JMLR: Workshop and Conference Proceedings 11 (2010) 96–102 Workshop on Applications of Pattern Analysis

Maximum Margin Learning with Incomplete Data: Learning Networks instead of Tables

Sandor Szedmak SS03V@ECS.SOTON.AC.UK
University of Southampton, United Kingdom

Yizhao Ni YN05R@ECS.SOTON.AC.UK
Univeristy of Southampton

Steve R. Gunn SRG@ECS.SOTON.AC.UK
Univeristy of Southampton

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

In this paper we address the problem of predicting when the available data is incomplete. We show that changing the generally accepted table-wise view of the sample items into a graph representable one allows us to solve these kind of problems in a very concise way by using the well known convex, one-class classification based, optimisation framework. The use of the one-class formulation in the learning phase and in the prediction as well makes the entire procedure highly consistent. The graph representation can express the complex interdependencies among the data sources. The underlying optimisation problem can be transformed into a on-line algorithm, e.g. a perceptron type one, and in this way it can deal with data sets of million items. This framework covers and encompasses supervised, semi-supervised and some unsupervised learning problems. Furthermore, the data sources can be chosen as not only simple binary variables or vectors but text documents, images or even graphs with complex internal structures.

1. Introduction

The data collected in the real world are very frequently noisy, incomplete and ambiguous. For example in weather prediction the malfunction of measuring devices, limitations of sensors lead to the loss of significant parts of the information making hard the predictions be reliable. Another area where the incompleteness has to be addressed is the clinical trials, e.g. follow-up studies, in survival analysis, where new patients enter into the trial and others fall out before the trial has been finished. The records of these patients are incomplete, see examples in Molenberghs and Kenward (2007). Restricted resources can also lead to incomplete data when some measurements are too expensive to carry out in all experiments.

The machine learning methods are mostly planned to work when all the objects observed and recorded in all experiments. To overcome on the difficulties caused by the incompleteness quite a few approaches have been developed since the importance was recognised. The well known and generally applied expectation-maximisation(EM) algorithm was developed partially too incorporate the potential sources of information from unobserved data by exploiting certain prior knowledge and assumptions, see in Dempster et al. (1977) and Beal

and Ghahramani (2002). A great part of the statistical literature, e.g. Little and Rubin (2002), is devoted to address the problem of missing data since the occurrence of them can not be in generally avoided in a real experiment.

Our approach to dealing with the incomplete data stems out of conjecture: there exists a network of relationships among the data sources and the connections between these sources can be expressed by real valued functions. By approximating these functions by multilinear functions the recovering of the missing items can be turned into convex one-class classification problems.

The contribution of this paper to missing value handling can be summarised in the following points:

- The proposed approach can accommodate missing objects with complex structure, e.g. documents, amino acid sequences of proteins, molecules, images, user profiles. The structured objects are represented in inner product spaces, thus the flexibility of the kernel based learning can be applied.

- The estimation of the inferences is based on the maximum margin principle which is exploited in the well known Support Vector Machine. The optimisation framework is convex, and it can be reformulated as a on-line method, e.g. a perceptron type algorithm, and in turn can handle very large datasets.

- It can incorporate prior knowledge about the possible interdependency between the objects by representing the relations by a graph.

The paper first defines the learning environment, then sets up the basic optimisation framework. Following that the prediction of the unobserved objects is discussed. At the end a simple experiment is presented to demonstrate the workability of the approach.

2. Setting

Our description of a sample $\mathcal{S}$ is built upon the procedure of the data collection. In this procedure we have a series of experiments indexed by the set $\mathcal{I} = \{1, \ldots, m\}$. In each experiment a set of object observed. These objects come from a given set of object classes $\{\mathcal{X}_1, \ldots \mathcal{X}_{n_r}\}$, but the classes represented by an object in an experiment can vary among the experiments, in other words the set of object observed in an experiment is incomplete, the classes not observed can be referred as missing items. Let $\mathcal{R} = \{1, \ldots, n_r\}$ be a set of indeces of all object classes. The object classes can cover a broad range of collections of different objects, e.g. vectors, class labels, strings, text documents, see examples on Shawe-Taylor and Cristianini (2004). They can be collections of graph represented objects, see examples in Astikainen et al. (2010).

To express the relationship between the classes we assume that there exist a set of unknown real valued functions $F_{rs} : \mathcal{X}_r \times \mathcal{X}_s \to \mathbb{R}$, for some $r, s \in \mathcal{R}$, which measures the strength of the similarity between the objects of all class pairs, where the greater value of the functions corresponds to the stronger similarity. We call these functions as class similarities.

The goal is to discover the functions $\{F_{rs}\}$ connecting $\mathcal{X}_1, \ldots \mathcal{X}_{n_r}$ based on a given series of experiments. Let $\mathcal{R}_i \subseteq \mathcal{R}$ be the index set of object classes observed in experiment i,

thus a sample $\mathcal{S}$ is given by as a set of tuples $(x_i^r),\ r \in \mathcal{R}_i$ where $i = 1, \ldots, m$. In the sequel we call the objects observed in an experiment as sample items too.

We can represent this kind of sample by a table, where the sample items are in the rows and the object classes correspond to the columns. An element in column r and in row i of this table is considered missing, or unknown, if there is no observed object of class r in experiment i. The next table shows an example for $n_r = |\mathcal{R}| = 4$, where $\varnothing$ denotes the missing observations.

$$\begin{array}{cccc} x_1^1 & \varnothing & x_1^3 & \varnothing \\ x_2^1 & \varnothing & x_2^3 & x_2^4 \\ x_3^1 & \varnothing & \varnothing & x_3^4 \\ \vdots & \vdots & \vdots & \vdots \\ \varnothing & x_m^2 & x_m^3 & x_m^4. \end{array} \tag{1}$$

This representation has a hidden disadvantage. The potential relationships among the classes is not represented explicitly, however an implicit conjecture is accepted, namely those objects which observed in an experiment relate to each other.

Our setting is similar to the graphical models. In graphical models the relationships between random variables are represented via a graph G. The vertices of G are labelled by a set of random variables and the edges in G correspond to the possible inferences between these variables, see further details in Wainwright and Jordan (2008). In our case the vertices of G are labelled by the object classes, but the potential inferences, the prior knowledge, is given similarly to the graphical models by the edge set $\mathcal{E}$ of G. In this paper the graph G is chosen as an undirected one in the sense such that for all pairs of object classes the following equality holds: $F_{rs}(x^r, x^s) = F_{sr}(x^s, x^r),$ for any $x^r \in \mathcal{X}_r,\ x^s \in \mathcal{X}_s$.

Experiment i provides a subgraph G_i of G with vertices labelled by the objects observed in this experiment, and an edge $e \in \mathcal{E}$ is an edge of G_i if both classes spanning e are observed in this experiment. The set of these edges is denoted by $\mathcal{E}_i$. Figure 1 demonstrates the configuration described above.

2.1. Learning framework

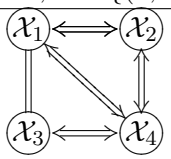
Let the set $\mathcal{I}_r \subseteq \mathcal{I}$ be the index set of experiments in which the class r is observed, and similarly $\mathcal{I}_e \subseteq \mathcal{I}$ be the index set where the edge e is observed. We assume every experiment provides at least two observed objects.

We assume that to every object class $\mathcal{X}_r,\ r \in \mathcal{R}$ there is a function ϕ_r which embeds the objects of this class into a linear vector space $\mathcal{H}_r$, and every vector space $\mathcal{H}_r$ is equipped with a positive definite inner product. The vector $\phi_r(x^r)$ of the corresponding space $\mathcal{H}_r$ for any $r \in \mathcal{R}$ is called feature vector of the objects x^r.

Suppose the following hypothesis that the class similarity function F_e between two classes $\mathcal{X}_r$ and $\mathcal{X}_s$ for any edge $e = (r, s) \in \mathcal{E}$ can be estimated by a function defined on the feature vectors, $\Psi_e : \mathcal{H}_r \times \mathcal{H}_s \to \mathbb{R}$, and these functions are linear in each variable. Exploiting the theory of the multilinear functions we have

$$\Psi_e(\phi_r(x^r), \phi_s(x^s)) = \langle \mathbf{W}_e, \phi_s(x^s) \otimes \phi_r(x^r) \rangle, \tag{3}$$

Graph of the class relations:$G, \mathcal{E} = \{(1,2),(1,3),(1,4),(2,4),(3,4)\}$



Sample items, observed objects in the experiments		
Sample Item 1	Sample Item 2	Sample Item 3
(x_1^1, x_1^3)	(x_2^1, x_2^3, x_2^4)	(x_3^1, x_3^4)
$G_1, \mathcal{E}_1 = \{(1,3)\}$	$G_2, \mathcal{E}_2 = \{(1,3),(1,4),(2,4)\}$	$G_3, \mathcal{E}_3 = \{(1,4)\}$
$x_1^1 \in \mathcal{X}_1$ $\quad \varnothing$ $x_1^3 \in \mathcal{X}_3$ $\quad \varnothing$	$x_2^1 \in \mathcal{X}_1$ $\quad \varnothing$ $x_2^3 \in \mathcal{X}_2$ $\quad x_2^4 \in \mathcal{X}_4$	$x_3^1 \in \mathcal{X}_1$ $\quad \varnothing$ $\varnothing$ $\quad x_3^4 \in \mathcal{X}_4$

(2)

Figure 1: An example of the graph of the classes and the structure of some sample items

for all $e = (r,s) \in \mathcal{E},$ and $x^r \in \mathcal{X}_r,\ x^s \in \mathcal{X}_s$, which states that a multilinear function can be expressed as an inner product between the tensor product of the variables and a tensor $\mathbf{W}_e$ representing the multilinear function itself.

We can show via some linear algebra that the multilinear functions Ψ_e corresponding to the edges can be reformulated to express directed relations

$$\langle \mathbf{W}_e, \phi_s(x^s) \otimes \phi_r(x^r) \rangle = \langle \phi_r(x^r), \mathbf{W}_e, \phi_s(x^s) \rangle_{\mathcal{H}_r} = \langle \phi_s(x^s), \mathbf{W}'_e, \phi_r(x^r) \rangle_{\mathcal{H}_s}, \tag{4}$$

where the subscript of the inner product $\langle\ \rangle_{\mathcal{H}}$ refers to the space on which the inner product is defined. These equations turn the tensor $\mathbf{W}_e$ into a linear transformation projecting one space into the other.

As a consequence, $\phi_r(x^r)$ can be predicted by a linear function $\mathbf{W}_e\phi_s(x^s)$ in the sense of angle minimisation, since if the norm of the linear operator $\mathbf{W}_e$ is fixed then the greater value of the inner product implies smaller angle between the vectors $\phi_r(x^r)$ and $\mathbf{W}_e\phi_s(x^s)$.

This tensor based formulation allows us to extend the framework into one dealing with relationships between more than two object classes. In this way the graph G can be generalised to hyper-graphs.

3. Optimisation problem, off-line case

Here we briefly outline the maximum margin based, one-class classification type, optimisation problem to learn all the functions ψ_{rs} simultaneously. It must be emphasise that the optimisation schema presented here is not the only one to solve this kind of tasks, it is only a possible prototype.

We apply edge relating Hinge-loss type loss functionals

$$\max\big(0, 1 - \langle \mathbf{W}_e, \phi_s(x_i^s) \otimes \phi_r(x_i^r) \rangle\big). \tag{5}$$

To enforce the hyperplane to achieve the maximum separating margin from the origin the Frobenious norm of the linear operators, the Euclidean norm of the linear vector space of these operators, have to be minimised. The optimisation problem realising this separation problem can be written up by

$$\begin{array}{lll} \min & \frac{1}{2}\sum_{e\in\mathcal{E}}\|\mathbf{W}_e\|_F^2 & \#\# \text{ regularisation} \\ & +C_R\sum_{e\in\mathcal{E}}\xi^e + C_I\sum_{i\in\mathcal{I}}\xi_i & \#\# \text{ edge wise and experiment wise loss} \\ \text{w.r.t.} & \{\mathbf{W}_e\},\ \{\xi^e\},\ e\in\mathcal{E},\ \{\xi_i\},\ i\in\mathcal{I}, & \\ \text{s.t.} & \langle \mathbf{W}_e, \phi_s(x_i^s)\otimes\phi_r(x_i^r)\rangle \geq 1-\xi^e-\xi_i, & \\ & \xi^e\geq 0,\ \xi_i\geq 0,\ i\in\mathcal{I},\quad e\in\mathcal{E} & \end{array} \tag{6}$$

where $C_R > 0$ and C_I are penalty constants. After assigning Lagrange multipliers to all constraints, and exploiting the Karush-Kuhn-Tucker conditions of optimality we obtain

$$\mathbf{W}_e = \sum_{i\in\mathcal{I}_e}\alpha_i^e\big(\phi_s(x_i^s)\otimes\phi_r(x_i^r)\big),\ e=(r,s). \tag{7}$$

By substituting back the formulas of the primal variables depending on the Lagrange multipliers into the Lagrangian functional and turning the maximisation with respect to the dual variables into minimisation we have the dual problem of (6)

$$\begin{array}{lll} \min & \frac{1}{2}\sum_{e,f\in\mathcal{E}}\sum_{i,j\in\mathcal{I}_e}\alpha_i^e\alpha_j^f\kappa_r(x_i^r,x_j^r)\kappa_s(x_i^s,x_j^s) - \sum_{i\in\mathcal{I}}\sum_{e\in\mathcal{E}_i}\alpha_i^e & \\ \text{w.r.t.} & \{\alpha_i^e\},\quad i\in\mathcal{I},\ e\in\mathcal{E}_i, & \\ \text{s.t.} & 0\leq\sum_{i\in\mathcal{I}_e}\alpha_i^e\leq C_R,\ e=(r,s)\in\mathcal{E},\ \#\# \text{ experiment wise coupling} & \\ & 0\leq\sum_{e\in\mathcal{E}_i}\alpha_i^e\leq C_I,\ i\in\mathcal{I}, \qquad \#\# \text{ edge wise coupling} & \end{array} \tag{8}$$

where we have the shorthand notations for the kernel functions to each object class

$$\kappa_r(x_i^r,x_j^r) = \langle\phi_r(x_i^r),\phi_r(x_j^r)\rangle,\ r\in\mathcal{R}, i,j\in\mathcal{I}_r. \tag{9}$$

In writing up the dual problem the following identity is applied

$$\langle\phi_s(x_i^s)\otimes\phi_r(x_i^r),\phi_s(x_j^s)\otimes\phi_r(x_j^r)\rangle = \langle\phi_s(x_i^s),\phi_s(x_j^s)\rangle\langle\phi_r(x_i^r),\phi_r(x_j^r)\rangle. \tag{10}$$

4. Estimating unknown items in an experiment

After computing the linear operators $\{\mathbf{W}_e\},\ e\in\mathcal{E}$ based on a given sample we are ready the predict the missing items both in a known experiment or a new one. A new experiment with no contribution in the computation of the linear operators has the index i_+, and $\tilde{\mathcal{I}} = \mathcal{I}\cup i_+$ is the index set referring both the known and the new included experiments.

We assumed that there is a multivariate real valued function Ψ such that if the similarity is stronger than the value of $\Psi(x^r,x^s)$ is greater. Therefore if $x_{\tilde{i}}^t,\ t\in\mathcal{R},\ t\notin\mathcal{R}_{\tilde{i}}$ is a missing item in experiment $\tilde{i}\in\tilde{I}$ then we expect that this missing item maximises the multilinear similarity functions with the observed ones. The reason for this approach is that because the edge wise predictions of a missing item can contradict, the solution with the possible smallest discrepancy should be accepted.

The prediction is obtained by the following optimisation problem

$$x_{\tilde{i}}^t = \left\{ \begin{array}{ll} \arg\max & g(\lambda) - D\sum_{r\in\mathcal{R}_{\tilde{i}}} \eta_r \qquad \#\# \text{ margin maximisation + loss} \\ \text{w.r.t.} & x^t \in \mathcal{X}^t, \eta_r \in \mathbb{R}, r \in \mathcal{R} \\ \text{s.t.} & \langle \phi_t(x^t), \mathbf{W}_e \phi_r(x_{\tilde{i}}^r)\rangle \geq \lambda - \eta_r \\ & \eta_r \geq 0,\ e = (r,t),\ r \in \mathcal{R}_{\tilde{i}},\ t \notin \mathcal{R}_{\tilde{i}}. \end{array} \right\} \tag{11}$$

In this optimisation problem $D > 0$ is penalty parameter, and the function g is a monotonically increasing real valued function on $\mathbb{R}_+$. To receive SVM type objective function g can be chosen as $g(u) = u^2$.

The optimisation problem (11) similarly to (6) implements a one-class classification problem where the data points are the predictions computed on the observed objects and the feature vector of the unknown object serves as normal vector in the separation from the origin, hence the optimisation problem (11) tries to maximise the margin against all possible relations expressible by the observed objects in the given experiment by exploiting the knowledge incorporated in the linear operators computed earlier. The slack variables $\{\eta^r\}$ provide the loss if the feature vectors of one or more observed objects depart too much from the feature vector of the unobserved item.

We can solve this problem in two steps, in the first one the optimum feature vector $\phi_t^*(x^t)$ is derived. Following a similar argument that was applied in the computations of the linear operators $\{\mathbf{W}_e\}$ the optimum feature vector can be computed based on the optimum value of dual variables $\{\beta_r\},\ r \in \mathcal{R}_{\tilde{i}}$ corresponding to the dual problem of (11), thus we have

$$\phi_t^*(x_{\tilde{i}}^t) \quad = \sum_{r\in\mathcal{R}_{\tilde{i}}} \beta_r \mathbf{W}_e \phi_r(x_{\tilde{i}}^r) = \sum_{r\in\mathcal{R}_{\tilde{i}}} \beta_r^* \sum_{j\in\mathcal{I}_e} \alpha_j \phi_t(x_j^t)\kappa(x_j^r, x_{\tilde{i}}^r). \tag{12}$$

From this formula we can estimate the optimum for $x_{\tilde{i}}^t$ by applying

$$\begin{array}{ll} x_{\tilde{i}}^{t*} & = \arg\max_{x^t\in\mathcal{X}_t} \langle x^t, \sum_{r\in\mathcal{R}_{\tilde{i}}} \beta_r \mathbf{W}_e \phi_r(x_{\tilde{i}}^r)\rangle \\ & = \arg\max_{x^t\in\mathcal{X}_t} \sum_{r\in\mathcal{R}_{\tilde{i}}} \beta_r^* \sum_{j\in\mathcal{I}_e} \alpha_j \kappa(x^t, x_j^t)\kappa(x_j^r, x_{\tilde{i}}^r). \end{array} \tag{13}$$

5. Experiments

In the experiments we used examples from the UCI Repository of machine learning datasets, the details about this repository are given by Blake and Merz (1998). The datasets included into the experiments are **Wisconsin Breast Cancer**, $N = 699$, and **Credit Card Application Approval**,$N = 690$.

The variables were uniformly randomly subsampled with probabilities running from 0.1 to 1.0 with step size 0.1 in ten different experiments. Two-fold cross validation is applied to avoid empty folds when the subsamples are very sparse. The subsampling was carried out on the training and the test sets as well. The implicit similarity functions are defined by linear kernels. The network graph connecting the classes was a star shaped one with the class label in the centre, since both datasets used in the experiments originally represent a binary classification problem. The accuracy of the prediction of the class labels of the incomplete data sources is demonstrated in Figure 2.

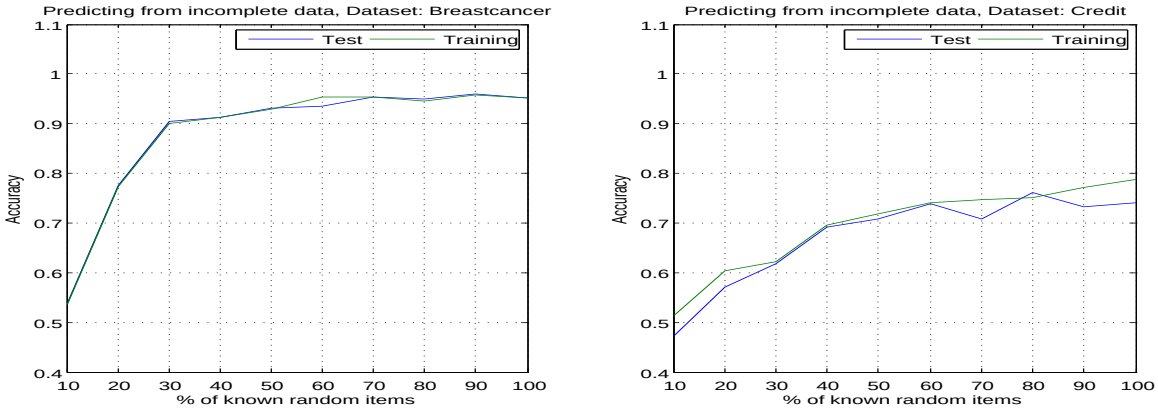


Figure 2: Predicting from incomplete data

6. Discussion

In this paper we outlined a general maximum margin based learning framework to handle incomplete data sources. The data sources can cover broad range of possible objects with reach internal structure. We exploit the inner product based geometric relationships between these complex objects to solve the prediction problem.

References

K. Astikainen, L. Holm, E. Pitkanen, J. Rousu, and S. Szedmak. Reaction kernels, structured output prediction approaches for novel enzyme function. In *Conference on Bioinformatics 2010, Valencia*. 2010. Best Paper Award.

M.J. Beal and Z. Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7, 2002.

C.L. Blake and C.J. Merz. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39 (1), 1977.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing data*. Wiley, second edition, 2002.

G. Molenberghs and M. G. Kenward. *Missing Data in Clinical Studies*. Wiley, 2007.

J. Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. In *Foundations and Trends in Machine Learning*, volume 1, pages 1–305. 2008.