Knowing When to Stop: Joint Heterogeneous Feature Selection and Classification

Imara Nazar*, Daphney–Stavroula Zois*, Charalampos Chelmis†
*Department of Electrical and Computer Engineering, †Department of Computer Science
University at Albany, SUNY, Albany, NY, USA
Emails: {imohamednazar, dzois, cchelmis}@albany.edu

2020 54th Asilomar Conference on Signals, Systems, and Computers | 978-0-7381-3126-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEEECONF51394.2020.9443288

***Abstract*—We consider the problem of joint heterogeneous feature selection and classification when multiple feature sets are present. Specifically, we want to identify which feature sets and features per set to review, and perform classification using this information. To this end, we formulate an optimization problem that considers the cost of reviewing individual features, the switching cost between feature sets, and the associated classification decision cost. The objective is to minimize the expected total cost of reviewing feature sets and associated features and the misclassification cost. We derive the optimum classification decision rule, and show that it minimizes the average misclassification cost. Additionally, we derive the optimum feature review rule, which determines both the feature sets and features per set to be reviewed. We illustrate the performance of the proposed methodology on the application of the automatic classification of civil issues reported on crowdsourcing platforms. We observe that an accurate classification decision can be reached by examining ~ 2.6 features on average.**

I. Introduction

In recent years, multi–view data has become increasingly available in the majority of real–world applications, where instances are described by multiple different sources (e.g., text, audio, image) and/or different feature subsets. Generally, such data tends to provide a complementary and more holistic understanding of the phenomenon of interest and can lead to more accurate prediction models [1], [2]. In fact, since the performance of machine learning algorithms heavily depends on the available data, integrating information from multiple views/modalities with the goal of predicting an outcome can improve the robustness of the prediction task, and even handle missing information [1], [2].

Integrating information from multiple independent feature sets has been studied within the context of various text classification applications. In [3]–[6], a final classification decision is determined by combining individual–level decisions generated from multiple feature selection algorithms. In contrast, [7], [8] focus on the design of different classifiers that use different feature set types. In all these cases, all features from the selected feature sets are used for classification. Finally, existing multi–view and multi–modal learning methods (see [1], [2], [9] and references therein) either seek for representations that maximize the mutual agreement between the distinct views of the data or combine outcomes in ad–hoc manner to improve learning performance.

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

In this work, we propose an alternative methodology for handling multi–view/modal data in prediction tasks that guarantees accurate classification using the least amount of available information. The proposed methodology has also the potential to improve the interpretability of the classification decision. Specifically, we define the problem of joint heterogeneous feature selection and classification with multiple feature sets. Our goal is to identify which feature sets and features per set to review, and perform classification using this information. To this end, we formulate an optimization problem that considers the cost of reviewing individual features, the switching cost between feature sets, and the associated classification decision cost. We derive both the optimum classification decision rule, which assigns an instance to the class with the minimum average misclassification cost, and the optimum feature review rule, which decides the feature sets and associated features per set to be reviewed. We evaluate the proposed methodology on the problem of automatically classifying civil issue reports on crowdsourcing platforms, and show its ability to achieve up to 94.1% classification accuracy using on average ~ 2.6 features. The current work extends our prior work [10]–[13] to the multi–view/modal data setting, while also dynamically deciding on the feature sets to be reviewed [14].

II. Problem Formulation

A. Setting

We consider a set $\mathcal{I}$ of data instances, where each data instance $i \in \mathcal{I}$ is described by a vector $\mathbf{f} \triangleq [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_Q]^\top$ of heterogeneous features. Vector $\mathbf{f}_q \triangleq [f_{q,1}, f_{q,2}, \ldots, f_{q,N_q}]^\top$ denotes the qth feature set, where $f_{q,n}$ represents the nth feature of the qth feature set, and N_q is the total number of features in the qth feature set. We assume that there are in total $N = \sum_{i=1}^{Q} N_q$ features available distributed across Q feature sets. Each data instance i may belong to one out of L possible classes with associated prior probability $p_l \triangleq P(C = C_l)$ for each assignment $C_l, l = 1, 2, \ldots, L$, of the class variable C. Additionally, the relationship between feature $f_{q,n}$ and class C_l is captured by $P(f_{q,n}|C_l)$, which denotes the conditional probability of the nth feature in the qth feature set under class C_l. We also denote by $c_{q,n} > 0, n \in \{1, 2, \ldots, N_q\}, q \in \{1, 2, \ldots, Q\}$ the effort required to extract and evaluate feature $f_{q,n}$. Since there are Q available features sets, we consider switching costs $s_{q,q'} > 0, q, q' \in \{1, 2, \ldots, Q\}$, to describe the cost of moving between the qth and the q'th feature set.

Finally, we define the misclassification cost $M_{ml} \geqslant 0, m, l \in \{1, 2, \ldots, L\}$, to represent the cost of assigning a particular data instance to class C_l, when the true class is $C_m, m \neq l$.

In order to get an accurate classification decision for each data instance i, we propose the following adaptive sequential review process. At each step, the decision maker selects between continuing to review features or not by considering the so far accumulated information, and the cost of reviewing additional features. Reviewing more features entails either staying in the current feature set, or moving to the next one. Thus, the cost of reviewing additional features includes either $c_{q,n}$ or $s_{q,q'}$. The decision maker can stop the review process at any time without considering all available feature sets, at which point a classification decision is reached.

B. Problem Statement

Consider a collection $(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)})$ of random variables. Random variable $\Gamma \in \{1, 2, \ldots, Q\}$ denotes the last feature set reviewed before the decision maker reaches a classification decision. Random variable $R_q, q \in \{1, 2, \ldots, \Gamma\}$, indicates the last feature the decision maker reviews before either moving to the next feature set or reaching a classification decision. Finally, $D_{(\Gamma,R_1,\ldots,R_\Gamma)}$ represents the classification decision of the decision maker after the end of the review process. Our goal is to jointly select the stopping feature set Γ, stopping features $R_1, \ldots, R_\Gamma$, and classification decision $D_{(\Gamma,R_1,\ldots,R_\Gamma)}$ to accurately classify each data instance i, while minimizing the cost incurred from reviewing individual features and switching between feature sets. The associated optimization problem is described as follows:

$$\min_{\Gamma,R_1,\ldots,R_\Gamma,D_{(\Gamma,R_1,\ldots,R_\Gamma)}} J(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)}) \quad (1)$$

where

$$J(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)}) \triangleq \mathbb{E}\left\{ \sum_{q=1}^{\Gamma} \sum_{n=1}^{R_q} c_{q,n} + \sum_{q=1}^{\Gamma-1} s_{q,q+1} + \sum_{l=1}^{L} \sum_{m=1}^{L} M_{ml} P\left(D_{(\Gamma,R_1,\ldots,R_\Gamma)} = l, C_m\right) \right\}. \quad (2)$$

The first term in Eq. (1) represents the expected total cost of reviewing features belonging to different feature sets. The second term denotes the expected cost of switching between different feature sets. The last term expresses the expected cost of the classification decision reached by the decision maker at the end of the review process. Table I outlines some of the most commonly used notation in this paper, while Fig. 1 graphically illustrates the proposed adaptive sequential review process.

The above problem statement requires the following assumptions on the ordering of features, feature sets, and the switching process between the feature sets.

(A1) The ordering of the feature sets is fixed and given. For simplicity, we begin our review process from the first available feature set, which corresponds to $\mathbf{f}_1$ without loss of generality. Additionally, during switching between different feature sets, the given ordering of feature sets is respected, i.e., $\mathbf{f}_1 \rightarrow \mathbf{f}_2 \rightarrow \ldots \mathbf{f}_Q$.

(A2) Features within a specific feature set q can be reviewed with respect to any ordering, resulting in $N_q!$ possible orderings (see Section IV for an efficient heuristic ordering). Determining the optimum feature ordering is out of the scope of the current work and part of our future research directions.

TABLE I: Notation overview.

Symbol	Explanation	Symbol	Explanation
Q	#feature sets	$P(f_{q,n}\|C_l)$	probability of $f_{q,n}$ given C_l
$f_{q,n}$	nth feature in qth feature set	$c_{q,n}$	cost of reviewing $f_{q,n}$
N_q	#features in qth feature set	$s_{q,q'}$	switching cost for qth and q'th feature sets
$\mathbf{f}_q$	qth feature set	M_{ml}	misclassification cost for C_m and C_l
N	total #features	Γ	stopping feature set
C_l	class l	R_q	stopping feature at qth feature set
P_l	probability of class l	$D_{(\Gamma,R_1,\ldots,R_\Gamma)}$	classification decision

Running Example: To facilitate the understanding of our methodology, we introduce the following simple example. Our goal is to reach a classification decision (i.e., C_1 : pothole, C_2 : code violation, C_3 : parking enforcement) for a civil issue report posted on a government 2.0 platform such as SeeClickFix [15] based on $Q = 3$ feature sets: textual description features $\mathbf{f}_1$, location features $\mathbf{f}_2$, and image features $\mathbf{f}_3$. There are various possible paths to reach a classification decision, e.g., $(\Gamma = 2, R_1 = 3, R_2 = 4)$, $(\Gamma = 3, R_1 = 1, R_2 = 2, R_3 = 6)$. For $M_{ml} = 1, m \neq l$, $M_{ll} = 0$, $c_{q,n} = 1$, and $s_{q,q'} = 1$, our proposed adaptive sequential review process sets forth to find the path that uses the least number of features and feature sets to reach an accurate classification decision. An interesting byproduct of the proposed approach is that different civil issue reports may end up requiring different values of Γ and R_q, thus improving the interpretability of the classification decision.

C. Reformulation of the Objective Function

Consider the posterior probability vector $\boldsymbol{\pi}_{q,n} \triangleq [\pi^1_{q,n}, \pi^2_{q,n}, \ldots, \pi^L_{q,n}]$, where the element $\pi^l_{q,n} \triangleq p(C_l|f_{1,1}, \ldots, f_{1,R_1}, f_{2,1}, \ldots, f_{q,n})$ denotes the posterior probability of an instance belonging to class C_l given that the decision maker has reviewed features $f_{1,1}, \ldots, f_{1,R_1}, f_{2,1}, \ldots, f_{q,n}$. The posterior probability vector $\boldsymbol{\pi}_{q,n}$ constitutes a sufficient statistic of the accumulated information until feature $f_{q,n}$, and can be recursively computed as new features are reviewed according to the following Bayesian update expression:

$$\boldsymbol{\pi}_{q,n} = \frac{\boldsymbol{\pi}_{q,n-1} \operatorname{diag}(\Delta_{q,n}(f_{q,n}))}{\boldsymbol{\pi}_{q,n-1} \Delta^{\top}_{q,n}(f_{q,n})}. \quad (3)$$

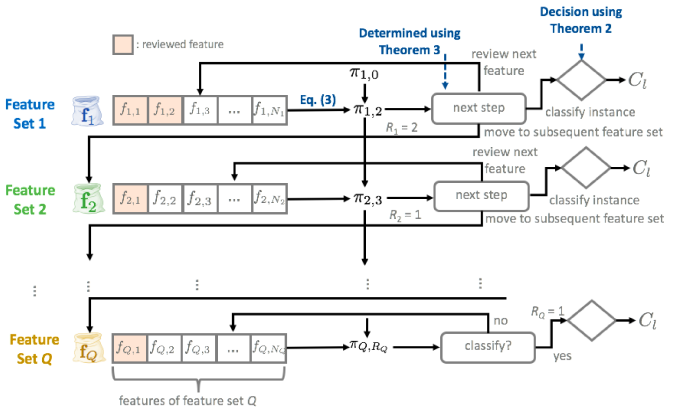

Fig. 1: Proposed adaptive sequential review process.

Note that $\Delta_{q,n}(f_{q,n}) \triangleq [P(f_{q,n}|C_1), P(f_{q,n}|C_2), \ldots, P(f_{q,n}|C_L)]$, and diag($\mathbf{v}$) represents a diagonal matrix of appropriate dimensions with diagonal elements being the elements in vector $\mathbf{v}$. Additionally, the posterior probability vector is initialized as $\boldsymbol{\pi}_{1,0} \triangleq [p_1, p_2, \ldots, p_L]$, and $\boldsymbol{\pi}_{q,0} \triangleq \boldsymbol{\pi}_{q-1,R_{q-1}}$ for $q \in \{2, \ldots, \Gamma\}$.

Next, consider the indicator function $\mathbb{1}_A$ for any event A, where $\mathbb{1}_A = 1$ if A occurs and zero, otherwise. Furthermore, recall that $\Gamma \in \{1, \ldots, Q\}$, $R_1 \in \{0, 1, 2, \ldots, N_1\}$, $\ldots$, $R_Q \in \{0, 1, 2, \ldots, N_Q\}$, all of which correspond to finite sets. Taking this fact into consideration, we then define:

$$x_{(\Gamma,R_1,\ldots,R_\Gamma)} \triangleq \sum_{q=1}^{Q} \sum_{n_1=0}^{N_1} \cdots \sum_{n_q=0}^{N_q} x_{(q,n_1,\ldots,n_q)} \mathbb{1}_{\{\Gamma=q, R_1=n_1, \ldots, R_q=n_q\}}. \tag{4}$$

Exploiting Eq. (4) and the posterior probability vector definition, the average cost in Eq. (2) can be rewritten in the form given in Lemma 1.

Lemma 1. *The objective function* $J(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)})$ *can be expressed as follows:*

$$J(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)}) = \mathbb{E}\left\{ \sum_{q=1}^{\Gamma} \sum_{n=1}^{R_q} c_{q,n} + \sum_{q=1}^{\Gamma-1} s_{q,q+1} + \sum_{l=1}^{L} \boldsymbol{\pi}_{\Gamma,R_\Gamma} \mathbf{M}_l^\top \mathbb{1}_{\{D_{(\Gamma,R_1,\ldots,R_\Gamma)}=l\}} \right\}, \tag{5}$$

where $\mathbf{M}_l \triangleq [M_{1l}, M_{2l}, \ldots, M_{Ll}]$.

III. Main Results

In this section, we derive the optimum solution of the optimization problem defined in Eq. (1). In Section III-A, we derive the optimum classification decision rule, while in Section III-B, we give the optimum feature review rule.

A. Optimum Classification Rule

Assume fixed stopping feature set Γ, and stopping features $R_1, \ldots, R_\Gamma$. In this case, we observe that the classification decision rule contributes only to the last term in the expression of Eq. (5). As shown in Theorem 2, the optimum classification decision rule assigns each instance i to the class that gives rise to the smallest misclassification cost. This optimum rule can be found by deriving a lower bound for the last term in Eq. (5).

Theorem 2. *Assuming fixed stopping feature set* Γ*, and stopping features* $R_1, \ldots, R_\Gamma$*, the optimum classification decision rule* $D^*_{(\Gamma,R_1,\ldots,R_\Gamma)}$ *is:*

$$D^*_{(\Gamma,R_1,\ldots,R_\Gamma)} = \arg\min_{1 \leqslant l \leqslant L} \left[\boldsymbol{\pi}_{\Gamma,R_\Gamma} \mathbf{M}_l^\top \right], \tag{6}$$

Since Theorem 2 derives the optimum classification decision rule, we note that:

$$\begin{aligned} &J(\Gamma, R_1, \ldots, R_\Gamma, D^*_{(\Gamma,R_1,\ldots,R_\Gamma)}) \\ &\leqslant J(\Gamma, R_1, \ldots, R_\Gamma, D_{(\Gamma,R_1,\ldots,R_\Gamma)}) \end{aligned} \tag{7}$$

This implies that we can simplify the objective function in Eq. (5) as follows:

$$J(\Gamma, R_1, \ldots, R_\Gamma, D^*_{(\Gamma,R_1,\ldots,R_\Gamma)}) = \mathbb{E}\left\{ \sum_{q=1}^{\Gamma} \sum_{n=1}^{R_q} c_{q,n} + \sum_{q=1}^{\Gamma-1} s_{q,q+1} + g(\boldsymbol{\pi}_{\Gamma,R_\Gamma}) \right\}, \tag{8}$$

where $g(\boldsymbol{\pi}_{\Gamma,R_\Gamma}) \triangleq \min_{1 \leqslant l \leqslant L} \left[\boldsymbol{\pi}_{\Gamma,R_\Gamma} \mathbf{M}_l^\top \right]$.

Running Example: Applying the preceding result to the running example introduced in Section II-B, we observe that:

$$D^*_{(\Gamma,R_1,\ldots,R_\Gamma)} = \arg\min \left[1 - \pi^1_{\Gamma,R_\Gamma}, 1 - \pi^2_{\Gamma,R_\Gamma}, 1 - \pi^3_{\Gamma,R_\Gamma} \right]. \tag{9}$$

B. Optimum Feature Review Rule

Determining the optimum feature review rule involves minimizing the expression in Eq. (8) with respect to both Γ and $R_1, \ldots, R_\Gamma$. The structure of the problem enables us to use dynamic programming [16] to find such a rule, as summarized in Theorem 3.

Theorem 3. *The optimum feature review rule is:*

$$(\Gamma^*, R_1^*, \ldots, R_{\Gamma^*}^*) = \begin{cases} \arg\min_{j \in \{sc,sw,cr\}} \left[\bar{J}^j_{q,n}(\boldsymbol{\pi}_{q,n}) \right], \\ q \in [[1, Q-1]], n \in [[0, N_q - 1]], \\ \arg\min_{j \in \{sc,sw\}} \left[\bar{J}^j_{q,n}(\boldsymbol{\pi}_{q,n}) \right], \\ q \in [[1, Q-1]], n = N_q, \\ \arg\min_{j \in \{sc,cr\}} \left[\bar{J}^j_{q,n}(\boldsymbol{\pi}_{q,n}) \right], \\ q = Q, n \in [[0, N_Q - 1]], \\ \arg\min \left[\boldsymbol{\pi}_{q,n} \mathbf{M}_l^\top \right], q = Q, n = N_q, \end{cases} \tag{10}$$

where $\bar{J}^{sc}_{q,n}(\boldsymbol{\pi}_{q,n})$ *represents the optimum average cost of stopping the feature review process at the* n*th feature of the* q*th feature set, and reaching a classification decision,* $\bar{J}^{sw}_{q,n}(\boldsymbol{\pi}_{q,n})$ *denotes the optimum average cost of switching from the* q*th to the* $q+1$*th feature set,* $\bar{J}^{cr}_{q,n}(\boldsymbol{\pi}_{q,n})$ *is the optimum average cost of continuing the feature review process in the same feature set, and* $[[a, b]]$ *represents the interval of all integers between*

a and b. These functions are represented by the following set of equations:

$$\bar{J}^{sc}_{q,n}(\boldsymbol{\pi}_{q,n}) = g(\boldsymbol{\pi}_{q,n}), \tag{11}$$

$$\bar{J}^{sw}_{q,n}(\boldsymbol{\pi}_{q,n}) = s_{q,q+1} + c_{q+1,1} + \sum_{f_{q+1,1}} \bar{J}_{q+1,1}(\boldsymbol{\pi}_{q+1,1})\boldsymbol{\pi}_{q,n} \times \Delta^{\mathsf{T}}_{q+1,1}(f_{q+1,1}) \tag{12}$$

$$\bar{J}^{cr}_{q,n}(\boldsymbol{\pi}_{q,n}) = c_{q,n+1} + \sum_{f_{q,n+1}} \bar{J}_{q,n+1}(\boldsymbol{\pi}_{q,n+1})\boldsymbol{\pi}_{q,n}\Delta^{\mathsf{T}}_{q,n+1}(f_{q,n+1}), \tag{13}$$

$$\bar{J}_{q,n}(\boldsymbol{\pi}_{q,n}) = \begin{cases} \min_{j\in\{sc,sw,cr\}} \left[\bar{J}^{j}_{q,n}(\boldsymbol{\pi}_{q,n})\right], \\ q \in [[1, Q-1]], n \in [[0, N_q - 1]], \\ \min_{j\in\{sc,sw\}} \left[\bar{J}^{j}_{q,n}(\boldsymbol{\pi}_{q,n})\right], \\ q \in [[1, Q-1]], n = N_q, \\ \min_{j\in\{sc,cr\}} \left[\bar{J}^{j}_{q,n}(\boldsymbol{\pi}_{q,n})\right], \\ q = Q, n \in [[0, N_Q - 1]], \\ \min\left[\boldsymbol{\pi}_{q,n}\mathbf{M}^{\mathsf{T}}_{l}\right], q = Q, n = N_q. \end{cases} \tag{14}$$

The optimum feature review rule in Eq. (10) has a very intuitive structure. Specifically, if the decision maker has not reached the Qth feature set and there are remaining features in the current feature set, there are three available options given posterior probability $\boldsymbol{\pi}_{q,n}$: (i) stop reviewing features and classify the instance, (ii) switch to the next feature set and review the first feature, or (iii) review the next feature of the current feature set. If there are no more features to be reviewed in the current feature set, the decision maker must decide between the first two options above. When the decision maker reaches the Qth feature set, there are only two available options given posterior probability $\boldsymbol{\pi}_{q,n}$: (i) stop reviewing features and classify the instance, or (ii) review the next feature of the Qth feature set. Of course, if all feature sets and features have been reviewed, the decision maker has no option but to classify the instance.

Running Example: Applying the preceding result to the running example introduced in Section II-B, we observe that the optimum feature review rule selects the least number of feature sets and features to review to reach an accurate classification decision.

IV. Experiments

In order to evaluate the proposed methodology, we consider the problem of automatically classifying civil issue reports posted on the SeeClickFix platform. Our dataset consists of 529 civil issue reports collected from SeeClickFix, between Jan. $5, 2010$ and Feb. $10, 2018$ for Albany, New York. In our experiments, we consider $L = 4$ hypotheses, i.e., C_1 : "Parking Enforcement", C_2 : "Code Violation", C_3 : "Traffic Signal Repair", and C_4 : "Signs (Missing, Needed, or Damaged)". Each civil issue report is described by two distinct feature sets, $N_1 = 99$ textual features extracted from the title of the report ($\mathbf{f}_1$) and $N_2 = 1,507$ textual features extracted from the description of the report ($\mathbf{f}_2$). All features in the two feature sets indicate the number of times a specific word appears in either the title or the description of the civil issue report. During feature extraction, sentences were tokenized into unigrams, and punctuation, stopwords, and digits were removed. Additionally, each word was stemmed to its root (e.g., "parking" was replaced with "park"). Last but not least, words present in $\geqslant 95\%$ or $\leqslant 2\%$ of the reports, respectively, were excluded.

We conduct our experiments for varying feature costs $c_{q,n} = c \in \{10^{-4}, 10^{-2}, 0.1, 0.2, 0.3\}$, and switching cost $s_{q,q'} = s \in [0.1, \infty)$, and misclassification costs $M_{ml} = 1, \forall m \neq l$ with $M_{ll} = 0$. The conditional probabilities $P(f_{q,n}|C_l)$ were estimated via a smoothed maximum likelihood estimator as follows:

$$\hat{P}(f_{q,n}|C_l) = \frac{N(f_{q,n}, l) + 1}{\sum_{f'_{q,n}} N(f'_{q,n}, l) + V}, \tag{15}$$

where $N(f_{q,n}, l)$ is the number of civil issue reports belonging to class C_l that yield outcome $f_{q,n}$ after reviewing the nth feature in the qth feature set, and V is the maximum outcome among all features. The prior probabilities for each class C_l are estimated as follows:

$$P(C_l) = \frac{N_l}{\sum_{m=1}^{L} N_m}, \ l = 1, 2, \ldots, L. \tag{16}$$

We numerically compute the values of $\bar{J}_{q,n}(\boldsymbol{\pi}_{q,n})$ in Eq. (14) by first quantizing the interval $[0, 1]$ in increments of 0.1 such that $\sum_{l=1}^{L} \pi^{l}_{n,q} = 1$ and then proceeding with the evaluation of the expressions given in Theorem 3. We perform this computation once offline and store the results in a $(N_q + 1) \times d$ matrix for each feature set q, where N_q corresponds to the number of features in the qth feature set and d is the number of possible $\boldsymbol{\pi}_{n,q}$ vectors. We observe that for each feature set q, there are $N_q!$ possible orders by which features can be reviewed. Instead of considering all possible orders, we propose to sort features in each feature set in increasing order of the sum of type I and type II errors, scaled by the associated feature cost. This approach tends to promote features that are informative and cost–efficient at the same time.

We compare our methodology with the following baselines: (i) standard Bayesian detection [17] that uses the top $1, 5, 10, 50$, and all available features from all feature sets ordered using our heuristic ordering approach, (ii) Support Vector Machine with feature selection (SVM–FS) [18] with linear (SVM–L) and Gaussian (SVM–G) kernels, and PCA (SVM–PCA), and (iii) Random Forest (RF) with maximum tree depths $d = 5, 10$, and XG Boosting (XG–B) [19], [20]. Results are reported with respect to micro–averaged accuracy, macro–averaged precision, macro–averaged recall, and average number of features reviewed. For all baselines, we consider the unweighted average of the performance index when the two feature sets are used independently ("Average"), and the performance index when the two feature sets are regarded as a single feature set ("Combined"). Table II summarizes the results we obtain.

TABLE II: Performance comparison with baselines.

	Parameters	Accuracy	Precision	Recall	Avg. # feat.	
					Set 1	Set 2
Proposed Methodology	$s = 0.1$, $c = 10^{-2}$	0.9205	0.9230	0.9244	3.9314	3.7735e-03
	$s = 0.1$, $c = 0.1$	0.9263	0.9271	0.9294	2.7910	0
	$s = 0.1$, $c = 0.2$	**0.9413**	**0.9352**	**0.9415**	**2.6344**	0
	$s = 0.1$, $c = 0.3$	0.3362	0.2209	0.3914	0.4585	0
	$s \in [0.2, 0.5]$, $c = 10^{-4}$	0.9205	0.9230	0.9244	4.0484	0
	$s \in [0.2, 0.5]$, $c = 0.01$	0.9205	0.9230	0.9244	3.9314	0
	$s \in [0.2, 0.5]$, $c = 0.1$	0.9263	0.9271	0.9294	2.7910	0
	$s \in [0.2, 0.5]$, $c = 0.2$	**0.9413**	**0.9352**	**0.9415**	**2.6344**	0
	$s \in [0.2, 0.5]$, $c = 0.3$	0.3362	0.2209	0.3914	0.4585	0
	$s \in [0.6, \infty)$, $c = 0.01$	0.9205	0.9230	0.9244	3.9314	0
	$s \in [0.6, \infty)$, $c = 0.1$	0.9263	0.9271	0.9294	2.7910	0
	$s \in [0.6, \infty)$, $c = 0.2$	**0.9413**	**0.9352**	**0.9415**	**2.6344**	0
	$s \in [0.6, \infty)$, $c = 0.3$	0.3362	0.2209	0.3914	0.4585	0
Bayesian Detection	All (Average)	0.9205	0.9230	0.9243	99	1507
	All (Combined)	0.9205	0.9230	0.9243	1606	
	Top 50 (Average)	0.9205	0.9230	0.9243	50	50
	Top 50 (Combined)	0.9205	0.9230	0.9243	50	
	Top 10 (Average)	0.9205	0.9230	0.9243	10	10
	Top 10 (Combined)	0.9205	0.9230	0.9243	10	
	Top 5 (Average)	0.9318	0.9260	0.9313	5	5
	Top 5 (Combined)	0.9111	0.9177	0.9143	5	
	Top 1 (Average)	0.6181	0.5049	0.5782	1	1
	Top 1 (Combined)	0.4539	0.3230	0.4757	1	
SVM	SVM-L (Average)	0.8875	0.8730	0.8837	99	1507
	SVM-L (Combined)	**0.9697**	0.9612	0.9663	**1606**	
	SVM-G (Average)	0.8487	0.8631	0.8503	99	1507
	SVM-G (Combined)	0.9678	**0.9639**	0.9688	**1606**	
	SVM-FS (Average)	0.6637	0.7334	0.6987	24	10
	SVM-Fs (Combined)	0.9470	0.9496	0.9467	6	
	SVM-PCA (Average)	0.8478	0.8633	0.8493	11	190
	SVM-PCA (Combined)	0.96	0.95	0.96	1606	
RF	d=5 (Average)	0.8497	0.8567	0.8530	99	1507
	d=5 (Combined)	0.9583	0.9452	0.9617	1606	
	d=10 (Average)	0.8771	0.8618	0.8710	99	1507
	d=10 (Combined)	0.9696	0.9627	**0.9708**	**1606**	
XG-B	All (Average)	0.8789	0.8653	0.8726	99	1507
	All (Combined)	0.96	0.96	0.96	1606	

Among all baselines, SVM–L (Combined) and SVM–G (Combined) achieve the highest accuracy and recall, respectively, using all $1,606$ features, but requiring ~ 609 times as many features as our proposed methodology for a mere 2.84% and 1.97% improvement, respectively. Similar is the case for RF (Combined) for $d = 10$, which achieves the highest recall. Our proposed methodology, on the other hand, achieves accurate classification by just examining on average $\lessapprox 5$ features from feature set $\mathbf{f}_1$. Additionally, it does not require reviewing any features from feature set $\mathbf{f}_2$.

V. Conclusions

We formulated the problem of joint heterogeneous feature selection and classification, in which a decision maker sequentially reviews features belonging to multiple feature sets until a classification decision is reached. We derived the optimum classification decision rule, which minimizes the average misclassification cost. We also derived the optimum feature review rule, which selects both the feature sets and features per set to be reviewed to achieve an accurate classification decision. We demonstrated the performance of the proposed methodology on the problem of civil issue reports classification using real–world SeeClickFix data, and showed that we can perform accurate classification by just reviewing ~ 2.6 features on average.

Extensions of the current work involve analyzing the structure of the optimum solution as in [21], theoretically deriving the average number and associated standard deviation of feature sets and features per set needed to reach an accurate classification decision, and identifying the optimum ordering of reviewing feature sets and associated features.

References

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[2] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[3] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," Stanford InfoLab, Tech. Rep., 1997.

[4] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Inf. Sci.*, vol. 181, pp. 1138–1152, 2011.

[5] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, April 2005.

[6] K. Chen, L. Wang, and H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 11, no. 03, pp. 417–445, 1997.

[7] S. Stüker, F. Metze, T. Schultz, and A. H. Waibel, "Integrating multilingual articulatory features into speech recognition," in *INTERSPEECH*, 2003.

[8] R. Rose and P. Momayyez, "Integration of multiple feature sets for reducing ambiguity in asr," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, April 2007, pp. 325–328.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[10] Y. Liyanage, M. Yao, C. Yong, D.-S. Zois, and C. Chelmis, "What Matters the Most? Optimal Quick Classification of Urban Issue Reports by Importance," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 106–110.

[11] D.-S. Zois, C. Yong, C. Chelmis, A. Kapodistria, and W. Lee, "Improving Monitoring of Participatory Civil Issue Requests through Optimal Online Classification," in *52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 2034–2038.

[12] Y. W. Liyanage, D.-S. Zois, C. Chelmis, and M. Yao, "Automating the Classification of Urban Issue Reports: an Optimal Stopping Approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3137–3141.

[13] Y. W. Liyanage, D.-S. Zois, and C. Chelmis, "On–The–Fly Feature Selection and Classification with Application to Civic Engagement Platforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3762–3766.

[14] I. Nazar, Y. Warahena Liyanage, D.-S. Zois, and C. Chelmis, "Automated Optimal Online Civil Issue Classification using Multiple Feature Sets," in *53rd Asilomar Conference on Signals, Systems, and Computers*, 2019.

[15] Mergel, I., "Distributed Democracy : SeeClickFix.Com for Crowdsourced Issue Reporting," Tech. Rep., 2012.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2005, vol. 1.

[17] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, 2004.

[18] S. Hirokawa, T. Suzuki, and T. Mine, "Machine Learning is Better Than Human to Satisfy Decision by Majority," in *International Conference on Web Intelligence*, 2017, pp. 694–701.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[21] Y. W. Liyanage, D.-S. Zois, and C. Chelmis, "On–the–Fly Joint Feature Selection and Classification," *arXiv preprint arXiv:2004.10245*, 2020.