Robust Integration of Multiple Information Sources by View Completion

Shankara Subramanya, Baoxin Li and Huan Liu
Department of Computer Science and Engineering
Arizona State University, Tempe AZ-85287 USA
{shankarbs, baoxin.li, huan.liu}@asu.edu

Abstract

There are many applications where multiple data sources, each with its own features, are integrated in order to perform an inference task in an optimal way. Researchers have shown that for many tasks like webpage classification, image classification, and pattern recognition, combining data from multiple information sources yields significantly better results than using a single source. In these tasks each of the multiple data sources can be thought of as providing one view of the underlying object. However in many domains not all of the views are available for the available instances; some of the views would be missing. This problem of missing views affects the performance of the machine learning task. In this paper we provide a method of view completion to heuristically predict the missing views. We show that with view completion we are able to achieve significantly better results. We also show that by considering the information at a higher level in terms of views rather than considering them at a lower level in terms of features we are able to achieve better results. We demonstrate this by comparing our method with existing methods which consider the missing views problem as a missing value problem.

1 Introduction

Combining data from multiple information sources has received significant attention in recent years. Many researchers have shown that using multiple information sources is significantly better than using a single information source. In an application of battle field surveillance, for example, information from infrared sensors, video feeds, and Laser range finder can be combined for object recognition. For webpage classification, anchor text, images, and body text of a webpage can be combined and used. For classifying images, images as well as the text occurring with each image can be combined. Multiple information sources yield complementary information about the underlying object which results in improved performance.

In all the above examples, the object being observed is characterized by multiple sets of features. In the webpage classification example, the terms from the body text yield one set of features and the terms from the anchor text yield another set of features. Even though the features from each source are different, the feature sets represent the same underlying object instances and hence would be semantically related. For example, features derived from the body text and the anchor text of a webpage will most likely be semantically related since both sets of features are representing the same webpage. The anchor text would most likely contains terms which describe the webpage, which in turn is defined by the body text present in it. Each of the multiple sets of features can be considered a *'View'* and each object is characterized by multiple independent views.

Researchers have devised various ways of combining multiple views to achieve higher accuracy. In [8], text and image views were combined using a fusion SVM classifier. In [2], multiple views of a webpage were combined using a density based method. [11] combined multiple text and image views of a webpage using LSI. All these methods use the fusion of multiple views to achieve higher accuracy. Multiple views have also been used to improve classification performance using co-training [3]. Co-training improves classification learning by enforcing internal consistency among predicted classes of unlabeled objects based on different views. The idea behind these methods of combining multiple views is that the views though semantically related and possibly overlapping, provide useful complementary information.

2 Missing Views

One problem in combining multiple views is that in many domains not all the views are available for the

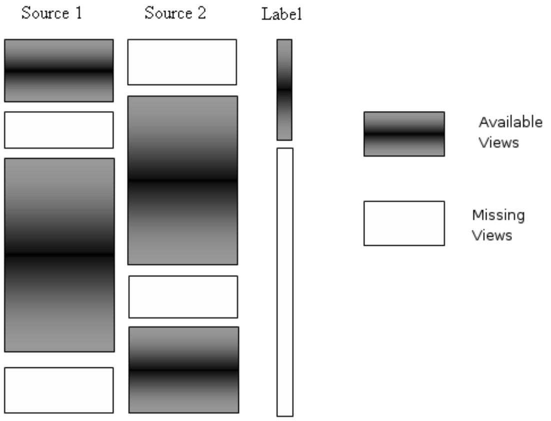

Figure 1. Schematic description - Missing view problem

available instances. Some of the instances will be incompletely characterized and only a subset of the views will be available for them. In the webpage classification example some of the webpages may not have any body text in them (for example, webpages containing images only). In this case the particular webpage will have the 'body text view' missing. Another example of missing views occurs in the domain of battle field surveillance. In many surveillance applications a large number of cheap sensors are usually employed to avoid high cost. These sensors could be prone to frequent failures. When a sensor fails, the view from that particular sensor would become unavailable. The problem of missing views is represented schematically in Figure 1.

This missing view problem is common to many domains. In all these scenarios the missing view will affect classification performance. In this work we address the problem of heuristically predicting missing views to obtain a complete characterization of the object. We then show experimentally that with this complete characterization we are able to achieve better classification performance.

3 Related Work

The analysis of data with missing values has for long been a well studied problem in statistics [7]. Imputation methods like maximum likelihood imputation and median imputation have been recommended to deal with the missing value problem. Schafer and Graham [9] provide a good survey of the methods for dealing with the missing value problem. The authors in [1] evaluate and compare the effect of different imputation methods which deal with missing values on classification accuracy. All these methods for handling missing values however work with the lower level features. The problem of missing views in a data integration scenario is different from the missing value problem. In the missing view problem, a complete feature set from a particular source would be unavailable. Hence, it would be not only possible but also desirable to deal with higher-level "features" in terms of views. We compare the view completion method with the missing value methods later in Section 7.

The problem of missing views has been considered previously by [6] in an active learning setting. They discussed the problem of deciding which additional features need to be acquired for an incompletely characterized object in order to improve the performance of the classifier. This approach of active learning is applicable only in the cases where active acquisition of features is possible. However in many real world scenarios it would be impossible to acquire additional features for some incompletely characterized objects. For example in an application using sensors, a missing view due to a failed or obstructed sensor would be impossible to acquire at a later stage as all the incoming data would be real time data. If a webpage does not have body text, anchor text, or images, then active learning would be of little use since there is no data in the source itself to acquire. So in data fusion such missed views are usually ignored though the missing views affect the final classification accuracy.

To overcome these problems, in the following sections we propose a method for completing the missing views based on canonical correlation analysis. We first define formally the problem of View Completion. We then propose an algorithm to select the closest matching value for the missing view using the views which are not missing and observed. We next show experimentally that with this view completion method we are able to achieve significant improvement in the classification accuracy.

4 View Completion

We define the multi-source model as follows. Each object is represented by two or more views. For any given object zero or more views may be missing. But for every object at least one view will be present. We also assume that typically there is a set of objects which have the class labels and there is a larger set of unlabeled data. Each view is represented by a set of feature vectors. This model is similar to the co-training model with the additional detail that some missing views of objects are missing.

An object with n views can be represented as an (n+1)-tuple. If f_1, f_2, ...f_n represent the n views and c the corresponding class label, then the instance i_n is represented as

$i_n = (f_1, f_2, ..., f_n, c)$

Specifically, an object instance with two views can be represented as a 3-tuple. If X and Y represent feature sets corresponding to two different views of an object. Each instance i is defined as follows.

$i = (x, y, c)$, where $x \in X$, $y \in Y$

Here $x \in X$ is a vector corresponding to features from first source, $y \in Y$ is a vector corresponding to features from second source and c is the class label. Either one of x or y can be $\emptyset$.

Let X_p and Y_p represent the feature sets corresponding to those instances which have features from both the views present. Let $X_{ym} \in X$ be the set of features corresponding to instances which have the other view missing, i.e., the corresponding $Y_{ym} = \varnothing$. Let I_{ym} be the set of all such instances. Our goal is now to find for each of such instances, the view $y_{ym} \in Y_{ym}$ using the available view $x_{ym} \in X_{ym}$ and the paired views from other instances X_p and Y_p. Similarly let I_{xm} represent the instances with the first view missing. Let $X_{xm} = \varnothing$ and $Y_{xm} \in Y$ represent the two views of these instances. We can then find $x_{xm} \in X_{xm}$ using $y_{xm} \in Y_{xm}$, X_p and Y_p.

To accomplish this, we develop a method of view completion which heuristically predicts the missing view(s) of the objects. Since this method uses only the available views and not the class label, it can be used on both the labeled and unlabeled data. To predict the missing view from the view which is available we first need to find the semantic relationship between the views. To find this semantic relationship we use the statistical technique of Canonical Correlation Analysis.

5 Canonical Correlation Analysis (CCA)

CCA attempts to find basis vectors for the two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized [5]. The correlation between the two sets of variables may not be visible in their original coordinate system. CCA tries to find a linear transformations for two sets variables such that in the transformed space they are maximally correlated.

The canonical correlation between any two data sets is defined as

$$\rho = max_{W_x,W_y}\ corr(F_x \cdot W_x, F_y \cdot W_y)$$

where F_x and F_y are the two sets of variables, and W_x and W_y are the basis vectors onto which F_x and F_y are projected, respectively. The equation for ρ can be rewritten as

$$\rho = max_{W_x,W_y}\ \frac{(F_x \cdot W_x, F_y \cdot W_y)}{(||F_x \cdot W_x|| \cdot ||F_y \cdot W_y||)}$$

The problem of finding ρ is therefore an optimization problem with respect to W_x and W_y. This optimization problem can be formulated as a standard Eigen problem [4] which can be easily solved. Since W_x and W_y are always calculated to maximize the correlation of the projections, CCA is independent of the original coordinate system unlike other correlation analysis techniques. There may be more than one canonical correlation, each representing orthogonally separate pattern of relationship between the two sets of variables. The correlation for the successively extracted canonical variates are smaller and smaller. When extracting the canonical correlation the eigen values are calculated. The square root of the eigen values can be interpreted as the canonical coefficients. Corresponding to each canonical correlation the canonical weights for each of the variable in the data set is calculated. The canonical weights represent the unique positive or negative contribution of each variable to the total correlation.

CCA has been used previously by researchers to find the semantic relationship between two multimodal inputs. In [4] kernel CCA was used to find correlation between image and text features obtained from a webpage and used it for content based image retrieval. [10] used CCA to find the language independent semantic representation of a text by using English text and its French translation as two views. When two multidimensional variables represent the two views of the same object, then the projections found by CCA can be thought of as capturing the underlying semantics of the object. In other words we can say that in the semantic feature space, the different views of the object are highly correlated. So to acquire a missing view of an object we can select the closest match from the observed views of other objects, such that it has the maximum correlation with the non-missing views of the current object, in the semantic feature space. In the next section we present the procedure for view completion using CCA.

6 View Completion Procedure

Let $CCA(p, q)$ denote the canonical correlation analysis of vectors p and q which gives the basis vectors and the projections of p and q on the basis vectors. The basis vectors can be considered as representing the lower dimensional semantic feature space which captures the underlying semantics of the object. Therefore to find y_{ym} we can select $y_p \in Y_p$, which has the highest correlation with x_{ym}

in the lower dimensional semantic feature space. Using these notations the procedure for the view completion is as follows.

1. Perform the canonical correlation analysis between X_p and Y_p and find the basis vectors.
   $[A, B, U, V] = CCA(X_p, Y_p)$,
   U and V are the matrices where the columns represent the basis vectors corresponding to X_p and Y_p respectively
   A and B are the projection of X_p and Y_p onto U and V respectively

2. For each instance $i \in I_{ym}$, $i = (x_{ym}, y_{ym}, c)$, where $x_{ym} \in X_{ym}$ and $y_{ym} \in Y_{ym}$,
   Project each $y_p \in Y_p$ onto V and the feature set x_{ym} onto U
   $p = y_p * V_k$
   $q = x_{ym} * U_k$
   where U_k and V_k are obtained by selecting top k basis vectors from U and V respectively.

3. The Pearson correlation cor between p and q is then calculated
   $cor = correlation(p, q)$

4. Select y_p with the maximum value for cor as y_{max}. Set $y_{ym} = y_{max}$ and update the instance $i = (x_{ym}, y_{max}, c)$

5. Repeat the procedure to find missing features x_{xm} for instances $i \in I_{xm}$, $i = (x_{xm}, y_{xm}, c)$

Though the above mentioned procedure is for object instances which have two views, it can be easily extended to instances with n views, $n > 2$ by doing a pairwise view completion. For example, in an instance of n views $i_n = (f_1, f_2, ..., f_n, c)$, if, say, f_1 is missing in the least number of object instances, then all the other views $f_2, ..., f_n$ could be completed using the above method by performing pairwise comparison with f_1.

7 Experiments and Results

The above procedure for view completion was run on the adult website classification data set used by [2]. The data set used contains seven different sources of data for classifying a webpage. The seven sources obtained based on the HTML tags of the webpages are BODY, Anchor Text and HREF(A), Image and ALT(Img), TITLE, METADATA, TABLE, Webpage URL(URL).

Data sources used	Body	Anchor	Body + Anchor	All 7 sources
Without View completion	0.38+-0.02	0.135+-0.017	0.085+-0.013	0.065+-0.017
With View completion	0.29+-0.027	0.135+-0.017	0.07+-0.016	0.055+-0.016

Table 1. Classification error rate on web dataset

7.1 Experiment with different numbers of features

We evaluated the classification accuracy on the original dataset with and without view completion. The experiments were performed on the following four cases.

1. Body text view alone: In this setting only body text view was used for classification. For the completion of body text view, the anchor text view was used.

2. Anchor text view alone: In this setting only anchor text view was used for classification. For the completion of anchor text view, the body text view was used.

3. Anchor text and body text features: In this setting both body and anchor text views were used for classification. And each of those views was used for the completion of the other.

4. All seven views:In this setting all the seven views were used for classification. The anchor view was used for the completion of all the other views except URL. URL was present for all the webpages and hence did not need to be completed). The completion for anchor view was done using body view.

Body and anchor views were selected among the seven views for the first three experiments, since these two are the most commonly used among the seven views for webpage classification and the most intuitive. The anchor view was used for view completion of all the other views since anchor view was missing among least number of object instances except for URL and it is more semantically related to all the other views compared to the URL view. The integration of the multiple views were done using the density based method proposed by [2]. The classifications were done using the SVM classifier with a linear kernel.

Table 1 gives the results of the experiment. This experiment is used to show the possibilities of view completion: using only a single view for classification (as given in first two columns) does not give good classification results, but completing them can reduce error rates significantly. Also

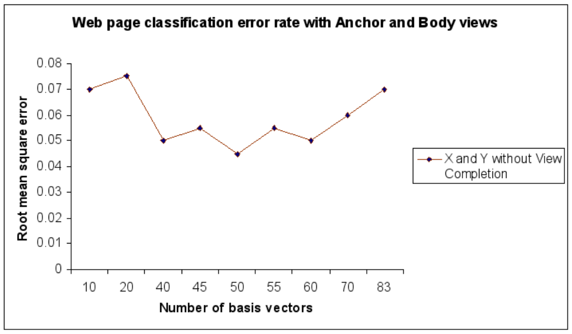

Figure 2. Effect of number of basis vectors

Number of basis vector	Error rates
10	0.07+-0.0186
20	0.075+-0.0112
30	0.05+-0.0129
40	0.055+-0.0138
50	0.045+-0.0117
60	0.055+-0.0138
70	0.05+-0.0129
83	0.06+-0.0145

Table 2. Effect of number of basis vectors

there may be scenarios where the multiple views are present only for a small amount of data used for training, but only one single view is present for the rest of the unlabeled data.

In the first column we see that after view completion the error rate for classification using body view has reduced by 9% with view completion. When anchor view was used, there was no change in the error rates. This was not surprising as the number of instances with anchor view missing was less in the beginning itself. So there was no additional gain with view completion. When both body and anchor views were used for classification a gain of 1.5% in accuracy was observed. Finally when all the seven sources were used for classification we observed that the error rate reduced by 1% and the accuracy increased from 93% to 94%. In all the above cases, the number of basis vectors was fixed to get the least error rates as described below.

7.2 Effect of numbers of basis vectors

In this experiment the number of basis vectors was varied as mentioned in Step 2 of Section 6. The effect of numbers of basis vectors on view completion was tested with different numbers of basis vectors. The experiment was carried out using the Body view + Anchor view setting. The results are presented in the table 2 and the figure 2. We see that the least error rate was obtained with k = 50.

7.3 Effect of missing views

In the third part of the experiment, the number of instances with missing views in *Anchor text* view was incrementally increased and the classification accuracy was evaluated. Since the anchor text and body text features are represented by term counts, if any of the view is missing, the term counts for that particular instance would be zero. So to get an additional mi instances with missing anchor text view, we randomly select mi instances which have anchor text view and then set the vector corresponding to anchor text view to zero. Table 3 and the Figure 3 give the results of the experiment. From the results we that over the whole range of instances with missing views applying view completion gives a consistently better classification accuracy compared to having no view completion at all. The figure also shows the results of KNN Imputation, Mean Imputation and EM Imputation. These imputation methods consider each missing value individually and substituting a value for it. From the figure it is clear that view completion which addresses the problem at a higher level in terms of views rather than at a lower level in terms of values, performs much better than the other methods.

8 Conclusions

We identify the need for studying the problem of missing views in the domain of multi-source data integration. We formally define the problem of missing views and we propose and efficient algorithm for view completion. By dealing with the missing view problem at a higher level in terms of *views* instead of a lower level representation of features we are able to achieve better results. By using the underlying existing semantic relationship between multiple views we propose a heuristic algorithm for view completion using CCA. Experiments on the web classification dataset demonstrates the advantages of our method.

The view completion method can be applied whenever the multiple views representing an instance are correlated. However in some application domains where the views are completely complementary and orthogonal, or when the views are completely uncorrelated, view completion would not be of much help. Our future work includes extending the algorithm using kernel methods to handle nonlinear correlation and developing algorithms to handle non-numeric attributes. We also plan to devise formal tests to decide when view completion procedure would be useful.

Acknowlegments

This work is in part sponsored by AFOSR via a grant (No. FA9550-08-1-0132).

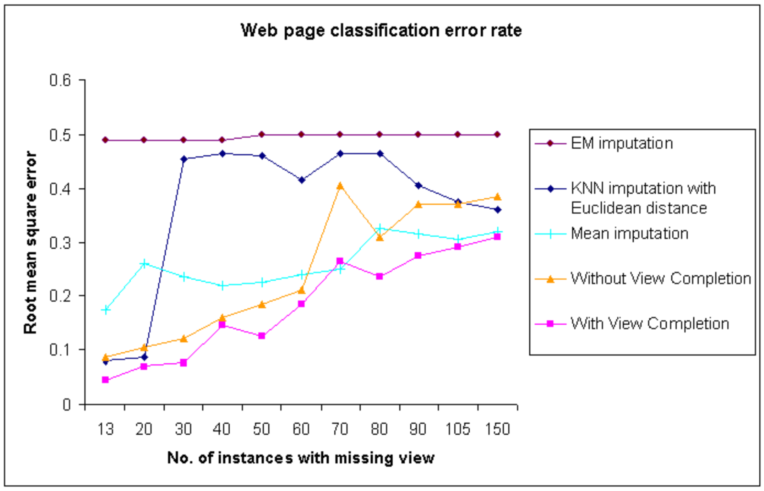

Figure 3. Effect of missing views

No. of instances missing anchor view	Error rate No view completion	Error rate With view completion	Error rate KNN Imputation	Error rate Mean Imputation	Error rate EM Imputation
13	0.085+-0.013	0.045+-0.012	0.08+-0.015	0.175+-0.028	0.49+-0.013
20	0.105+-0.019	0.07+-0.0152	0.085+-0.015	0.26+-0.058	0.49+-0.013
30	0.12+-0.019	0.075+-0.017	0.455+-0.051	0.235+-0.056	0.49+-0.013
40	0.16+-0.014	0.145+-0.028	0.465+-0.044	0.22+-0.027	0.49+-0.013
50	0.185+-0.026	0.125+-0.029	0.46+-0.052	0.225+-0.054	0.5+-0.033
60	0.21+-0.026	0.185+-0.018	0.415+-0.05	0.24+-0.031	0.5+-0.033
70	0.405+-0.033	1.0.265+-0.039	0.465+-0.044	0.25+-0.035	0.5+-0.033
80	0.31+-0.047	0.235+-0.029	0.465+-0.044	0.325+-0.039	0.5+-0.033
90	0.37+-0.03	0.275+-0.038	0.405+-0.051	0.315+-0.04	0.5+-0.033
105	0.37+-0.05	0.29+-0.026	0.375+-0.052	0.305+-0.037	0.5+-0.033
150	0.385+-0.049	0.31+-0.029	0.36+-0.043	0.32+-0.034	0.5+-0.033

Table 3. Webpage classification error rate - Varying number of instances with missing views

References

[1] E. Acuna and C. Rodriguez. The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering and Data Mining Applications*, pages 639–648, 2004.

[2] N. Agarwal, H. Liu, and J. Zhang. Blocking objectionable web content by leveraging multiple information sources. *SIGKDD Explor. Newsl.*, 8(1):17–26, 2006.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1998.

[4] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[5] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(312-377), 1936.

[6] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Active learning of features and labels. *Workshop on learning with multiple views at the 22nd International Conference on Machine Learning (ICML-05)*, pages 43–50, 2005.

[7] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[8] B. Rafkind, M. Lee, S.-F. Chang, and H. Yu. Exploring text and image features to classify images in bioscience literature. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 73–80, New York, New York, June 2006. Association for Computational Linguistics.

[9] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7:147–77, June 2002. PMID: 12090408.

[10] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis, 2002.

[11] R. Zhao and W. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *Multimedia, IEEE Transactions on*, 4(2):189–200, Jun 2002.