USING DEPENDENCIES TO PAIR SAMPLES FOR MULTI-VIEW LEARNING

Abhishek Tripathi

University of Helsinki
Department of Computer Science
P.O.Box 68, 00014 UH, Finland

Arto Klami, Samuel Kaski

Helsinki University of Technology
Department of Information and Computer Science
P.O.Box 5400, 02015 TKK, Finland

ABSTRACT

Several data analysis tools such as (kernel) canonical correlation analysis and various multi-view learning methods require paired observations in two data sets. We study the problem of inferring such pairing for data sets with no known one-to-one pairing. The pairing is found by an iterative algorithm that alternates between searching for feature representations that reveal statistical dependencies between the data sets, and finding the best pairs for the samples. The method is applied on pairing probe sets of two different microarray platforms.

Index Terms— canonical correlation, co-occurrence data, dependency, multi-view learning

1. INTRODUCTION

Multi-view learning considers the task of learning from two or more data sets with co-occurring observations. Increased performance compared to traditional single-view learning has been reported in various applications, including semi-supervised classification [1], cross-lingual text mining and machine translation [2], and multimodal information retrieval [3]. The improvement comes from utilizing statistical dependencies between the data sources, either in form of maximizing a consensus between models learned based on each data source [4], through explicit maximization of a measure of dependency between the views [5], or by building generative latent variable models that capture the dependency [6].

The traditional multi-view learning methods require strict co-occurrence. That is, the views must have known one-to-one pairing for the samples. In some applications, such as machine translation or analysis of microarray data, there are, however, cases where such pairing in principle exists but is unknown. For example, two different microarray platforms attempt to measure activities of the same genes, yet the exact probes on the chips are different and hence not directly paired between platforms. Traditional multi-view learning methods cannot be directly applied to such problems.

All authors belong to Helsinki Institute for Information Technology HIIT and Adaptive Informatics Research Centre. The work was partly supported by PASCAL2, EU Network of Excellence, and by a grant from University of Helsinki's Research Funds.

We would like to be able to apply multi-view learning methods also for non-paired data, typically in cases where some kind of vague pairing information or prior information on possible pairs is available. In this paper we study an approach where we explicitly find one-to-one pairing between data sets, using the actual measurements for defining the pairing. This complements approaches that would infer the pairing based on additional data sources, such as sequence information in the case of microarray platforms. Given the solution of the proposed algorithm, any multi-view learning method can be applied on the data.

The pairing is based on statistical dependency between the data sources. We want to find such a pairing that the dependency between the two sources becomes maximal. Given a sample we should be able to predict with the measurement values of its true pair in the other data set, and dependency is a justified measure for two-way prediction accuracy. Maximizing the dependency should hence find the true pairs. Also note that completely random pairing necessarily leads to independency between the sources.

In the remaining paper, we first introduce an algorithm for finding the pairing and then demonstrate it in a practical application of finding corresponding probe sets in two different microarray platforms.

2. METHOD

Given two data sets, $\mathbf{X} \in \mathbb{R}^{N \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D_y}$, $M \geq N$, we want to find a permutation $\mathbf{p}$ of the samples in $\mathbf{Y}$ such that the ith sample in $\mathbf{X}$ is paired with the sample $\mathbf{p}(i)$ in $\mathbf{Y}$. The pairing will be primarily based on the actual data vectors, though prior information on pairings can be included as explained later.

We propose a generally applicable algorithm that can be used to pair any two data sets that are supposed to have a one-to-one pairing between the samples. The underlying assumption is that a pairing that reveals statistical dependency between the two data sets is more likely to be correct. We consider lower-dimensional mappings $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{y})$ and learn them and the pairing to maximize the dependency

$$\max_{\mathbf{p},\mathbf{f},\mathbf{g}} \quad \mathrm{Dep}\left(\mathbf{f}(\mathbf{X}), \mathbf{g}(\mathbf{Y}(\mathbf{p}))\right),$$

where $\mathrm{Dep}(\cdot,\cdot)$ denotes any measure of dependency between the two arguments, and $\mathbf{Y}(\mathbf{p}) \in \mathbb{R}^{N \times D_y}$ is a matrix obtained by picking the rows indicated by $\mathbf{p}$.

The dependency measure and parameterization of the mappings $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ can be chosen freely. Here we resort to Pearson correlation and linear projections $\mathbf{f}(\mathbf{x}) = \mathbf{x}\mathbf{W}_x$ and $\mathbf{g}(\mathbf{y}) = \mathbf{y}\mathbf{W}_y$. This gives a simple and computationally efficient method with sufficient accuracy, but it is worth mentioning that these assumptions can be relaxed if needed. With these assumptions, the optimization problem becomes

$$\max_{\mathbf{p},\mathbf{W}_x,\mathbf{W}_y} \mathrm{corr}\left(\mathbf{X}\mathbf{W}_x, \mathbf{Y}(\mathbf{p})\mathbf{W}_y\right).$$

For solving the problem, we propose an iterative algorithm that alternates between learning the pairing and learning the projections. Given fixed projections, we can write the cost with the sample estimate of correlation as

$$\max_{\mathbf{p}} \frac{\mathbf{W}_x^T \mathbf{X}^T \mathbf{Y}(\mathbf{p}) \mathbf{W}_y}{\|\mathbf{X}\mathbf{W}_x\| \|\mathbf{Y}(\mathbf{p})\mathbf{W}_y\|}.$$

The numerator can equivalently be expressed as constant minus the sum of Euclidean distances between the projections of paired samples. The denominator, in turn, is constant with respect to $\mathbf{p}$ if $N = M$, and it can be safely assumed constant even for M slightly larger than N. Hence, it can be ignored in optimization and we get the task

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \|\mathbf{x}_i \mathbf{W}_x - \mathbf{y}_{\mathbf{p}(i)} \mathbf{W}_y\|^2. \quad (1)$$

This is a classical *assignment problem* where the cost of assignments is defined by the distance in the projection space. The assignment problem can be solved exactly with e.g. the Hungarian algorithm.

The projections, in turn, are solved by assuming a fixed pairing. Then the task reduces to the canonical correlation analysis (CCA) problem, which can be solved exactly with linear algebra (see e.g. [7]). These two steps can be combined into a simple alternating algorithm, which finds the pairing but can also be considered as a way to compute CCA for non-paired data. First a random pairing is given as an initialization and CCA is used to find the optimal projections for that pairing. Then a new pairing is solved via the assignment problem using distances computed in the feature space. After this, the algorithm repeats the CCA and assignment problem steps.

Notice that if the features of $\mathbf{X}$ and $\mathbf{Y}$ are paired, the sample pairing can alternatively be inferred by directly solving the assignment problem $\min_{\mathbf{p}} \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_{\mathbf{p}(i)})$, where $d(\mathbf{x}_i, \mathbf{y}_{\mathbf{p}(i)})$ is some conventional distance measure between the samples. This approach assumes that direct comparison of samples is sensible, whereas the proposed algorithm learns feature representations that can be compared in all cases.

2.1. Details and related work

Due to scale-invariance of correlation, CCA does not fix the scales of the dimensions. Hence, we can choose the scales of different dimensions in the optimization problem in (1) to maximize the accuracy. In the experimental section we empirically compare to choices. The first choice is to give equal scale for each dimension. The other choice utilizes the fact that dimensions with high correlation are more likely to contain useful information and weights each dimension with the corresponding canonical correlation. The latter choice is shown to be better in practice.

Possible prior information on pairing, typically obtained from yet another data source, can be taken into account in numerous ways. Here we consider a simple method that excludes sets of possible pairs from consideration if we know they definitely are not the true pairs. We formulate this through the concept of *candidate sets*. Instead of allowing any pairing, we use the prior information to create a subset of samples in $\mathbf{Y}$ for each $\mathbf{x}$. Samples not belonging to the candidate set are given infinite cost in the assignment problem. This helps in avoiding overlearning. In all the experiments, the algorithm is run for the maximum of 20 iterations, however, it converged in less than 10 iterations in most cases.

Recently, [8] studied a similar approach for learning bilingual lexicons from monolingual corpora. They introduce a latent variable model for the task, and optimize it with an EM algorithm resembling our alternating algorithm. In the E-step of their algorithm they solve the assignment problem to maximize the sum of pointwise mutual informations, but comment that a heuristic using Euclidean distances between the projections is in practice more accurate. In our formulation, (1) follows directly from the cost function. They also consider the paired corpus as the main result, whereas we solve the pairing to enable the use of further multi-view learning methods.

3. MICROARRAY PLATFORM PAIRING

One application area for the method is combining microarray measurements done on different measurement platforms. We can infer pairing between different brands of microarrays aiming to measure the same activities, and also for example between arrays designed for different measurement types (RNA or DNA) or species. As a demonstration, we apply the method for pairing the probe sets of two different versions of Affymetrix oligonucleotide arrays, HG-U95 and HG-U133.

As measurement data we use gene expression profiles of pediatric acute lymphoblastic leukemia (ALL) patients from [9, 10]. The data consists of expression measurements of the same 131 patients on both HG-U95 and HG-U133 platforms, providing an excellent test bed for the algorithm. Typically the probes of different array platforms would be paired primarily based on sequence information, which is available in this special case of pairing problems. Here we demonstrate

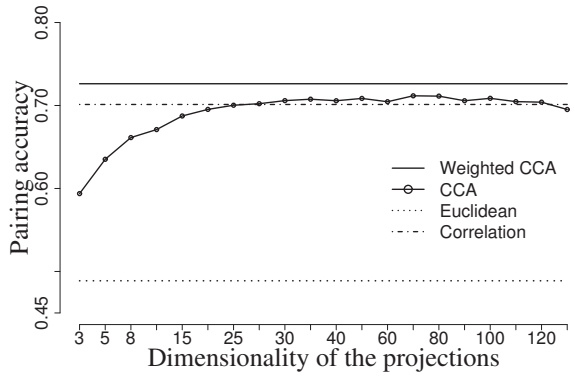

Fig. 1. Accuracy in finding Affymetrix best matches amongst all possible matches in the task of pairing HGU-95 probe sets with HGU-133 probe sets. CCA-based pairing methods have the best accuracy, with correlation-weighted distance in CCA projection space providing the best result. All methods clearly exceed the baseline accuracy 0.31 of random pairing.

how it is possible to improve the pairing based on the expression measurements. Sequences are only used as prior information to define the candidate sets.

To evaluate the accuracy, we use Affymetrix's comparison sheet between HGU-95 and HG-U133 as the ground truth. For each probe set in HG-U95, the sheet lists a set of potential matches in HG-U133, defined based on sequence information. There are generally more than one match for each HG-U95 probe set, and the quality of each match is characterized as "match", "good match", or "best match". We create each candidate set as the collection of all matches for the probe set, and measure the accuracy by checking how often the found pair has the "best match" label. It is not possible to obtain a perfect score due to the best matches not providing a one-to-one mapping, and hence we report accuracies normalized so that 1 means the best possible score.

In total we have $N = 11728$ HG-U95 and $M = 17857$ HG-U133 probe sets. 2171 of the HG-U95 probe sets have only one match in HG-U133, and hence the pairing is fixed for those. Since M is here so much larger than N, we apply a slight heuristic to remedy the fact that $\mathbf{Y}(\mathbf{p})$ is not constant with respect to $\mathbf{p}$. In all experiments with the microarray data, we normalize the distances in (1) by $\|\mathbf{x}_i\mathbf{W}_x\|\|\mathbf{y}_i\mathbf{W}_y\|$ to prioritize pairing probe sets with higher total activity.

4. RESULTS AND TECHNICAL VALIDATION

4.1. Pairing of Affymetrix probe sets

We study the accuracy of the algorithm with two different ways to compute the distance in the projection space. The first solution finds a fixed-dimensional CCA subspace and treats each dimension with equal weight. The other approach finds the full CCA subspace, but weights each dimension with the corresponding canonical correlation. Since the data has paired features, we can compare the accuracy of the proposed algorithm with the simple method of directly solving the assignment problem in the original data space. We use both correlation and Euclidean distance for comparison.

Table 1. Pairing accuracies in different feature settings. All differences in accuracy are statistically significant (t-test, all p-values below 10^{-14}).

Feature setting	Method	Mean Accuracy	Std Dev
1. Paired	CCA	0.64	0.04
	Corr	0.56	0.04
2. Permuted	CCA	0.64	0.04
	Corr	0.29	0.01
3. Different	CCA	0.54	0.03
	Corr	0.30	0.01
4. Paired + Noise	CCA	0.60	0.03
	Corr	0.34	0.02

The pairing accuracies are shown in Figure 1. The proposed CCA-based method outperforms the comparison methods, and the weighted version provides better accuracy compared to using lower-dimensional subspaces. As the weighted solution also sidesteps the need to choose the dimensionality, it is to be preferred. For the comparison method correlation seems to be clearly better choice of distance, indicating scale differences between the data sets.

This application, chosen to enable comparison with the naive method, is a special case as the features are known to be paired. In practice, however, we typically will not have data sets with paired features. Next, we demonstrate on modified versions of the data how the results would change in scenarios where the features are not paired. The studied scenarios are:

1. 35 paired features (patients)
2. The same, but the features in $\mathbf{Y}$ are randomly permuted
3. 35 different features in each data
4. 20 paired features, along with 15 noise features created by permuting the values for genes randomly

We repeated the same experiment for 20 randomly subsampled versions for each scenario, each having around 1000 samples, and used the better variants of the two approaches (correlation-weighted CCA and correlation-based method) for pairing the samples. The results are shown in Table 1. The main observation is that the performance of the CCA-based method is comparable for all scenarios, whereas the comparison method works only with the paired features, as expected. The proposed method works well even with the noisy dimensions showing that it is able to ignore them.

4.2. Technical validation on artificial data

To study the accuracy of the method as a function of dependency between the data sources, we created artificial data sets

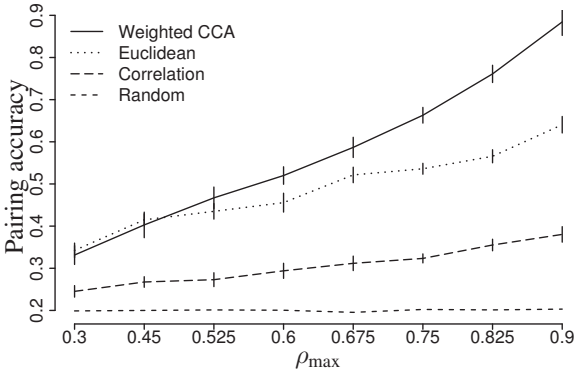

Fig. 2. Pairing accuracy as a function of maximal correlation ($\rho_{\max}$) between the data sets. The advantage of the CCA-based method increases with higher dependency. The bars represent one standard deviation (over 20 runs).

with varying degree of dependency. The toy data is generated by drawing 1000 samples from a 10-dimensional multivariate normal distribution with zero mean and a given covariance matrix. The data is split into two 5-dimensional matrices, $\mathbf{X}$ and $\mathbf{Y}$, so that the true canonical correlations are of the form $\lambda \times (0.3, 0.25, 0.15, 0.05, 0)$, where $\lambda \in [1, 3]$. The candidate sets for each sample in $\mathbf{X}$ were created so that each set includes the true pair and 4 random samples.

Figure 2 shows the pairing accuracy for different degrees of dependency. With low dependency CCA is comparable to directly using Euclidean distance, but with higher dependency CCA clearly outperforms the comparison methods. In the microarray application the largest correlations were close to 0.9, so it falls into the region where CCA helps most. In this case the Euclidean distance works better than correlation, whereas for microarray data the order was opposite. This demonstrates how the choice of the distance measure plays an important role for the comparison method, while the CCA-based method overperforms both choices in both problems.

5. CONCLUSIONS

Multi-view learning tasks require co-occurring observations in the different views. In many applications no clear one-to-one mapping is known, but we may have some information on possible pairs. To enable running multi-view learning algorithms in such applications, we presented a method for finding co-occurring samples based on the actual measurements. The method uses an alternating iterative algorithm to find such a pairing that statistical dependency between the data sets is maximized. The method was demonstrated in an application of pairing probe sets of two microarray platforms.

In addition to pairing the samples, the method can be used to compute CCA for unpaired data sets. Even in cases where the pairing accuracy is not perfect, the projection vectors seem to converge towards the true projection vectors. On the toy data the cosine similarity between the true and estimated projections was above 0.98 for all CCA components even in cases where the pairing accuracy was only 65% (results not shown due to lack of space).

The proposed algorithm, available on request from the authors as R code, is an example of a wider family of pairing algorithms. It uses linear projections to find a feature representation that allows comparison of samples, and correlation for measuring the dependency. Methods relaxing these assumptions could be devised, e.g. by maximizing a non-parametric estimate of mutual information as in [5].

6. REFERENCES

[1] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," In *Proc. of 11th Annual Conference on Learning Theory*, pp. 92–100. 1998.

[2] Y. Li and J. Shawe-Taylor, "Using KCCA for Japanese-English cross-language information retrieval and document classification," *Journal of intelligent information systems*, vol. 27, no. 2, pp. 117–133, 2006.

[3] J.D.R. Farquhar, D.R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," In *Advances in Neural Information Processing Systems 18*, pp.355–362, MIT Press, 2006.

[4] S. Bickel and T. Scheffer, "Estimation of mixture models using Co-EM," In *Proceedings of the European Conference on Machine Learning*, pp. 35–46, 2005.

[5] A. Klami and S. Kaski, "Non-parametric dependent components," In *ICASSP'05*, pp. V–209–V–212, 2005.

[6] A. Klami and S. Kaski, "Generative models that discover dependencies between data sets," In *Machine learning for signal processing XVI*, pp.123–128, 2006.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[8] A. Haghighi, P. Liang, T. Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," In *Proc. of Association for Computational Linguistics*, pp. 771–779, 2008.

[9] Yeoh et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[10] Ross et al., "Classification of pediatric acute lymphoblastic leukemia by gene expression profiling," *Blood*, vol. 102, no. 8, pp. 2951–2959, 2003.