Data Fusion Techniques and Application

Guangyu Zhou

Reference paper: Zheng Yu: Methodologies for Cross-Domain Data Fusion: An Overview
Agenda

- Introduction
- Related work
- Data fusion techniques & applications
  - Stage-based methods
  - Feature level-based methods
  - Semantic meaning-based data fusion methods
- Summary
What is data fusion?

- **Data fusion** is the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source —— Wikipedia
Why data fusion?

- In the big data era, we face a diversity of datasets from different sources in **different domains**, consisting of **multiple modalities**:
  - Representation, distribution, scale, and density.
- How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets?
  - Treating different datasets equally or simply concatenating the features from disparate datasets?
In the big data era, we face a diversity of datasets from different sources in **different domains**, consisting of **multiple modalities**: 
- Representation, distribution, scale, and density.

How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets?
- Treating different datasets equally or simply concatenating the features from disparate datasets
- Use advanced data fusion techniques that can fuse **knowledge** from various datasets organically in a machine learning and data mining task
Related Work

- Relation to Traditional Data Integration
Related Work

- Relation to Heterogeneous Information Network
  - It only links the object in a single domain:
    - Bibliographic network, author, papers, and conferences.
    - Flickr information network: users, images, tags, and comments.
  - Aim to fuse data across different domains:
    - Traffic data, social media and air quality
  - Heterogeneous network may not be able to find explicit links with semantic meanings between objects of different domains.
Data fusion methodologies

- Stage-based methods
- Feature level-based methods
- Semantic meaning-based data fusion methods
  - multi-view learning-based
  - similarity-based
  - probabilistic dependency-based
  - and transfer learning-based methods.
Stage-based data fusion methods

- Different datasets at different stages of a data mining task.
- Datasets are loosely coupled, without any requirements on the consistency of their modalities.
- Can be a meta-approach used together with other data fusion methods
Map partition and graph building for taxi trajectory

A) Map segmentation

B) Region graph
**Friend recommendation**

- **Stages**
  - I. Detect stay points
  - II. Map to POI vector
  - III. Hierarchical clustering
  - IV. Partial tree
  - V. Hierarchical graph
    - -> comparable (from same tree)
Data fusion methodologies

- Stage-based methods
- **Feature level-based methods**
- Semantic meaning-based data fusion methods
  - multi-view learning-based
  - similarity-based
  - probabilistic dependency-based
  - and transfer learning-based methods.
Feature-level-based data fusion

- **Direct Concatenation**
  - Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector

- **Limitations:**
  - **Over-fitting** in the case of a small size training sample, and the specific statistical property of each view is ignored.
  - Difficult to discover highly non-linear relationships that exist between low-level features across different modalities.
  - **Redundancies and dependencies** between features extracted from different datasets which may be correlated.
Feature-level-based data fusion

- Direct Concatenation + sparsity regularization:
  - handle the feature redundancy problem
    \[
    P(\Psi|\Omega) = P(\omega|0, \beta^2) \prod_{m} N(\omega_m|0, \beta_m^2) \prod_{m} \text{Inverse-Gamma}(\beta_m^2|a, b);
    \]

- Dual regularization (i.e., zero-mean Gaussian plus inverse-gamma):
  - Regularize most feature weights to be zero or close to zero via a Bayesian sparse prior
  - Allow for the possibility of a model learning large weights for significant features
Feature-level-based data fusion

- DNN-Based Data Fusion
  - Using supervised, unsupervised and semi-supervised approaches, Deep Learning learns multiple levels of representation and abstraction
  - Unified feature representation from disparate dataset
DNN-Based Data Fusion

- Deep Autoencoder Models of feature representation between 2 modalities (audio + video)
The multimodal DBM is a generative and undirected graphic model.

- Enables bi-directional search.

To learn $P(v_{img}, v_{text}; \theta)$
Limitations of DNN-based fusion model

- Performance heavily depend on parameters

- Finding optimal parameters is a labor intensive and time-consuming process given a large number of parameters and a non-convex optimization setting.

- Hard to explain what the middle-level feature representation stands for.
  - We do not really understand the way a DNN makes raw features a better representation either.
Semantic meaning-based data fusion

- Unlike feature-based fusion, semantic meaning-based methods understand the **insight** of each dataset and **relations** between features across different datasets.

- 4 groups of semantic meaning methods:
  - multi-view-based, similarity-based, probabilistic dependency-based, and transfer-learning-based methods.
Data fusion methodologies

- Stage-based methods
- Feature level-based methods
- **Semantic meaning-based data fusion methods**
  - multi-view learning-based
    - co-training, multiple kernel learning (MKL), subspace learning
  - similarity-based
  - probabilistic dependency-based
  - and transfer learning-based methods.
Multi-View Based Data Fusion

- Different datasets or different feature subsets about an object can be regarded as different views on the object.
  - Person: face, fingerprint, or signature
  - Image: color or texture features

- Latent consensus & complementary knowledge

- 3 subcategories:
  - 1) co-training
  - 2) multiple kernel learning (MKL)
  - 3) subspace learning
Co-training considers a setting in which each example can be partitioned into two distinct views, making three main assumptions:

- **Sufficiency**: each view is sufficient for classification on its own
- **Compatibility**: the target functions in both views predict the same labels for co-occurring features with high probability
- **Conditional independence**: the views are conditionally independent given the class label. (Too strong in practice)
Multi-View Based Data Fusion: Co-training

- **Original Co-training**

  Given:
  
  - a set $L$ of labeled training examples
  - a set $U$ of unlabeled examples

  Create a pool $U'$ of examples by choosing $u$ examples at random from $U$

  Loop for $k$ iterations:

  Use $L$ to train a classifier $h_1$ that considers only the $x_1$ portion of $x$
  Use $L$ to train a classifier $h_2$ that considers only the $x_2$ portion of $x$
  Allow $h_1$ to label $p$ positive and $n$ negative examples from $U'$
  Allow $h_2$ to label $p$ positive and $n$ negative examples from $U'$
  Add these self-labeled examples to $L$
  Randomly choose $2p + 2n$ examples from $U$ to replenish $U'$
Co-training-based air quality inference model

A) Philosophy of the inference model
- A location with AQI labels
- A location to be inferred

B) Procedure of co-training
- Co-Training
  - Temporal Classifier
  - Spatial Classifier
- Temporal Features
- Spatial Features
- Labeled Data
- Unlabeled Data

Temporal dependency 
Spatial correlation

Traffic  Human Mobility  Road Networks
Meteorology  Time of Day  POIs  AQI Labels
Multi-View Based Data Fusion: MKL

2. Multi-Kernel Learning

- A kernel is a **hypothesis** on the data
- MKL refers to a set of machine learning methods that uses a predefined set of kernels and learns an optimal linear or non-linear combination of kernels as part of the algorithm.
  - Eg: Ensemble and boosting methods, such as Random Forest, are inspired by MKL.
MKL-based framework for forecasting air quality.
Multi-View Based Data Fusion: MKL

- The MKL-based framework outperforms a single kernel-based model in the air quality forecast example
  - Feature space:
    - The features used by the spatial and temporal predictors do not have any overlaps, providing different views on a station’s air quality.
  - Model:
    - The spatial and temporal predictors model the local factors and global factors respectively, which have significantly different properties.
  - Parameter learning:
    - Decomposing a big model into 3 coupled small ones scales down the parameter spaces tremendously.
Multi-View Based Data Fusion: subspace learning

- Obtain a latent subspace shared by multiple views by assuming that input views are generated from this latent subspace,
- Subsequent tasks, such as classification and clustering
- Lower dimensionality
Multi-View Based Data Fusion: subspace learning

- Eg: PCA ->
  - Linear case: Canonical correlation analysis (CCA)
    - maximizing the correlation between 2 views in the subspace
      \[
      x \rightarrow \langle w_x, x \rangle \quad \rho = \max_{w_x,w_y} \text{corr}(S_x w_x, S_y w_y) \\
      = \max_{w_x,w_y} \frac{\langle S_x w_x, S_y w_y \rangle}{\|S_x w_x\| \|S_y w_y\|}.
      \]
  - Non-linear: Kernel variant of CCA (KCCA)
    - map each (non-linear) data point to a higher space in which linear CCA operates.
Multi-View Based Data Fusion

- Summary of Multi-View Based methods
  1) co-training: maximize the mutual agreement on two distinct views of the data.
  2) multiple kernel learning (MKL): exploit kernels that naturally correspond to different views and combine kernels either linearly or non-linearly to improve learning.
  3) subspace learning: obtain a latent subspace shared by multiple views, assuming that the input views are generated from this latent subspace.
Data fusion methodologies

- Stage-based methods
- Feature level-based methods
- **Semantic meaning-based data fusion methods**
  - multi-view learning-based
  - similarity-based
    - Coupled Matrix Factorization
    - Manifold Alignment
  - probabilistic dependency-based
  - and transfer learning-based methods.
Recall: Matrix decomposition by SVD

\[
\begin{pmatrix}
 x_{11} & x_{12} & \cdots & x_{1n} \\
 x_{21} & x_{22} & \cdots & \vdots \\
 \vdots & \vdots & \ddots & \vdots \\
 x_{m1} & x_{m2} & \cdots & x_{mn}
\end{pmatrix} =
\begin{pmatrix}
 u_{11} & \cdots & u_{1r} \\
 \vdots & \ddots & \vdots \\
 u_{m1} & \cdots & u_{mr}
\end{pmatrix}
\begin{pmatrix}
 s_{11} & \cdots & 0 \\
 \vdots & \ddots & \vdots \\
 0 & \cdots & s_{rr}
\end{pmatrix}
\begin{pmatrix}
 v_{11} & \cdots & v_{1n} \\
 \vdots & \ddots & \vdots \\
 v_{r1} & \cdots & v_{rn}
\end{pmatrix}
\]

Problems of single matrix decomposition on different datasets:
- Inaccurate complementation of missing values in the matrix.
Similarity-Based: Coupled Matrix Factorization

- Solution by coupled (context-aware) matrix factorization:
  - To accommodate different datasets with different matrices (distribution, meaning), which share a common dimension between one another.
  - By decomposing these matrices collaboratively, we can transfer the similarity between different objects learned from a dataset to another one, therefore complementing the missing values more accurately.
Estimate the travel speed on each road segment in an entire city, based on the GPS trajectory of a sample of vehicles.
Coupled Matrix Factorization Application

- Coupled matrix factorization

\[
\begin{bmatrix}
  g_1 & g_2 & \ldots & g_{16} \\
  g_1 & g_2 & \ldots & g_{16}
\end{bmatrix}
\begin{bmatrix}
  r_1 & r_2 & \ldots & r_n \\
  r_1 & r_2 & \ldots & r_n
\end{bmatrix}
\begin{bmatrix}
  f_1 & f_2 & \ldots & f_k
\end{bmatrix}
\]

- Objective function:

\[
L(T, R, G, F) = \frac{1}{2} \left( \| Y - T(G; G)^T \|_2^2 + \frac{\lambda_1}{2} \| X - T(R; R)^T \|_2^2 \right) + \frac{\lambda_2}{2} \| Z - RF^T \|_2^2 + \frac{\lambda_3}{2} (\| T \|_2^2 + \| R \|_2^2 + \| G \|_2^2 + \| F \|_2^2),
\]

Algorithm TSE

**Input:** Incomplete matrix \( X \), context matrices \( Y \) and \( Z \)

**Output:** Complete matrix \( X \).

1. \( t = 1 \);
2. While \((t < N \text{ and } L_t - L_{t+1} > \epsilon)\) // \( N \) is \#(max iterations)
3. Get the gradients \( \nabla_{T_t}, \nabla_{R_t}, \nabla_{G_t}, \text{ and } \nabla_{F_t} \) by Eq.(6);
4. \( \gamma = 1 \);
5. While
6. \( L(T_t - \gamma \nabla_{T_t}, R_t - \gamma \nabla_{R_t}, G_t - \gamma \nabla_{G_t}, F_t - \gamma \nabla_{F_t}) \geq L(T_t, R_t, G_t, F_t) \)
7. \( \gamma = \gamma / 2; \) // search for the maximal step size
8. \( T_{t+1} = T_t - \gamma \nabla_{T_t}, R_{t+1} = R_t - \gamma \nabla_{R_t}, G_{t+1} = G_t - \gamma \nabla_{G_t}, F_{t+1} = F_t - \gamma \nabla_{F_t}; \)
9. \( t = t + 1; \)
10. Return \( X \);
Similarity-Based: Manifold Alignment

- Utilizes the **relationships** of instances **within** each dataset to strengthen the knowledge of the **relationships between** the datasets, thereby ultimately **mapping** initially disparate datasets to a joint latent space.

- Maps two datasets \((X, Y)\) to a new joint latent space \((f(X);g(Y))\),
Similarity-Based: Manifold Alignment

- Preserves 2 similarities:
  - The local similarity within a dataset,
    \[ C_\lambda(F^a) = \sum_{i,j} \| F^a(i,.) - F^a(j,.) \|^2 \cdot W^a(i,j), \]
  - The correspondences across different datasets.
    \[ C_k(F^a, F^b) = \sum_{i,j} \| F^a(i,.) - F^b(j,.) \|^2 \cdot W^{a,b}(i,j), \]

- C, cost function; F, embedding of data; W, similarity matrix; a, the a\(^{th}\) dataset
Similarity-Based: Manifold Alignment

- Manifold alignment assumes the disparate datasets to be aligned have the same underlying manifold structure.
- The second loss function is simply the loss function for Laplacian Eigen-maps using the joint adjacency matrix: \( L = D - W \)

\[
C_2(F) = \sum_{i,j} \sum_k \| F(i,k) - F(j,k) \|^2 \cdot W^{a,b}(i,j)
\]

\[
= \sum_{k} \sum_{i,j} \| F(i,k) - F(j,k) \|^2 \cdot W^{a,b}(i,j)
\]

\[
= \sum_k \text{tr}(F(.,k)'LF(.,k)) = \text{tr}(F'LF),
\]
Example: Infer the fine-grained noise situation by using complaint data together with social media, road network data, and POIs.
Data fusion methodologies

- Stage-based methods
- Feature level-based methods
- **Semantic meaning-based data fusion methods**
  - multi-view learning-based
  - similarity-based
  - probabilistic dependency-based
  - and transfer learning-based methods.
Probabilistic Dependency-Based Fusion

- This category of approaches bridges the gap between different datasets by the probabilistic dependency, which emphasize more about the interaction rather than the similarity between two objects.

- Two branches of graphical representations of distributions are commonly used:
  - Bayesian Networks
  - Markov Networks (a.k.a. Markov Random Field)
Probabilistic Dependency-Based Fusion Model

- The graphical structure of traffic volume inference model based on POIs, road networks, travel speed and weather.
  - A gray node denotes a hidden variable and white nodes are observations.
    - $\theta$: road hidden variable
    - $\alpha$: POI hidden variable
    - $N_a$: Traffic volume hidden variable
Data fusion methodologies

- Stage-based methods
- Feature level-based methods
- **Semantic meaning-based data fusion methods**
  - multi-view learning-based
  - similarity-based
  - probabilistic dependency-based
  - transfer learning-based methods.
Transfer learning-based methods

- An assumption in many machine learning algorithms is that the training and test data must be in the **same feature space** and have the **same distribution**.

- Transfer learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different.

- Examples:
  - A user’s transaction records in Amazon -> application of travel recommendation.
  - The knowledge learned from one city’s traffic data -> another city.
Taxonomy of Transfer learning

<table>
<thead>
<tr>
<th>Learning settings</th>
<th>Source and target domains</th>
<th>Source and target tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional ML</td>
<td>the same</td>
<td>the same</td>
</tr>
<tr>
<td>Inductive learning / unsupervised TL</td>
<td>the same</td>
<td>different but related</td>
</tr>
<tr>
<td>Transductive learning</td>
<td>different but related</td>
<td>different but related</td>
</tr>
</tbody>
</table>

Diagram:
- Transfer Learning
  - Same Task
    - Y: Transductive Transfer Learning
    - N: Unsupervised Transfer Learning
      - N: Label data in target domain
        - Y: Inductive Transfer Learning
        - N: Multi-task Learning
          - Y: Label in source domain
            - N: Self-taught Learning
Transfer between the Same Type of Datasets

- Examples of multi-task transfer learning

A) Book-travel interests co-estimation

- Physical location history
- Books browsed online

- Joint Task Learner
- Travel Packages {A, B, C}
- Book categories {war, romantic, sci-fi}

B) Air quality-traffic co-prediction

- Air quality data
- Traffic Data

- Joint Task Learner
- {Good, moderation, Unhealthy}
- {fast, normal, congestion}
Transfer Learning among Multiple Datasets

A) Complete Datasets and instances

B) Some datasets missing

C) Datasets complete but instance sparse

D) Datasets and instances missing
### Comparison of Different Data Fusion Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Meta</th>
<th>Labels</th>
<th>Goals</th>
<th>Train</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Vol</td>
<td>Pos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage-based</td>
<td>Y</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>L</td>
<td>Flex</td>
<td>F,P,C,O</td>
<td>S</td>
</tr>
<tr>
<td>Direct</td>
<td>N</td>
<td>L</td>
<td>Flex</td>
<td>F,P,A,O</td>
<td>U/S</td>
</tr>
<tr>
<td>DNN</td>
<td>N</td>
<td>L</td>
<td>Flex</td>
<td>F,P,A,O</td>
<td>U/S</td>
</tr>
<tr>
<td>Semantic</td>
<td>Y</td>
<td>S</td>
<td>Fix</td>
<td>F,P,O</td>
<td>S, SS</td>
</tr>
<tr>
<td>Multiview</td>
<td>Y</td>
<td>S</td>
<td>Fix</td>
<td>F,P,C,O,A</td>
<td>S/U</td>
</tr>
<tr>
<td>Similarity</td>
<td>N</td>
<td>S</td>
<td>Flex</td>
<td>F,A,O</td>
<td>U</td>
</tr>
<tr>
<td>Transfer</td>
<td>Y</td>
<td>S</td>
<td>Fix</td>
<td>F,P,A</td>
<td>S/U</td>
</tr>
</tbody>
</table>

Filling Missing Values (of a sparse dataset), Predict Future, Causality Inference, Object Profiling, and Anomaly Detection.
Thank you!

Q & A