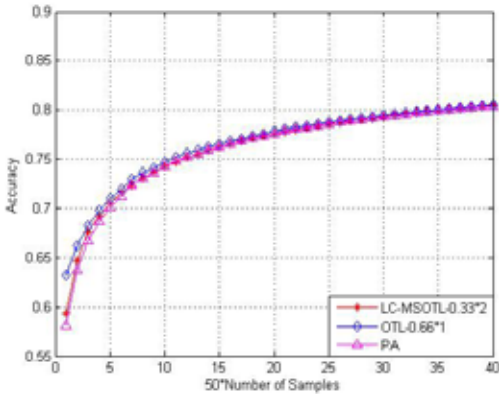
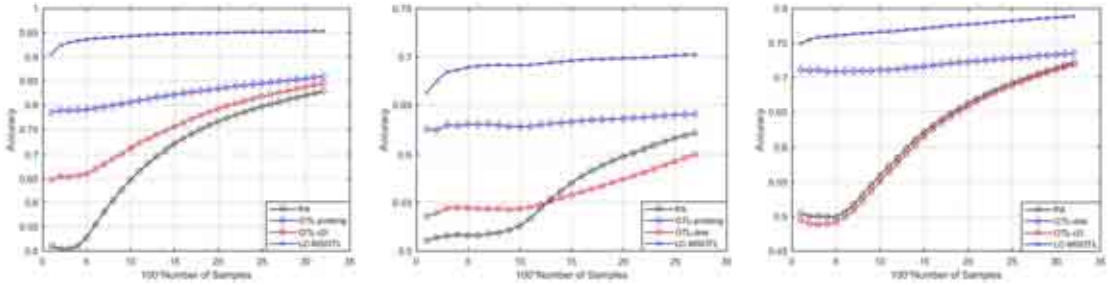


Fig.10 Classification accuracy change for Waveform data set with LC-MSOTL-0.33*2 and OTL-0.66

图 10 LC-MSOTL-0.33*2 与 OTL-0.66 对 Waveform 数据集的分类准确率变化

3.2.3 LC-MSOTL 对 Intrusion detection 数据集的实验

在本节实验中,PA 的参数为σ=16,惩罚系数 C=2.从图 11 可以看出,在不清楚源领域和目标领域相似程度及相似的局部区域的情形下,LC-MSOTL 的分类准确率仍然比 PA 和 OTL 要好.更进一步地,从图 11(a)还可以看到,当两个源对目标领域都有较好的迁移学习效果时,LC-MSOTL 在分类初期相对于 PA 较大幅度地提高了分类准确率;从图 11(b)、图 11(c)还可以看到,当两个源领域中只有 1 个对目标领域显示出较好的迁移学习效果时,LC-MSOTL 仍然在分类初期相对于 PA 有效地提高了分类准确率.



(a) 以 DOS 作为目标领域 (b) 以 R2L 作为目标领域 (c) 以 Probing 作为目标领域

Fig.11 Classification accuracy change of each method for different target domains

图 11 各算法在不同目标领域下的分类准确率变化

3.2.4 LC-MSOTL 参数敏感性实验

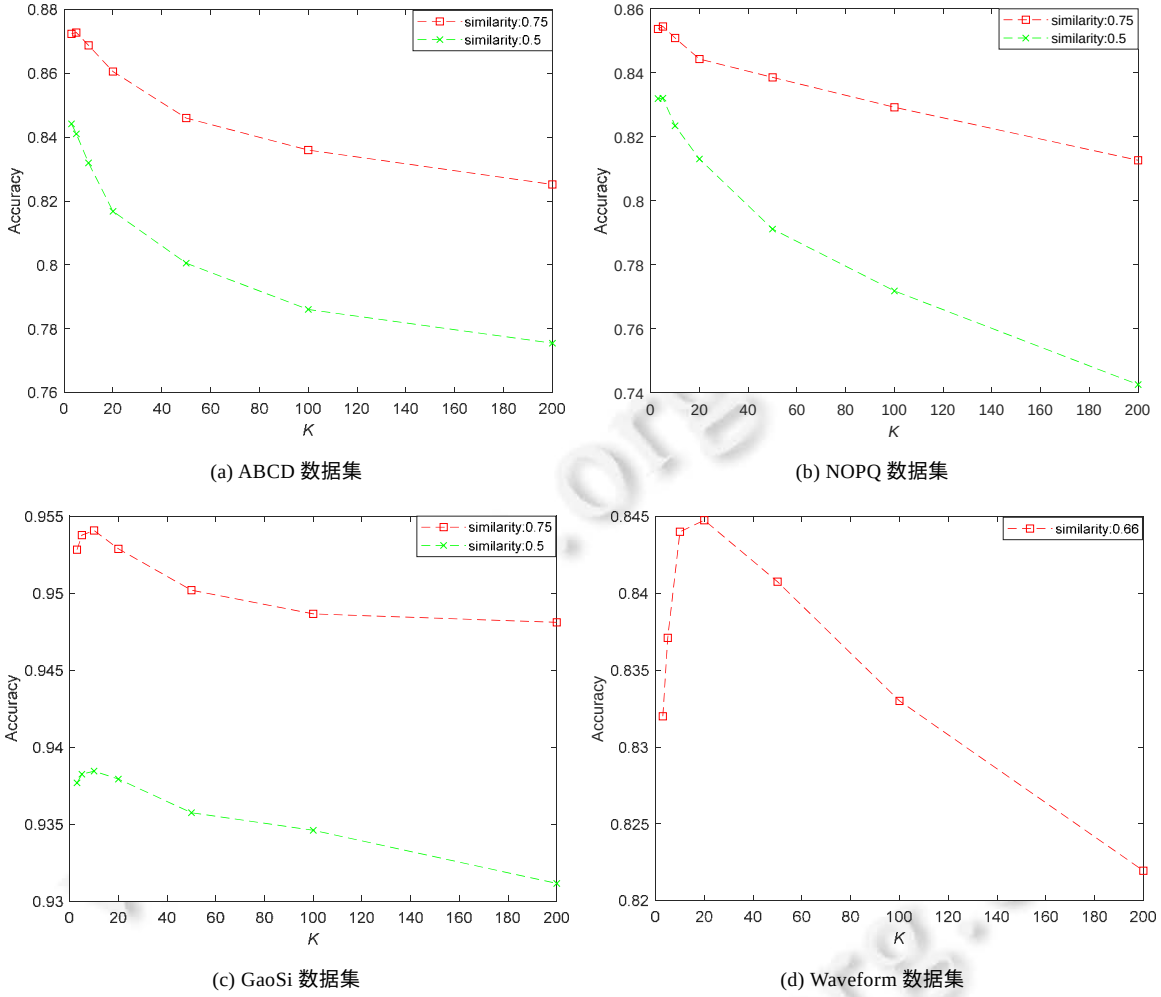
在不同的参数设置下,采用第 3.2.1 节中的方式进行实验,以分析各参数对 LC-MSOTL 分类准确率的影响.

(1) 近邻个数 K 对 LC-MSOTL 算法的影响

为了验证近邻个数 K 对 LC-MSOTL 算法的影响,在这部分实验中,将固定放大系数ζ设置为 1,以测试 K 对实验效果的影响.在该实验中,K 的取值依次是 3,5,10,20,50,100,200.

从图 12(a)可以看出,在 ABCD 数据集上,不管相似度是 0.75 还是 0.5,随着 K 值的不断增大,分类准确率越来越低.这说明在 ABCD 数据集上,K=3 是比较好的近邻个数取值;且随着 K 值的增大,分类器选择的越来越不准确.从图 12(b)可以看出,NOPQ 数据集在相似度 0.75 时,随着 K 的增大,分类准确率先轻微上升,之后逐步下降;在相似度为 0.5 时,NOPQ 数据集随着 K 值的不断增大,准确率下降很明显.从图 12(c)可以看出,在 GaoSi 数据集上,当 K=10 时准确率达到最大,说明此时分类器选择较为准确;随着 K 值的增大,准确率越来越低.从图 12(d)可以看出,在 Waveform 数据集上,当 K=20 时,此时的分类准确率达到最大;且当 K<20 时,随着 K 值的增大,准确率

提高的速度较快;当 K>20 时,随着 K 值的增大,分类准确率越来越低.这些都说明,在 LC-MSOTL 算法中,K 的值表现与 K 近邻分类算法中的 K 的表现基本一致.因此,K 的取值与数据集相关.



(a) ABCD 数据集 (b) NOPQ 数据集

(c) GaoSi 数据集 (d) Waveform 数据集

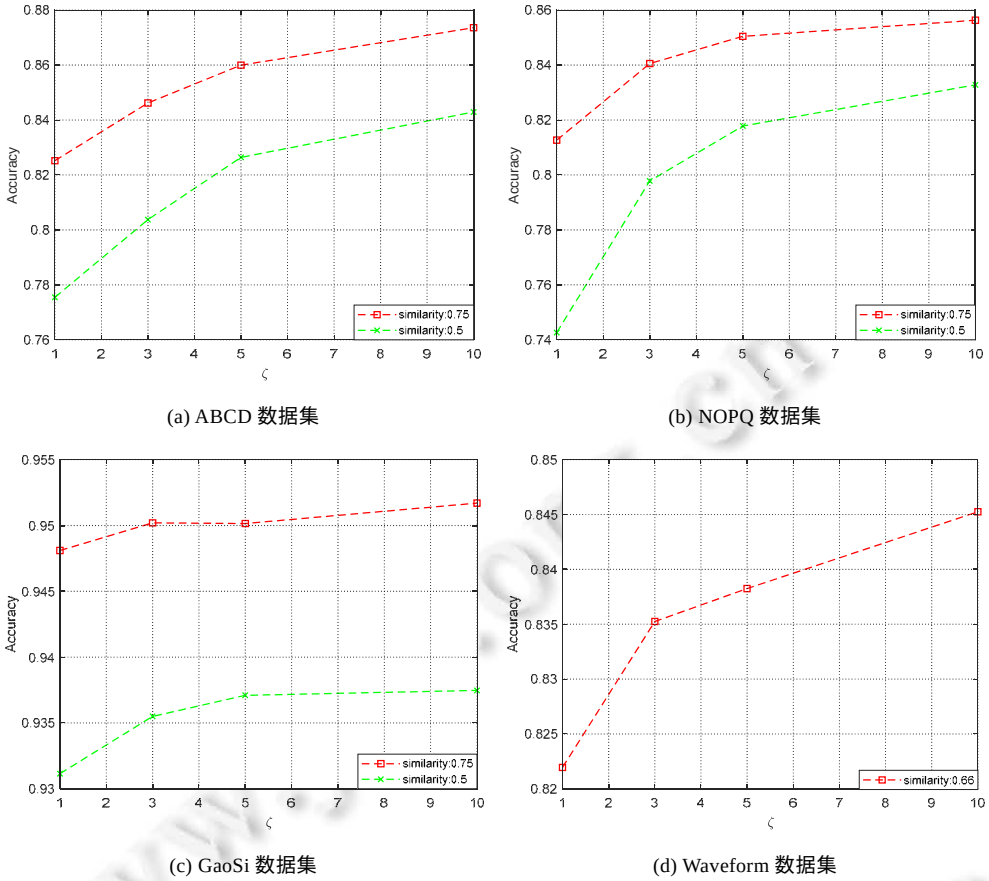
Fig.12 Classification accuracy change for different data sets with different K

图 12 各数据集在不同 K 上的准确率变化图

(2) 放大系数ζ对 LC-MSOTL 分类准确率的影响

为了分析放大系数ζ对 LC-MSOTL 分类准确率的影响,在这部分实验中,固定近邻个数 K,以测试放大系数ζ对分类准确率的影响.由于近邻个数 K 较小时ζ对分类准确率的影响较小,故在该实验中,为了显示出放大系数ζ对分类准确率的影响,将近邻个数 K 的取值设置为 200,在该实验中,ζ的取值依次是 1,3,5,10.

从图 13 中可以看出,当近邻个数 K 取值较大且放大系数ζ较小时,LC-MSOTL 分类准确率相对较低;在各个数据集上,随着放大系数ζ的不断增大,LC-MSOTL 的分类准确率逐渐提高,充分表明了放大系数ζ可以缓解当近邻个数 K 较大时,LC-MSOTL 分类准确率相对较低的问题.如表 14~表 16 所示,尽管 K 较大,但通过增大ζ,可以得到与 K 较小时相当的分类准确率.换句话说,ζ具有放松对 K 的选择的功能.其中原因在于,从 LC 的定义可知,若 K 比较大,则 $\boldsymbol{x}_t$ 的近邻中可能存在离它较远的样本,但这些离 $\boldsymbol{x}_t$ 较远的样本则由于放大系数ζ的作用而减弱对 LC 值的不良影响.需指出的是,表 14~表 16 中之所以没有提供 GaoSi 数据集上类似的实验结果,是因为在 GaoSi 数据集上调节近邻个数 K 以及放大系数ζ时,分类准确率波动很小.



(a) ABCD 数据集 (b) NOPQ 数据集

(c) GaoSi 数据集 (d) Waveform 数据集

Fig.13 Classification accuracy change for different data sets with different amplification factor

图 13 各数据集在不同放大系数上的准确率变化图

Table 14 Comparison of ABCD data sets

表 14 ABCD 数据集对比

相似度 0.75			相似度 0.5		
K	ζ	分类准确率	K	ζ	分类准确率
5	1	0.872 794 118	5	1	0.841 042 781
200	10	0.873 596 257	200	10	0.842 914 439

Table 15 Comparison of NOPQ data sets

表 15 NOPQ 数据集对比

相似度 0.75			相似度 0.5		
K	ζ	分类准确率	K	ζ	分类准确率
5	1	0.854 467 806	5	1	0.832 063 075
200	10	0.856 307 49	200	10	0.832 785 808

Table 16 Comparison of Waveform data sets

表 16 Waveform 数据集对比

相似度 0.66		
K	ζ	分类准确率
20	1	0.844 75
200	10	0.845 25

4 总 结

本文结合多源迁移学习和在线学习,提出了一种适用于数据流环境的多源在线迁移学习算法——LC-MSOTL.本文首先提出了一种局部分类精度的计算方法,并从理论上初步分析了引入局部分类精度选取源领域分类器方法的合理性.LC-MSOTL 利用局部分类精度动态地从源领域分类器集合选取局部分类精度最高的分类器,将该分类器和目标领域分类器加权集成.进一步的实验结果表明,LC-MSOTL 可以从多个源领域中选取合适的源领域分类器,能够有效地利用多个源领域提高对目标领域样本的分类准确率.因此,当单个源领域与目标领域的相似性不高时,通过增加源领域数量,同时确保源领域之间的区别度,可以有效地提高迁移学习效果.未来的工作包括:研究其他选取源领域的方法;研究确定源领域与目标领域局部相似性的方法;研究如何将 LC-MSOTL 用于概念漂移数据流分类等.

References:

[1] Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46(4):1−37. [doi: 10.1145/2523813]

[2] Wen YM, Qiang BH, Fan ZG. A survey of the classification of data streams with concept drift. CAAI Trans. on Intelligent Systems, 2013,8(2):95−104 (in Chinese with English abstract).

[3] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(10):1345−1359. [doi: 10.1109/TKDE.2009.191]

[4] Zhuang FZ, Luo P, He Q, Shi ZZ. Survey on transfer learning research. Ruan Jian Xue Bao/Journal of Software, 2015,26(1):26−39 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4631.htm [doi: 10.13328/j.cnki.jos.004631]

[5] Pan W, Zhong H, Xu CF, Ming Z. Adaptive Bayesian personalized ranking for heterogeneous implicit feedbacks. Journal of Knowledge-Based Systems, 2015,73(1):173−180. [doi: 10.1016/j.knosys.2014.09.013]

[6] Pan W, Yang Q. Transfer learning in heterogeneous collaborative filtering domains. Journal of Artificial Intelligence, 2013,197(4): 39−55. [doi: 10.1016/j.artint.2013.01.003]

[7] Zhuang FZ, Luo P, Yin PF, He Q, Shi ZZ. Concept learning for cross-domain text classification: A general probabilistic framework. In: Proc. of the 23th Int'l Joint Conf. on Artificial Intelligence. Palo Alto: AAAI, 2013. 1960−1966.

[8] Zhuang FZ, Luo P, Du CY, He Q, Shi ZZ. Triplex transfer learning: Exploiting both shared and distinct concepts for text classification. IEEE Trans. on Cybernetics, 2014,44(7):1191−1203. [doi: 10.1109/TCYB.2013.2281451]

[9] Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proc. of the 24th Int'l Conf. on Machine Learning. New York: ACM Press, 2007. 193−200. [doi: 10.1145/1273496.1273521]

[10] Pan SJL, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans. on Neural Networks, 2011, 22(2):199−210. [doi: 10.1109/TNN.2010.2091281]

[11] Zhang Q, Li M, Wang XS, Cheng YF, Zhu MQ. Instance-Based transfer learning for multi-source domains. Acta Automatica Sinica, 2014,40(6):1176−1183 (in Chinese with English abstract).

[12] Yao Y, Doretto G. Boosting for transfer learning with multiple sources. In: Proc. of the Computer Vision and Pattern Recognition. New York: IEEE, 2010. 1855−1862. [doi: 10.1109/CVPR.2010.5539857]

[13] Eaton E, Desjardins M. Selective transfer between learning tasks using task-based boosting. In: Proc. of the AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2011. 337−342.

[14] Huang PP, Wang G, Qin SY. Boosting for transfer learning from multiple data sources. Pattern Recognition Letters, 2012,33(5): 568−579. [doi: 10.1016/j.patrec.2011.11.023]

[15] Duan LX, Tsang IW, Xu D, Chua TS. Domain adaptation from multiple sources via auxiliary classifiers. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. New York: ACM Press, 2009. 289−296. [doi: 10.1145/1553374.1553411]

[16] Chattopadhyay R, Ye JP, Panchanathan S, Fan W, Davidson I. Multi-Source domain adaptation and its application to early detection of fatigue. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2011. 717−725. [doi: 10.1145/2020408.2020520]

[17] Zhao PL, Hoi SCH, Wang JL, Li B. Online transfer learning. Artificial Intelligence, 2014,216(16):76−102. [doi: 10.1016/j.artint.2014.06.003]

[18] Zhao PL, Hoi SCH. OTL: A framework of online transfer learning. In: Proc. of the Int'l Conf. on Machine Learning. New York: ACM Press, 2010. 1231−1238.

[19] Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y. Online passive-aggressive algorithms. Journal of Machine Learning Research, 2006,7(3):551−585.

[20] Li ZJ, Li YC, Wang F, He GL, Kuang L. Online learning algorithms for big data analytics: A survey. Journal of Computer Research and Development, 2015,52(8):1707−1721 (in Chinese with English abstract).

[21] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958,65(6):386−408. [doi: 10.1037/h0042519]

[22] Cesa-Bianchi N, Conconi A, Gentile C. A second-order perceptron algorithm. SIAM Journal on Computing, 2002,34(3):640−668. [doi: 10.1137/S0097539703432542]

[23] Dredze M, Grammer K, Pereira F. Confidence-Weighted linear classification. In: Proc. of the 25th Int'l Conf. on Machine Learning. New York: ACM Press, 2008. 264−271. [doi: 10.1145/1390156.1390190]

[24] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient. Journal of Machine Learning Research, 2008,10(2): 777−801.

[25] Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Journal of Machine Learning, 1999,37(3): 277−296. [doi: 10.1023/A:1007662407062]

[26] Fan HJ, Song Q, Shrestha SB. Online learning with kernel regularized least mean square algorithms. Knowledge-Based Systems, 2014,59(2):21−32. [doi: 10.1016/j.knosys.2014.02.005]

[27] Yang HQ, Lyu MR, King I. Efficient online learning for multitask feature selection. ACM Trans. on Knowledge Discovery from Data, 2013,7(2):1−27. [doi: 10.1145/2499907.2499909]

[28] Domingos P, Hulten G. Mining high-speed data streams. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2000. 71−80. [doi: 10.1145/347090.347107]

[29] Gama J, Rocha R, Medas P. Accurate decision trees for mining high-speed data streams.In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2003. 523−528. [doi: 10.1145/956750.956813]

[30] Guo GD, Huang J, Chen LF. KNN model based incremental learning algorithm. Pattern Recognition and Artificial Intelligence, 2010, 23(5):701−707 (in Chinese with English abstract).

[31] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. Journal of Big Data, 2016,3:9. https://doi.org/10.1186/s40537-016-0043-6

[32] Ling X, Dai WY, Xue GR, Yang Q, Yu Y. Spectral domain-transfer learning. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. 488−496. [doi: 10.1145/1401890.1401951]

[33] Jiang W, Zavesky E, Chang SF, Loui A. Cross-Domain learning methods for high-level visual concept classification. In: Proc. of the IEEE Int'l Conf. on Image Processing. IEEE, 2008. 161−164. [doi: 10.1109/ICIP.2008.4711716]

[34] Dai WY, Xue GR, Yang Q, Yu Y. Co-Clustering based classification for out-of-domain documents. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 210−219. [doi: 10.1145/1281192.1281218]

[35] Dai WY, Xue GR, Yang Q, Yu Y. Transferring naive Bayes classifiers for text classification. In: Proc. of the National Conf. on Artificial Intelligence. Palo Alto: AAAI, 2007. 540−545.

[36] Hong JM, Yin J, Huang Y, Liu YB, Wang JH. TrSVM: A transfer learning algorithm using domain similarity. Journal of Computer Research and Development, 2011,48(10):1823−1830 (in Chinese with English abstract).

[37] Davis J, Domingos P. Deep transfer via second-order Markov logic. In: Proc. of the Int'l Conf. on Machine Learning. New York: ACM Press, 2009. 217−224. [doi: 10.1145/1553374.1553402]

[38] Sun SL, Shi HL, Wu YB. A survey of multi-source domain adaptation. Information Fusion, 2015,24(C):84−92. [doi: 10.1016/j.inffus.2014.12.003]

[39] Eaton E, desJardins M, Lane T. Modeling transfer relationships between learning tasks for improved inductive transfer. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2008. 317−332. [doi: 10.1007/978-3-540-87479-9_39]

[40] Luo P, Zhuang FZ, Xiong H, Xiong YH, He Q. Transfer learning from multiple source domains via consensus regularization. In: Proc. of the ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 103−112. [doi: 10.1145/1458082.1458099]

[41] Wang XS, Pan J, Chen YH, Cao G. Self-Adaptive transfer for decision trees based on similarity metric. Acta Automatica Sinica, 2013,39(12):2186−2192 (in Chinese with English abstract).

[42] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification. In: Proc. of the IEEE Int'l Conf. on Data Mining. New York: IEEE, 2009. 159−168. [doi: 10.1109/ICDM.2009.32]

[43] Tommasi T, Orabona F, Caputo B. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. New York: IEEE, 2010. 3081−3088. [doi: 10.1109/CVPR.2010.5540064]

[44] Cawley GC. Leave-One-Out cross-validation based model selection criteria for weighted LS-SVMs. In: Proc. of the Int'l Joint Conf. on Neural Networks. New York: IEEE, 2006. 1661−1668. [doi: 10.1109/IJCNN.2006.246634]

[45] Pan SSJ, Ni XC, Sun JT, Yang Q, Chen Z. Cross-Domain sentiment classification via spectral feature alignment. In: Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 751−760. [doi: 10.1145/1772690.1772767]

[46] Gao J, Fan W, Jiang J, Han JW. Knowledge transfer via multiple model local structure mapping. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. 283−291. [doi: 10.1145/1401890.1401928]

[47] Ge L, Gao J, Zhang AD. OMS-TL: A framework of online multiple source transfer learning. In: Proc. of the ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2013. 2423−2428. [doi: 10.1145/2505515.2505603]

[48] Zhou ZH, Li M. Semi-Supervised regression with co-training style algorithms. IEEE Trans. on Knowledge and Data Engineering, 2007,19(11):1479−1493. [doi: 10.1109/TKDE.2007.190644]

[49] Zhou ZH, Li M. Semi-Supervised regression with co-training. In: Proc. of the 19th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2005. 908−916.

[50] Woods K, Bowyer K, Jr WPK. Combination of multiple classifiers using local accuracy estimates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(4):405−410. [doi: 10.1109/34.588027]

[51] Liu M, Yuan ZB, Miao ZJ, Tang XF, Li KL. Transformation from local accuracy to classification confidence. Journal of Computer Research and Development, 2008,45(9):1612−1619 (in Chinese with English abstract).

附中文参考文献:

[2] 文益民,强保华,范志刚.概念漂移数据流分类研究综述.智能系统学报,2013,8(2):95−104.

[4] 庄福振,罗平,何清,史忠植.迁移学习研究进展.软件学报,2015,26(1):26−39. http://www.jos.org.cn/1000-9825/4631.htm [doi: 10.13328/j.cnki.jos.004631]

[11] 张倩,李明,王雪松,程玉虎,朱美强.一种面向多源领域的实例迁移学习.自动化学报,2014,40(6):1176−1183.

[20] 李志杰,李元香,王峰,何国良,匡立.面向大数据分析的在线学习算法综述.计算机研究与发展,2015,52(8):1707−1721.

[30] 郭躬德,黄杰,陈黎飞.基于 KNN 模型的增量学习算法.模式识别与人工智能,2010,23(5):701−707.

[36] 洪佳明,印鉴,黄云,刘玉葆,王甲海.TrSVM:一种基于领域相似性的迁移学习算法.计算机研究与发展,2011,48(10):1823−1830.

[41] 王雪松,潘杰,程玉虎,曹戈.基于相似度衡量的决策树自适应迁移.自动化学报,2013,39(12):2186−2192.

[51] 刘明,袁保宗,苗振江,唐晓芳,李昆仑.从局部分类精度到分类置信度的变换.计算机研究与发展,2008,45(9):1612−1619.

唐诗淇(1990－),男,湖南长沙人,硕士,主要研究领域为机器学习,迁移学习,数据流分类.

秦一休(1992－),男,学士,主要研究领域为机器学习,数据挖掘.

文益民(1969－),男,博士,教授,CCF 高级会员,主要研究领域为机器学习,迁移学习,数据流分类,教育数据挖掘.