Label Correlation Propagation for Semi-supervised Multi-label Learning

Aritra Ghosh(✉) and C. Chandra Sekhar

Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India
aritrag94@gmail.com, chandra@cse.iitm.ac.in

Abstract. Many real world machine learning tasks suffer from the problem of scarce labeled data. In multi-label learning, each instance is associated with more than one label as in semantic scene understanding, text categorization and bio-informatics. Semi-supervised multi-label learning has attracted recent interest as gathering labeled data is both expensive and requires manual effort. Further, many of the labels have semantic correlation which manifests as co-occurrence and this information can be used to build effective classifiers in the multi-label scenario. In this paper, we propose two different graph based transductive methods, namely, the label correlation propagation and the k-nearest neighbors based label correlation propagation. Extensive experimentation on real-world datasets demonstrates the efficacy of the proposed methods and the importance of using the label correlation information in semi-supervised multi-label learning.

Keywords: Semi-supervised learning · Multi-label learning · Graph based learning

1 Introduction

In supervised learning based approaches to multi-class pattern classification, a training example represented by a corresponding feature vector is related to a distinct class (label) describing its semantics. However, for many real-world objects, the single label assumption may not be appropriate. In the task of image annotation, an image can have multiple labels, referred to as a relevant label set. Likewise, in the text categorization task, a news article can be associated with a number of topics like “military”, “business” and “international”. Multi-label learning has been used for a number of applications like automatic multimedia content annotation [1,2] and bioinformatics [3,4]. The multi-label learning task involves building models which can predict the relevant label set for a test example.

The existing techniques for multi-label learning are predominantly supervised learning based approaches. These techniques require huge amount of labeled data for building a classifier. As labeling the data is both time-consuming and costly,

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 52–60, 2017.
https://doi.org/10.1007/978-3-319-69900-4_7

it is undesirable to use only labeled data. However, unlabeled data is easily available and cheap, and the information from it can be used to build better classifiers. In the recent times, semi-supervised learning techniques have been found to be effective in building classifiers.

There have been a number of techniques for semi-supervised multi-label learning (SSMLL) [5–7]. Most of these methods are transductive in nature and they aim at predicting the label set for the existing unlabeled data. It is important to exploit the inherent label correlation present among the labels to boost the performance of the multi-label classifier. For example, it is common to have the natural scene images that contain both "hill" and "tree".

In [5], a graph-based learning framework that accounts for label consistency in the graph and the correlation among labels is presented. After optimising an objective function, a closed form solution for prediction of labels for the unlabeled data is obtained. In [6], the TRAnsductive Multi-label classification (TRAM) is formulated as an optimization problem of estimating label compositions and a closed-form solution is obtained. This method is extended for estimation of the cardinality of the predicted label set for the unlabeled examples based on the estimated label compositions. In [8], the non-negative matrix factorization algorithm is used to solve the problem where the basic hypothesis is that *"two examples which have high similarity in the input space should have similar label memberships"*. All the above methods build a graph with nodes representing the labeled and unlabeled examples, and a similarity measure used as weight of an edge between two nodes. Graph based methods form the majority of semi-supervised learning [9] because of their effectiveness and efficacy.

The rest of the paper is organized as follows. We present the graph based methods for semi-supervised learning from multi-label data in Sect. 2. In Sect. 3, we propose two methods to propagate the label correlation in multi-label learning. Studies on benchmark datasets that demonstrate the effectiveness of the proposed methods are presented in Sect. 4.

2 Graph Based Methods for Semi-supervised Learning with Multi-label Data

In semi-supervised multi-label learning, the training set $\mathcal{D} = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_i, Y_i), \ldots, (\mathbf{x}_L, Y_L), \mathbf{x}_{L+1}, \ldots, \mathbf{x}_{L+j} \ldots, \mathbf{x}_{L+U}\}$ consists of L labeled examples $\mathcal{D}_l = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_i, Y_i) \ldots, (\mathbf{x}_L, Y_L)\}$, and U unlabeled examples $\mathcal{D}_u = \{\mathbf{x}_{L+1}, \ldots, \mathbf{x}_{L+j} \ldots, \mathbf{x}_{L+U}\}$. The total number of examples is $N = L + U$. The task involves learning a family of K functions, $f_k : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$. Here, $f_k(\mathbf{x}_i, y_k)$ is measure of the confidence of the k^{th} label $y_k \in \mathcal{Y}$ being a label of $\mathbf{x}_i$. The label vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})^T$ is represented as a K-dimensional vector with $y_{ij} \in \{0, 1\}$, 0 indicating that the j^{th} label is not associated with the example $\mathbf{x}_i$ and 1 indicating that the label y_j belongs to the label set of the example $\mathbf{x}_i$. Let $\mathbf{W}$ denote the $N \times N$ weight matrix where w_{ij} represents similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. The matrix $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$ is called the combinatorial graph Laplacian matrix where $\mathbf{D}$ is a $N \times N$ diagonal matrix with entries $d_{ii} = \sum_{j=1}^{N} w_{ij}$.

The normalized combinatorial Laplacian is $L = \mathbf{D^{-1/2} \Delta D^{-1/2}}$. We define $\mathbf{\Lambda}$, a $N \times N$ diagonal matrix with $\lambda_{ii} = \infty$ for $i \leq L$, and $\lambda_{ii} = 0$ otherwise. The vector $\mathbf{f} = [f_1 \;\; f_2 \cdots f_N]^T$ has the confidence scores for each of the N examples. This setting is for a binary setting that can be extended to multi-label setting.

In [9], the objective function in the Gaussian Random Field (GRF) method for graph based semi-supervised learning in the single label setting is formulated as follows:

$$E(\mathbf{f}) = E_l(\mathbf{f}) + \alpha E_s(\mathbf{f}) \text{ where} \tag{1}$$

$$E_l(\mathbf{f}) = \infty \sum_{i \in \mathcal{L}} (f_i - y_i)^2 = (\mathbf{f} - \mathbf{y})^T \Lambda (\mathbf{f} - \mathbf{y}) \tag{2}$$

$$E_s(\mathbf{f}) = \frac{1}{2} \sum_{i,j \in \mathcal{L} \cup \mathcal{U}} w_{ij} (f_i - f_j)^2 = \mathbf{f}^T \Delta \mathbf{f} \tag{3}$$

Here, $E_l(\mathbf{f})$ is the term that corresponds to deviation from the already assigned labels (labeled data) and $E_s(\mathbf{f})$ is the penalty term that corresponds to smoothness of labels over the graph. In the second term, if the two examples $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar, the predictions f_i and f_j should be close as well.

In [10], the authors propose the Local and Global Consistency (LGC) method where the two terms are modified given below.

$$E_l(\mathbf{f}) = \sum_{i \in \mathcal{L} \cup \mathcal{U}} (f_i - y_i)^2 = (\mathbf{f} - \mathbf{y})^T (\mathbf{f} - \mathbf{y}) \tag{4}$$

$$E_s(\mathbf{f}) = \frac{1}{2} \sum_{i,j \in \mathcal{L} \cup \mathcal{U}} w_{ij} \left(\frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}}\right)^2 = \mathbf{f}^T L \mathbf{f} \tag{5}$$

In the multi-label setting, the matrix $\mathbf{Y}$ is an $N \times K$ matrix such that y_{ik} is equal to 1 if a labeled example $\mathbf{x}_i$ has label k associated with it, and 0 otherwise. This corresponds to the given ground truth. Similarly, the predicted matrix $\mathbf{F}$ is an $N \times K$ matrix where f_{ik} indicates the confidence of the example $\mathbf{x}_i$ in the label y_k.

The approach in [5] introduces a term $E_c(\mathbf{F})$ corresponding to regularizer for the label correlation. In the $K \times K$ label correlation matrix $\mathbf{C}$, the entry c_{kl} represents the correlation between label y_k and label y_l that can be estimated using the label based co-occurrence. The term $E_c(\mathbf{F})$ is defined as follows:

$$E_c(\mathbf{F}) = \sum_{i=1}^{N} \sum_{k,l=1}^{K} c_{kl} (f_{ik} - f_{il})^2 = -tr(\mathbf{F C^{'} F^T}) \tag{6}$$

where $\mathbf{C}^{'} = \mathbf{C} - \mathbf{D}_c$ and $\mathbf{D}_c$ is a diagonal matrix with diagonal entries $d_{c^{'}_{ii}} = \sum_{j=1}^{K} c_{ij}$. Here, $tr(\mathbf{M})$ is the trace of the matrix $\mathbf{M}$. The term $E_c(\mathbf{F})$ quantifies the smoothness in the label space rather than in the input space. If the correlation between two labels y_k and y_l is high, the predictions f_{ik} and f_{il} should be similar.

In the multi-label case, the terms $E_l(\mathbf{F})$, $E_s(\mathbf{F})$, and $E_c(\mathbf{F})$ are computed as follows:

$$E_l(\mathbf{F}) = tr((\mathbf{F} - \mathbf{Y})^{\mathbf{T}} \mathbf{\Lambda} (\mathbf{F} - \mathbf{Y})) \tag{7}$$

$$E_s(\mathbf{F}) = tr(\mathbf{F}^{\mathbf{T}} \mathbf{\Delta} \mathbf{F}) \tag{8}$$

$$E_c(\mathbf{F}) = -tr(\mathbf{F}\mathbf{C}^{'} \mathbf{F}^{\mathbf{T}}) \tag{9}$$

$$E(\mathbf{F}) = tr((\mathbf{F} - \mathbf{Y})^{\mathbf{T}} \mathbf{\Lambda} (\mathbf{F} - \mathbf{Y})) + \alpha.tr(\mathbf{F}^{\mathbf{T}} \mathbf{\Delta} \mathbf{F}) - \beta.tr(\mathbf{F}\mathbf{C}^{'} \mathbf{F}^{\mathbf{T}}) \tag{10}$$

where α and β are the trade-off parameters. The formulation in (10) is referred to as Multi-Label Correlation Gaussian Random Field (MLC-GRF) and the solution turns out to be the Sylvester Equation [11].

Similarly, the objective function in the Multi-Label Correlation Local and Global Consistency (MLC-LGC) method is formulated in Eq. (11) where μ, ν are hyper-parameters. The solution to the optimization problem in (11) is given by Eq. (12)

$$E(\mathbf{F}) = tr((\mathbf{F} - \mathbf{Y})^{\mathbf{T}} (\mathbf{F} - \mathbf{Y})) + \mu.tr(\mathbf{F}^{\mathbf{T}} \mathbf{L} \mathbf{F}) - \nu.tr(\mathbf{F}\mathbf{C}^{'} \mathbf{F}^{\mathbf{T}}) \tag{11}$$

$$(\mu \mathbf{L} + \mathbf{I})\mathbf{F} - \nu \mathbf{F}\mathbf{C}^{'} = \mathbf{Y} \tag{12}$$

This equation also turns out to be a Sylvester equation similar to the solution of the MLC-GRF method.

3 Proposed Methods for Semi-supervised Multi-label Learning

3.1 Label Correlation Propagation-GRF (CP-GRF)

As discussed in previous sections, incorporating label correlation information can help improve the performance of the classifier. Our fundamental hypothesis here is that the predictions for a given example should be label correlation consistent i.e., if two labels are correlated and the prediction score for one of the labels is high, the score for the other label should also be high. Thus, the labels of one example are propagated to the correlated labels of that example as follows:

$$f_{ik}^{(t+1)} = f_{ik}^{initial} + \alpha \sum_{l=1}^{K} f_{il}^{t} c_{lk} \quad i = 1, 2...N \quad k = 1, 2...K \tag{13}$$

Here, f^t represents the prediction for a given example at iteration t. The prediction score of a label for a given example is obtained both from the initial prediction ($f^{initial}$) and the other labels based on the correlation between the labels. The parameter α balances the two terms and is chosen by cross-validation. This iterative update is repeated till convergence. The Eq. (13) in the matrix form is given below.

$$\mathbf{F}^{(t+1)} = \mathbf{F}^{inital} + \alpha \mathbf{F}^{t} \mathbf{C} \tag{14}$$

The CP-GRF method is similar to the page rank approach [12] where the page rank of a webpage is proportional to the page rank of its incoming neighbours. The CP-GRF method takes the higher order label correlations into consideration by iteratively propagating the second order correlations. This method involves the propagation on label correlation graph for each example. The CP-GRF method is expected to perform well when the number of labels is large and the labels have significant correlations. The initial predictions can in principle be obtained from any typical multi-label classifier. In order to validate our method, $\mathbf{F}^{initial}$ is obtained from the GRF method.

3.2 Weighted Label Correlation Propagation-GRF (WCP-GRF)

The WCP-GRF method is an extension of the CP-GRF method. In this method, the correlation is propagated not only from the other correlated labels but also based on the examples close to the particular example. The hypothesis here is that the label correlation is a local effect i.e., the predictions for a given label will be influenced by the predictions for correlated labels of neighbors. Let $kNN(\mathbf{x_i})$ represent the k-nearest neighbours of $\mathbf{x_i}$. The update equations in the proposed WCP-GRF method are given by:

$$f_{ik}^{t+1} = f_{ik}^{initial} + \alpha \sum_{\mathbf{x}_j \in kNN(\mathbf{x}_i)} w_{ij} f_{jl}^{t} c_{lk} \tag{15}$$

$$\mathbf{F}^{(t+1)} = \mathbf{F}^{inital} + \alpha \mathbf{W} \mathbf{F}^{t} \mathbf{C} \tag{16}$$

As seen in (15), in each iteration there is a contribution from the initial prediction (any multi-label classifier) as well as from the correlated labels of the k nearest neighbors in the feature space. Since the contribution is only from the nearest neighbors, the WCP-GRF method uses the k-NN based weight matrix $\mathbf{W}$ i.e., the weight entries are 0 if the two nodes are not in the k-nearest neighbors of each other. The matrix update in WCP-GRF method is given by Eq. (16). Here, the initial predictions $\mathbf{F}^{initial}$ are obtained from the GRF method and therefore this is called WCP-GRF method. We use the iterative method and terminate at convergence or after a sufficient number of iterations.

The rest of the proposed methods are the same as the previously discussed methods but use the normalized graph Laplacian instead of the usual graph Laplacian. The other two proposed methods use the LGC method for the initial predictions. The Label Correlation Propagation-LGC (CP-LGC) is the same as Section in 3.1 except for the fact the initial predictions come from the LGC method. Similarly, the Weighted Label Correlation Propagation (WCP-LGC) is the same as in Sect. 3.2 except for the fact that the initial predictions are taken from the LGC method.

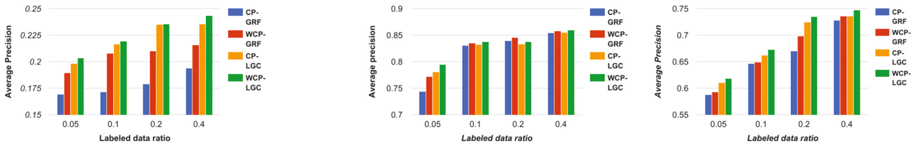
4 Experiments and Results

Details of benchmark datasets used for comparison of the various methods are given in Table 1. The evaluation metrics used to compare the various methods are: One error, Coverage, Average precision, Hamming loss and Ranking loss [13]. The label correlation matrix **C** is calculated using the cosine similarity on the labeled data. For most of the datasets, the label correlation for several pairs of labels is low. However, there are a few pairs of labels with high correlation.

Table 1. Details of the datasets used for experimentation

Dataset	Domain	Examples	Attributes	Labels	Cardinality
Yeast	Biology	2417	103	14	4.237
Image	Image	2000	135	5	1.24
Scene	Image	2407	294	6	1.074
MSRC-v2	Image	591	630	21	2.394
Corel-5k	Image	5000	499	374	3.522

The dataset has been divided into 10% labeled data, 70% unlabeled data and the rest as test data. All hyper-parameters were selected by choosing the maximum average-precision on the validation data over 5 runs. All the results correspond to average of 5 runs and the standard deviation of each metric has also been recorded. The weight matrix used is k-nearest neighbours with a Gaussian function where the width parameter is chosen based on performance on the validation data. The number of nearest neighbours (k) is fixed to 15 as it does not affect the results much. For Hamming loss, the number of labels for a given example is chosen based on the average cardinality of the dataset. All the techniques have been implemented in MATLAB and run on 32 GB RAM 8-core machine. The parameters β, μ and ν were chosen based on cross-validation. The performance of different methods for multi-label classification on different datasets is presented in Table 2. In Fig. 1, the average precision for different sizes of the labeled dataset and for various datasets is plotted. The test ratio is fixed at 20% and the labeled data size is varied. In all the datasets, it is seen that the average precision increases with increase in size of labeled dataset.

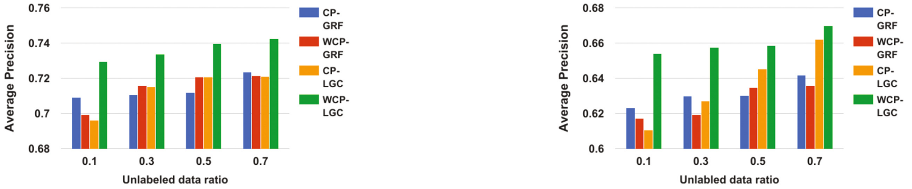


(a) Corel5K Dataset (b) Scene Dataset (c) MSRCV-2 Dataset

Fig. 1. Variation of average precision with size of the labeled data for different datasets

Table 2. Performance of the transductive methods for multi-label classification on different datasets

Dataset	Method	HLoss ↓	RLoss ↓	OneEr ↓	Cover ↓	AvePrec ↑
Corel-5k	GRF	0.0182	0.1705	0.8760	0.3571	0.1431
	MLC-GRF	0.0175	0.1699	0.8534	0.3592	0.1675
	CP-GRF	0.0179	0.1695	0.8423	0.3609	0.1715
	WCP-GRF	**0.0162**	0.1782	**0.7762**	0.3606	**0.2077**
	LGC	0.0161	0.1599	0.765	0.3498	0.2187
	MLC-LGC	0.0165	0.1594	0.7687	0.3521	0.2147
	CP-LGC	0.0162	0.1592	0.7697	0.3430	0.2168
	WCP-LGC	0.0163	**0.1554**	**0.7593**	0.3455	**0.2193**
Yeast	GRF	0.2217	0.1910	0.2526	0.4633	0.7304
	MLC-GRF	0.2148	0.1842	0.2601	0.4571	0.7389
	CP-GRF	0.2149	0.1819	0.2410	0.4605	0.7426
	WCP-GRF	**0.2082**	**0.1792**	**0.2373**	**0.4547**	**0.7467**
	LGC	0.2208	0.1933	0.2588	0.4639	0.7262
	MLC-LGC	0.2247	0.1949	0.3052	0.4662	0.7290
	CP-LGC	0.2096	0.1804	0.2422	0.4590	0.7421
	WCP-LGC	0.2133	0.1846	**0.2398**	0.4615	**0.7431**
Scene	GRF	0.1173	0.1066	0.3168	0.1017	0.8132
	MLC-GRF	0.1104	0.1002	0.2906	0.0988	0.8267
	CP-GRF	0.1078	0.0974	0.2865	0.0952	0.8310
	WCP-GRF	**0.1038**	**0.0958**	**0.2761**	**0.0932**	**0.8351**
	LGC	0.1107	0.1004	0.2919	0.0985	0.8274
	MLC-LGC	0.1066	0.0952	0.2869	0.092	0.8306
	CP-LGC	0.1058	0.0964	0.2798	0.0945	0.8326
	WCP-LGC	**0.1040**	**0.0916**	**0.2728**	**0.0910**	**0.8378**
Image	GRF	0.3336	0.2937	0.5275	0.2854	0.6628
	MLC-GRF	0.3355	0.2836	0.4920	0.2795	0.6783
	CP-GRF	0.3245	0.2664	0.4935	0.2656	0.6844
	WCP-GRF	**0.3215**	0.2695	**0.4795**	**0.2634**	**0.6898**
	LGC	0.3232	0.2699	0.4769	0.2678	0.6898
	MLC-LGC	0.3224	0.2687	0.4775	0.2662	0.6901
	CP-LGC	0.3188	0.2627	0.4761	0.2634	0.6945
	WCP-LGC	0.3235	**0.2626**	**0.4655**	**0.2622**	**0.6974**
MSRCv-2 DATASET	GRF	0.1085	0.1883	0.401	0.3227	0.6157
	MLC-GRF	0.1111	0.1917	0.4227	0.3178	0.6225
	CP-GRF	0.1090	0.1750	0.3672	0.3055	0.6468
	WCP-GRF	**0.1074**	**0.1585**	**0.3589**	**0.2810**	**0.6492**
	LGC	0.1316	0.1845	0.3728	0.3029	0.6464
	MLC-LGC	0.1246	0.1543	0.3985	0.2679	0.6525
	CP-LGC	0.1232	0.1609	0.3680	0.2786	0.6625
	WCP-LGC	0.1239	**0.1484**	**0.3611**	**0.2641**	**0.6732**



(a) Yeast Dataset

(b) Image Dataset

Fig. 2. Variation of average precision with the size of the unlabeled dataset for different datasets

In Fig. 2, the variation of average precision with size of unlabeled dataset for two of the datasets is plotted. We keep the labeled ratio fixed at 5% and test ratio at 20%. We vary the unlabeled dataset size and observe the effect on average precision. Again, we observe that the average precision increases with increase in the unlabeled data thus showing the importance of unlabeled data.

We observe the following for the transductive methods:

- In most of the cases, the weighted label correlation based method perform better than the other methods.
- In general, the LGC based methods perform better than the GRF based methods. This is expected as the normalized combinatorial Laplacian is a better representative than the conventional combinatorial graph Laplacian.
- Increase in labeled and unlabeled (to a certain extent) data results in an increase in performance

5 Summary and Conclusion

In this paper, we introduced the problem of semi-supervised multi-label learning and discussed some of the recent graph-based semi-supervised methods. We proposed the label correlation based propagation methods to improve the predictions. The proposed methods outperform the state-of-art methods. Extensive experiments validate our hypothesis of the importance of accounting for label correlation and also show the importance of using labeled data and unlabeled data. In our future work, we would like to incorporate higher order correlation directly.

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Patt. Recogn. **37**(9), 1757–1771 (2004)
2. Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., Zhang, H.-J.: Correlative multi-label video annotation. In: Proceedings of the 15th ACM International Conference on Multimedia, MM 2007, pp. 17–26. ACM, New York (2007)

3. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: De Raedt, L., Siebes, A. (eds.) PKDD 2001. LNCS, vol. 2168, pp. 42–53. Springer, Heidelberg (2001). doi:10.1007/3-540-44794-6_4
4. Elisseeff, A., Weston, J.: A kernel method for multi-labeled classification. In: Advances in Neural Information Processing Systems, pp. 681–687 (2001)
5. Zha, Z.-J., Mei, T., Wang, J., Wang, Z., Hua, X.-S.: Graph-based semi-supervised learning with multiple labels. J. Vis. Commun. Image Representation **20**(2), 97–103 (2009)
6. Kong, X., Ng, M.K., Zhou, Z.-H.: Transductive multilabel learning via label set propagation. IEEE Trans. Knowl. Data Eng. **25**(3), 704–719 (2013)
7. Chen, G., Song, Y., Wang, F., Zhang, C.: Semi-supervised multi-label learning by solving a Sylvester equation. In: SDM, SIAM, pp. 410–419 (2008)
8. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21(1), p. 42. AAAI Press, MIT Press, MenloPark, Cambridge, London (1999/2006)
9. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: ICML, vol. 3, pp. 912–919 (2003)
10. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Scholkopf, B.: Learning with local and global consistency. Adv. Neural Inf. Process. Syst. **16**, 321–328 (2004)
11. Hu, Q., Cheng, D.: The polynomial solution to the Sylvester matrix equation. Appl. Math. Lett. **19**(9), 859–864 (2006)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford Info Lab, Technical Report 1999-66, November 1999, Previous number = SIDL-WP-1999-0120
13. Zhang, M.-L., Zhou, Z.-H.: A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. **26**(8), 1819–1837 (2014)