DW-ML-kNN: A Dual Weighted Multi-label kNN Algorithm

Duoqian Miao, Zhifei Zhang*, Zhihua Wei, Chuanyan Wang

Department of Computer Science and Technology, Tongji University, No. 4800, Cao'an Highway Shanghai, 201804, China
{dqmiao, zhihua_wei}@tongji.edu.cn, zhifei.zzh@gmail.com, chunyanfriend@163.com

Multi-label learning is a continually developing field in machine learning and has wide applications in many domains. ML-kNN is a simple but effective method to learn from multi-label data. But when data are imbalanced, its performance is not encouraging. Therefore, we propose a dual weighted ML-kNN (for short, DW-ML-kNN) algorithm, which utilizes distances of neighbors as weights and considers the impact of neighbors without one certain label. The experiment results on three multi-label datasets show that the new algorithm achieves better performance than ML-kNN.

Keywords: Multi-label learning; k nearest neighbors; distance weight.

1. Introduction

Multi-label data are popular in our daily life. For example, an image may be classified into semantic classes *sunset* and *beach* simultaneously; a news report is about *politics*, as well as *military*. Recently, the issue of learning from these data, which is called multi-label learning, has drawn significant attention of researchers. Multi-label learning is widely used in many domains, such as text classification[1,2,3], semantic annotation of videos[4], image annotation[5,6] and functional genomics[7,8]. There are two major tasks in multi-label learning: multi-label classification and label ranking[9]. The methods of solving this problem are grouped into two categories: problem transformation and algorithm adaptation[10].

Problem transformation is to transform the multi-label learning problem into one or more single-label learning problem, for which traditional learning algorithms are used. Ranking by pairwise comparison proposed by Hüllermeier et al.[11] constructs multiple binary classifiers, and determines the label ranking of a new instance by voting. Calibrated label ranking[12] extends it by introducing an additional virtual label, which assigns the labels ranking before the virtual label to unseen example. Zhang et al.[13] gives an algorithm called INSDIF, in which each prototype for each label is computed with the all instances belonging to this label as the training set. This group of methods poses an important complexity problem, especially for large value of the number of labels. PPT[14] and RAkE[15] methods solve the problem to some extent.

* Corresponding Author

The comparison results on Emotions dataset are described in Fig. 4. The meaning of the coordinates is the same with that in Fig. 2. The performance of DW-ML-kNN on Precision and Recall is better than ML-kNN, but not obviously on other criteria, especially worse on Coverage. The reason is similar to that of Scene dataset.

The overall result is obtained from the above three results. Table 3 reports the average performance of the two algorithms across 11 values of the parameter K for each dataset. The best result on each criterion is shown with bold typeface. The last line means the number of criteria on which each algorithm achieves the best result.

Table 3. Overall Comparison Results of Two Algorithms

Criterion	Yeast		Scene		Emotions	
	ML-kNN	DW-ML-kNN	ML-kNN	DW-ML-kNN	ML-kNN	DW-ML-kNN
Hamming loss	**0.1973**	0.2028	**0.0978**	0.1013	**0.2900**	0.2981
Precision	**0.7273**	0.6880	0.6583	**0.7106**	0.4928	**0.5468**
Recall	0.5653	**0.6275**	0.6471	**0.7625**	0.3510	**0.4781**
One-error	0.2405	**0.2398**	0.2481	**0.2462**	**0.3896**	0.4012
Coverage	6.4015	**6.3956**	**0.5804**	0.5828	2.4980	**2.4888**
Average precision	0.7576	**0.7595**	0.8472	**0.8490**	0.6898	**0.6994**
Win(s)	2(6)	4(6)	2(6)	4(6)	2(6)	4(6)

The results show that DW-ML-kNN is better than ML-kNN on more than half of the six criteria on three benchmark datasets. Average precision and Recall are always improved, while Hamming loss is still worse. The performance on other three criteria is unstable, but close. In a word, the dual weighted ML-kNN (DW-ML-kNN) algorithm achieves promising and better performance.

5. Conclusion

To solve the problem of imbalanced data, we propose a new algorithm-dual weighted multi-label k nearest neighbors, named DW-ML-kNN. The experiment results show that our algorithm is better than ML-kNN. We will carry out research on the criterion of Hamming loss, and try to improve its performance further. What's more, the tuning factor in our algorithm is also a further research topic.

Acknowledgments

The work is partially supported by the National Natural Science Foundation of China (No. 60970061, No. 61075056, and No. 61103067), the Opening Project of Shanghai Key Laboratory of Digital Media Processing and Transmission (No.2011KF03), and the Fundamental Research Funds for the Central Universities.