Dynamic Feature Selection for Classification in Structured Environments

Sachini Piyoni Ekanayake*, Yasitha Warahena Liyanage*, Daphney–Stavroula Zois*
*Department of Electrical and Computer Engineering
University at Albany, SUNY, Albany, NY, USA
Emails: {sekanayake, yliyanage, dzois}@albany.edu

2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723311

***Abstract*—In many real–world applications, e.g., medical diagnosis, behavioral analysis, Bayesian networks are used to describe relationships between variables. However, such variables are not directly observable, but can be inferred through noisy but costly features. In this paper, our previously proposed framework of dynamic instance–wise feature selection and classification is extended to work with structured data instances, i.e., data instances where relationships between classification variables are represented using a known Bayesian network. The objective is to maximize classification accuracy while minimizing the total cost of selected features. To this end, starting from lowest degree nodes, the proposed method sequentially selects features for each variable in the Bayesian network and performs classification. The resulting classification decisions are propagated through the Bayesian network and used during the classification process of the remaining variables. The performance of the proposed method is illustrated on two datasets and its effectiveness is compared with existing methods.**

***Index Terms*—Sequential feature selection, inference, Bayesian networks, noisy observations, instance–wise decision making**

I. Introduction

In many application domains, Bayesian networks represented by directed acyclic graphs (DAG), are used to describe relationships between variables of interest [1]. For example, in medical diagnosis, the cancer Bayesian network [2] with five nodes, i.e., "*pollution*", "*smoker*", "*cancer*", "*X–ray*" and "*dyspnoea*", represents the factors that might affect a patients' chance of having cancer. Nonetheless, the Bayesian network variables are not directly observed, but instead can be inferred through access to noisy but costly features. For instance, emotion and personality traits can be extracted from various physiological signals collected via biomedical sensors (e.g., galvanic skin response (GSR), electrocardiogram (ECG), electroencephalogram (EEG)) and self–reported questionnaires [3]–[5].

The wide applicability of Bayesian networks in various domains (e.g., medical diagnosis [2], [6], [7], behavioral analysis [8]) has led to the design of various classifiers for such networks [9]–[13]. However, such methods typically assume that the Bayesian network represents the relationships between a single classification variable and a set of noisy features, and thus, the goal is to infer the value of such variable. On the other hand, prior work on Bayesian networks for supervised learning consider more than one classification variables [14]–[16], the relationships between which are represented through a known Bayesian network structure. The objective in this case is to infer the values of all classification variables in the network, exploiting their relationships. Still, such classification variables are assumed to be fully observable (i.e., not observed through noisy features). However, in many real–world applications, this assumption does not hold; in fact, classification variables are observed via noisy features, while at the same time, they exhibit relationships between them. For instance, the affective state and personality traits of an individual, which are *only observable* through noisy GSR, ECG, and EEG sensor data [3], are related and thus, can be represented by a Bayesian network with two classification variables. Finally, in time–sensitive applications like medical diagnosis, accessing features comes at a cost while exhibiting different informativeness (e.g., magnetic resonance imaging is a costly but informative operation). Thus, a mechanism to dynamically select features in a sequential manner to infer the values of classification variables in such settings is necessary.

In this paper, a method is proposed to dynamically select features for classification of structured data instances, i.e., data instances where relationships between classification variables are represented using a known Bayesian network. Building upon our prior work [17], which considers a single classification variable, the proposed approach dynamically selects a subset of the available features and reaches an appropriate classification decision for each variable in the Bayesian network during testing. Classification decisions are propagated through the Bayesian network starting from the lowest degree nodes and used during the classification of the remaining variables. The performance of the proposed approach is validated in a synthetic and a real–world dataset, and compared with established classification and feature selection methods with respect to accuracy, number of features selected, and time. The obtained results indicate the superiority of the proposed approach compared to such existing methods in the context of structured environments.

II. Preliminaries

In this section, the process of dynamically selecting features and reaching a classification decision is summarized for each variable in the Bayesian network. For more details, the reader is referred to our prior work [17].

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

Authorized licensed use limited to: UNIVERSITY AT ALBANY SUNY. Downloaded on May 17,2022 at 16:30:32 UTC from IEEE Xplore. Restrictions apply.

A. Description

Consider a known Bayesian network structure $\mathcal{G} = (X, E)$ described by a DAG. Here, $X \triangleq \{X_1, X_2, \ldots, X_n\}$ is the set of n nodes corresponding to n variables in $\mathcal{G}$, where each $X_i, i = 1, 2, \ldots, n$, is a categorical variable that can take multiple values. E is the set of directed edges that represent relationships between variables. A set $F \triangleq \{F_1^{X_1}, \ldots, F_{K_1}^{X_1}, F_1^{X_2}, \ldots, F_{K_2}^{X_2}, \ldots, F_1^{X_n}, \ldots, F_{K_n}^{X_n}\}$ of features is available, where $F_k^{X_i}, k = 1, 2, \ldots, K_i$, denotes the kth feature associated with variable X_i. Our objective is to infer the values of the variables in X by balancing classification accuracy with the total cost of selected features. Next, we state the following two important assumptions:

(A1) Features $F_k^{X_i}, k = 1, 2, \ldots, K_i$, associated with variable X_i are assumed conditionally independent given the variable X_i.

(A2) The ordering of the features $F_k^{X_i}, k = 1, 2, \ldots, K_i$, is fixed and given for each variable X_i.

B. Optimization Setup

Consider the pair of random variables (R_i, D_{R_i}) associated with class variable $X_i, i = 1, 2, \ldots, n$. Random variable $R_i \in \{0, 1, \ldots, K_i\}$ represents the last feature selected from the ordered set $F^{X_i} \triangleq \{F_1^{X_i}, \ldots, F_{K_i}^{X_i}\}$ associated with variable X_i. Random variable D_{R_i} represents the classification decision associated with variable X_i. Since X_i is a categorical variable, X_i belongs to one of N_i possible classes. As a result, D_{R_i} takes values in the set $\{1, 2, \ldots, N_i\}$. Associated with each feature $F_k^{X_i}$ is a cost $e_k^i, k = 1, 2, \ldots, K_i$. Furthermore, associated with a classification decision is cost M_{lm}^i, which represents the cost of selecting class $C_l^{X_i}, l = 1, 2, \ldots, N_i$, while the true class is $C_m^{X_i}, m = 1, 2, \ldots, N_i$. The objective is to sequentially select the minimum number R_i of features to reach a classification decision D_{R_i} by minimizing the following cost function:

$$J(R_i, D_{R_i}) = \mathbb{E}\left[\sum_{k=1}^{R_i} e_k^i\right] + \sum_{l=1}^{N_i}\sum_{m=1}^{N_i} M_{lm}^i P(D_{R_i} = l, C_m^{X_i}), \quad (1)$$

where $P(D_{R_i} = l, C_m^{X_i})$ represents the joint probability of selecting class $C_l^{X_i}$ while the true class is $C_m^{X_i}$. Here, the first expression denotes the total cost of evaluating R_i features in the sequential process, while the second expression penalizes missclassification errors.

C. Optimum Solution

To minimize the cost function in Eq. (1), we first find the optimum decision $D_{R_i}^*$ for a given R_i. Then, the reduced cost function, $J(R_i)$, depends only on R_i. Finally, we find the optimum R_i^* by minimizing $J(R_i)$ [17]. We refer to $D_{R_i}^*$ and R_i^* as optimum classification and feature selection strategies, respectively.

Consider the posterior probability $\pi_k \triangleq [\pi_k^1, \pi_k^2, \ldots, \pi_k^{N_i}]^T$, after selecting k out of K_i features associated with variable X_i. The probability $\pi_k^m \triangleq P(C_m^{X_i} | F_1^{X_i}, \ldots, F_k^{X_i})$ denotes the posterior probability of the m^{th} class, $m = 1, 2, \ldots, N_i$. At stage $k = 0$, $\pi_0 \triangleq [p_1, p_2, \ldots, p_{N_i}]^T$, where $P(C_m^{X_i}) = p_m, m = 1, 2, \ldots, N_i$. From Bayes' rule, as more features are sequentially selected, the posterior probability π_k^m is recursively updated as follows:

$$\pi_k^m = \frac{P(F_k^{X_i} | C_m^{X_i}) \pi_{k-1}^m}{P(F_k^{X_i} | C_1^{X_i}) \pi_{k-1}^1 + \ldots + P(F_k^{X_i} | C_{N_i}^{X_i}) \pi_{k-1}^{N_i}}. \quad (2)$$

Eq. (1) can be rewritten in terms of the posterior probability and the indicator function $\mathbb{1}_A$ (i.e., $\mathbb{1}_A \triangleq 1$ when event A occurs, and 0 otherwise) as follows:

$$J(R_i, D_{R_i}) = \mathbb{E}\left[\sum_{k=1}^{R_i} e_k^i + \sum_{l=1}^{N_i}\sum_{m=1}^{N_i} M_{lm}^i \pi_{R_i}^m \mathbb{1}_{D_{\{R_i = l\}}}\right]. \quad (3)$$

The optimum classification strategy $D_{R_i}^*$ for any given feature selection strategy R_i can be shown to be [17]:

$$D_{R_i}^* = \underset{1 \le l \le N_i}{\operatorname{argmin}} \left[(\mathbf{M}_l^i)^T \pi_{R_i}\right], \quad (4)$$

where $\mathbf{M}_l^i \triangleq [M_{1l}^i, M_{2l}^i, \ldots, M_{N_i l}^i]^T$. As a result, the cost function in Eq. (3) can be written as:

$$J(R_i) = \mathbb{E}\left[\sum_{k=1}^{R_i} e_k^i + g(\pi_{R_i})\right], \quad (5)$$

where $g(\pi_{R_i}) \triangleq \min_{1 \le l \le N_i}[(\mathbf{M}_l^i)^T \pi_{R_i}]$.

Finally, the optimum feature selection strategy R_i^* can be found by minimizing the cost function in Eq. (5) via dynamic programming [17]. Specifically, since there are K_i available features associated with variable X_i, there are maximum $K_i + 1$ stages for the associated dynamic programming equations:

$$L_k(\pi_k) = \min\left[g(\pi_k), \tilde{L}_k(\pi_k)\right], k = 0, \ldots, K_i - 1, \quad (6)$$

where,

$$\tilde{L}_k(\pi_k) = e_{k+1}^i + \sum_{F_{k+1}^{X_i}} L_{k+1}(\pi_{k+1}) \left(\Delta_{k+1}^T (F_{k+1}^{X_i}) \pi_k\right), \quad (7)$$

with $\Delta_k(F_k^{X_i}) \triangleq [P(F_k^{X_i} | C_1^{X_i}), \ldots, P(F_k^{X_i} | C_{N_i}^{X_i})]^T$ and $L_{K_i}(\pi_{K_i}) = g(\pi_{K_i})$.

III. Proposed Approach

In this section, the proposed approach to identify the values of all variables in the Bayesian network is described exploiting the results of Section II. Specifically, for each variable $X_i, i = 1, 2, \ldots, n$, in the Bayesian network $\mathcal{G}$, features associated with that particular variable are sequentially selected based on Eq. (6), and a final classification decision is reached using Eq. (4). The process begins by initializing the posterior probability π_0 for each variable X_i. If the cost of continuing the feature selection process is less than the cost of reaching a classification decision, the first feature in the set F^{X_i} is selected and the posterior probability is updated based on Eq. (2). The process is repeated until either a subset of features is selected or all of the available features are reviewed. In either case, a classification decision is reached using the updated posterior probability along with the optimum classification strategy of Eq. (4). The proposed approach consists of a training and a testing phase, as discussed next.

Authorized licensed use limited to: UNIVERSITY AT ALBANY SUNY. Downloaded on May 17,2022 at 16:30:32 UTC from IEEE Xplore. Restrictions apply.

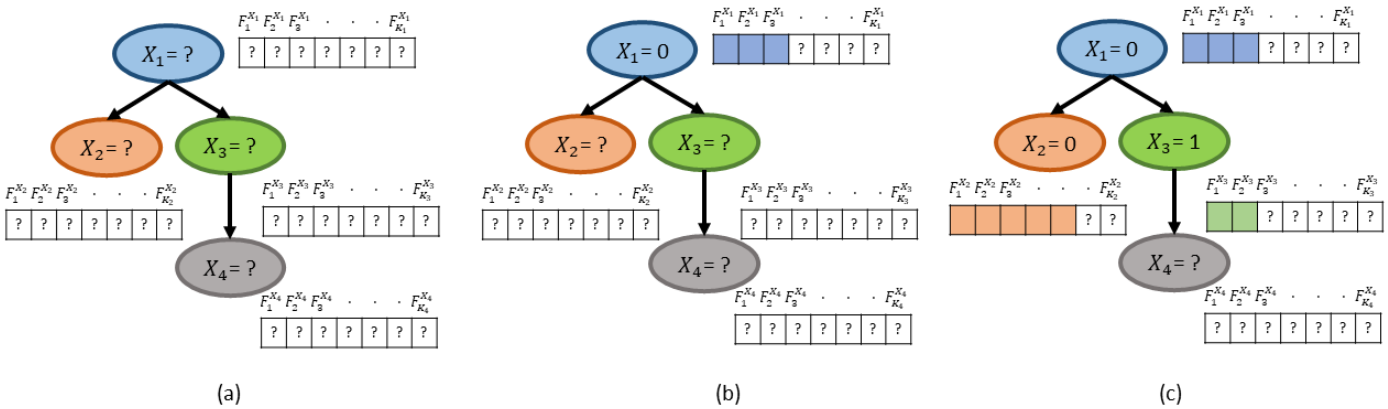

Fig. 1. Graphical representation of proposed approach for a data instance during testing. (a) Original Bayesian network, (b) Feature selection and classification for variables of in–degree 0, (c) Feature selection and classification for variables of in–degree greater than 0.

During training, the optimum classification and feature selection strategies described by Eqs. (4) and (6) are solved offline for each variable $X_i, i = 1, 2, ..., n$. Specifically, quantizing the interval $[0, 1]$ such that $\sum_{m=1}^{N_i} \pi_k^m = 1$, a $K_i \times d$ matrix is generated for each variable X_i, where d is the number of possible π_k vectors, and used to numerically solve Eqs. (4) and (6). This procedure is done for all variables in the Bayesian network $\mathcal{G}$.

During testing, the numerical solutions determined during training are used to dynamically select features and reach a classification decision for each variable $X_i, i = 1, 2, \ldots, n$, in the Bayesian network $\mathcal{G}$. Specifically, starting with nodes of in–degree 0 in the Bayesian network, the proposed approach sequentially selects features based on Eq. (6) and reaches a classification decision based on Eq. (4). Next, considering the structure of the Bayesian network, the proposed approach moves on to nodes with in–degree greater than 0 for which classification decisions for their parents have been reached. The classification decisions of the parents are incorporated into the posterior probability computation in Eq. (2) for each of their children. This process is repeated until all variables in the Bayesian network $\mathcal{G}$ have been assigned a classification decision. Fig. 1 shows a graphical representation of the proposed approach for an example Bayesian network $\mathcal{G}$ consisting of four binary random variables. Selected features at each round of the proposed method are highlighted. As illustrated, variables X_1, X_2, and X_3 are classified using $3, 5$ and 2 features, respectively.

IV. Numerical Results

In this section, experiments are conducted to illustrate the performance of the proposed approach. Before proceeding with presenting and discussing the relevant results, some practical considerations are reviewed. Specifically, for each variable $X_i, i = 1, 2, \ldots, n$, a maximum likelihood estimator is used to estimate $P(F_k^{X_i} | C_m^{X_i}) = \frac{S_{k,m}+1}{S_m+B}, k = 1, 2, \ldots, K_i, m = 1, 2, \ldots, N_i$ during training. Here $S_{k,m}$ denotes the number of instances that belong to class $C_m^{X_i}$ and $F_k^{X_i}$ takes a specific value, while S_m denotes the total number of instances that belong to class $C_m^{X_i}$. Prior probabilities $P(C_m^{X_i})$ are estimated as $\frac{S_m}{\sum_{m=1}^{N_i} S_m}, m = 1, \ldots, N_i, i = 1, \ldots, n$ during training. Features are ordered for each variable as per the increasing order of the sum of type I and II errors scaled by the cost coefficient of the kth feature of each variable to promote low cost features. As a result, feature orderings differ for each classification variable.

TABLE I
Accuracy and average number of features for same cost $e = 0.0001$ for all features in the synthetic dataset.

Variable	Accuracy	Average No. of features
X_1	0.9030	4.4119
X_2	0.8850	4.2400
$X3$	0.8830	4.3740
$X4$	0.8770	4.4930
X_5	0.8890	4.4074

Initially, a synthetic dataset containing five binary random variables $X \triangleq \{X_1, X_2, X_3, X_4, X_5\}$ is considered. Each variable X_i is associated with five features, i.e., $n = 5$ and $K_i = 5, i = 1, \ldots, 5$. Thus, the total size of the feature space is $|F| = 5 \times 5 = 25$. Each feature $F_k^{X_i}, k = 1, 2, \ldots, 5$, takes random discrete values in the set $\{1, 2, 3\}$. A dataset of $1,000$ data instances was created, where variables were generated as a linear combination of features plus noise. Specifically, $X_i \triangleq \sum_{k=1}^{5} c_k^{X_i} F_k^{X_i} + \sigma$, where $c_k^{X_i}$ are real–valued constants and $\sigma \in \{0, \ldots, 10\}$. Constants $c_k^{X_i}$ represent the relative importance of each feature to the corresponding variable, i.e., $c^{X_1} = [2, 2, 2, 0.2, 3], c^{X_2} = [2, 2, 0.1, 0.2, 3], c^{X_3} = [2, 2, 0.1, 0.2, 3], c^{X_4} = [2, 2, 0.1, 2, 3]$ and $c^{X_5} = [10, 10, 10, 10, 10]$. Then, X_i was converted to a binary variable using its median as the threshold. The five–fold cross validated results are reported in Table I for feature costs $e_k^i = 0.0001$ when all features have the same cost, i.e., $e_k^i = e, \forall k, i$. The results indicate that high accuracy can be achieved with less than five features. Further, we suspect that differences in accuracy and number of features arise due to wrong classification decisions propagated through the Bayesian network and different feature orderings per variable. Assigning larger cost values results in selecting less number of features, affecting accuracy and vise versa.

Next, the performance of the proposed approach is illustrated on a real–world dataset of student performance [18]. The dataset includes information about 649 students described by

Authorized licensed use limited to: UNIVERSITY AT ALBANY SUNY. Downloaded on May 17,2022 at 16:30:32 UTC from IEEE Xplore. Restrictions apply.

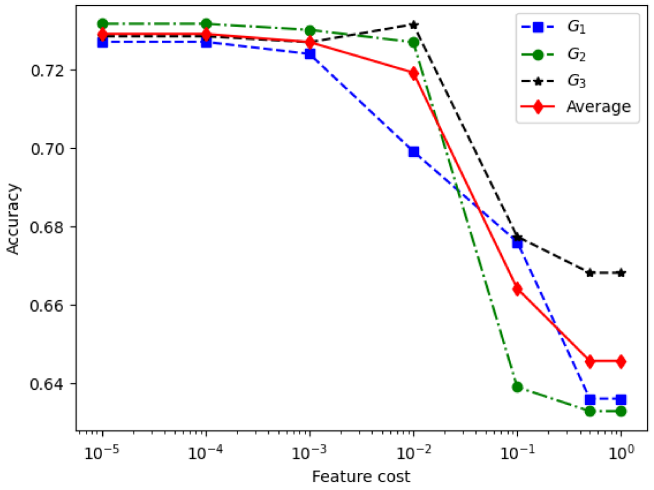

Fig. 2. Variation for accuracy of each variable G_1, G_2, G_3 as a function of feature cost $e \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0\}$ and average accuracy of the three variables.

30 features (e.g., demographic, social, school–related), and the goal is to infer three period grades, denoted as G_1, G_2, and G_3, respectively. To generate the Bayesian network for this dataset, a correlation–based analysis is followed and directed edges are drawn considering the immediate effect of cause variables [14]. Specifically, G_3 exhibits strong correlation with G_1 and G_2 [18], and thus, the resulting Bayesian network consists of the variables $X \triangleq \{G_1, G_2, G_3\}$, with two directed edges, $G_1 \rightarrow G_3$ and $G_2 \rightarrow G_3$. We treat G_1, G_2 and G_3 as binary random variables by setting them equal to one if $G_i \geqslant 11, i = 1, 2, 3$, and zero otherwise. The feature set is $F \triangleq \{F_1^{G_1}, \ldots, F_{30}^{G_1}, F_1^{G_2}, \ldots, F_{30}^{G_2}, F_1^{G_3}, \ldots, F_{30}^{G_3}\}$, where $K_i = 30, i = 1, 2, 3$. All experiments are conducted on a PC with Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz with 16 GB memory, and five–fold cross validated results are reported.

Fig. 2 illustrates the trade–off between accuracy and feature cost for different cost values under the assumption that features incur the same cost, i.e., $e_k^i = e, \forall k, i$. The accuracy for each variable as well as the average accuracy of variables G_1, G_2 and G_3 are reported. Results are reported in the case where misclassification costs are assumed the same, i.e., $M_{lm}^i = 1, \forall l \neq m, M_{ll}^i = 0, l, m \in \{1, \ldots, N_i\}$. As expected, using different feature costs leads to different accuracy levels, while acquiring more features typically leads to better accuracy, as long as such features are informative. In the case $e_k^i = e = 0.0001$, accuracy levels for G_1, G_2, and G_3 are 0.7271, 0.7317, and 0.7284, respectively, using on average 12.7470, 12.3127 and 6.6724 features, respectively. The number of features selected to classify variables G_1 and G_2 do not differ significantly. However, variable G_3 requires less features to be classified, since apart from feature information, the proposed approach exploits parent classification decisions. From here onwards, results are reported for $e_k^i = e = 0.0001$.

Fig. 3 shows the distribution of number of features used by data instances during testing. We observe that the majority of data instances use less than the available 30 features to reach a classification decision. Most of the time, G_1 and G_2 use 6 and 7 features respectively, while G_3 uses only 4 features.

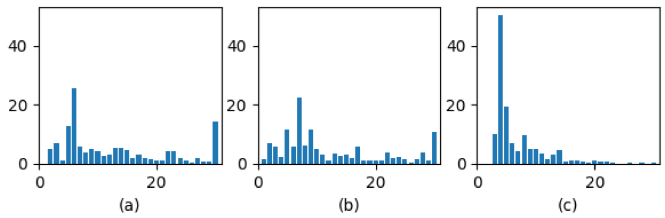

Fig. 3. Distribution of number of features used by data instances during testing for $e = 0.0001$. (a) G_1, (b) G_2, (c) G_3.

The frequency of features selected during testing (Fig. 4) demonstrates the importance of different features on classification decisions. For example, features "school" and "failures" are most often selected to classify all variables. Intuitively, student's school and number of past class failures have higher impact on student's grades compared to other features. To better understand the effect of each feature, different costs $e_k^i, k = 1, \ldots, 30$, were considered for each feature. Specifically, features were ranked based on the difficulty (i.e., easy, medium, hard) of collecting them. For instance, "absences" is easily accessible from school reports [18] (i.e., assigned rank $= 1$), while private information like "romantic" which describes the romantic relationship status of a student, was harder to acquire (i.e., assigned rank = 3). Cost was then computed as $e_k^i = \text{rank} \times 0.0001$ for each $k = 1, \ldots, 30$. This results in a mere 0.67% difference in accuracy using 27.88% more features on average compared to constant feature costs, which illustrates that the accuracy is robust for different feature costs. Furthermore, in order to preserve accuracy levels, in this case, the proposed approach tends to select more lower rank features (i.e., rank 1 or 2).

Finally, the proposed approach is compared with i) two widely used classifiers, Logistic Regression (LR) and Naive Bayes (NB), ii) one offline feature selection method, L–1 norm based feature selection (Lasso), and iii) our own prior work, ETANA [17]. Since none of these methods can explicitly handle multiple variables related through a Bayesian network, G_1, G_2, and G_3 are combined to create a single variable $G \triangleq (G_1, G_2, G_3)$ with 2^3 possible assignments. Table II reports five–fold cross validated average accuracy, average number of features selected, and training and testing time. We observe that both the proposed approach and ETANA use less than the 30 available features contrary to LR and NB. Lasso, on the other hand, even though it performs feature selection offline, still uses most of the available features. With respect to accuracy, we observe that the proposed method achieves the best accuracy (even with respect to ETANA). This is expected since the proposed method takes advantages of the relationships between variables in the Bayesian network. With respect to time, the proposed method runs much faster than ETANA, but slower than the baselines. Putting our findings into perspective, the proposed approach is able to achieve the best accuracy by selecting the most informative or low–cost features, while significantly decreasing the time requirements of ETANA.

Authorized licensed use limited to: UNIVERSITY AT ALBANY SUNY. Downloaded on May 17,2022 at 16:30:32 UTC from IEEE Xplore. Restrictions apply.

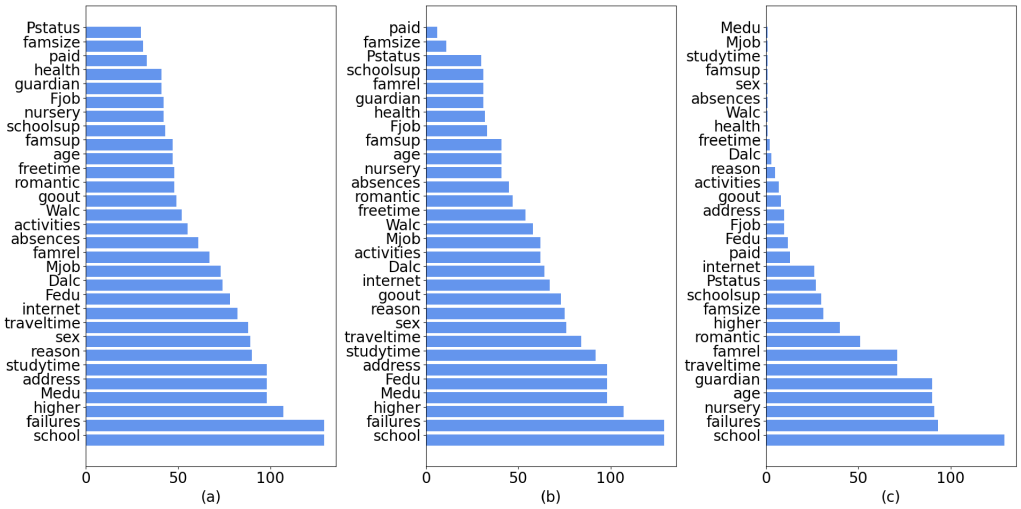

Fig. 4. Frequency of features selected during testing for different data instances. (a) G_1, (b) G_2, (c) G_3.

TABLE II
COMPARISON BETWEEN PROPOSED APPROACH AND BASELINES.

	Proposed approach	LR	NB	Lasso	ETANA
Accuracy	0.7291	0.6021	0.0816	0.6237	0.6191
Training (sec)	0.5726	0.0650	0.0012	0.0638	3093.9127
Testing (sec)	0.6398	0.0002	0.0007	0.0002	1.5129
Avg. # features	10.5774	30.0000	30.0000	29.2000	18.2159

V. CONCLUSIONS AND FUTURE WORK

In this paper, a dynamic feature selection method for classification of structured data instances is proposed. Relationships between variables of a data instance are described by a known Bayesian network. The proposed method dynamically selects features in a sequential manner during testing, and reaches a classification decision for each variable in the Bayesian network using a subset of the available features. Classification decisions are propagated through the Bayesian network during this process and used during the decision–making process of the remaining variables. The proposed approach is shown to outperform existing classification and feature selection methods in terms of accuracy and average number of features used. In future work, we plan to carefully analyze and address the effect of misclassifications as well as extend the proposed approach to account for correlated feature spaces.

REFERENCES

[1] J. Pearl, "Probabilistic reasoning in intelligent systems; networks of plausible inference," Tech. Rep., 1988.

[2] K. B. Korb and A. E. Nicholson, *Bayesian artificial intelligence*. CRC press, 2010.

[3] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[4] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[5] A. Rossi, E. Da Pozzo, D. Menicagli, C. Tremolanti, C. Priami, A. Sîrbu, D. A. Clifton, C. Martini, and D. Morelli, "A public dataset of 24-h multi-levels psycho-physiological responses in young healthy adults," *Data*, vol. 5, no. 4, p. 91, 2020.

[6] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

[7] S. Andreassen, R. Hovorka, J. Benn, K. G. Olesen, and E. R. Carson, "A model-based approach to insulin adjustment," in *AIME 91*. Springer, 1991, pp. 239–248.

[8] E. Nazerfard and D. J. Cook, "Using bayesian networks for daily activity prediction." in *AAAI workshop: plan, activity, and intent recognition*. Citeseer, 2013.

[9] N. Friedman and M. Goldszmidt, "Building classifiers using bayesian networks," in *Proceedings of the national conference on artificial intelligence*, 1996, pp. 1277–1284.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997.

[11] J. Su and H. Zhang, "Full bayesian network classifiers," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 897–904.

[12] J. Cheng and R. Greiner, "Comparing bayesian network classifiers," *arXiv preprint arXiv:1301.6684*, 2013.

[13] C. Bielza and P. Larranaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1–43, 2014.

[14] D. Heckerman, "A tutorial on learning with bayesian networks," *Innovations in Bayesian networks*, pp. 33–82, 2008.

[15] ——, "Bayesian networks for data mining," *Data mining and knowledge discovery*, vol. 1, no. 1, pp. 79–119, 1997.

[16] Y. Liu, W. Shi, and W. Czika, "Building bayesian network classifiers using the hpbnet procedure," in *Proceedings of SAS Global Forum*, 2017.

[17] Y. W. Liyanage, D.-S. Zois, and C. Chelmis, "Dynamic Instance–Wise Joint Feature Selection and Classification," *IEEE Transactions on Artificial Intelligence, arXiv preprint arXiv:2004.10245*, 2020.

[18] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Authorized licensed use limited to: UNIVERSITY AT ALBANY SUNY. Downloaded on May 17,2022 at 16:30:32 UTC from IEEE Xplore. Restrictions apply.