Multi-Label Manifold Learning

Peng Hou, Xin Geng*, Min-Ling Zhang
MOE Key Laboratory of Computer Network and Information Integration,
School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
{hpeng, xgeng, zhangml}@seu.edu.cn

Abstract
This paper gives an attempt to explore the manifold in the label space for multi-label learning. Traditional label space is logical, where no manifold exists. In order to study the label manifold, the label space should be extended to a Euclidean space. However, the label manifold is not explicitly available from the training examples. Fortunately, according to the smoothness assumption that the points close to each other are more likely to share a label, the local topological structure can be shared between the feature manifold and the label manifold. Based on this, we propose a novel method called ML², i.e., Multi-Label Manifold Learning, to reconstruct and exploit the label manifold. To our best knowledge, it is one of the first attempts to explore the manifold in the label space in multi-label learning. Extensive experiments show that the performance of multi-label learning can be improved significantly with the label manifold.

1 Introduction
In multi-label learning, there are multiple labels associated to the same instance simultaneously (Tsoumakas, Katakis, and Vlahavas 2009; Zhang and Zhou 2014). Formally speaking, let $\mathcal{X} = \mathbb{R}^d$ be the $d$-dimensional feature space and $\mathcal{Y} = \{y_1, \ldots, y_q\}$ be the label set with $q$ possible labels. Given a training set $\mathcal{D} = \{(x_i, y_i)|1 \leq i \leq n\}$, where $x_i \in \mathcal{X}$ is the feature vector and $y_i \in \{0, 1\}^q$ is the label vector, the task of traditional multi-label learning is to learn a predictor which maps from the space of feature vectors to the space of label vectors. Each element of the label vector $y_i$ is a logical indicator of whether the corresponding label is relevant or irrelevant to the instance $x_i$. During the past decade, multi-label learning has been applied successfully to learn from the data with rich semantics, such as text (Rubin et al. 2012; Yang et al. 2009), image (Cabral et al. 2011; Wang, Huang, and Ding 2009), audio (Lo et al. 2011; Sanden and Zhang 2011), video (Wang et al. 2011), etc.

In this paper, we give an attempt to improve the multi-label learning performance with the manifold in the label space. To our best knowledge, it is one of the first attempts

*Corresponding author.
Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
For the scribe the image more apparently than the latter. Similarly, the value than the label sky, the label instance label variance boat within-sky both annotated with the labels values. An example is shown in Fig. 2. Image (a) and (b) are vant to different instances may also have different numerical 2.

In further detail, the difference of the relative impor-
tance revealed by the numerical label could be two-fold: 1. within-instance label variance, i.e., different labels relevant to the same instance may have different numerical values; 2. between-instance label variance, i.e., the same label relevant to different instances may also have different numerical values. An example is shown in Fig. 2. Image (a) and (b) are both annotated with the labels water, boat, mountain, and sky. Once extended to the numerical labels, for the within-instance label variance, the label boat should have larger value than the label sky in (b), because the former can describe the image more apparently than the latter. Similarly, for the between-instance label variance, the value of the label boat in (b) should be larger than the one in (a).

The rest of this paper is organized as follows. First, existing work related to our proposed approach is discussed in Section 2. The details of ML2 are proposed in Section 3. After that, the results of comparative studies are reported in Section 4. Finally, conclusions are drawn in Section 5.

2 Related Work

Existing multi-label approaches can be roughly grouped into three categories based on the thought of order of label correlations (Zhang and Zhou 2014). The simplest ones are the first-order approaches which assume independence among class labels (Boutell et al. 2004; Zhang and Zhou 2007). Then the multi-label classification becomes a series of binary classification problems. On the contrary, second-order approaches consider the correlations between pairs of class labels (Elisseeff and Weston 2001; Frankranz et al. 2008), and the high-order approaches consider the correlations among label subsets or all the class labels (Tsoumakas, Kakatis, and Vlahavas 2011). For all of them, the common modeling strategy is to treat each label in a crisp manner, i.e., being either relevant or irrelevant to an instance. In contrast, ML2 explores the manifold in the label space and treats the label as numerical. The label manifold contains more semantic information, which is beneficial for the learning process. There have been some multi-label works which transform the logical label space to the Euclidean label space. For example, (Tai and Lin 2012) tries to reduce the computational effort by seeking the principle correlations between labels, especially for the data sets with large numbers of labels. The bases of the Euclidean space are the combinations of the logical label vectors. Another work (Sun, Ji, and Ye 2011) projects the feature space and the label space to a new space where the correlation between the projections of the two spaces are maximized. In both cases, the dimensionality of the label space is reduced. However, ML2 differs from them without the dimensionality reduction. Besides, the former cases both produce a new space projected from the original label space, however, ML2 extends the original logical label space to a Euclidean space where the meaning of each dimension still remains.

Another more related work is Label Distribution Learning (LDL) (Geng, Yin, and Zhou 2013), which is a new machine learning paradigm where each instance is annotated by a label distribution. The label distribution covers a certain number of labels, representing the degree to which each label describes the instance. Thus the value of each label is numerical. However, LDL requires the availability of the label distributions in the training set, which is not always satisfiable for the real applications. On the contrary, ML2 can reconstruct the label manifold automatically from the logical multi-label data.

It is worthy to emphasize the difference between the manifold learning and ML2. Manifold learning assumes that the data of interest actually lie on an embedded non-linear manifold within the higher-dimensional space. Thus manifold learning is mostly used for the dimensionality reduction and visualization. The three famous local approaches in manifold learning are Locally Linear Embedding (LLE)
(Roweis and Saul 2000), Laplacian Eigenmaps (LE) (Belkin and Niyogi 2003) and Locality Preserving Projection (LPP) (Niyogi 2004). The reconstruction process of the label manifold in ML$^2$ is similar to LLE. However, the relation between the feature manifold and the label manifold is not embedding or dimensionality reduction. They are in two different spaces that merely share the local topological structure according to the smoothness assumption.

Note that the local topological structure is transferred from the feature space to the label space in ML$^2$, but it is different from transfer learning (Pan and Yang 2010). Transfer learning deals with two problems of different fields or distributions. The target of the transference mainly refers to the domain knowledge. However, ML$^2$ transfers the topological structure from the feature space to the label space.

3 The ML$^2$ Algorithm

As shown in Section 1, the training set of multi-label learning can be expressed as $D = \{(x_i, y_i) | 1 \leq i \leq n\}$. Given any instance $x_i \in \mathbb{R}^d$ and the logical label vector $y_i \in \{-1, 1\}^q$, we use $\mu_i \in \mathbb{R}^q$ to denote the numerical label vector. Note that here we use $-1$ instead of 0 in the logical label vector to represent irrelevant to the example. As many graph based learning methods do, the topological structure can be represented by a graph $G = < \mathcal{V}, \mathcal{E}, W >$, where $\mathcal{V}$ is the vertex set, $\mathcal{E}$ is the edge set in which each edge $e_{ij}$ represents the relationship between the data $x_i$ and $x_j$, and $W$ is the weight matrix with each element $W_{ij}$ representing the weight of the edge $e_{ij}$.

According to the smoothness assumption, the topological structure of the feature space can be transferred to the numerical label space local by local. In order to keep the locality, we need to use the local neighborhood information of each point to construct $G$. For computational convenience, we assume that each data point can be optimally reconstructed using a linear combination of its neighbors (Roweis and Saul 2000; Wang and Zhang 2008). Then the approximation of the feature manifold is to induce the minimization of

$$\mathcal{E}(W) = \sum_{i=1}^{n} \|x_i - \sum_{j \neq i} W_{ij} x_j\|^2,$$  \hspace{1cm} (1)

where $W_{ij} = 0$ unless $x_j$ is one of $x_i$’s $K$-nearest neighbors. Note that under most conditions $W_{ij} \neq W_{ji}$. Further for translation invariance, we constrain $1^T W_i = 1$, where $W_i = [W_{i1}, \ldots, W_{in}]^T$, and 1 is the vector of all ones.

Then the approximation can be solved by the following $n$ standard least square programming problems

$$\min_{W_i} W_i^2 \mathcal{E}(W), \hspace{1cm} \text{s.t.} \hspace{0.5cm} 1^T W_i = 1,$$  \hspace{1cm} (2)

where $G_i$ is the local Gram matrix at point $x_i$ with $G_{ik}^i = (x_i - x_k)^T (x_i - x_k)$.

With the transferred topological structure, the reconstruction of the label manifold can infer to the minimization of

$$\Phi(\mu) = \sum_{i=1}^{n} \|\mu_i - \sum_{j \neq i} W_{ij} \mu_j\|^2.$$  \hspace{1cm} (3)

Note that we are now minimizing with respect to the numerical label vector $\mu$ rather than $W$.

Besides, we add a constraint that makes the sign of the numerical label represent whether the corresponding label is relevant or irrelevant to the example that

$$\forall 1 \leq i \leq n, 1 \leq l \leq q : \mu_i \mu_l \geq \lambda,$$  \hspace{1cm} (4)

where $\lambda > 0$. The optimization for (3) with constraint (4) is a constrained quadratic programming process, and it can be solved efficiently.

There are three advantages for the constraint (4): 1. It is convenient to judge whether a label is relevant or irrelevant to the example by the sign of it; 2. It guarantees that the relevant numerical labels are larger than the irrelevant ones; 3. The minimum of the relevant numerical labels will be equal to $\lambda$ or the maximum of the irrelevant numerical labels will be equal to $-\lambda$. This makes the scale of the reconstructed numerical labels on the control.

The reconstructed numerical labels are real and the problem can not be treated as a classification but rather a regression problem. In the multi-label case, it is actually a multi-output regression problem. There have been some efficient algorithms proposed such as multi-output support vector regression (MSVR) (Prez-Cruz et al. 2002; Tuia et al. 2011; Chung et al. 2014), k-nearest neighbor regression (KNR) (Burba, Ferraty, and Vieu 2009) and structured output-associative regression (SOAR) (Bo and Smichesescu 2009). Here we propose a regressor based on the MSVR.

Similar to the MSVR, we generalize the 1-D SVR to solve the multi-dimensional case. In addition, our regressor not only concerns the distance between the predicted and the real values, but also the sign consistency of them. It leads to the minimization of

$$L(\Theta, b) = \frac{1}{2} \sum_{j=1}^{n} \|\theta^j\|^2 + C_{\lambda} \sum_{i=1}^{n} L_1(r_i) + C_2 \sum_{j=1}^{n} \sum_{l=1}^{q} L_2(t_{ij}),$$  \hspace{1cm} (5)

where $r_i = ||e_i|| = \sqrt{e_i^T e_i}$, $e_i = \mu_i - \varphi(x_i)^T \Theta - b$, $t_{ij} = y_i^j (\varphi(x_i)^T \theta^j + b^j)$, $\Theta = [\theta^1, ..., \theta^q]$, $b = [b^1, ..., b^q]$, and $\varphi(x)$ is a nonlinear transformation of $x$ to a higher-dimensional feature space $\mathbb{R}^d$.

To consider all dimensions into a unique restriction and yield a single support vector for all dimensions, the $L_1$ loss is set as

$$L_1(r) = \begin{cases} 0, & r < \varepsilon \\ \frac{r^2 - 2\varepsilon + \varepsilon^2}{\varepsilon}, & r \geq \varepsilon. \end{cases}$$  \hspace{1cm} (6)

This will create an insensitive zone determined by $\varepsilon$ around the estimate, i.e., the loss of $r$ less than $\varepsilon$ will be ignored.

To make the signs of the numerical label and the logical label same as much as possible, the $L_2$ loss is set as

$$L_2(t) = -t \sigma(-t) = \begin{cases} 0, & t > 0 \\ -t, & t \leq 0. \end{cases}$$  \hspace{1cm} (7)

where $\sigma(t)$ is an activation function where the value will be equal to 0 if $t$ is negative, otherwise the value will be equal to 1. The meaning of Eq. (7) is that if the signs of the predicted numerical label and the logical label are different, there will be some positive loss, otherwise the loss will be zero.
To minimize $L(\Theta, b)$, we use an iterative quasi-Newton method called Iterative Re-Weighted Least Square (IRWLS) (Prez-Cruz et al. 2000). Firstly, $L_1(\Theta, b)$ is approximated by its first order Taylor expansion at the solution of the current $k$-th iteration, denoted by $\Theta^{(k)}$ and $b^{(k)}$:

$$L_1'(r_i) = L_1(r_i) + \frac{dL_1(r)}{dr}_{r_i} (e_i - e_i^{(k)})$$

where $e_i^{(k)}$ and $r_i^{(k)}$ are calculated from $\Theta^{(k)}$ and $b^{(k)}$. Then a quadratic approximation is further constructed as

$$L_1''(r_i) = L_1(r_i) + \frac{dL_1(r)}{dr}_{r_i} r_i^2 - (r_i^{(k)})^2$$

where

$$a_i = \frac{1}{r_i^{(k)}} \frac{dL_1(r)}{dr}_{r_i}$$

and $\tau$ is a constant term that does not depend on either $\Theta^{(k)}$ or $b^{(k)}$. Combining Eq. (5), (7) and (9) can get

$$L''(\Theta, b) = \frac{1}{2} \sum_{j=1}^{n} \|\theta_j\|^2 + \frac{1}{2} C_1 \sum_{i=1}^{n} a_i r_i^2 - C_2 \sum_{i=1}^{n} \frac{1}{2} \sum_{j=1}^{n} t_i^j \sigma(-t_i^j) + \tau.$$

It is a piecewise quadratic problem whose optimum can be integrated as solving a system of linear equations for $j = 1, \ldots, n$:

$$C_1 \Phi^T D_n \Phi + I \quad C_1 \Phi^T a \quad \Theta^j = \left[ C_1 \Phi^T D_n \mu + C_2 \Phi^T D_n y^j \right]$$

where

$$(\Phi_i)^j = \varphi(x_1^j), \ldots, \varphi(x_n^j)^T, \quad a = [a_1, \ldots, a_n]^T,$$

and $D_n^j = \sigma(-t_i^j)$ is the Kronecker’s delta function, $(D_n)^j = \sigma(-t_i^j)$ and $\sigma^j = [\sigma(-t_1^j), \ldots, \sigma(-t_n^j)]^T$, $y^j = [y_1^j, \ldots, y_n^j]^T$. Then, the direction of the optimal solution of Eq. (12) is used as the descending direction for the optimization of $L(\Theta, b)$, and the solution for the next iteration

$$(\Theta^{(k+1)} and b^{(k+1)})$$

is obtained via a line search algorithm along this direction.

According to the representer’s theorem (Smola and Schlkopf 1998), under fairly general conditions, a learning problem can be expressed as a linear combination of the training examples in the feature space, i.e., $\Theta^j = \sum_i \varphi(x_i) \beta^j = \Phi^T \beta^j$. If we replace this expression into E. (12), it will generate the inner product $\langle \varphi(x_1), \varphi(x_j) \rangle$, then the kernel trick can be applied. After that the line search algorithm can be expressed in terms of $\beta^j$ and $b^j$.

## 4 Experiments

### 4.1 Experiment Configuration

#### Data Sets

For comprehensive performance evaluation, we collect one toy data set and fourteen real data sets for experimental studies. The toy data is 3-dimensional, where the third dimension is calculated as the Gaussian distribution of the first two dimensions. The mean of the Gaussian distribution is 0, and the variance on each dimension is 1. The label space is 2-dimensional, where the numerical label vector is calculated by $\mu = M^T x$, and

$$M = \begin{bmatrix} 0.48 & 0.21 \\ 0.87 & -0.76 \\ -0.12 & 0.3 \end{bmatrix}.$$

Fig. 3(a) shows the toy data points in the feature space and (b) shows the real numerical label points. The two figures demonstrate the rationality of the smoothness assumption.
Table 2: Predictive performance of each comparing algorithm (mean ± std. deviation) on the regular-scale data sets.

<table>
<thead>
<tr>
<th>Comparing algorithm</th>
<th>Hamming loss ↓</th>
<th>Coverage ↓</th>
<th>Ranking loss ↓</th>
<th>Average precision ↑</th>
<th>Macro-averaging AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAL500</td>
<td>llog</td>
<td>enron</td>
<td>image</td>
<td>scene</td>
</tr>
<tr>
<td>ML²</td>
<td>0.188±0.0016</td>
<td>0.705±0.014</td>
<td>0.714±0.021</td>
<td>0.861±0.005</td>
<td>0.948±0.004</td>
</tr>
<tr>
<td>BR</td>
<td>0.137±0.002</td>
<td>0.017±0.001</td>
<td>0.060±0.001</td>
<td>0.185±0.004</td>
<td>0.111±0.003</td>
</tr>
<tr>
<td>CLR</td>
<td>0.137±0.002</td>
<td>0.018±0.001</td>
<td>0.055±0.001</td>
<td>0.186±0.005</td>
<td>0.112±0.003</td>
</tr>
<tr>
<td>ECC</td>
<td>0.182±0.005</td>
<td>0.025±0.001</td>
<td>0.056±0.001</td>
<td>0.218±0.027</td>
<td>0.096±0.003</td>
</tr>
<tr>
<td>RAKEL</td>
<td>0.138±0.002</td>
<td>0.017±0.001</td>
<td>0.058±0.001</td>
<td>0.173±0.004</td>
<td>0.096±0.004</td>
</tr>
</tbody>
</table>

where the points close to each other in the feature space (a) are also close in the label space (b).

Table 1 summarizes detailed characteristics of the real data sets, which are roughly organized in ascending order of the number of examples |S|, with seven of them being regular-scale, i.e., |S| < 5,000 and seven of them being large-scale, i.e., |S| ≥ 5,000. As shown in Table 1, the fourteen data sets cover a broad range of cases with diversified multi-label properties and thus serve as a solid basis for thorough comparative studies.

Comparing Algorithms In this paper, we choose to compare the performance of ML² against four well-established multi-label learning algorithms: Binary Relevance (BR) (Boutell et al. 2004), Calibrated Label Ranking (CLR) (Frnkranz et al. 2008), Ensemble of Classifier Chains (ECC) (Read et al. 2011) and RAndom k-labelsets (RAKEL) (Tsoumakas, Katakis, and Vlahavas 2011), which learn from multi-label data based on various correlation orders among labels.

The number of neighbors K for ML² is set to q + 1, because it is necessary that K is larger than q to generate a q-dimensional space using K vectors. The parameters λ, C₁ and C₂ are set to 1, 1 and 10, respectively. The ensemble size for RAKEL is set to 2q with k = 3.

Evaluation Metrics We use six evaluation metrics widely-used in multi-label learning in this paper, i.e., Hamming loss, One-error, Coverage, Ranking loss, Average precision and AUC (Zhang and Zhou 2014). Note that for all the six multi-label metrics, their values vary between [0,1]. Furthermore, for average precision and AUC, the larger the values the better the performance; While for the other four metrics, the smaller the values the better the performance.
Table 3: Predictive performance of each comparing algorithm (mean ± std. deviation) on the large-scale data sets.

<table>
<thead>
<tr>
<th>Comparing algorithm</th>
<th>Hamming loss ↓</th>
<th>One-error ↓</th>
<th>Ranking loss ↓</th>
<th>Coverage ↓</th>
<th>Average precision ↑</th>
<th>Macro-averaging AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>corel5k</td>
<td>rcv1-s1</td>
<td>rcv1-s2</td>
<td>bibtex</td>
<td>corel16k-s1</td>
<td>corel16k-s2</td>
</tr>
<tr>
<td>ML²</td>
<td>0.010 ± 0.001</td>
<td>0.026 ± 0.001</td>
<td>0.023 ± 0.001</td>
<td>0.013 ± 0.001</td>
<td>0.021 ± 0.001</td>
<td>0.019 ± 0.001</td>
</tr>
<tr>
<td>BR</td>
<td>0.012 ± 0.001</td>
<td>0.031 ± 0.001</td>
<td>0.028 ± 0.001</td>
<td>0.015 ± 0.001</td>
<td>0.020 ± 0.001</td>
<td>0.019 ± 0.001</td>
</tr>
<tr>
<td>CLR</td>
<td>0.011 ± 0.001</td>
<td>0.029 ± 0.001</td>
<td>0.025 ± 0.001</td>
<td>0.014 ± 0.001</td>
<td>0.019 ± 0.001</td>
<td>0.018 ± 0.001</td>
</tr>
<tr>
<td>ECC</td>
<td>0.015 ± 0.001</td>
<td>0.030 ± 0.001</td>
<td>0.024 ± 0.001</td>
<td>0.017 ± 0.001</td>
<td>0.030 ± 0.001</td>
<td>0.018 ± 0.001</td>
</tr>
<tr>
<td>RAKEL</td>
<td>0.012 ± 0.001</td>
<td>0.031 ± 0.001</td>
<td>0.027 ± 0.001</td>
<td>0.015 ± 0.001</td>
<td>0.020 ± 0.001</td>
<td>0.019 ± 0.001</td>
</tr>
</tbody>
</table>

These metrics serve as good indicators for comprehensive comparative studies as they evaluate the performance of the learned models from various aspects.

4.2 Experimental Results

Fig. 3(c) shows the numerical labels reconstructed by ML² using the topological structure transferred from the feature space in (a) and the signs of the real labels in (b) (note that only the signs of the label vectors in (b) are used in ML² to simulate the real multi-label data). We can see from (b) and (c) that ML² indeed effectively recover the label manifold.

Table 2 and 3 report the detailed experimental results of each comparing algorithm on the regular-scale and large-scale data sets respectively. On each data set, 50% samples are randomly sampled without replacement to form the training set, and the rest 50% examples are used to form the test set. The sampling process is repeated for ten times and the average predictive performance across ten training/testing trials are recorded. For each evaluation metric, ↓ indicates the smaller the better while ↑ indicates the larger the better. Furthermore, the best performance among the five comparing algorithms is shown in boldface.

From the result table we can see, on the regular-scale data sets (Table 2), across all the evaluation metrics, ML² ranks 1st in 73.8% cases and ranks 2nd in 21.4% cases, and on the large-scale data sets (Table 3), across all the evaluation metrics, ML² ranks 1st in 57.1% cases and ranks 2nd in 40.5% cases. Thus ML² achieves competitive performance against the well-established multi-label learning algorithms across extensive benchmark data sets and diverse evaluation metrics, which validate the effectiveness of the label manifold for multi-label learning.
5 Conclusion

This paper explores the manifold in the label space for multi-label learning. Because the label manifold is not explicitly available from the training examples, we propose a novel method called ML² to reconstruct and exploit the label manifold based on the smoothness assumption. Extensive comparative studies clearly validate the advantage of ML² against the state-of-the-art multi-label learning approaches. In the future, we will explore if there exists better ways to estimate and make use of the label manifold for multi-label learning.

Acknowledgements

This work was partly supported by the National Science Foundation of China (61273300, 61232007, 61175049, 61222309), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), and the MOE Program for New Century Excellent Talents in University (NCET-13-0130).

References


