA}}}^* = \underset{1 \le j \le N}{\text{argmin}} [\mathbf{M}_j^T \boldsymbol{\pi}_{S_{\text{FA}}}], \quad (6)$$

where $\mathbf{M}_j \triangleq [M_{1j}, \ldots, M_{Nj}]^T$. To obtain the optimum feature acquisition strategy S_{FA}^*, Eq. (5) is written in the reduced form [6] as: $J_{\text{FA}}(S_{\text{FA}}) = \mathbb{E}\left[\sum_{f=1}^{S_{\text{FA}}} e_f + g(\boldsymbol{\pi}_{S_{\text{FA}}})\right]$, where $g(\boldsymbol{\pi}_{S_{\text{FA}}}) \triangleq \min_{1 \le j \le N} [\mathbf{M}_j^T \boldsymbol{\pi}_{S_{\text{FA}}}]$. Finally, S_{FA}^* can be obtained by minimizing the reduced cost function via dynamic

[1] $S_{\text{FA}} = 0$ represents the case where no features have been acquired.

[2] $\mathbb{1}_A \triangleq 1$ when event A occurs and 0 otherwise.

programming [14]. Specifically, since there are F available features, there exists a maximum of $F+1$ stages for the associated dynamic programming equation:

$$A_f(\boldsymbol{\pi}_f) = \min\left[g(\boldsymbol{\pi}_f), \tilde{A}_f(\boldsymbol{\pi}_f)\right], f = 0, \ldots, F-1, \quad (7)$$

where $\tilde{A}_f(\boldsymbol{\pi}_f) = e_{f+1} + \sum_{x_{f+1}} A_{f+1}(\boldsymbol{\pi}_{f+1})\left(\boldsymbol{\Delta}_{f+1}^T(x_{f+1})\boldsymbol{\pi}_f\right)$, with $\boldsymbol{\Delta}_f(x_f) \triangleq [P(X_f = x_f|Y=1), \ldots, P(X_f = x_f|Y = N)]^T$ and $A_F(\boldsymbol{\pi}_F) = g(\boldsymbol{\pi}_F)$. The term $g(\boldsymbol{\pi}_f)$ in Eq. (7) represents the cost of stopping FA when f features have already been acquired, while $\tilde{A}_f(\boldsymbol{\pi}_f)$ is the cost of continuing this process. Thus, $S^*_{\text{FA}} = f$ if $g(\boldsymbol{\pi}_f) < \tilde{A}_f(\boldsymbol{\pi}_f)$ for $f < F$ or $S^*_{\text{FA}} = F$ when all the features are acquired.

EA Stage. We use posterior probability $\boldsymbol{\pi}_{S^*_{\text{FA}}}$ from FA stage to drive EA. First, we discuss the optimum final decision strategy D^*_{EA}. Specifically, Eq. (2) in the EA stage can be written in terms of the posterior probability vector $\boldsymbol{\phi}_{S_{\text{EA}}}$ and indicator function $_{\text{A}}$ for a fixed S_{EA} as follows:

$$J_{\text{EA}}(S_{\text{EA}}, D_{S_{\text{EA}}}) = \left[\sum_{z=1}^{S_{\text{EA}}} c_z + \sum_{j=1}^{N}\sum_{i=1}^{N} W_{ij}\phi^i_{S_{\text{EA}}} \quad _{D\{S_{\text{EA}}=j\}}\right]. \quad (8)$$

The optimum decision strategy D^*_{EA} can be obtained by finding an appropriate lower bound of the second term of Eq. (8). Specifically, for a given S_{EA}, for any $D_{S_{\text{EA}}}$, $\sum_{j=1}^{N} \mathbf{W}_j^T \boldsymbol{\phi}_{S_{\text{EA}}} \mathbb{I}_{\{D_{S_{\text{EA}}}=j\}} \geq h(\boldsymbol{\phi}_{S_{\text{EA}}})$ where $h(\boldsymbol{\phi}_{S_{\text{EA}}}) \triangleq \min_{1\leq j\leq N}[\mathbf{W}_j^T \boldsymbol{\phi}_{S_{\text{EA}}}]$ and $\mathbf{W}_j \triangleq [W_{1j}, \ldots, W_{Nj}]^T$. Therefore, the optimum decision strategy is given by:

$$D^*_{S_{\text{EA}}} = \underset{1\leq j\leq N}{\operatorname{argmin}}[\mathbf{W}_j^T \boldsymbol{\phi}_{S_{\text{EA}}}]. \quad (9)$$

Next, to find the optimum expert acquisition strategy S^*_{EA}, the cost function in Eq. (8) is reduced to: $J_{\text{EA}}(S_{\text{EA}}) = \mathbb{E}\left[\sum_{z=1}^{S_{\text{EA}}} c_z + h(\boldsymbol{\phi}_{S_{\text{EA}}})\right]$. This enables us to find S^*_{EA} using dynamic programming [14]. In particular, we derive the following dynamic programming equation:

$$B_z(\boldsymbol{\phi}_z) = \min\left[h(\boldsymbol{\phi}_z), \tilde{B}_z(\boldsymbol{\phi}_z)\right], z = 0, \ldots, Z-1, \quad (10)$$

where $\tilde{B}_z(\boldsymbol{\phi}_z) = c_{z+1} + \sum_{\hat{x}_{z+1}} B_{z+1}(\boldsymbol{\phi_{z+1}})\left(\boldsymbol{\Theta}_{z+1}^T(\hat{x}_{z+1})\boldsymbol{\phi}_z\right)$, with $\boldsymbol{\Theta}_z(\hat{x}_z) \triangleq [P(\hat{X}_z = \hat{x}_z|Y=1), \ldots, P(\hat{X}_z = \hat{x}_z|Y = N)]^T$ and $B_Z(\boldsymbol{\phi}_Z) = h(\boldsymbol{\phi}_z)$. In Eq (10), the term $h(\boldsymbol{\phi}_z)$ represents the cost associated with stopping EA after obtaining z number of expert decisions. Conversely, $\tilde{B}_z(\boldsymbol{\phi}_z)$ denotes the cost of continuing the EA process. Thus, if $h(\boldsymbol{\phi}_z) < \tilde{B}_z(\boldsymbol{\phi}_z)$ for $z < Z$, the optimum EA strategy is $S^*_{\text{EA}} = z$. If the cost of stopping is always greater than the cost of acquiring more experts, the EA continues until all experts have been acquired. At that point, the optimum EA strategy is $S^*_{\text{EA}} = Z$, and the final decision is made considering the decisions of all experts.

SAFE Algorithm[3]. *Training:* The interval $[0, 1]$ is quantized to generate all possible posterior probability vectors $\boldsymbol{\pi}_f$ and $\boldsymbol{\phi}_z$ such that $\boldsymbol{\pi}_f \mathbb{1}^T = 1$ and $\boldsymbol{\phi}_z \mathbb{1}^T = 1$. Here, $\mathbb{1}$ is a N dimensional vector of all ones. With a precision of discretization η, d possible vectors $\boldsymbol{\pi}_f$ and $\boldsymbol{\phi}_z$ are generated. For all $\boldsymbol{\pi}_f$ and $\boldsymbol{\phi}_z$, Eqs. (6) and (7) for FA stage and Eqs. (9) and (10) for EA stage are numerically solved to determine the optimum feature and expert acquisition strategies. Experts $C_z, \forall z$, are trained for each number $f = 1, \ldots, F$, of features and probabilities $P(X_f = x_f|Y = i)$ and $P(\hat{X}_f = \hat{x}_z|Y = i)$ are estimated (c.f. Section 4). *Testing:* The process begins in the FA stage by initializing posterior $\boldsymbol{\pi}_0 \triangleq [\pi_0^1, \ldots, \pi_0^N]^T$ where $\pi_0^i = P(Y = i)$. Then features are sequentially acquired based on the numerical solutions determined during training and an intermediate decision is reached based on Eq. (6). Assuming f features have already been acquired, if the stopping cost is greater than the continuing cost, next feature X_{f+1} is acquired, and the posterior probability gets updated using Eq. (3). This continues until all or a subset of features are acquired. After that, S^*_{FA} and $\boldsymbol{\pi}_{S^*_{\text{FA}}}$ from FA stage are forwarded to EA stage and we set $\boldsymbol{\phi}_0 = \boldsymbol{\pi}_{S^*_{\text{FA}}}$. Then, experts are sequentially queried based on the numerical solutions found during training, and an expert decision is acquired based on features S^*_{FA}. Assuming z expert decisions have already been acquired, if the stopping cost is greater than the continuing cost, the next expert decision $\hat{X}_{z+1}$ is acquired. This continues until a subset or all the experts are acquired. Either way, a label is assigned to the current instance based on Eq. (9).

4. NUMERICAL RESULTS

To illustrate the performance of the SAFE algorithm, we consider 6 real–world datasets: MONKS [15] (601, 6, 2), STUDENT [15] (649, 31, 2), GENDER [16] (4746, 20, 3), SPAMBASE [15] (4601, 57, 2), MADELON [15] (2600, 500, 2), and CANCER [17] (569, 30, 2)[4]. STUDENT dataset is preprocessed such that the classification variable final grade G_3 is binary [8]. We use accuracy, average number of acquired features and average number of acquired experts (when relevant) to compare SAFE's performance with the baselines: (i) SDFA–DT[5] [8], a recent instance–wise feature and classifier selection algorithm, (ii) SVM with Gaussian kernel, a standard often used supervised learning algorithm, (iii) Lasso, an offline feature selection algorithm, (iv) FIRE–DES++[6] [9], a recent standard DES algorithm, and (v) AdaBoost and XGBoost with five decision stumps [19], often used ensemble learning algorithms. For diversity reasons [10], we consider five widely used [20] standard classifiers as the experts for SAFE and FIRE–DES++: (i) Naive Bayes, (ii) SVM with Gaussian kernel, (iii) Decision Tree, (iv) K–Nearest Neighbours ($k = 7$), and (v) Logistic Regression.

During training, $P(X_f = x_f|Y = i) = \frac{R_{f,i}+1}{R_i+\beta}$ is estimated. Here $R_{f,i}$ denotes the number of instances that have label i and x_k takes a specific value, while R_i denotes the total number of instances that have label i. β is the number of bins considered. The training dataset is used to obtain $\hat{X}$ first and then $P(\hat{X}_z =$

[3] Sequential Acquisition of Features and Experts.

[4] (# instances, # features, # classes).

[5] $B = 10, c = 0.0001$, and $M_{i\varepsilon} = 0.4$.

[6] KNE [18] for $k = 7$ implemented with DESlib [13].

Table 1. Accuracy ("Acc") and average number of acquired features ("Feat") for SAFE and baselines. Highest and next highest Acc values are bold and gray–shaded, and gray–shaded, respectively. Smallest and next smallest Feat values are bold and gray–shaded, and gray-shaded, respectively.

Method	MONKS		STUDENT		GENDER		SPAMBASE		MADELON		CANCER	
	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat
SAFE	**0.900**	5.730	0.880	**3.723**	**1.000**	7.543	0.899	40.884	**0.766**	198.572	0.953	7.960
SDFA–DT	0.795	5.722	0.869	4.656	0.979	**3.439**	0.849	**31.177**	0.624	**68.822**	0.910	**2.827**
SVM	0.657	6.000	0.871	32.000	0.588	20.000	0.690	57.000	0.617	500.000	0.891	30.000
Lasso	0.654	**4.800**	0.886	19.200	0.928	17.800	**0.909**	50.600	0.560	492.000	0.947	10.600
FIRE–DES++	0.800	6.000	0.793	32.000	0.996	20.000	0.895	57.000	0.641	500.000	**0.967**	30.000
AdaBoost	0.577	6.000	0.908	32.000	0.839	20.000	0.874	57.000	0.613	500.000	0.928	30.000
XGBoost	0.657	6.000	**0.909**	32.000	0.964	20.000	0.852	57.000	0.593	500.000	0.933	30.000

$\hat{x}_z|Y=i) = \frac{\hat{R}_{z,i}}{\hat{R}_i}$ is estimated. Here $\hat{R}_{z,i}$ denotes the number of instances that have true label i and predicted label $\hat{x}_z$ when expert z is used, while $\hat{R}_i$ denotes the total number of instances that have true label i. Equal prior probability $P(Y=i)=\frac{1}{N}$ is considered assuming equally likely scenarios. Features are ranked in ascending order of the total of type I and type II errors, which is scaled by the cost coefficient of the fth feature to prioritize low–cost features. Experts are ranked based on increasing order of training errors to promote the most competent experts. The ranked feature set is used in both training and testing phases, and the experts are trained for $f=1,\ldots,F$, features. We consider $\eta=100, \beta=10, M_{ij}=W_{ij}=1, \forall i\neq j$ and $M_{ii}=W_{ii}=0, \forall i,j\in\{1,\ldots,N\}$. We assume feature and expert costs to be the same for all features and experts, i.e., $e_f=e, \forall f$ and $c_z=c, \forall z$. Five–fold cross validated results are reported in Table 1.

The average number of acquired experts for SAFE is 1.000, 1.168, 1.922, 1.945, 2.484, and 1.227 for datasets in the order shown in Table 1. Comparing SAFE with ensemble methods FIRE–DES++, AdaBoost, and XGBoost, we observe that SAFE achieves better accuracy (up to 56% improvement) using fewer experts (up to 80% less). Importantly, SAFE uses between 4% and 88% fewer features. For the cases where SAFE's accuracy is less (up to 3%), SAFE uses significantly fewer features (between 73% to 88%) and experts (between 75% to 77%). Comparing with Lasso, we observe that SAFE's performance is better in accuracy (up to 38%) for the majority of the datasets using fewer features (between 19% to 81%). For the cases where Lasso performs better in terms of accuracy, SAFE uses fewer features (19% and 81%) resulting in a small accuracy drop. Comparing SAFE with SVM, SAFE's performance is always better in terms of accuracy (up to 70%) using fewer features (4% to 88%). Finally, we compare SAFE with SDFT–DT, a recently proposed instance–wise feature and classifier selection algorithm. SAFE's accuracy is always better (up to 23%) than SDFT–DT, although the latter one uses fewer features in most cases with a single expert. In some cases, SAFE saves on features (20% fewer) by expending more on experts (17% more). This is essential in applications like medicine where we can do less medical tests and rely more on

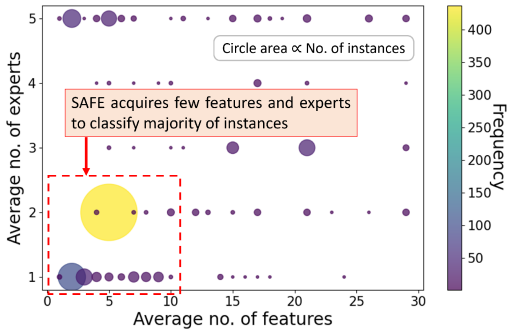

Fig. 2. Frequency of the number of acquired features and experts during testing for the SPAMBASE dataset.

experts' decisions of the same tests. SAFE's performance is due to a combination of both features and experts. We note that when the FA posterior is inaccurate due to few acquired features, SAFE attempts to acquire more experts to get a better posterior value. For the majority of data instances, accurate classification can happen with just a few features and experts (Fig. 2). This is important in real–world applications where the acquisition of features and experts is costly and/or feature and expert space is large.

5. CONCLUSION

We proposed a supervised learning algorithm, SAFE, that sequentially acquires features and experts for instance–wise classification in a costly environment. We devised a two–stage approach that involves sequential feature acquisition followed by an expert decision acquisition stage. SAFE's performance is validated on a set of experiments. Observations confirm that SAFE achieves a good balance between accuracy, the average number of acquired features, and the average number of acquired experts. It uses varying features and experts per instance; thus, its decisions can be explained [21]. SAFE forwards all instances from the first to the second stage, no matter the intermediate decision. Other limitations include training complexity and ordering assumptions. Forwarding only most ambiguous instances [22] and training a subset of experts using heuristics can potentially reduce complexity without hurting accuracy.

6. REFERENCES

[1] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry, ""hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making," *Proceedings of the ACM on Human-computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[2] Anika Radkowitsch, Martin R Fischer, Ralf Schmidmaier, and Frank Fischer, "Learning to diagnose collaboratively: Validating a simulation for medical students," *GMS journal for medical education*, vol. 37, no. 5, 2020.

[3] Vijay S Mookerjee and Michael V Mannino, "Sequential decision models for expert system optimization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 5, pp. 675–687, 1997.

[4] Tianshi Gao and Daphne Koller, "Active classification based on value of classifier," *Advances in neural information processing systems*, vol. 24, 2011.

[5] David P Kao, James D Lewsey, Inder S Anand, Barry M Massie, Michael R Zile, Peter E Carson, Robert S McKelvie, Michel Komajda, John JV McMurray, and JoAnn Lindenfeld, "Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response," *European journal of heart failure*, vol. 17, no. 9, pp. 925–935, 2015.

[6] Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis, "Dynamic instance-wise joint feature selection and classification," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 169–184, 2021.

[7] Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis, "Dynamic instance-wise classification in correlated feature spaces," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 6, pp. 537–548, 2021.

[8] Sachini Piyoni Ekanayake, Daphney–Stavroula Zois, and Charalampos Chelmis, "Sequential datum–wise joint feature selection and classification in the presence of external classifier," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] Rafael MO Cruz, Dayvid VR Oliveira, George DC Cavalcanti, and Robert Sabourin, "Fire-des++: Enhanced online pruning of base classifiers for dynamic ensemble selection," *Pattern Recognition*, vol. 85, pp. 149–160, 2019.

[10] Zhi-Hua Zhou and Zhi-Hua Zhou, *Ensemble learning*, Springer, 2021.

[11] Aritra Ghosh and Andrew Lan, "Difa: Differentiable feature acquisition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 7705–7713, Jun. 2023.

[12] Rafael MO Cruz, Robert Sabourin, and George DC Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Information Fusion*, vol. 41, pp. 195–216, 2018.

[13] Rafael M. O. Cruz, Luiz G. Hafemann, Robert Sabourin, and George D. C. Cavalcanti, "Deslib: A dynamic ensemble selection library in python," *Journal of Machine Learning Research*, vol. 21, no. 8, pp. 1–5, 2020.

[14] Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, 2nd edition, 2000.

[15] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo, "Openml: networked science in machine learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013.

[16] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[17] "Kaggle," https://www.kaggle.com.

[18] Albert HR Ko, Robert Sabourin, and Alceu Souza Britto Jr, "From dynamic classifier selection to dynamic ensemble selection," *Pattern recognition*, vol. 41, no. 5, pp. 1718–1731, 2008.

[19] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014.

[20] Iqbal H Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, pp. 160, 2021.

[21] Christoph Molnar, *Interpretable machine learning*, Lulu. com, 2020.

[22] Feng Nan and Venkatesh Saligrama, "Adaptive classification for prediction under a budget," *Advances in neural information processing systems*, vol. 30, 2017.