计算机应用 2013, 33(08) 2204-2207 DOI: ISSN: 1001-9081 CN: 51-1307/TP

数据库技术

基于数据集特点的增强聚类集成算法

侯勇[1,2],郑雪峰[1]

1. 北京科技大学 计算机与通信工程学院，北京 100083;
2. 山东经贸职业学院 科学与人文学院，山东 潍坊 261011

摘要： 当前流行的聚类集成算法无法依据不同数据集的不同特点给出恰当的处理方案，为此提出一种新的基于数据集特点的增强聚类集成算法,该算法由基聚类器的生成、基聚类器的选择与共识函数构成。该算法依据数据集的特点，通过启发式方法，选出合适的基聚类器，构建最终的基聚类器集合，并产生最终聚类结果。实验中，对ecoli，leukaemia与Vehicle三个基准数据集进行了聚类，所提出算法的聚类误差分别是0.014，0.489，0.479，同基于Bagging的结构化集成(BSEA)、异构聚类集成(HCE)和基于聚类的集成分类(COEC)算法相比，所提出算法的聚类误差始终最低；而在增加候基聚类器的情况下，所提出算法的标准化互信息(NMI)值始终高于对比算法。实验结果表明，同对比的聚类集成算法相比，所提出算法的聚类精度最高，可伸缩性最强。

关键词： 基聚类器 共识函数 聚类集成算法 聚类误差 自适应性 标准化互信息

Enhanced clustering ensemble algorithm based on characteristics of data sets

HOU Yong[1,2],ZHENG Xuefeng[2]

1. College of Humanities and Science, Shandong Vocational College of Economics and Business, Weifang Shandong 61011, China
2. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract: The popular clustering ensemble algorithms cannot give the appropriate treatment program in the light of the different characteristics of the different data sets. A new clustering ensemble algorithm — Enhanced Clustering Ensemble algorithm based on Characteristics of Data sets (ECECD) was proposed for overcoming this defect. ECECD was composed of generation of base clustering, selection of base clustering and consensus function. It selected a special range of ensemble members to form the final ensemble and produced the final clustering based on the characteristic of the data set. Three Benchmark data sets including ecoli, leukaemia and Vehicle were clustered in the experiment, and the clustering errors gained by the proposed algorithm were 0.014, 0.489 and 0.361 respectively, which were always the minimum compared with that of the other algorithms such as Bagging based Structure Ensemble Approach (BSEA), Hybrid Cluster Ensemble (HCE) and Cluster-Oriented Ensemble Classifier (COES). The Normalized Mutual Information (NMI) values of the proposed algorithm were also always higher than that of these algorithms when increasing candidate base clusterings. Therefore, compared with these popular clustering ensemble algorithms, the proposed algorithm has the highest clustering precision and the strongest scalability.

Keywords: base clustering consensus function clustering ensemble algorithm clustering error adaptivity Normalized Mutual Information (NMI)

收稿日期 2013-02-04 修回日期 2013-03-12 网络版发布日期 2013-09-11

DOI:

基金项目:

山东省企业培训与职工教育课题资助项目;潍坊市社科规划重点课题资助项目;山东省高校人文社科研究计划项目

通讯作者: 侯勇

作者简介: 侯勇(1978-)，男，山东蓬莱人，讲师，博士研究生，主要研究方向：数据挖掘、网络安全、机器学习;
郑雪峰(1951-)，男，福建福州人，教授，主要研究方向：网络安全。
作者Email: aspnetcs@163.com

参考文献：

［1］GIOTIS I, PETKOV N. Cluster-based adaptive metric classification ［J］. Neurocomputing, 2012, 81:33-40.

［2］ ANDREW S, KHALED A. Clustering sentence-level text using a novel fuzzy relational clustering algorithm ［J］. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1):62-75.

［3］KANNAN S R, RAMATHILAGAM S, CHUNG P C, et al. Effective fuzzy c-means clustering algorithms for data clustering problems［J］. Expert Systems with Application,2012,39(7):6292-6300.

［4］WOLOSZYNSKI T, KURZYNSKI M, PODSIADLO P, et al. A measure of competence based on random classification for dynamic ensemble selection ［J］. Information Fusion, 2012, 13(3):207-213.

［5］CHEN J C, WU C-C, CHEN C-W, et al. Flexible job shop scheduling with parallel machines using genetic algorithm and grouping genetic algorithm［J］. Expert Systems with Application, 2012, 39(11): 10016-10021.

［6］KHALEGHI M, FARSANGI M M, NEZAMABADI-POUR H, et al. Pareto-optimal design of damping controllers using modified artificial immune algorithm ［J］.IEEE Transactions on Systems, Man and Cybernetics: Part C,Applications and Reviews,2011,41(2):240-250.

［7］PARTALAS I, TSOUMAKAS G, VLAHAVAS I, et al. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning ［J］. Machine Learning, 2010, 81(3):257-282.

［8］MAHAJAN M, NIMBHORKAR P, VARADARAJAN K, et al. The planar k- means problem is NP-hard ［J］. Theoretical Computer Science, 2012, 442: 13-21.

［9］ZHANG S, WONG H S, SHEN Y, et al. Generalized adjusted rand indices for cluster ensembles ［J］. Pattern Recognition, 2012, 45(6): 2214-2226.

［10］QING C, JIANG J, YANG Z. Normalized co-occurrence mutual information for facial pose detection inside videos ［J］. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(12): 1898-1902.

［11］YU Z, YOU J, WONG H S, et al. From cluster ensemble to structure ensemble ［J］. Information Sciences, 2012, 198:81-99.

［12］YU Z, WONG H S, YOU J, et al. Hybrid cluster ensemble framework based on the random combination of data transformation operators ［J］. Pattern Recognition, 2012, 45(5): 1826-1837.

［13］VERMA B, RAHMAN A. Cluster-oriented ensemble classifier: impact of multicluster characterization on ensemble classifier learning ［J］.IEEE Transactions on Knowledge and Data Engineering, 2012, 24(4):605-618.

［14］ANKARA H, YEREL S. Determination of sampling errors in natural stone plates through single linkage cluster method ［J］.Journal of Materials Processing Technology, 2009, 209(5):2483-2487.

［15］DOSEA M, SILVA L, SILVA M A, et al. Adaptive mean-linkage with penalty: a new algorithm for cluster analysis ［J］.Chemometrics and Intelligent Laboratory System, 2008, 94(1): 1-8.

［16］OKUMOTO K, FUKUNAGA T, NAGAMOCHI H, et al. Divide-and-conquer algorithms for partitioning hypergraphs and submodular systems ［J］. Algorithmica, 2012, 62(3/4):787-806.

［17］SEVILLANO X, ALIAS F, SOCORO J C, et al. Positional and confidence voting-based consensus functions for fuzzy cluster ensembles ［J］.Fuzzy Sets and Systems, 2012,193:1-32.

本刊中的类似文章

1．曾宪权 裴洪文.支持扩展的自适应移动中间件模型及其设计[J]. 计算机应用, 2009,29(09): 2559-2561

2．陈昌涛 朱勤 周圣毅 张家铭.核函数带宽自适应的Mean-Shift跟踪算法[J]. 计算机应用, 2009,29(06): 1680-1682