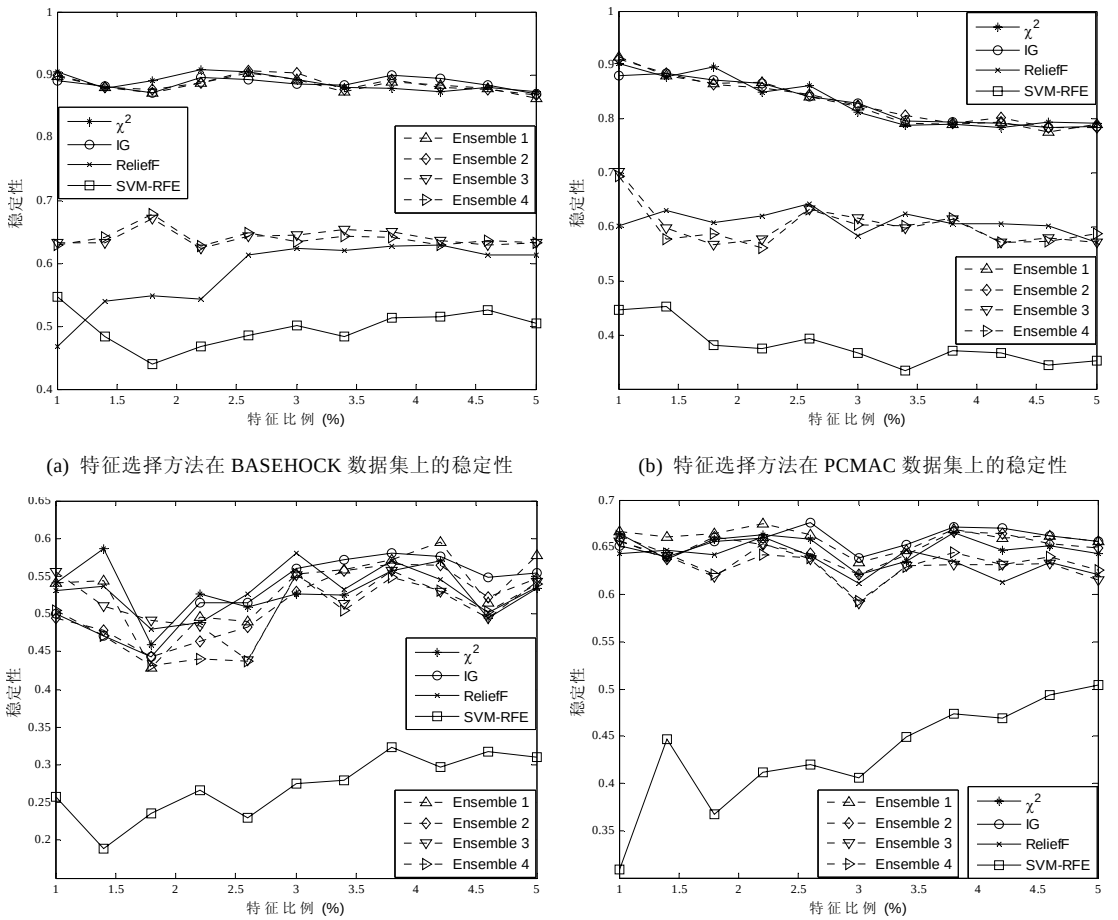

(a) 特征选择方法在 BASEHOCK 数据集上的稳定性

(b) 特征选择方法在 PCMAC 数据集上的稳定性

(c) 特征选择方法在 COLON 数据集上的稳定性

(d) 特征选择方法在 ALLAML 数据集上的稳定性

Fig.6 Stability comparisons among ensemble feature selection methods

图 6 集成特征选择方法稳定性比较

综上所述,对于稳定性较强的特征选择方法(如χ^2 和 IG),采用集成方法对其稳定性的提升效果并不显著.而对于稳定性较弱的特征选择方法(如 ReliefF 和 SVMFS),采用集成方法能够在一定程度上提高特征选择稳定性.因此,采用集成方法能够在一定程度上综合保证特征选择结果的稳定性,即,能够在不同的数据集上确保特征选择稳定性的提升.进一步观察可以看出,单变量方法χ^2和 IG 的稳定性要好于多变量方法 ReliefF 和 SVMFS.这也是显然的,单变量方法采用特定内在的度量方式独立评估每个特征,而多变量方法在评估特征的同时也会考虑该特征与其他特征的关联.因此,由于高维数据特征间复杂关系的存在,导致多变量特征选择方法的稳定性也在一定程度上受到了影响.

表 4~表 7 给出了在 4 个数据集上,4 种基本特征选择方法和 4 种集成特征选择方法在 3 种不同分类器上的分类正确率,特征比例仍然设置为 1%~5%.

通过表 4~表 7 可以看出:除了 COLON 数据集,使用 SVM 分类器时,4 种基本特征选择方法和 4 种集成特征选择方法在多数情况下能够取得较好的分类正确率.因此从分类器的角度,SVM 分类性能要好于 *K*NN 和 NB 分类器.其次,从集成方法的角度看,与基本的单变量和多变量特征选择方法相比,4 种集成特征选择方法在多数情况下都能够获得较好的分类性能,特别是在 BASEHOCK 数据集上,4 种集成方法提供了全面较好的结果.而针对具体的集成方法而言,4 种集成方式并无明显的差异.这说明在单变量和多变量方法同时存在的情况下,集成选

择能够有效提高分类性能,而集成的算法对分类性能并无明显的影响.最后,从分类性能的角度,单变量方法与多变量方法相比,在分类性能上并无明显的差异.这说明多变量方法能够获得在分类性能上较好的特征子集,但是由于多变量方法稳定性较弱,因此其生成特征子集与单变量方法相比变化程度较强,即可信度较低.

Table 4 Classification accuracy of feature selection methods in BASEHOCK

表 4 特征选择方法在 BASEHOCK 数据集上的分类正确率

方法	分类器	特征比例				
		1%	2%	3%	4%	5%
Ensemble1	SVM	**0.945 8**	0.952 3	0.959 4	0.960 9	0.965 4
	*K*NN	0.871 1	0.858 5	0.867 0	0.872 1	0.863 5
	NB	0.907 2	0.916 7	0.928 2	0.926 7	0.926 7
Ensemble2	SVM	**0.945 8**	**0.953 3**	0.960 4	0.965 4	**0.966 4**
	*K*NN	0.872 1	0.862	0.866 5	0.869 0	0.867 6
	NB	0.906 7	0.915 7	0.927 2	0.928 7	0.926 2
Ensemble3	SVM	0.940 3	0.950 3	**0.966 9**	0.970 4	0.961 9
	*K*NN	0.878 1	0.870 0	0.885 1	0.883 6	0.877 6
	NB	0.898 1	0.905 7	0.923 7	0.925 2	0.926 7
Ensemble4	SVM	0.945 3	0.952 3	0.964 4	**0.972 4**	0.965 4
	*K*NN	0.888 6	0.876 6	0.887 6	0.888 1	0.884 1
	NB	0.898 1	0.904 7	0.921 2	0.926 7	0.924 7
χ^2	SVM	0.942 3	0.952 3	0.960 9	0.965 4	0.964 9
	*K*NN	0.867 0	0.858 5	0.865 0	0.865 0	0.862 5
	NB	0.907 2	0.918 7	0.927 2	0.930 8	0.927 8
IG	SVM	0.944 8	**0.953 3**	0.957 9	0.962 9	0.964 4
	*K*NN	0.876 1	0.855 0	0.864 0	0.870 6	0.871 1
	NB	0.906 2	0.914 7	0.925 7	0.924 7	0.925 2
ReliefF	SVM	0.709 5	0.750 1	0.828 4	0.841 9	0.892 1
	*K*NN	0.619 7	0.676 3	0.739 1	0.779 2	0.774 7
	NB	0.596 6	0.645 7	0.719 0	0.757 1	0.802 8
SVM-RFE	SVM	**0.945 8**	0.952 8	0.954 8	0.962 9	0.955 9
	*K*NN	0.911 7	0.904 2	0.881 1	0.891 1	0.895 1
	NB	0.892 6	0.897 6	0.918 2	0.915 7	0.919 2

Table 5 Classification accuracy of feature selection methods in PCMAC

表 5 特征选择方法在 PCMAC 数据集上的分类正确率

方法	分类器	特征比例				
		1%	2%	3%	4%	5%
Ensemble1	SVM	0.876 0	0.883 7	0.895 5	0.907 4	0.904 3
	*K*NN	0.823 0	0.804 4	0.793 1	0.803 4	0.796 2
	NB	0.734 4	0.749 9	0.767 9	0.773 6	0.777 7
Ensemble2	SVM	0.876 0	0.883 7	**0.899 1**	**0.909 4**	0.908 4
	*K*NN	0.823 0	0.802 4	0.794 6	0.798 7	0.793 6
	NB	0.734 4	0.749 9	0.767 9	0.772 5	0.776 6
Ensemble3	SVM	0.877 5	**0.894 5**	0.896 0	0.899 1	0.896
	*K*NN	0.843 0	0.825 0	0.797 7	0.795 1	0.805 5
	NB	0.724 6	0.763 2	0.755 6	0.750 9	0.777 7
Ensemble4	SVM	0.881 6	0.892 9	0.897 1	0.896 6	0.900 1
	*K*NN	0.848 2	0.833 2	0.804 9	0.807 5	0.809 6
	NB	0.726 2	0.763 2	0.749 4	0.752 4	0.772 5
χ^2	SVM	0.872 9	0.883 2	0.896 0	0.902 2	**0.910 4**
	*K*NN	0.814 7	0.800 8	0.789 0	0.797 2	0.797 2
	NB	0.728 8	0.751 9	0.768 9	0.770 5	0.776 1
IG	SVM	0.876 0	0.880 1	0.894 0	0.905 3	0.908 4
	*K*NN	0.823 0	0.808 0	0.791 5	0.796 2	0.804 9
	NB	0.729 8	0.748 8	0.764 8	0.772 0	0.779 2
ReliefF	SVM	0.598 0	0.658 8	0.807 0	0.820 4	0.833 3
	*K*NN	0.560 0	0.620 7	0.735 4	0.732 9	0.740 6
	NB	0.529 6	0.579 0	0.662 3	0.656 7	0.696 9
SVM-RFE	SVM	**0.887 3**	0.890 9	0.880 1	0.884 7	0.894 5
	*K*NN	0.879 1	0.863 1	0.829 1	0.829 1	0.823 5
	NB	0.739 6	0.739 6	0.750 4	0.762 7	0.754 5

Table 6 Classification accuracy of feature selection methods in COLON

表 6 特征选择方法在 COLON 数据集上的分类正确率

方法	分类器	特征比例				
		1%	2%	3%	4%	5%
Ensemble1	SVM	0.694 9	0.724 4	0.757 7	0.757 7	0.757 7
	*K*NN	0.807 7	0.757 7	0.773 1	0.838 5	0.821 8
	NB	0.807 7	0.742 3	0.806 4	0.823 1	0.803 8
Ensemble2	SVM	0.693 6	0.707 7	0.803 8	0.774 4	0.774 4
	*K*NN	0.825 6	0.757 7	0.789 7	**0.855 1**	**0.821 8**
	NB	0.793 6	0.742 3	0.791 0	0.838 5	0.803 8
Ensemble3	SVM	0.693 6	0.741	0.803 8	0.773 1	0.773 1
	*K*NN	**0.839 7**	**0.803 8**	0.823 1	0.838 5	0.806 4
	NB	0.792 3	0.756 4	0.806 4	0.806 4	0.803 8
Ensemble4	SVM	0.660 3	0.756 4	0.739 7	0.759	0.773 1
	*K*NN	0.839 7	**0.803 8**	0.821 8	0.823 1	0.791 0
	NB	0.775 6	0.756 4	0.774 4	0.806 4	0.803 8
χ^2	SVM	0.743 6	0.738 5	0.725 6	0.821 8	0.741 0
	*K*NN	0.807 7	0.773 1	0.821 8	0.838 5	**0.821 8**
	NB	0.807 7	0.742 3	0.789 7	0.823 1	0.803 8
IG	SVM	0.628 2	0.741	0.788 5	0.788 5	0.757 7
	*K*NN	0.824 4	**0.803 8**	0.789 7	0.823 1	0.805 1
	NB	0.793 6	0.725 6	0.791 0	0.823 1	0.803 8
ReliefF	SVM	0.694 9	0.706 4	0.787 2	0.726 9	0.821 8
	*K*NN	0.824 4	0.756 4	**0.838 5**	0.823 1	0.807 7
	NB	0.792 3	0.725 6	0.821 8	0.838 5	0.803 8
SVM-RFE	SVM	0.793 6	0.752 6	0.788 5	0.791 0	0.820 5
	*K*NN	0.807 7	0.721 8	0.805 1	0.838 5	0.774 4
	NB	0.776 9	0.709	0.787 2	0.838 5	0.771 8

Table 7 Classification accuracy of feature selection methods in ALLAML

表 7 特征选择方法在 ALLAML 数据集上的分类正确率

方法	分类器	特征比例				
		1%	2%	3%	4%	5%
Ensemble1	SVM	0.945 7	**0.973 3**	0.958 1	0.944 8	0.972 4
	*K*NN	0.945 7	0.960 0	0.957 1	**0.973 3**	0.942 9
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
Ensemble2	SVM	0.945 7	**0.973 3**	0.958 1	0.944 8	**0.985 7**
	*K*NN	0.960 0	0.960 0	0.957 1	**0.973 3**	0.929 5
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
Ensemble3	SVM	0.959 0	**0.973 3**	0.943 8	0.959 0	**0.985 7**
	*K*NN	0.916 2	0.945 7	0.900 0	0.918 1	0.915 2
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
Ensemble4	SVM	0.959 0	**0.973 3**	0.943 8	0.959 0	**0.985 7**
	*K*NN	0.959 0	0.945 7	0.900 0	0.918 1	0.915 2
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
χ^2	SVM	0.931 4	**0.973 3**	0.958 1	0.944 8	0.972 4
	*K*NN	0.959 0	0.960 0	0.942 9	**0.973 3**	0.929 5
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
IG	SVM	0.945 7	0.959 0	0.958 1	0.958 1	0.972 4
	*K*NN	0.945 7	**0.973 3**	**0.971 4**	**0.973 3**	0.957 1
	NB	0.959 0	0.959 0	0.958 1	0.958 1	0.958 1
ReliefF	SVM	0.959 0	**0.973 3**	0.957 1	0.959 0	0.971 4
	*K*NN	0.930 5	0.932 4	0.914 3	0.889 5	0.928 6
	NB	0.945 7	**0.973 3**	**0.971 4**	0.958 1	0.971 4
SVM-RFE	SVM	**0.986 7**	0.960 0	0.957 1	0.972 4	0.971 4
	*K*NN	0.929 5	0.945 7	0.873 3	0.931 4	0.901 9
	NB	0.930 5	0.916 2	0.944 8	0.943 8	0.958 1

综上,与基本特征选择方法相比,使用结合单变量与多变量方法的集成方法能够确保选择的特征子集在不同数据集上具有良好稳定性,同时也具有优越的分类性能;其次,集成方法在分类性能上的提升效果与分类器并无显著关联性,采用 SVM 分类器能够获得较好的分类性能.

4 结束语

本文总结了特征选择稳定性提升方法的研究进展,概要阐述了演化算法在特征选择稳定性中的应用,归纳特征选择稳定性中的评估,通过实验分析典型的子集法稳定性度量指标的性能,并验证了结合单变量与多变量算法的集成方法能够同时提高算法的稳定性和分类性能.

尽管特征选择稳定性在近两年得到了学术界的重视和发展,但其仍属于起步阶段,还有一些亟待解决的问题:在高维数据中,除了特征维度较高之外,还有一些常常被忽略的因素,如样本的不平衡、数据分布的漂移和噪声数据等,而目前的提升特征选择稳定性的方法并未考虑这些情况的存在,因此结合高维数据蕴含的特点,提高特征选择方法的稳定性是一项值得深入研究的课题;特征选择稳定性度量指标是特征选择稳定性研究的基础,虽然研究人员提出或借鉴了一些度量指标,但由于在稳定性度量指标应当具备的性质方面并未有统一的标准,造成不同指标度量的结果可能存在差异性,导致我们不能客观全面地评价特征选择稳定性的研究成果,因此对特征选择稳定性度量指标的研究仍然任重道远;目前,多数特征选择稳定性提升方法的研究成果仍然是建立在集成或扰动的机械方法之上,虽然特征法在特征层面对提高稳定性做了进一步的探索,但其泛化能力也是值得商榷的,是否可以针对特征选择稳定性发展出专用的特征选择算法,也是值得探讨的问题;当前,对特征选择稳定性的研究主要聚焦于独立于分类器的过滤式特征选择方法,而作为重要分支的基于进化算法的特征选择方法,在稳定性方面的研究还存在较多的空白.基于进化算法的特征选择方法的稳定性是否与采用的进化算法相关,其与分类器和评价准则之间是否具有关联性,如何提高基于进化算法特征选择的稳定性,也是需要进一步探索的研究方向;对影响特征选择稳定性因素的深入研究和探索,这是从根本上解决特征选择稳定性问题的出发点和落脚点.对不同的数据集或不同的应用而言,造成特征选择不稳定的因素不尽相同,如特征规模、样本数量、数据分布等,然而目前鲜有研究成果对其进行深入探讨.对导致特征选择不稳定的因素以及这些因素之间相互的影响做判断及分析,并以此作为依据提出对应的解决方案,是特征选择稳定性研究的重要内容.

References:

[1] Emani CK, Cullot N, Nicolle C. Understandable big data: A survey. Computer Science Review, 2015,17:70–81. [doi: 10.1016/j.cosrev.2015.05.002]

[2] Fakhraei S, Soltanian-Zadeh H, Fotouhi F. Bias and stability of single variable classifiers for feature ranking and selection. Expert Systems with Applications, 2014,41(15):6945–6958. [doi: 10.1016/j.eswa.2014.05.007]

[3] Li JD, Liu H. Challenges of feature selection for big data analytics. IEEE Intelligent Systems, 2016,32(2):9–15. [doi: 10.1109/MIS.2017.38]

[4] Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. Feature selection for high dimensional data. Progress in Artificial Intelligence, 2016,5(2):65–75. [doi: 10.1007/s13748-015-0080-y]

[5] Goh WW, Wong L. Evaluating feature selection stability in next generation proteomics. Journal of Bioinformatics and Computational Biology, 2016,14(5):1650029. [doi: 10.1142/S0219720016500293]

[6] Du W, Cao ZB, Song TC, Li Y, Liang YC. A feature selection method based on multiple kernel learning with expression profiles of different types. BioData Mining, 2017,10:4. [doi: 10.1186/s13040-017-0124-x]

[7] Chlis NK, Bei ES, Zervakis M. Introducing a stable bootstrap validation framework for reliable genomic signature extraction. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2016,PP(99):1–1. [doi: 10.1109/TCBB.2016.2633267]

[8] Yu K, Wu XD, Ding W, Pei J. Scalable and accurate online feature selection for big data. ACM Trans. on Knowledge Discovery from Data, 2016,11(2):Article 16. [doi: 10.1145/2976744]

[9] Iglesias F, Zseby T. Analysis of network traffic features for anomaly detection. Machine Learning, 2015,101(1):59–84. [doi: 10.1007/s10994-014-5473-9]

[10] Wang YL, Li ZQ, Wang YF, Wang XN, Zheng JJ, Duan XJ, Chen HF. A novel approach for stable selection of informative redundant features from high dimensional fMRI data. Computer Science, 2016,146:191–208. [doi: arXiv:1506.08301]

[11] Park CH, Kim SB. Sequential random *K* nearest neighbor feature selection for high dimensional data. Expert Systems with Applications, 2015,42(5):2336–2342. [doi: 10.1016/j.eswa.2014.10.044]

[12] Aldehim GN. Heuristic ensembles of filters for accurate and reliable feature selection [Ph.D. Thesis]. Norwich: University of East Anglia, 2015.

[13] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: A study on high dimensional spaces. Knowledge and Information Systems, 2007,12(1):95−116. [doi: 10.1007/s10115-006-0040-8]

[14] Fan M, Chou CA. Exploring stability based voxel selection methods in MVPA using cognitive neuroimaging data: A comprehensive study. Brain Informatics, 2016,3(3):193−203. [doi: 10.1007/s40708-016-0048-0]

[15] Tohka J, Moradi E, Huttunen H. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics, 2016,14(3):1−18. [doi: 10.1007/s12021-015-9292-3]

[16] Tommasel A, Godoy D. Short text feature construction and selection in social media data: A survey. Artificial Intelligence Review, 2016:1−38. [doi: 10.1007/s10462-016-9528-0]

[17] Alkuhlani A, Nassef M, Farag I. Multistage feature selection approach for high dimensional caner data. Soft Computing, 2016:1−12. [doi: 10.1371/journal.pone.0117988]

[18] Gangeh MJ, Zarkoob H, Ghodsi A. Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2017,14(1):167−181. [doi: 10.1109/TCBB.2016.2631164]

[19] Schirra LR, Lausser L, A.Kestler H. Selection stability as a means of biomarker discovery in classification. Studies in Classification, Data Analysis, and Knowledge Organization, 2016:79−89. [doi: 10.1007/978-3-319-25226-1_7]

[20] Zhou QF, Ding JC, Ning YP, Luo LK, Li T. Stable feature selection with ensembles of multi ReliefF. In: Proc. of the 2014 10th Int'l Conf. on Natural. 2014. 742−747. [doi: 10.1109/ICNC.2014.6975929]

[21] Pes B, Dessi N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: A case study on high dimensional genomic data. Information Fusion, 2017,35(C):132−147. [doi: 10.1016/j.inffus.2016.10.001]

[22] Saeys Y, Abeel T, Peer YVD. Robust feature selection using ensemble feature selection techniques. In: Proc. of the ECML/PKDD. 2008. 313−325. [doi: 10.1007/978-3-540-87481-2_21]

[23] Abeel T, Helleputte T, Peer YVD, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 2010,26(3):392−398. [doi: 10.1093/bioinformatics/btp630]

[24] Yang P, Ho JW, Yang YH, Zhou BB. Gene-Gene interaction filtering with ensemble of filters. Bmc Bioinformatics, 2011,12 Suppl 1(S1):S10. [doi: 10.1186/1471-2105-12-S1-S10]

[25] Rondina JM, Hahn T, Oliveira LD, Marquand A, Dresler T, Leitner T, Fallgatter AJ, Shawe-Taylor J, Mourao-Miranda J. SCoRS a method based on stability for feature selection and apping in neuroimaging. IEEE Trans. on Medical Imaging, 2014,33(1):85−98. [doi: 10.1109/TMI.2013.2281398]

[26] Kim HJ, Choi BS, Huh MY. Booster in high dimensional data classification. IEEE Trans. on Knowledge and Data Engineering, 2016,28(1):29−40. [doi: 10.1109/TKDE.2015.2458867]

[27] He ZY, Yu WC. Stable feature selection for biomarker discovery. Computational Biology and Chemistry, 2010,34(4):215−225. [doi: 10.1016/j.compbiolchem.2010.07.002]

[28] Kamker I, Gupta SK, Phung D, Venkatesh S. Stabilizing l_1-norm prediction models by supervised feature grouping. Journal of Biomedical Informatics, 2016,59(C):149−168. [doi: 10.1016/j.jbi.2015.11.012]

[29] Moayedikia A, Ong KL, Boo YL, Yeoh WGS, Jensen R. Feature selection for high dimensional imbalanced class data using harmony search. Engineering Applications of Artificial Intelligence, 2017,57(C):38−49. [doi: 10.1016/j.engappai.2016.10.008]

[30] Fahad A, Tari Z, Khalil I, Almalawi A, Zomaya A. An optimal and stable feature selection approach for traffic classification based on multi criterion fusion. Future Generation Computer Systems, 2014,36(7):156−169. [doi: 10.1016/j.future.2013.09.015]

[31] Aldehim G, Wang WJ. Weighted heuristic ensemble of filters. In: Proc. of the SAI Intelligent Systems Conf. 2015. 609−615. [doi: 10.1109/IntelliSys.2015.7361203]

[32] Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. Data classification using an ensemble of filters. Neurocomputing, 2014, 135:13−20. [doi: 10.1016/j.neucom.2013.03.067]

[33] Lior R, Barak C. A methodology for improving the performance of non-ranker feature selection filters. Int'l Journal of Pattern Recognition and Artificial Intelligence, 2007,21(5):809−830. [doi: 10.1142/S0218001407005727]

[34] Yang F, Ma KZ. Robust feature selection for microarray data based on multi criterion fusion. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2011,8(4):1080−1092. [doi: 10.1109/TCBB.2010.103]

[35] Boucheham A, Batouche M. Massively parallel feature selection based on ensemble of filters and multiple robust consensus functions for cancer gene identification. In: Proc. of the Intelligent Systems in Science and Information. 2014. 93−108. [doi: 10.1007/978-3-319-14654-6_6]

[36] Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano A. Comparing two new gene selection ensemble approaches with the commonly used approach. In: Proc. of the 11th Int'l Conf. on Machine Learning and Applications. 2012. 184−191. [doi: 10.1109/ICMLA.2012.175]

[37] Kuncheva L, Smith CJ, Syed Y, Phillips CO, Lewis KE. Evaluation of feature ranking ensembles for high dimensional biomedical data: A case study. In: Proc. of the IEEE Int'l Conf. on Data Mining Workshops. 2013. 49−56. [doi: 10.1109/ICDMW.2012.12]

[38] Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. In: Proc. of the ACM Conf. on Knowledge Discovery and Data Mining. 2009. 567−575. [doi: 10.1145/1557019.1557084]

[39] Garcia-Torres M, Gomez-Vela F, Melian-Batista B, Moreno-Vega JM. High dimensional feature selection via feature grouping: A variable neighborhood search approach. Information Sciences, 2016,326(C):102−118. [doi: 10.1016/j.ins.2015.07.041]

[40] Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. In: Proc. of the 14th ACM Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 803−811. [doi: 10.1145/1401890.1401986]

[41] Huang J, Horowitz JL, Ma SG. Asymptotic properties of bridge estimators in sparse high dimensional regression models. Annals of Statistics, 2008,36(2):587−613. [doi: 10.1214/009053607000000875]

[42] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 2005,67(2): 301−320. [doi: 10.1111/j.1467-9868.2005.00503.x]

[43] Gauraha N. Stability feature selection using cluster representative lasso. In: Proc. of the Int'l Conf. on Pattern Recognition Applications and Methods. 2016. 381−386. [doi: 10.5220/0005827003810386]

[44] Silva B, Marques N. Feature clustering with self-organizing maps and an application to financial time-series for portfolio selection. In: Proc. of the 6th Int'l Conf. on Neural Computation. 2010. 301−309.

[45] Wang LP, Chu F, Xie W. Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2007,4(1):40−53. [doi: 10.1109/TCBB.2007.1006]

[46] Dettling M, Buhlmann P. Finding predictive gene groups from microarray data. Journal of Multivariate Analysis, 2004,90(1): 106−131. [doi: 10.1016/j.jmva.2004.02.012]

[47] Song QB, Ni JJ, Wang GT. A fast clustering based feature subset selection algorithm for high dimensional data. IEEE Trans. on Knowledge and Data Engineering, 2013,25(1):1−14. [doi: 10.1109/TKDE.2011.181]

[48] Shu L, Ma TY, Latecki LJ. Stable feature selection with minimal independent dominating sets. In: Proc. of the ACM Int'l Conf. on Bioinformatics. 2013. 450−457. [doi: 10.1145/2506583.2506600]

[49] Beinrucker A, Dogan U, Blanchard G. Extensions of stability selection using subsamples of observations and covariates. Statistics and Computing, 2016,26(5):1059−1077. [doi: 10.1007/s11222-015-9589-y]

[50] Jerbi W, Brahim AB, Essoussi N. A hybrid embedded filter method for improving feature selection stability of random forests. In: Proc. of the 16th Int'l Conf. on Hybrid Intelligent Systems. 2016. 370−379. [doi: 10.1007/978-3-319-52941-7_37]

[51] Gabriel P, Belanche LA. Improved stability of feature selection by combining instance and feature weighting. In: Proc. of the Research and Development in Intelligent Systems XXXI. 2014. 35−49. [doi: 10.1007/978-3-319-12069-0_3]

[52] Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing, 2016,173:346−354. [doi: 10.1016/j.neucom.2014.12.123]

[53] Li Y Si J, Zhou GJ, Huang SS, Chen SC. FREL: A stable feature selection algorithm. IEEE Trans. on Neural Networks and Learning Systems, 2015,26(7):1388−1402. [doi: 10.1109/TNNLS.2014.2341627]

[54] Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B Chemical, 2015,212:353−363. [doi: 10.1016/j.snb.2015.02.025]

[55] Lin XH, Wang XM, Xiao NY, Huang X, Wang J. A feature selection method based on feature grouping and genetic algorithm. In: Proc. of the Int'l Conf. on Intelligent Science and Big Data Engineering. 2015. 150−158. [doi: 10.1007/978-3-319-23862-3_15]

[56] Soufan O, Kleftogiannis D, Kalnis P, Bajic VB. DWFS: A wrapper feature selection tool based on a parallel genetic algorithm. Plos One, 2015,10(2):e0117988. [doi: 10.1371/journal.pone.0117988]

[57] Liu QJ, Zhao ZM, Li YX, Yu XL. Ensemble feature selection method based on neighborhood information and pso algorithm. Acta Electronica Sinica, 2016,44(4):995−1002 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.2016.04.034]

[58] Xue B, Zhang MJ, Brownw WN, Yao X. A survey on evolutionary computation approaches to feature selection. IEEE Trans. on Evolutionary Computation, 2016,20(4):606−626. [doi: 10.1109/TEVC.2015.2504420]

[59] Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics, 2008,24(2):258−264. [doi: 10.1093/bioinformatics/btm550]

[60] Bloulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. Briefings in Bioinformatics, 2009,10(5):556−568. [doi: 10.1093/bib/bbp034]

[61] Guzman-Martinez R, Alaiz-Rodriguez R. Feature selection stability assessment based on the Jensen-Shannon divergence. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2011. 597−612. [doi: 10.1007/978-3-642-23780-5_48]

[62] Nogueria S, Brown G. Measuring the stability of feature selection with applications to ensemble methods. In: Proc. of the 12th Int'l Workshop on Multiple Classifier Systems. 2015. 135−146. [doi: 10.1007/978-3-319-20248-8_12]

[63] Kuncheva LI. A stability index for feature selection. In: Proc. of the 25th ACM Conf. on Int'l Multi-Conf. Artificial Intelligence and Applications. 2007. 390−395.

[64] Somol P, Novovicovaa J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(11):1921−1939. [doi: 10.1109/TPAMI.2010.34]

[65] Ning YP. Research on feature selection and stability analysis for high dimensionality small sample size data [MS. Thesis]. Xiamen: Xiamen University, 2014 (in Chinese with English abstract).

[66] Ji JS. Feature selection and its stability for typical geoobjects of high resolution remote sensing image [MS. Thesis]. Shanghai: Shanghai Jiao Tong University, 2015 (in Chinese with English abstract).

[67] Gulgenzen G, Cataltepe Z, Yu L. Stable and accurate feature selection. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. 2009. 455−468. [doi: 10.1007/978-3-642-04180-8_47]

[68] Nogueira S, Brown G. Measuring the stability of feature selection. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2016. 442−457. [doi: 10.1007/978-3-319-46227-1_28]

[69] Kamkar I, Gupta SK, Phung D, Venkatesh S. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-LASSO. Journal of Biomedical Informatics, 2015,53:277−290. [doi: 10.1016/j.jbi.2014.11.013]

[70] Drotar P, Smekal Z. Stability of feature selection algorithms and its influence on prediction accuracy in biomedical datasets. In: Proc. of the TENCON IEEE Region 10th Conf. 2014. 1−5. [doi: 10.1109/TENCON.2014.7022309]

[71] Wang H, Khoshgoftaar TM, Seliya N. On the stability of feature selection methods in software quality prediction: an empirical investigation. Int'l Journal of Software Engineering and Knowledge Engineering, 2015,25(9n10):1467−1490. [doi: 10.1142/S0218194015400288]

[72] Wang H, Khoshgoftaar T, Napolitano A. Stability of three forms of feature selection methods on software engineering data. In: Proc. of the Int'l Conf. on Software Engineering and Knowledge Engineering. 2015. 385−390. [doi: 10.18293/SEKE2015-198]

[73] Hassan SS, Ruusuvuori P, Latonen L, Huttunen H. Flow cytometry based classification in cancer research: A view on feature selection. Cancer Informatics, 2016,14(5):75. [doi: 10.4137/CIN.S30795]

[74] Wang HJ, Khoshgoftaar TM, Napolitano A. Stability of filter- and wrapper- based software metric selection techniques. In: Proc. of the IEEE Int'l Conf. on Information Reuse and Integration. 2015. 309−314. [doi: 10.1109/IRI.2014.7051905]

[75] Dessi N, Pes B. Similarity of feature selection methods: An empirical study across data intensive classification tasks. Expert Systems with Applications, 2015,42:4632−4642. [doi: 10.1016/j.eswa.2015.01.069]

[76] Dernoncourt D, Hanczar B, Zucker JD. Analysis of feature selection stability on high dimension and small sample data. Computational Statistics and Data Analysis, 2014,71(C):681−693. [doi: 10.1016/j.csda.2013.07.012]

[77] Tohka J, Moradi E, Huttunen H. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics, 2016,14(3):1−18. [doi: 10.1007/s12021-015-9292-3]

[78] Perthame E, Friguet C, Causeur D. Stability of feature selection in classification issues for high dimensional correlated data. Statistics and Computing, 2016,26(4):783−796. [doi: 10.1007/s11222-015-9569-2]

[79] Drotar P, Smekal Z. Comparison of stability measures for feature selection. In: Proc. of the IEEE 13th Int'l Symp. on Applied Machine Intelligence and Informatics. 2015. 71−75. [doi: 10.1109/SAMI.2015.7061849]

附中文参考文献:

[57] 刘全金,赵志敏,李颖新,俞晓磊.基于近邻信息和 PSO 算法的集成特征选取.电子学报,2016,44(4):995−1002. [doi: 10.3969/j.issn.0372-2112.2016.04.034]

[65] 宁永鹏.高维小样本数据的特征选择研究及其稳定性分析[硕士学位论文].厦门:厦门大学,2014.

[66] 季金胜.高分辨率遥感影像典型地物目标的特征选择及其稳定性研究[硕士学位论文].上海:上海交通大学,2015.

刘艺(1990－),男,安徽蚌埠人,博士生,主要研究领域为数据治理,演化算法.

刁兴春(1964－),男,研究员,博士生导师,主要研究领域为数据工程.

曹建军(1975－),男,博士,副研究员,CCF 高级会员,主要研究领域为数据治理,演化算法.

周星(1988－),男,博士,工程师,主要研究领域为数据挖掘,数据工程.

www.jos.org.cn