Learning a Feature Transformation to Improve Performance of Clustering and Classification

Xinlei Zhou

Big Data Technology and Application Institute
ShenZhen University

2019.03.13

Outlines

Outlines

Outlines

1. Introduction
   - What's a feature transformation?
   - How to transform?
2. What are our contributions?
3. Have we improved the performance?

What's a feature transformation:
Feature: the characteristic or prominent aspect of each data
Transformation: a series of operations, which make something presented in another form

The feature transformation is a data preprocessing technology performing a series of operations to reform the way of feature presented.

How to transform?
What's the series of operations?

Outlines

1. Introduction
   - What's a feature transformation?
   - How to transform?
2. What are our contributions?
3. Have we improved the performance?

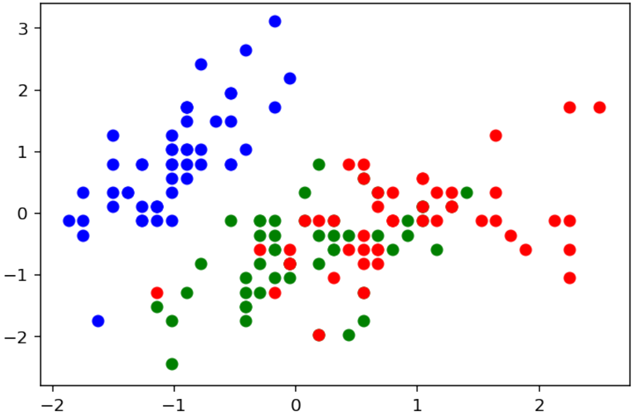

Figure 1: Distribution of Iris data set

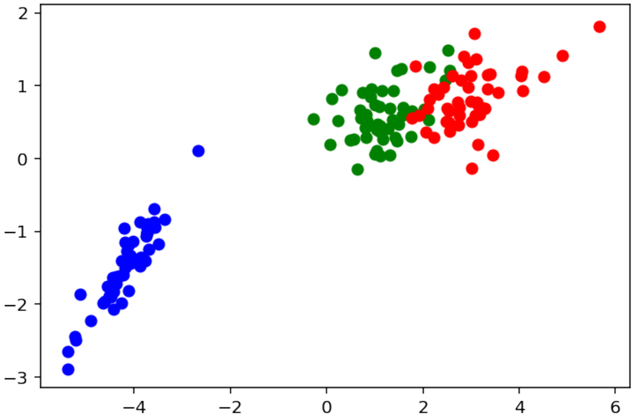

Figure 2: Distribution of Iris data set after transform

The distance between different categories increases, while the distance between the same categories decreases, and the separability of data increases significantly.

The **main idea** of our method is to learn a matrix **W** which maps the data onto a new feature space. The data in the new feature space will have better representation for clustering or classification tasks, and we call it **weight-matrix learning (WML)**.

Outlines

1. Introduction
   - What's a feature transformation?
   - How to transform?
2. What are our contributions?
3. Have we improved the performance?

Contributions and advantages of our method:

1. The mapping between new and old feature spaces is linear.

Suppose that $\boldsymbol{S} \subset \boldsymbol{R^n}$ is a data set containing N n-dimensional column vectors, and is represented as

$$S = \{\vec{x}_i | \vec{x}_i \subset R^n, i = 1, 2, \cdots, N\}. \tag{1}$$

The transformed data set is defined as

$$S_W = \{\vec{y}_i | \vec{y}_i = W\vec{x}_i, i = 1, 2, \cdots, N\}, \tag{2}$$

where $\boldsymbol{W} = (\boldsymbol{w_{ij}})_{\boldsymbol{n \times n}}$ is a full rank matrix to be determined.

2. The similarity in WML is based on a pseudo-distance, i.e., the square of weighted distance, rather than the distance itself. This improvement reduces the computational complexity of the similarity matrix to some extent.

$$\rho_{pq}^{(W)} = \frac{1}{1 + \beta \cdot d_{pq}^{(W)}} \tag{3}$$

where

$$d_{pq}^{(W)} = d^2(\vec{y_p}, \vec{y_q}) = (\vec{x_p} - \vec{x_q})^T (W^T W)(\vec{x_p} - \vec{x_q}), \tag{4}$$

β is a positive parameter determined by solving the following Equation (5)

$$\frac{2}{N(N-1)} \sum_{q>p} \rho_{pq}^{(I)} = 0.5 \tag{5}$$

where $\boldsymbol{N}$ is the number of objects, $\boldsymbol{\rho_{pq}^{(I)}}$ is the value of $\boldsymbol{\rho_{pq}^{(W)}}$ at $\boldsymbol{W} = \boldsymbol{I}$, indicating the similarity of the original data.

For the purpose of reducing uncertainty of the similarity matrix, we consider the minimization of the following evaluation function (objective function):

$$\begin{aligned} E(W) &= \frac{1}{N(N-1)} \sum_{q<p} E_{pq}(W) \\ &= \frac{1}{N(N-1)} \sum_{q<p} (\rho_{pq}^{(W)}(1-\rho_{pq}^{(I)}) + \rho_{pq}^{(I)}(1-\rho_{pq}^{(W)})) \end{aligned} \tag{6}$$

in which $\boldsymbol{N}$ is the number of objects, $\boldsymbol{W}$ represents the feature weight matrix, $\boldsymbol{\rho_{pq}^{(W)}}$ specified by Equation (3) is the similarity between objects $\vec{\boldsymbol{x}}_{\boldsymbol{p}}$ and $\vec{\boldsymbol{x}}_{\boldsymbol{q}}$, and $\boldsymbol{\rho_{pq}^{(I)}}$ is defined in Equation (5).

This evaluation function $\boldsymbol{E}(\boldsymbol{W})$, is constructed based on a simple function

$$f(x,y) = x(1-y) + y(1-x)(0 \leq x, y \leq 1) \tag{7}$$

Noting that $\boldsymbol{\frac{\partial f}{\partial x} = 1 - 2y}$:
$\boldsymbol{\frac{\partial f}{\partial x} > 0}$ if $\boldsymbol{y < 0.5}$, $\boldsymbol{\frac{\partial f}{\partial x} < 0}$ if $\boldsymbol{y > 0.5}$.

Therefore, the function $\boldsymbol{f(x, y)}$ with respect to $\boldsymbol{x}$:
is a strictly monotonically increasing function under the condition of fixed $\boldsymbol{y < 0.5}$ and
is a strictly monotonically decreasing function under the condition of fixed $\boldsymbol{y > 0.5}$.

3. The objective function we designed has ensured the interpretability of the data mapping process, which is impossible for most methods.

We can explain the optimization process of the objective function as follows:
We take 0.5 as a reference center of similarity value for a data set. During the optimization process the similarity deviates from 0.5 and approaches 0 or 1 gradually. That is, **the similarity before the transformation (which is larger than 0.5) larger, and the similarity before the transformation (which is smaller than 0.5) smaller.**

4. We place the WML into a feed-forward neural network in which the stochastic gradient descent or batch gradient descent algorithm or other gradient-based training techniques can be well used.

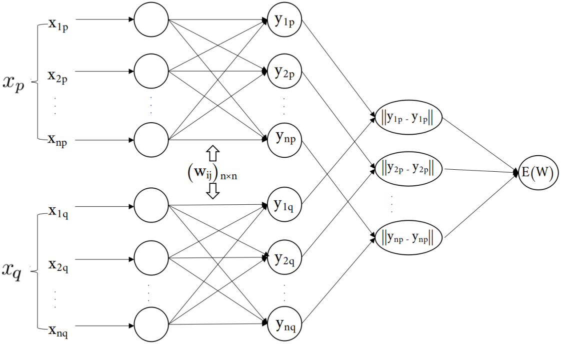

Figure 3: Network representation WML linear transformation

Outlines

1. Classification tasks

We take RWN(Random Weighted Network) and C4.5 to validate the performance of our method in classification tasks.

Data set	$RWN_{original}$	RWN_{FWL}	RWN_{WML}	$C4.5_{original}$	$C4.5_{FWL}$	$C4.5_{WML}$
1	0.9428	0.9360	**0.9693**	1.0000	1.0000	1.0000
2	0.8983	0.8922	**0.9857**	1.0000	1.0000	1.0000
3	0.9600	0.9590	**0.9694**	1.0000	1.0000	1.0000
4	0.9525	0.9435	**0.9719**	1.0000	1.0000	1.0000
5	0.7890	**0.7910**	0.7806	1.0000	1.0000	1.0000
6	0.8679	**0.8703**	0.8692	1.0000	1.0000	1.0000
7	**0.7327**	0.7216	0.7211	1.0000	1.0000	1.0000
8	0.7946	0.8082	**0.8154**	1.0000	1.0000	1.0000
9	0.7080	0.7130	**0.7184**	1.0000	1.0000	1.0000
10	**0.9437**	0.9421	0.9212	0.9976	0.9976	0.9976
11	**0.9083**	0.9077	0.8885	1.0000	1.0000	1.0000
12	0.7741	0.7805	**0.7840**	1.0000	1.0000	1.0000
13	0.7583	0.7711	**0.7939**	1.0000	1.0000	1.0000
14	0.8278	0.8293	**0.8384**	1.0000	1.0000	1.0000
15	0.9866	0.9894	**0.9937**	1.0000	1.0000	1.0000

Note: $\boldsymbol{RWN_{original}}$ is the result of RWN on the original data set; $\boldsymbol{RWN_{FWL}}$ is the result of RWN on data set transformed by FWL; $\boldsymbol{RWN_{WML}}$ is the result of RWN on data set transformed by WML.

Figure 4: Training Accuracy

Data set	$RWN_{original}$	RWN_{FWL}	RWN_{WML}	$C4.5_{original}$	$C4.5_{FWL}$	$C4.5_{WML}$
1	0.9248	0.9333	**0.9589**	**1.0000**	**1.0000**	0.9695
2	0.8759	0.8724	**0.9741**	**1.0000**	**1.0000**	0.9655
3	0.9583	0.9563	**0.9658**	**0.9698**	**0.9698**	0.9623
4	0.9447	0.9447	**0.9509**	**0.9316**	**0.9316**	0.9272
5	**0.741**	**0.741**	0.7077	**0.7154**	**0.7154**	0.6256
6	**0.8709**	0.8687	0.8619	**0.8179**	**0.8179**	0.7978
7	0.6535	**0.6651**	0.6465	**0.6628**	**0.6628**	0.6465
8	0.7887	**0.8113**	0.8056	**0.8732**	**0.8732**	0.8648
9	0.7065	0.7069	**0.7104**	0.5922	0.5922	**0.6355**
10	**0.9412**	0.9404	0.9185	**0.9611**	**0.9611**	0.9392
11	0.8692	0.8795	**0.8846**	**0.8821**	**0.8821**	0.7821
12	0.7656	0.7526	**0.7701**	**0.6974**	**0.6974**	0.6643
13	0.7615	0.7678	**0.7931**	0.7388	0.7388	**0.8085**
14	0.8285	0.8252	**0.834**	0.7576	0.7576	**0.7623**
15	0.9667	0.9667	**0.9833**	**0.9056**	**0.9056**	0.8361

Note: $\boldsymbol{RWN_{original}}$ is the result of RWN on the original data set; $\boldsymbol{RWN_{FWL}}$ is the result of RWN on data set transformed by FWL; $\boldsymbol{RWN_{WML}}$ is the result of RWN on data set transformed by WML.

Figure 5: Testing Accuracy

Performance analysis:

1. RWN is a **network structure** that reflects the mapping relationship between the feature space of the data and category space. **The complexity of the feature space of the data set directly affects the performance of the RWN.** Our WML performs a linear transformation on the data, which can reduce the uncertainty of the similarity matrix of the data. **Noting that the uncertainty of the similar matrix may be closely related to the complexity of the feature space, decreasing the uncertainty of the similar matrix may reduce the complexity of the feature space.** That is, WML may reduce the learning difficulty of RWN. Therefore, our WML can improve the training accuracy and testing accuracy of RWN algorithm.

2. C4.5 is a rule learning algorithm that reduces the information entropy of data by continuously selecting feature segmentation points with high information gain rate. C4.5 does not use the information of the similarity matrix of the data, so theoretically WML can not improve the performance of C4.5, which is consistent with our experimental results.

2. Clustering tasks

We take K-means to validate the performance of our method in clustering tasks. In order to evaluate the clustering results, we select four internal indexes DBI, DUNN, CHI, and SI as the evaluation criteria for clustering.

Data set	DBIo	DBIv	DBIm	DUNNo	DUNNv	DUNNm	SIo	SIv	SIm	CHIo	CHIv	CHIm
1	1.988	1.988	**0.532**	0.996	0.996	**3.396**	0.177	0.181	**0.625**	145.445	159.052	**1648.193**
2	2.405	2.403	**0.786**	0.774	0.776	**2.365**	0.131	0.142	**0.492**	46.08	48.425	**388.516**
3	1.124	1.124	**0.942**	**1.411**	**1.411**	1.352	0.309	0.329	**0.375**	1183.248	1262.696	**2777.321**
4	1.136	1.136	**0.509**	1.492	1.492	**2.830**	0.339	0.385	**0.698**	313.506	364.093	**1288.987**
5	1.823	1.823	**0.829**	1.076	1.076	**2.140**	0.007	0.189	**0.45**	4.501	51.434	**223.459**
6	1.795	1.795	**0.362**	1.059	1.059	**3.036**	0.147	0.221	**0.952**	114.667	192.183	**1271.013**
7	0.963	1.181	**0.777**	**0.804**	0.750	0.706	-0.05	0.382	**0.567**	24.443	87.751	**258.49**
8	1.506	1.506	**0.898**	1.025	1.025	**1.744**	0.153	0.297	**0.473**	17.062	120.158	**319.602**
9	0.933	0.933	**0.732**	2.027	2.027	**2.397**	0.016	0.453	**0.544**	22.062	875.234	**1324.064**
10	1.021	1.020	**0.443**	0.937	**0.939**	0.330	0.293	0.362	**0.8**	243.433	2900.386	**16357.606**
11	1.388	1.388	**0.907**	1.313	1.313	**1.596**	0.102	0.276	**0.574**	28.893	84.217	**173.415**
12	1.608	1.608	**0.919**	1.154	1.154	**1.917**	0.261	0.405	**0.635**	47.515	237.111	**615.153**
13	2.247	2.247	**1.172**	0.879	0.879	**1.691**	0.056	0.13	**0.292**	461.969	845.81	**4561.295**
14	1.497	1.497	**0.722**	1.306	1.306	**2.686**	0.102	0.233	**0.458**	905.872	1901.813	**11125.043**
15	1.305	1.309	**0.611**	1.360	1.359	**1.912**	0.292	0.301	**0.524**	79.804	83.351	**525.932**

Figure 6: clustering result

Summary:

Through the above analysis, it can be concluded that WML has the ability to improve the performance of similarity-based learning algorithm significantly, which could be proved by accuracy of classification tasks or several clustering indexes.

Thanks!