Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 1625 – 1636

doi:10.3906/elk-1301-181

Online feature selection and classification with incomplete data

Habil KALKAN*
Department of Computer Engineering, Faculty of Engineering, Süleyman Demirel University, Isparta, Turkey

Received: 24.01.2013 • **Accepted:** 06.03.2013 • **Published Online:** 07.11.2014 • **Printed:** 28.11.2014

Abstract: This paper presents a classification system in which learning, feature selection, and classification for incomplete data are simultaneously carried out in an online manner. Learning is conducted on a predefined model including the class-dependent mean vectors and correlation coefficients, which are obtained by incrementally processing the incoming observations with missing features. A nearest neighbor with a Gaussian mixture model, whose parameters are also estimated from the trained model, is used for classification. When a testing observation is received, the algorithm discards the missing attributes on the observation and ranks the available features by performing feature selection on the model that has been trained so far. The developed algorithm is tested on a benchmark dataset. The effect of missing features for online feature selection and classification are discussed and presented. The algorithm easily converges to the stable state of feature selection with similar accuracy results as those when using the complete and incomplete feature set with up to 50% missing data.

Key words: Online feature selection, classification, missing data, incremental learning

1. Introduction

In pattern recognition, the complete set of training data is given before system construction. Therefore, most of the learning algorithms encountered in the literature operate on a batch mode that requires the entire set of training data for learning. However, in many real-world applications for data mining, satellite imaging, time series prediction, etc., the data are provided progressively and the characteristics of the data and the environmental conditions (illumination, sensor failures, mechanical problems, etc.) may change during the acquisition. Consequently, batch-trained learning systems may not respond to the changing characteristics of the features. Therefore, learning should also be performed in an online fashion.

An online learning algorithm, which is also called incremental or continuous learning, processes one of the observations each time and removes it from the database after usage. Online learning algorithms have recently received a great deal of attention in the area of pattern recognition and they are mostly used in tracking and recognition [1,2]. Lehtonen et al. [3] proposed an online training system for the classification of electroencephalogram trials. They started with batch learning and then used a fixed-sized memory for the training samples of each class. The oldest samples in the training sets were then replaced with the newest incoming samples and the classifier was then trained with these updated sets. However, the replacement of the oldest samples may lead to forgetting the initial, possibly highly relevant, information for learning. Slavakis et al. [4] proposed an adaptive projection method in reproducing kernel Hilbert space for online classification. Hall

*Correspondence: habilkalkan@sdu.edu.tr

et al. [5] incrementally updated the eigenvectors and eigenvalues for principle component analysis. Moreover, Pang et al. [6] proposed an incremental local discriminant analysis algorithm by incrementally updating the between-scatter and within-scatter matrix. Similarly, Yang and Zhou [7] considered the evaluation of class priors in an online manner using the expectation-maximization (EM) algorithm.

The selection of relevant features is highly important in classification algorithms. However, many of the developed online learning algorithms operate on fixed selected features. The selection of features in an online manner was proposed to increase the accuracy of the online classification algorithms in [8,9]. Collins et al. [8] updated the relevant features for tracking moving objects. They adapted the best features at every frame of the image sequence. In addition, Grabner and Bischof [9] used the online-AdaBoost algorithm for selecting the relevant features at the time of testing. Recently, we proposed an online feature extraction and selection algorithm for the classification of 1-dimensional signals [10].

Problems in pattern recognition may be due to missing data. With the usage of a real dataset with, possibly, missing data, more attention has been given to handle this problem of missing data. Many approaches have been developed to deal with this problem. These approaches can be grouped into 2 main categories: deletion or imputation [11–13]. Deletion is basically discarding the observations with incomplete data. However, this approach may significantly decrease the number of available training data, causing rank-deficient covariance matrices that negatively affect the learning performance. Moreover, in the testing phase, one cannot discard the observation with missing features. On the other hand, imputation is defined as the process of completing the missing attributes with zeroes, statistical mean values or time-series predicted values, or the values obtained by the EM algorithm [14]. Mustafa et al. [15] adapted the EM algorithm to estimate the missing values in multispectral images for detecting forest growth. Kaya et al. [16] successfully used the EM, neural network (NN), and multiple imputation methods for imputing the missing values. Salberg [17] regarded the pixels of clouds and snow images as missing features and classified vegetation types using a modified version of the K-nearest neighbor, maximum likelihood, and Parzen classifiers. Marlin [18] used a subspace classification scheme by modifying the classifier input representation after performing single and multiple imputations. In contrast, Williams et al. [12] avoided single or multiple imputations by performing analytical integration procedures with a conditional density function, which was estimated with the Gaussian mixture model (GMM) and EM algorithm. However, these proposed estimation-based imputation algorithms require significant statistics about the dataset and this may cause a bias for the testing data. In addition to deletion and imputation, more comprehensive machine learning methods were used for handling missing data, such as NN ensembles [19,20], decision trees [21], fuzzy approaches [22,23], and support vector machines [24,25]. Although machine learning methods give high performance in estimating missing values, they cannot be used for online classification problems due to their computational complexity. To the best of our knowledge, online feature selection and classification with missing data has not yet been performed.

In this study, we propose a framework that performs learning, feature selection, and classification in an online fashion, even with a high level of incomplete data. Learning is performed by updating a predefined model by imputing the missing features with the corresponding mean values. Testing is performed after eliminating the missing attributes in the observed prototypes. The proposed feature selection is conducted before the classification phase using the updated model parameters. A nearest neighbor classifier with Gaussian mixture density is used and the robustness of this online learning framework for missing data is compared with the results obtained when using the complete dataset.

This paper is organized as follows. Section 2 describes the proposed online feature selection and clas-

sification algorithm for incomplete data. Experimental results and conclusions are given in Sections 3 and 4, respectively.

2. Online feature selection and classification algorithm

2.1. Online learning with missing features

Online learning (training) algorithms process each of the observations one by one and remove the processed observations from the database after the updating of predefined model parameters. The model approach presented in Figure 1 is used in this study. Assume that we have a set of class labeled incomplete data in a K class problem:

$$D = \left\{ (\mathbf{x_t},\, y_t, \rho_t) :\, \mathbf{x_t} \epsilon R^d, \qquad y_t \in \{1, 2, ..K\} \right\} , \tag{1}$$

where $\mathbf{x}_t$ denotes the feature vector of the tth observation, labeled as y_t. The observation index t is used for the design of the learning system in an online manner. The parameter ρ_t denotes the index vector, indicating the missing features where the values of 0 and 1 correspond to missing and existing features, respectively. For instance, the $\rho_t = [01001]$ vector corresponds to feature vector of $\mathbf{x}_t \epsilon R^5$, where the 1st, 3rd, and 5th features are missing.

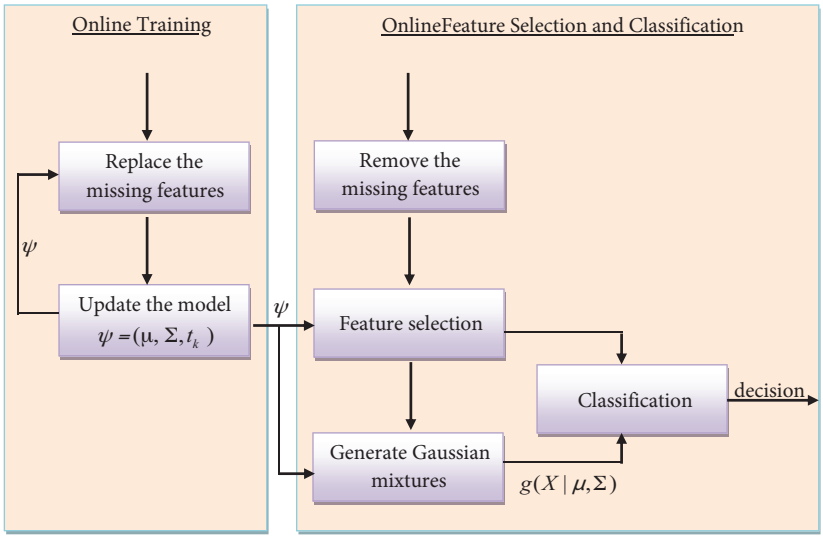

Figure 1. Flow chart of the proposed online learning, feature selection, and classification algorithm. Training includes the steps of substituting the missing features and then updating the model, whereas feature selection and testing include the elimination of missing features and then, using the trained model, selecting the features and classify the testing observation.

Let $\varphi = (\mu_k, \sum_k, t_k)_{k=1}^{K}$ be the proposed model, including the parameters of each K class, which are the mean vector μ_k, covariance matrix, $\sum_k$, and number of observations used, t_k. Assume that specifically $t-1$ number of training samples $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{t-1})$ has been given so far for class k and they were already discarded after updating the model parameters. When a new observation vector $\mathbf{x}_t \in R^d$ is received with a missing feature index vector, ρ_t, from the same class, the missing values in the feature vector are replaced by the corresponding values in the mean vector as follows:

$$x_t^{'}(i) = \begin{cases} x_t(i), & if\ \rho_t(i) = 1 \\ \mu_k(i), & if\ \rho_t(i) = 0 \end{cases} , \; i = 1, 2, \ldots, d\,. \tag{2}$$

This will allow only the nonmissing attributes in the feature vector to affect the learning model parameters. This replacement in the feature vector can also be validated by the EM algorithm [17].

After imputing the missing values in the feature vector, $\mathbf{x}_t^{'}$, the model parameter μ_k can be updated by:

$$\mu_k^t = \frac{1}{t}\left((t-1)\,\mu_k^{t-1} + \mathbf{x}_t^{'}\right). \tag{3}$$

As in [5], the new covariance matrix $\sum_k^t$ in the model could easily be updated as:

$$\sum\nolimits_k^t = \frac{t-1}{t}\sum\nolimits_k^{t-1} + \frac{t-1}{t^2}\mathbf{x}_t^{''}\mathbf{x}_t^{''T}, \tag{4}$$

where

$$\mathbf{x}_t^{''} = \mathbf{x}_t^{'} - \mu_k^t. \tag{5}$$

This incremental learning process is repeated whenever a new observation is received for training and the processed observation is removed from the database.

2.2. Online feature selection

In online classification algorithms, selection of the features should also be performed in an online fashion. Indeed, the relevance of the features may change during the incremental learning phase. Therefore, before testing, the most discriminative features should be detected to achieve better classification.

Recall that the incrementally updated model, $\varphi = (\mu_k, \sum_k, t_k)_{k=1}^K$, includes the class-specific mean vectors and covariance matrices. This information not only provides a statistic about the spread of features in the R^d space, but also provides the within-class and between-class scatter values of these features. Therefore, the relevance of these features could easily be obtained using the model when required. The relevance of these features for classification depends on how well they are separated from the corresponding features in each class. Various types of dissimilarity metrics can be used to get the relevance of features, such as Fisher, Mahalanobis, etc. [26]. However, the Fisher distance is used for its simplicity. The Fisher distance d of any variables between 2 classes, i and j, is described as:

$$d = \frac{|\mu_i - \mu_j|}{\sigma_i^2 + \sigma_j^2}, \tag{6}$$

where μ_i and σ_i are the mean and standard deviation values of the corresponding feature of the ith class. However, these metrics are mostly used for 2-class systems. Therefore, systems with more than 2 classes, say K classes, should be transformed into $h = \begin{pmatrix} K \\ 2 \end{pmatrix}$ 2-class problems. Next, the discrimination power of a specific feature in a K-class problem can be identified by evaluating the distances in these h number of class pairs.

Alternatively, in evaluating the discrimination power of the jth specific feature, the Fisher distance metric in Eq. (6) can be modified into a joint form to eliminate the pair-wise solution as follows:

$$d_j = \sum_{k=1}^{K} \frac{|\mu_{jk} - \tilde{\mu}_j|}{\sigma_{jk}^2}, \tag{7}$$

where $\tilde{\mu}_j$ is the common mean value of all of the classes for the jth feature:

$$\tilde{\mu}_j = \frac{1}{K}\sum_{k=1}^{K} \mu_{jk}, \tag{8}$$

and σ_{jk}^2 is the variance of the jth feature in class k, and can be obtained from model covariance matrix parameter $\sum_k$:

$$\sigma_{jk}^2 = \sum_k (j,j)\,. \tag{9}$$

The discrimination powers of each feature are identified using Eqs. (7) through (9) and sorted in decreasing order to be used in the classification, starting from the most discriminative ones. The vector of f_{index} in Figure 1 represents the discrimination rank of the nonmissing features in the test vector. Next, considering the rank in f_{index}, selection of the best features among the available ones can be possible.

2.3. Online classification with missing data

The suggested method is developed for those datasets that can be modeled with a Gaussian distribution. However, in many cases, Gaussian distributions may not efficiently model the data when compared to nonparametric methods. Instead, GMMs are capable of modeling a wide range of distributions. A GMM $f(\mathbf{X})$ is a weighted sum of G component Gaussian distributions and it is defined as:

$$f(\mathbf{X}) = \sum_{i=1}^{G} w_i g(\mathbf{X}|\,\mu_i, \sum_i) \text{ and } \sum_{i=1}^{G} w_i = 1\,, \tag{10}$$

where $g(\mathbf{X}|\,\mu_i \sum_i)$ is a multivariate Gaussian distribution, and w_i is the mixing coefficient.

Assuming a diagonal covariance matrix, Eq. (10) can be rewritten as:

$$f(\mathbf{X}) = \sum_{i=1}^{G} w_i \prod_{j=1}^{d} g(\mathbf{X}_j|\,\mu_{ij}, \sigma_{ij}^2), \tag{10}$$

where $\mathbf{X}_j$ denotes the jth feature of the training data, and μ_{ij} and σ_{ij}^2 denote the mean and variance of a univariate Gaussian distribution, respectively. The dataset under consideration may show various distributions (e.g., Poisson or binary) rather than normal. In such cases, a mixture Poisson model or mixture logistic model may be used instead of the GMM.

Assume that a test vector $\mathbf{x} \in R^d$ is received with missing feature index vector $\rho_{\mathbf{x}}$, and an observed feature vector $\tilde{x} \in R^{r \leq d}$ is obtained by discarding the missing features. The classification can be performed with r features only. However, instead of using all of the r features, the relevant features for classification are detected using a feature selection step, detailed in Section 2.2. The sorted features are incrementally included in the feature vector and tested by the classifier to obtain the best classification accuracy with a lower number of features. The features in the subset that give the best accuracy are regarded as more relevant features and the remaining features are dismissed.

When the missing and irrelevant features are ignored, Eq. (10) can be computed from partial data only [27] by:

$$f(\tilde{x}) = \sum_{i=1}^{G} w_i \prod_{j \in \rho_{\mathbf{x}}} g(\mathbf{X}_j|\,\mu_{ij}, \sigma_{ij}^2). \tag{11}$$

The developed GMM is trained by the EM algorithm [28] by considering the model parameters $\varphi = (\mu_k, \sum_k, t_k)_{k=1}^{K}$. For testing, a NN classifier based on Mahalanobis distance is used and the sample is assigned to the nearest class of mixture density. This distance metric is chosen for classification since it includes both the first- and second-order statistics and is well adapted for multinormal distributions.

3. Experimental results

Several standard benchmark datasets from the University of California-Irvine (UCI) Machine Learning Repository and the Statlog project are used. Brief descriptions about these datasets are given below.

- Statlog German Credit (*German*) Dataset: It classifies the people into low risk or high risk (2 classes) for credit. The set includes 1000 observations with 24 numerical attributes. The dataset is randomly divided into training and testing sets with 750 and 250 prototypes, respectively.
- Statlog Heart (*Heart*) Dataset: This set includes 270 prototypes collected from patients with and without heart disease (2 classes). There are 13 attributes in numerical and categorical forms. The categorical attributes are replaced by numerical values. The set is randomly divided into training and testing sets with 200 and 70 prototypes, respectively.
- UCI Ionosphere (*Ionosphere*) Dataset: This is a radar dataset and classifies the radar signals into 'good' or 'bad' depending on the existence of the free electrons in the ionosphere. The set includes 351 prototypes with 32 numerical attributes, of which 251 prototypes are assigned to a training set and 100 are assigned to a testing set, randomly.
- UCI Wisconsin (*Wisconsin*) Breast Cancer Dataset: This classifies biopsy samples into benign or malignant classes (2 classes). It includes a total of 699 prototypes with 9 attributes, where 16 prototypes include some missing attributes. These prototypes are ignored and the remaining 683 prototypes are randomly divided into training and testing sets with 400 and 283 prototypes, respectively.

Some of the attributes in the feature vector are randomly corrupted and considered as missing for testing the robustness of the developed algorithm for incomplete data. The frequency of missing parameters is controlled by a probabilistic parameter p, which takes the values of 0.75, 0.5, 0.25, and 0.0. A value of $p=$ 0.75 means that the attributes in both the training and testing datasets are missing with a probability of 0.75, while $p=$ 0.0 corresponds to the dataset with no missing information. The robustness of the proposed learning system for the presence of missing data is analyzed from 3 different perfectives, as follows.

3.1. Incremental learning with missing attributes

In an online learning phase, which is also called incremental or continuous learning, each incoming observation is processed one by one and removed from the databases after usage. In the case of missing attributes in the coming observation, the algorithm discards the missing attributes and trains the system with the valid attributes of the feature vector. In the testing phase, the classification is just performed with the valid attributes in the prototypes. It should be noted that the missing attributes may be different for each prototype due to randomness. The missing data rate affects the number of available attributes of each dataset. For example, for the *German* dataset with 24 complete attributes, missing rates of 25%, 50%, and 75% decrease the maximum available attribute number to 18, 12, and 6, respectively. In other words, it is not possible to represent the prototypes in the *German* dataset with more than 6 features if 75% of the attributes are missing.

In order to figure out the effect of observation for the stabilization of the learning algorithm, we test the learned system at times when 10% of the training data is processed. However, the testing is performed while using all of the prototypes reserved for the testing set. The procedure is repeated for different missing rates. Because of the random missing data generation, the algorithm may produce different results. Therefore, we randomly repeat the analysis 10 times and the average results are provided below (Figure 2).

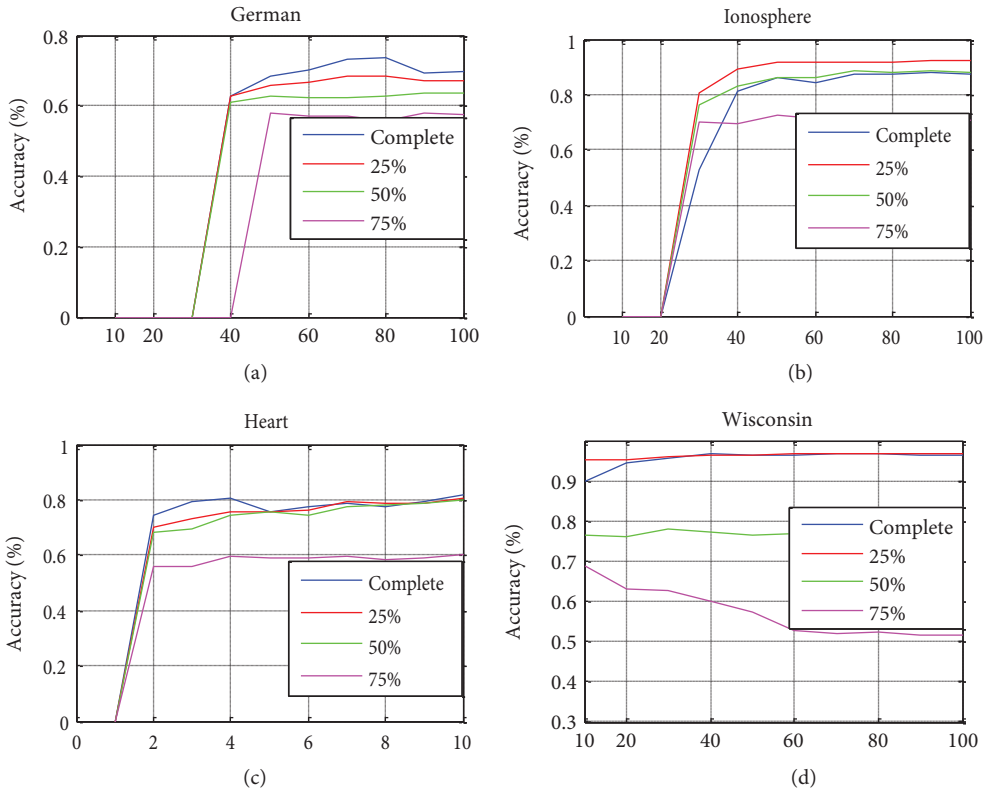

Figure 2. Classification accuracy curves using the online learning algorithm. The x-axis shows the percentage of the prototypes in the training set used for learning and the y-axis shows the classification accuracy. There are a total of 750, 251, 200, and 400 learning prototypes for the a) *German*, b) *Ionosphere*, c) *Heart*, and d) *Wisconsin* datasets, respectively.

To avoid the covariance matrix singularity at the beginning of the training, some experiments require at least 30%, 20%, or 10% of the prototype assigned to training sets, as observed in Figure 2. Using the proposed algorithm, it is possible to obtain similar classification accuracies with up to a 50% missing rate (Figures 2a–2c). For the *Wisconsin* dataset, we observe exactly the same accuracies with a complete and missing rate of 25% (Figure 2d). However, we observe poor accuracies when 75% of the attributes are missing in all of the datasets. It is observed, however, that the accuracy gap between the different missing rates (less than 50%) decreases when more prototypes are processed in the learning phase. An interesting result is observed for the *Ionosphere* dataset, where the accuracy curve obtained with a 25% missing rate is much better than the accuracy curve obtained with the complete dataset. This may be caused by the random elimination of the possible outliers in the *Ionosphere* dataset.

It should be noted that the missing values in the observation are removed for testing but are imputed with the corresponding mean values for training. However, the imputation of the missing attributes with the corresponding mean values may decrease the variance of these imputed features and the decrease in variance may negatively affect the classification performance. This decrease in variance will not be a significant problem if the missing attributes are equally encountered in the observations of each class. If only one class includes missing attributes and the rest have complete data, a regularization term should be included in the variance of these missing attributes.

3.2. Online feature selection with missing attributes

The forward feature selection algorithms in batch learning approaches incrementally construct the relevant feature subset from the available features. However, in online learning algorithms, especially when there are missing data, one cannot mention a globally valid discriminative feature subset, because the absence of an attribute that is also a member of the selected feature subset may completely change the order of these selected features. Therefore, the feature selection should be performed considering only the available attributes of the incoming observation. In that respect, we analyze the behavior of the feature selection algorithm at various missing data rates. Although we may have a higher number of features, we focus on selecting the best 4 features. The features included in the feature subset of the 4 dimensions clearly present the behavior of the online feature selection with missing data. Figure 3 shows the selected 4 best features at each learning phase with complete, 25%, and 50% data rate scenarios for the *German* (Figures 3a–3c), *Ionosphere* (Figures 3d–3f), *Heart* (Figures 3g–3i), and *Wisconsin* (Figures 3j–3l) datasets.

The first column of Figure 3 shows the selected features with complete datasets during the learning phase. In all of the datasets with complete data, fluctuations are observed at the beginning of the learning phase. However, after an adequate number of observations during the learning phase, the system is stabilized at feature selection. The adequate number depends on the dataset. The feature selection of the complete *German* dataset is stabilized after about 200 observations are processed. However, the stable states at feature selection for the *Ionosphere*, *Heart*, and *Wisconsin* datasets are achieved after about 100 observations are processed. Considering the data completeness, the difference with the sufficient prototype for stabilization at feature selection is related to the discrimination potential of the attributes. The poor accuracy curves, respectively, in the *German* dataset also explain the requirement of more prototypes for stabilization of feature selection for the *German* dataset.

It is observed from Figure 3 that the presence of missing features results in some instability for feature selection at the beginning of the learning phase. This is clearly observed from Figure 3 in the second column, obtained with 25% missing data. Compared to the selection with complete data, a higher number of observations is required to attain stable behavior with missing data. For example, for the *Wisconsin* dataset, the stable behaviors at feature selection are achieved at about 100, 120, and 200 observations with complete, 25%, and 50% missing data (fourth row in Figure 3), respectively. The inconsistency at feature selection is, respectively, higher in the scenario with 50% missing data for all of the datasets. However, we can get statistically significant information if we continue to learn with new observations. Figure 3 shows that the developed system is robust to the missing features and can converge to the system generated with complete data after a sufficient number of training samples are processed in the learning phase. The sufficient number depends on the missing data rate and the data themselves, and it seems that a higher missing rate leads to the requirement of more samples in the learning phase.

3.3. Online feature selection for classification

In the training phase, the algorithm discards the missing attributes and trains the system with the valid attributes of the feature vector. However, during the testing phase, a feature selection algorithm is performed to select the best features among the valid features in the test sequences. Assume that 6 of the 24 features in the *German* dataset are missing in the testing data and the classification can only be performed with the remaining 18 features. Instead of using these 18 features, the feature selection step ranks these features by considering their discrimination potential and uses the optimum number of features for classifying the test prototype.

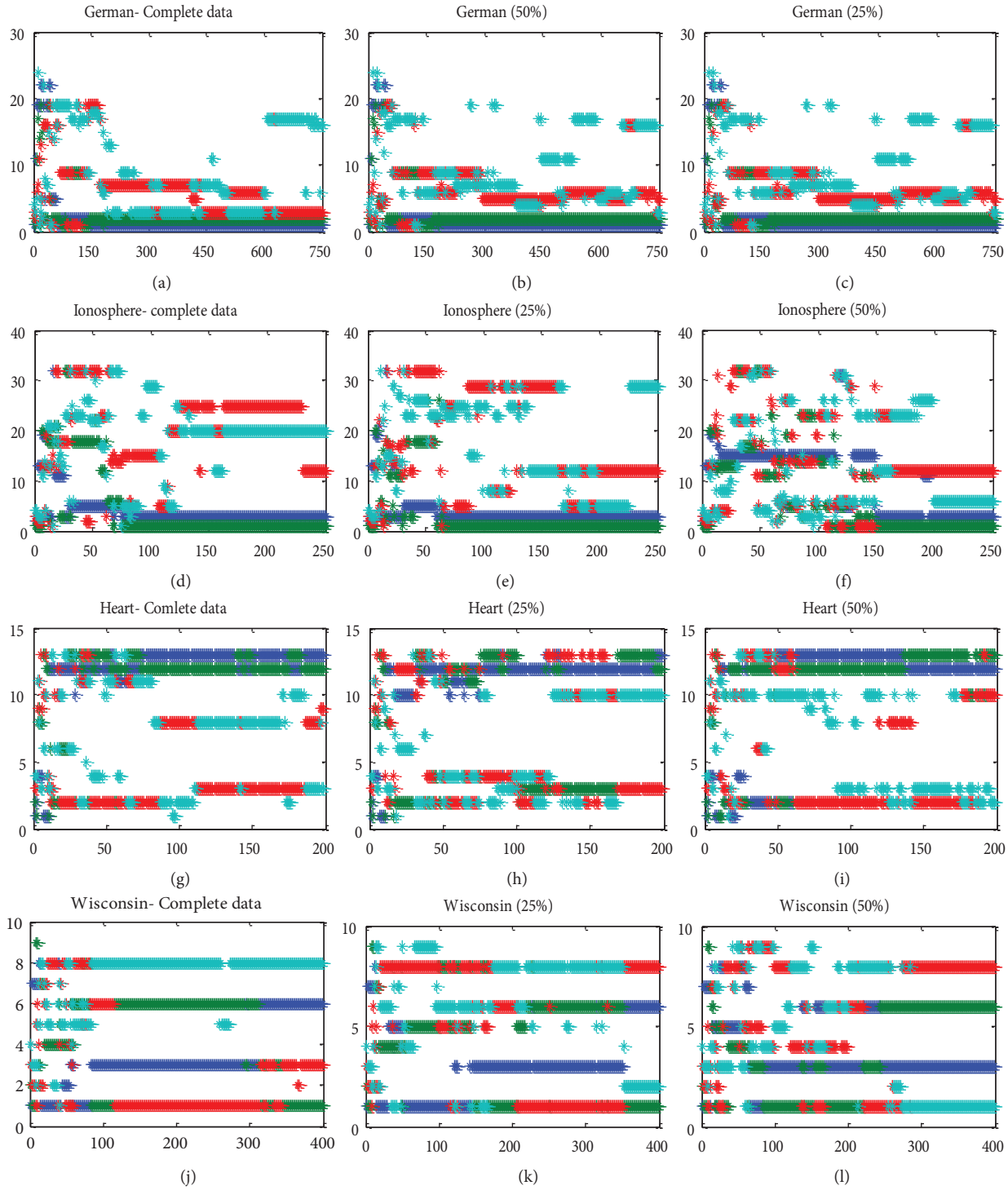

Figure 3. The selection of the best 4 features (blue = first, green = second, red = third, turquoise = fourth) during the learning phase. The x-axis and y-axis show the number of observations used at training and the features of these datasets, respectively. The experiment is repeated for complete (first column), 25% missing (second column), and 50% missing (third column) rates.

To analyze the effect of feature selection for classification with missing attributes, we evaluate the classification accuracies at each level of the feature selection. The random elimination of the attributes may change the behaviors of the proposed system. Therefore, we repeat the experiment and present the mean of the results obtained from 10 different runs. Figure 4 shows the effect of feature selection for classification at different missing data rates. Similar classification accuracy curves can be obtained with complete and missing data rates of up to 50% (Figures 4a–4c), except for the *Wisconsin* dataset (Figure 4d). In the *Ionosphere* dataset with 25% missing data, we obtain a similar accuracy curve compared to the curve obtained with complete data. This statement is also valid for the *Wisconsin* dataset. However, it should be noted that, in the case of missing data, a higher number of features are required to obtain the same performance in terms of accuracy.

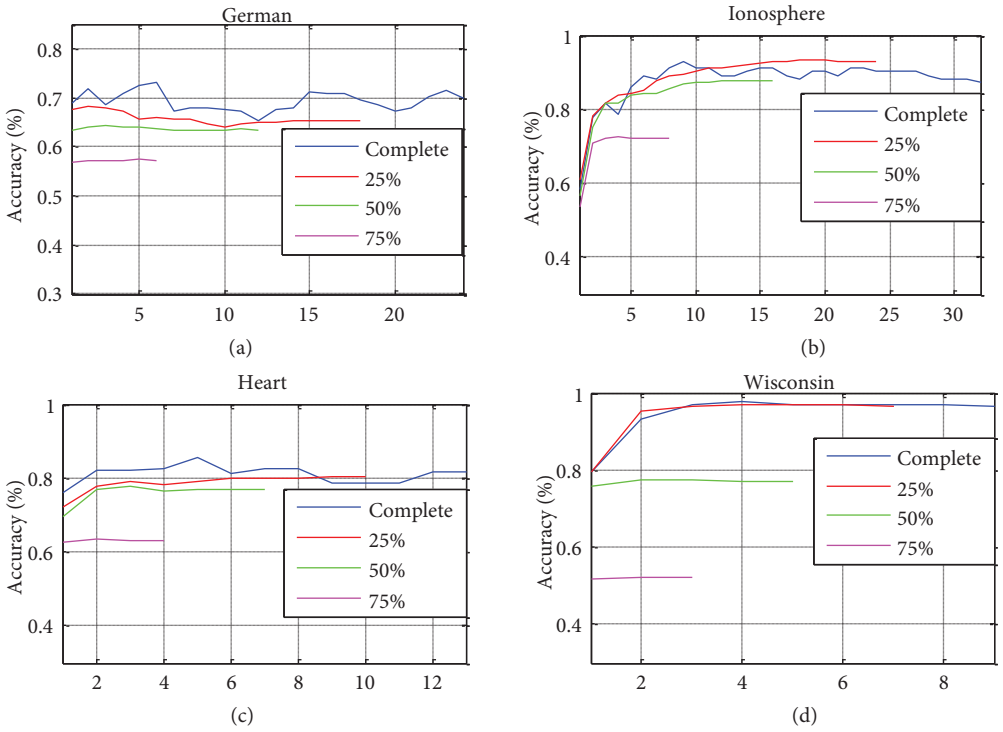

Figure 4. Classification accuracies obtained with various numbers of selected features at different missing data rates. The features are ranked using forward feature selection and are incrementally used for classification. The x-axis shows the number of features used for testing and the y-axis shows the obtained classification accuracy.

The highest accuracies achieved for each dataset are given in Table 4. For the *Ionosphere* and W*isconsin* datasets, we achieve nearly the same results with complete and 25% missing data. However, a higher number of features is required to obtain these results with 25% missing data compared to complete data. For the *Ionosphere* dataset, we obtain a slightly better accuracy curve with 25% missing data (with 19 features). These results show the ability of the proposed algorithm to adapt to the changing structure of the observation samples in an online manner. For achieving the optimal accuracies with missing data, the proposed algorithm compensates for the classification accuracy by using a higher number of features and more prototypes. The need for more features and prototypes depends on the rate of missing data and the discrimination power of each attribute of the dataset.

Table. Maximum classification rates at various missing data rates. The feature subset sizes giving these accuracies are given in parentheses.

Missing rates				
Dataset	Complete	25%	50%	75%
German	0.73 (6)	0.68 (2)	0.64 (3)	0.57 (5)
Ionosphere	0.93 (9)	0.93 (19)	0.87 (11)	0.73 (4)
Heart	0.86 (5)	0.80 (10)	0.77 (3)	0.63 (2)
Wisconsin	0.98 (4)	0.97 (4)	0.77 (2)	0.52 (3)

4. Conclusions

In this paper, an online feature selection and classification algorithm based on correlation analysis is proposed for the presence of missing data. Instead of discarding the observation with missing data, the developed algorithm discards the missing attributes of each observation and considers the valid attributes for learning and testing. The learned system behaves in an unstable manner at the beginning of the training phase but converges to stability when an adequate number of observations are used in training.

In addition to online learning, an online feature selection procedure is also performed at any time during the learning phase before classification. The selected features may change during the learning phase depending on the learned observation and missing data rate of this observation. The data distributions are modeled with a GMM and the Mahalanobis distance between the test sample and the underlying distribution is selected as classification criteria.

The learning, feature selection, and classification ability of the proposed algorithm are tested with the data at different levels of missing features and promising classification performance is achieved, even with the data that have a high percent of missing attributes. The missing data in the observation cause instability in the feature selection and classification. However, after processing an adequate number of observations, the system reaches stable behavior in the feature selection and classification. It is observed that the adequate number of features is related to the missing data rate and the relevance of the attributes in data.

The proposed algorithm, which is applied to 4 benchmark datasets, is robust to the changing characteristics of the observations with complete and incomplete attributes, and it can also be applied to other kinds of pattern recognition problems.

References

[1] Le KC, Kriegman D. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In: IEEE 2005 Computer Society Conference on Computer Vision and Pattern Recognition; 20–26 June 2005; San Diego, CA, USA. pp. 852–859.

[2] Jepson A, Fleet D, El-Maraghi T. Robust online appearance models for visual tracking. IEEE T Pattern Anal 2008; 25: 1296–1311.

[3] Lehtonen J, Jylänki P, Kauhanen L, Sams M. Online classification of single EEG trials during finger movements. IEEE T Bio-Med Eng 2008; 55: 713–720.

[4] Slavakis K, Theodoridis S, Yamada I. Online kernel-based classification using adaptive projection algorithms. IEEE T Signal Process 2008; 56: 2781–2796.

[5] Hall P, Marshall D, Martin R. Incremental Eigenanalysis for classification. In: 1998 British Machine Vision Conference; Southampton, UK. pp. 286–295.

[6] Pang S, Ozawa S, Kasabov N. One-pass incremental membership authentication by face classification. In: 2004 International Conference on Bioinformatics and its Applications; Fort Lauderdale, FL, USA. pp. 155–161.

[7] Yang C, Zhou J. Non-stationary data sequence classification using online class priors estimation, Pattern Recogn 2008; 41: 2656–2664.

[8] Collins RT, Liu Y, Leordeanu M. Online selection of discriminative tracking features. IEEE T Pattern Anal 2005; 27: 1631–1643.

[9] Grabner H, Bischof H. On-line boosting and vision. In: IEEE 2006 Computer Society Conference on Computer Vision and Pattern Recognition; 17–22 June 2006; New York, NY, USA. pp. 260–267.

[10] Kalkan H, Cetişli B. Online feature selection and classification. In: IEEE 2011 International Conference on Acoustics, Speech and Signal Processing; 22–27 May 2011; Prague, Czech Republic. pp. 2124–2127.

[11] Wang LL, Qu JJ, Siong X, Xianjun H, Yong X, Nianzeng C. A new method for retrieving band 6 of Aqua MODIS. IEEE Geosci Remote 2006; 3: 267–270.

[12] Williams D, Liao X, Xue Y, Carin L, Krishnapuram B. On classification with incomplete data, IEEE T Pattern Anal 2007; 29: 427–436.

[13] Laencina PJG, Gómez JLS, Vidal ARF. Pattern classification with missing data: a review. Neural Comput Appl 2010; 19: 263–282.

[14] Tsuda K, Akaho S, Asai K. The EM algorithm for kernel matrix completion with auxiliary data. J Mach Learn Res 2003; 4: 67–81.

[15] Mustafa YT, Stein A, Tolpekin V. Improving forest growth estimates using a Bayesian network approach. Photogramm Eng Rem S 2012; 78: 45–45.

[16] Kaya Y, Yesilova A, Almali MN. An application of expectation and maximization, multiple imputation and neural network methods for missing value. WASJ 2010; 9: 561–566.

[17] Salberg AB. Land Cover classification of cloud-contaminated multitemporal high-resolution images. IEEE T Geosci Remote 2011; 49: 377–387.

[18] Marlin BM. Missing data problems in machine learning, PhD, University of Toronto, Toronto, Canada, 2008.

[19] Sharpe PK, Solly RJ. Dealing with missing values in neural network-based diagnostic systems. Neural Comput Appl 1995; 3: 73–77.

[20] Juszczak P, Duin RPW. Combining one-class classifiers to classify missing data. Lecture Notes Comput Sc 2004; 3077: 92–101.

[21] Clark P, Niblett T. The CN2 induction algorithm. Mach Learn 1989; 3: 261–283.

[22] Ishibuchi H, Miyazaki A, Kwon K, Tanaka H. Learning from incomplete training data with missing values and medical application. In: IEEE International Joint Conference on Neural Networks; 25–29 October 1993; Nagoya, Japan. pp. 1871–1874.

[23] Lim CP, Leong JH, Kuan MM. A hybrid neural network system for pattern classification tasks with missing features. IEEE T Pattern Anal 2005; 27: 648–65.

[24] Shivaswamy PK, Bhattacharyya C, Smola AJ. A second order cone programming formulation for classifying missing data. J Mach Learn Res 2006; 7:1283–1314.

[25] Chechik G, Heitz G, Elidan H, Abbeel P, Koller D. Max-margin classification with incomplete data. J Mach Learn Res 2008; 9: 1–21.

[26] Webb AR. Statistical Pattern Recognition. 2nd ed. London, UK: Wiley, 2002.

[27] Drygajlo A, El-Maliki M. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. In: IEEE 1998 International Conference on Acoustics, Speech and Signal Processing; 12–15 May 1998; Seattle, WA, USA. pp. 121–124.

[28] Lin TI. On fast supervised learnin.g for normal mixture models with missing information. Pattern Recogn 2006; 39: 1177–1187.