FUSION OF INCOMPLETE MULTI-SENSOR DATA FOR CLASSIFICATION

Lawrence Carin, David Williams, Xuejun Liao, Ya Xue

Duke University
Department of Electrical and Computer Engineering
Box 90291, Durham, NC 27708-0291

ABSTRACT

When multiple sensors are employed for a remote-sensing task, some sensors may not be deployed to all data points. This scenario results in incomplete data points caused by missing sensor data (features). This work addresses the problem of classification in such an incomplete-data framework. We develop a classification algorithm that combines the data from all sensors in a principled manner. An estimated density function that expresses the relationships among features from different sensors is used to integrate out missing sensor data. Promising experimental results are shown for a challenging land mine detection problem involving real (measured) data from four sensors.

1. INTRODUCTION

For some remote-sensing tasks, a desired level of performance may be unattainable by using any one sensor. By exploiting multiple sensors to capture different physics of a problem, the limitations of single-sensor approaches can be overcome [1]. For example, in a land mine detection task, an infrared sensor may be better suited for detecting buried mines, while radar sensors may be more useful for detecting surface mines.

When multiple sensors are employed, the possibility exists that some sensors will not be deployed to all data points. This scenario results in incomplete data points that are missing sensor data (features). This work addresses the problem of classification when faced with such incomplete data points.

Several (unappealing) approaches can be used to avoid dealing with missing sensor data. For example, one may choose to build a classifier using only those data points that are not missing any data. A serious flaw with this approach is that testing data points to be classified may not have data from all sensors. Such a method would be forced to resort to heuristic methods to handle incomplete testing data points in the classification stage. Moreover, there may be very few data points that actually possess data from all sensors.

An alternative approach could instead build a classifier using only data from any one single sensor. This method suffers from a drawback similar to that of the first approach: testing data points may not have data from the chosen sensor. Another issue with using the data from a single sensor is that it is unknown *a priori* which sensor is best. Moreover, a major motivating reason for using multiple sensors is that some sensors may possess better discriminating power for different types of data; that is, there may be no uniformly best sensor.

Another displeasing aspect of these heuristic approaches is that (a potentially significant amount of) data is simply ignored. In contrast, imputation approaches can exploit all of the available data (information). In unconditional mean imputation, missing features are imputed, or "filled in," with the mean values of the data that is present. After the missing features have been imputed, the data set would be "complete." From a theoretical point of view, this approach is unsatisfactory because the imputed data is treated identically to the truly present data. The variance or uncertainty of the previously missing data is not taken into account [2].

This work avoids these heuristic approaches by instead proposing a principled method to deal with incomplete data. In our approach, missing data is analytically integrated out. This integration is made possible by exploiting the relationships of features from different sensors via a density function that is estimated from the observed data.

The remainder of this paper is organized as follows. Our proposed incomplete-data classification algorithm is presented in Section 2. Section 3 contains details of a four-sensor land mine data set for which experimental results are shown in Section 4. Concluding remarks appear in Section 5.

2. CLASSIFICATION WITH INCOMPLETE DATA

Assume we have an incomplete labeled data set

$$\mathcal{D}_l = \{(\mathbf{x}_i, y_i, m_i) : \mathbf{x}_i \in \mathbb{R}^d, x_{ij} \text{ missing } \forall j \in m_i\}_{i=1}^N \quad (1)$$

where $\mathbf{x}_i$ is the i-th data point, labeled as $y_i \in \{-1, 1\}$; the features in $\mathbf{x}_i$ indexed by m_i (*i.e.*, $x_{ij}, j \in m_i$) are missing. Each $\mathbf{x}_i$ has its own (possibly unique) set of missing features, m_i. In multi-sensor applications, missing features are the result of undeployed sensors that have not collected data.

In logistic regression, the probability of label y_i given $\mathbf{x}_i$ is $p\left(y_i|\mathbf{x}_i\right) = \sigma(y_i \mathbf{w}^T \mathbf{x}_i)$, where $\sigma(\nu) = (1 + \exp(-\nu))^{-1}$ is the sigmoid (or logistic) function, and $\mathbf{w}$ constitutes a classifier. We partition $\mathbf{x}_i$ into its observed and missing parts,

$\mathbf{x}_i = [\mathbf{x}_i^{o_i}; \mathbf{x}_i^{m_i}]$ where $\mathbf{x}_i^{o_i} = [x_{ij}, j \in o_i]^T$, $\mathbf{x}_i^{m_i} = [x_{ij}, j \in m_i]^T$, and $o_i = \{1, \cdots, d\} \setminus m_i$ is the (complementary) set of observed features in $\mathbf{x}_i$. We apply the same partition to $\mathbf{w}$ to obtain $\mathbf{w} = [\mathbf{w}_{o_i}; \mathbf{w}_{m_i}]$, yielding

$$p\left(y_i|\mathbf{x}_i\right) = \sigma(y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \nu_i)) \tag{2}$$

where $\nu_i = \mathbf{w}_{m_i}^T \mathbf{x}_i^{m_i}$. Because $\mathbf{x}_i^{m_i}$ (and hence ν_i) is missing, (2) cannot be evaluated. Instead the needed probability of y_i given the observed features $\mathbf{x}_i^{o_i}$ can be written as

$$\begin{aligned} p\left(y_i|\mathbf{x}_i^{o_i}\right) &= \int p\left(y_i|\mathbf{x}_i^{m_i}, \mathbf{x}_i^{o_i}\right) p\left(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}\right) d\mathbf{x}_i^{m_i} \quad (3) \\ &= \int \sigma(y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \nu_i)) p\left(\nu_i|\mathbf{x}_i^{o_i}\right) d\nu_i \quad (4) \end{aligned}$$

It is important to note that the integral in (3) is in general multi-dimensional, while the integral in (4) is one-dimensional. To perform the integration in (4), $p\left(\nu_i|\mathbf{x}_i^{o_i}\right)$ must be known. We first assume that $p\left(\mathbf{x}_i\right)$ is a (K-component) Gaussian mixture model (GMM):

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\left[\begin{array}{c} \mathbf{x}_i^{o_i} \\ \mathbf{x}_i^{m_i} \end{array}\right]; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{5}$$

where $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, and

$$\boldsymbol{\mu}_k = \left[\begin{array}{c} \boldsymbol{\mu}_k^{o_i} \\ \boldsymbol{\mu}_k^{m_i} \end{array}\right], \qquad \boldsymbol{\Sigma}_k = \left[\begin{array}{cc} \boldsymbol{\Sigma}_k^{o_i o_i} & \left(\boldsymbol{\Sigma}_k^{m_i o_i}\right)^T \\ \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{array}\right]. \tag{6}$$

Because of the linear relation $\nu_i = \mathbf{w}_{m_i}^T \mathbf{x}_i^{m_i}$, $p\left(\nu_i|\mathbf{x}_i^{o_i}\right)$ is also a GMM,

$$p\left(\nu_i|\mathbf{x}_i^{o_i}\right) = \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right), \tag{7}$$

with the parameters

$$\delta_k^i = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_\ell^{o_i}, \boldsymbol{\Sigma}_\ell^{o_i o_i}\right)} \tag{8}$$

$$\zeta_k^i = \mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i \tag{9}$$

$$\alpha_k^i = \sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i}} \tag{10}$$

$$\boldsymbol{\xi}_k^i = \boldsymbol{\mu}_k^{m_i} + \boldsymbol{\Sigma}_k^{m_i o_i} \left(\boldsymbol{\Sigma}_k^{o_i o_i}\right)^{-1} \left(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}\right) \tag{11}$$

$$\boldsymbol{\Omega}_k^i = \boldsymbol{\Sigma}_k^{m_i m_i} - \boldsymbol{\Sigma}_k^{m_i o_i} \left(\boldsymbol{\Sigma}_k^{o_i o_i}\right)^{-1} \left(\boldsymbol{\Sigma}_k^{m_i o_i}\right)^T \tag{12}$$

where

$$\mathcal{G}(\nu_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\nu_i^2}{2}\right\} \tag{13}$$

denotes a standard univariate Gaussian density function with zero mean and unit variance.

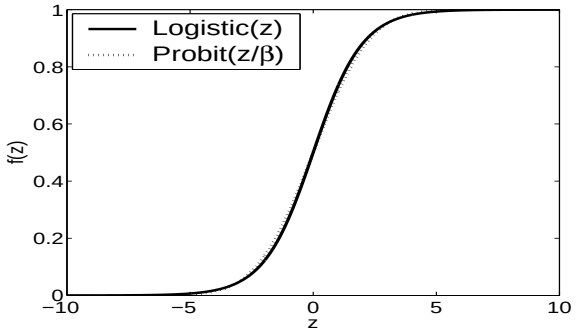

Fig. 1. Illustration of the accuracy of the approximation made between the logistic function and the (scaled) probit function.

We approximate the sigmoid (logistic) function as the cdf of a Gaussian (*i.e.*, a probit function)

$$\sigma(\nu) \approx \int_{-\infty}^{\nu} \mathcal{G}\left(\frac{z}{\beta}\right) dz \tag{14}$$

where $\beta = \frac{\pi}{\sqrt{3}}$. The accuracy of this approximation is shown in Figure 1.

Substituting (7) and (14) into (4), we obtain

$$\begin{aligned} & p\left(y_i|\mathbf{x}_i^{o_i}\right) \\ \approx & \int\int_{-\infty}^{y_i(\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \nu_i)} \mathcal{G}\left(\frac{z}{\beta}\right) dz \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i \\ \overset{a}{=} & \int\int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i \nu_i}{\beta}\right) dz' \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\nu_i - \zeta_k^i}{\alpha_k^i}\right) d\nu_i \\ \overset{b}{=} & \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \int \mathcal{G}\left(\frac{z' + y_i \nu_i}{\beta}\right) \\ & \qquad \times \mathcal{G}\left(\frac{y_i \nu_i - y_i \zeta_k^i}{y_i \alpha_k^i}\right) d\nu_i \, dz' \\ \overset{c}{=} & \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i \zeta_k^i}{\sqrt{(y_i \alpha_k^i)^2 + \beta^2}}\right) dz' \\ \overset{d}{=} & \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i}} \mathcal{G}\left(\frac{z' + y_i \zeta_k^i}{\beta} \frac{\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}\right) dz' \\ \overset{e}{=} & \sum_{k=1}^{K} \delta_k^i \int_{-\infty}^{\frac{y_i (\mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} + \zeta_k^i)\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}} \mathcal{G}\left(\frac{z}{\beta}\right) dz \\ \overset{f}{\approx} & \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\zeta_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{(\alpha_k^i)^2 + \beta^2}}\right) \end{aligned} \tag{15}$$

where equation a results from the change of variable $z' = z - y_i \nu_i$, equation b is due to exchanging the order of integrals and summation, equation c results because the convolution of two Gaussians is a Gaussian, equation d holds because $y_i^2 = 1$, equation e results from the change of variable

$z = \frac{(z' + y_i \zeta_k^i)\beta}{\sqrt{(\alpha_k^i)^2 + \beta^2}}$, and equation f is obtained by reverting to sigmoid representation. Thus we have expressed $p\left(y_i | \mathbf{x}_i^{o_i}\right)$ as a mixture of sigmoids. Substituting (9) and (10) into (15), we obtain the probability of y_i given only the observed portion of $\mathbf{x}_i$:

$$p\left(y_i | \mathbf{x}_i^{o_i}\right) \approx \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}} \right). \quad (16)$$

For the data set in (1), assuming the data points are independent of each other, we have the log-likelihood function

$$\begin{aligned} \ell(\mathbf{w}) &= \ln p\left(\{y_i\}_{i=1}^N | \{\mathbf{x}_i^{o_i}\}_{i=1}^N\right) \\ &\approx \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\mathbf{w}_{m_i}^T \boldsymbol{\xi}_k^i + \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i})}{\sqrt{\mathbf{w}_{m_i}^T \boldsymbol{\Omega}_k^i \mathbf{w}_{m_i} + \beta^2}} \right). \end{aligned} \quad (17)$$

Since the objective function (17) to be maximized is not concave, the solution may be trapped in local maxima. A good initialization is important, so we initialize $\mathbf{w}$ as follows. We "complete" the data set by replacing the missing features $\mathbf{x}_i^{m_i}$ with the conditional mean $\mathbb{E}(\mathbf{x}_i^{m_i} | \mathbf{x}_i^{o_i}) = \sum_{k=1}^{K} \delta_k^i \boldsymbol{\xi}_k^i$, where δ_k^i and $\boldsymbol{\xi}_k^i$ are defined in (8) and (11), respectively. This "completed" data set is then submitted to the standard logistic regression to obtain $\mathbf{w}_0$, which is the maximizer of

$$\sum_{i=1}^{N} \ln \sigma\left(y_i \mathbf{w}_{m_i}^T \sum_{k=1}^{K} \delta_k^i \boldsymbol{\xi}_k^i + y_i \mathbf{w}_{o_i}^T \mathbf{x}_i^{o_i} \right).$$

We then use $\mathbf{w}_0$ as the initialization of $\mathbf{w}$ in maximizing (17) by gradient ascent.

In summary, with only two assumptions—that $p(\mathbf{x}_i)$ is a GMM and that the sigmoid function can be approximated as the cdf of a Gaussian—the requisite integral in (3) can be computed analytically. (The GMM can be estimated from incomplete data using the variational Bayesian expectation-maximization algorithm, as outlined in [3].) As a result, the log-likelihood can be easily maximized to find the logistic regression classifier $\mathbf{w}$ in the presence of missing data. Thereafter, the class predictions of an unlabeled testing data point with incomplete (missing) features can also be computed trivially using (16).

3. MULTI-SENSOR DATA SET

The objective for the multi-sensor data set used in this work is to discriminate between land mines and clutter. The task is particularly challenging because the data set contains both metal and plastic mines, and also both buried and surface mines. The data set is comprised of data collected via four airborne sensors. The four sensors are a ground-penetrating radar (GPR) sensor, an electro-optic infrared (EOIR) sensor, a *Ku*-band synthetic aperature radar (SAR) sensor, and an *X*-band SAR sensor. Because each sensor is mounted on a different aircraft, the flight paths for each sensor are unique. This results in different areas of land — and hence different data points — missing sensor data where the flight path did not contain the location of the data point.

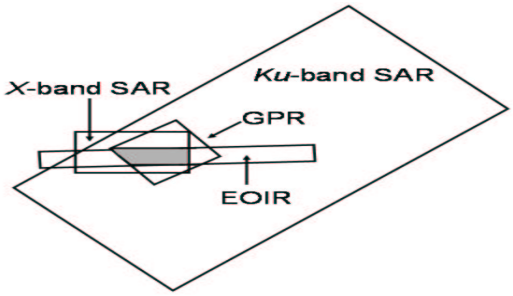

Fig. 2. Area interrogated by each of the four sensors. Data points located in the shaded region are characterized by data from all four sensors.

Table 1. Details of the multi-sensor data set.

Sensors	Number of Features	Number of Data Points		
		Total	Mines	Clutter
GPR	17	1754	150	1604
EOIR	6	1509	223	1286
Ku-band SAR	9	10058	259	9799
X-band SAR	9	1962	154	1808
At least one	—	10219	284	9935
All four	—	713	91	622

The result of the data collection (*i.e.*, sensor deployment) is one image per sensor, covering a large area of land. (The total area of land covered by the union of the four images was approximately 6×10^4 m^2.) Figure 2 shows the area interrogated by each of the sensors. As can be seen from the figure, the flight paths of the four sensors partially overlap with each other. An image registration algorithm (that is beyond the scope of this paper) is used to properly align the images; this allows a particular physical location to be established in multiple images.

For each of the sensor-produced images, a simple energy pre-screener is used to flag potential regions (points) of interest that may contain land mines. All points of interest declared by the pre-screeners are then combined into a single pool. For each physically unique point of interest — hereafter referred to as a data point — an image "chip" is extracted from each image that contains the given data point. Different features are then extracted from the chips of different sensors. All subsequent classification work is based solely on these features. Details of the data set are summarized in Table 1.

4. EXPERIMENTAL RESULTS

We present the results of our classification algorithm in terms of the area under a receiver operating characteristic curve (AUC). The AUC is given by the Wilcoxon statistic [4]

$$\text{AUC} = (MN)^{-1} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}_{x_m > y_n} \tag{18}$$

where $x_1, \ldots, x_M$ are the classifier outputs of data belonging to class 1, $y_1, \ldots, y_N$ are the classifier outputs of data belonging to class -1, and $\mathbf{1}$ is an indicator function. The AUC is a more informative measure of performance than accuracy when there exists a substantial class imbalance in the data (*e.g.*, when the proportion of data points that are mines is significantly smaller than the proportion that are clutter).

We seek to evaluate our proposed classification algorithm, which employs a form of sensor fusion. To this end, we compare our proposed method with six alternative methods. Each of the six alternative methods — which avoid the issue of missing data — uses a standard complete-data logistic regression classifier. The particular data each method uses to build the classifier will be different, however. The first alternative approach uses all data points, with the incomplete data handled via unconditional mean imputation. If $\mathbf{x}_i$ is missing feature a (*i.e.*, $a \in m_i$), unconditional mean imputation makes the substitution

$$x_{ia} \leftarrow \mathbb{E}[x_{ia}] = \frac{\sum_{j=1}^{N} x_{ja} \mathbf{1}_{a \in o_j}}{\sum_{\ell=1}^{N} \mathbf{1}_{a \in o_\ell}}. \tag{19}$$

The second alternative method builds a classifier using only those (complete) data points that are characterized by data from all four sensors. Each of the final four methods builds a classifier using data from only a single sensor.

To permit comparisons to the alternative approaches, we must restrict ourselves to the case where all testing data is complete, characterized by features from all four sensors. It should be emphasized, however, that our proposed approach can easily handle incomplete testing data. The alternative methods — with the exception of the mean imputation method — cannot handle incomplete testing data.

Since only a limited number of data points (713) have data from all four sensors, we perform "leave-one-out" testing in the following manner. All incomplete data points are treated as training data. One of the 713 data points with data from all four sensors will be treated as testing data, while the remaining complete-data points will be used as additional training data. For each of the seven methods, a classifier will be learned using the relevant training data. The single held-out testing data point will subsequently be submitted to each classifier. This entire process will then be repeated for each of the 713 complete data points. The AUC can then be computed for each method by collecting each method's classifier outputs for each of the 713 data points.

Table 2. Experimental results for the multi-sensor data set.

METHOD	AUC
PROPOSED METHOD (ALL DATA)	0.78476
MEAN IMPUTATION (ALL DATA)	0.74801
ALL FOUR SENSORS (ONLY COMPLETE DATA)	0.75427
SINGLE SENSOR (ONLY GPR DATA)	0.77229
SINGLE SENSOR (ONLY EOIR DATA)	0.76503
SINGLE SENSOR (ONLY *Ku*-BAND SAR DATA)	0.74543
SINGLE SENSOR (ONLY *X*-BAND SAR DATA)	0.75287

Table 2 summarizes the results of applying this training and testing procedure. As can be seen from the table, the proposed method outperformed all of the competing methods. Interestingly, the approach employing only the *Ku*-band SAR sensor, which had the most training data points of any sensor, performed the worst. Despite the large quantity of ostensibly "weak" data from this sensor, our proposed approach still performed better than each individual sensor.

5. CONCLUSION

The main contribution of this work is the introduction of a principled algorithm to fuse data from multiple sensors when some data points are missing data from some sensors. Our proposed approach also easily handles incomplete testing data, a situation where several common heuristic methods encounter problems. Moreover, our algorithm exploits all data (information) in a principled manner. Preliminary results have demonstrated the advantage of the proposed approach on a challenging four-sensor land mine detection task. Future experiments will be conducted as more multi-sensor data becomes available.

6. REFERENCES

[1] Y. Zhang, L. Collins, H. Yu, C. Baum, and L. Carin, "Sensing of Unexploded Ordnance with Magnetometer and Induction Data: Theory and Signal Processing," *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 41, No. 5, pp. 1005-1015, 2003.

[2] S. Rässler, *The Impact of Multiple Imputation for DACSEIS* (DACSEIS Research Paper Series 5), University of Erlangen-Nürnberg, Nürnberg, Germany, 2004.

[3] D. Williams, X. Liao, Y. Xue, and L. Carin, "Incomplete-Data Classification using Logistic Regression," *Proc. 22nd Int'l Conf. Machine Learning (ICML)*, pp. 977-984, 2005.

[4] J. Hanley and B. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology 143*, pp. 29-36, 1982.