[文章编号] 1007－7405(2021)02－0146－06 DOI:10.19715/j.jmuzr.2021.02.09

ECOC多分类算法在慕课数据挖掘中的应用

潘丽芳[1]，谢书童[2]，曹秀娟[2]

（1. 集美大学理学院，福建 厦门 361021；2. 集美大学计算机工程学院，福建 厦门 361021）

[摘要] 收集并整合多所高校学生的慕课学习行为数据，设计基于数据复杂度的纠错输出编码（ECOC）多分类算法。该算法利用数据复杂度降低多类之间的分类难度，从而提高算法的预测准确度。实验结果表明，在不同高校的慕课数据集的测试中，所设计基于数据复杂度的ECOC分类算法比传统的ECOC算法具有更高的分类准确度和鲁棒性，实现了学生学习成绩多等级的有效预测，为个性化教学奠定了基础。

[关键词] ECOC；多分类；慕课；成绩预测；教育数据挖掘

[中图分类号] TP 391

Application of Multi-Classification Algorithms Based on ECOC in MOOC Data Mining

PAN Lifang[1], XIE Shutong[2], CAO Xiujuan[2]

(1. School of Science, Jimei University, Xiamen 361021, China;
2. College of Computer Engineering, Jimei University, Xiamen 361021, China)

Abstract: This research collected and integrated the data of learning behaviors of students from many colleges through MOOC platform, and designed ECOC multi-classification algorithms based on data complexity. The algorithms use data complexity to reduce the classification complexity between multiple categories, thereby improving the prediction accuracy of the algorithms. The experimental results showed that the ECOC classification algorithms based on data complexity proposed in this paper have higher classification accuracy and robustness than the classic ECOC algorithms in MOOC data sets of different colleges. The proposed algorithms perform effective prediction of students' academic performance, which lays a foundation to realize the personalized teaching for the students.

Keywords: ECOC; multi-classification; MOOC; grade prediction; educational data mining

0 引言

近年来，随着慕课在线学习课程迅速普及，更多的学生得到了便捷高效的学习体验[1]。然而，慕课动则成千上万名学生在一起大规模的学习模式，与现代教育理念中学生学习的个性化需求形成了尖锐的矛盾。在慕课规模化教学条件下，合理引导并实现学生的个性化学习是一个教育和学术并重的

[收稿日期] 2020－04－11

[基金项目] 福建省自然科学基金项目（2018J01538，2020J01707，2020R0066）；福建省中青年教师教育科研项目（JAT200266）

[作者简介] 潘丽芳（1981—），女，硕士，讲师，从事应用数学、教育数据分析方向研究。通信作者：谢书童（1982—），男，博士，副教授，从事机器学习与数据挖掘方向研究。E-mail:shutong@jmu.edu.cn

问题，具有重要研究价值[2-3]。

蒋卓轩等[4]对北京大学在 Coursera 平台上开设的 6 门慕课的学习行为数据进行分析与挖掘。结果表明，基于学习行为特征的数据分析能有效地判别一个学生最后能否获得课程证书。Qiu 等[5]基于清华大学学堂在线的慕课数据，提出用潜在动态因子图模型（latent dynamic factor graph，LadFG）来预测学生作业情况，以及能否通过考试获得证书。Xu 等[6]通过分析学生在慕课平台上的活动日志，归纳出不同学生的学习动机，设计出基于支持向量机（SVM）的分类算法预测学生能否取得证书。Zhang 等[7]对北大《数据结构与算法》与《计算机导论》两门慕课，汇聚多源异构数据，分析学生的学习内容，识别课程中的重要概念，通过学生测验数据来评估学生的知识状态，并设计出算法预测学生是否中途退课。Yu 等[8]根据学生的视频点击流日志，识别出学生的七种认知参与模型，设计了基于 K 最近邻（KNN）、SVM、人工神经网络（ANN）的分类算法，预测学生能否通过课程考试。

为了提高传统线下课程与线上慕课的教学效率，Meier 等[9]利用课程的历史教学数据（主要包括作业、小测、期中考试等），预测学生在后继学习中的可能表现（好/差），为提前对学习不佳的学生进行教学干预赢得时间。Xu 等[10]设计了一个双层结构的集成分类系统，对学生不断变化的学习状态进行动态预测，并提出了一种基于潜在因子模型和概率矩阵分解的数据驱动方法，发现了课程的相关性，从而提高预测准确度。Ulloa-Cazarez 等[11]提出遗传规划（genetic programming，GP）算法预测学生能否通过期末考试。为了识别出学习《数字设计》课程有困难的学生，Hussain 等[12]使用 ANN 与 SVM 等算法，利用学习系统上的学习行为数据，预测出学习困难的学生，方便提前教学干预。

上述研究工作大多根据慕课中的学习行为数据，设计分类预测算法来预测后续学习成效，即预测学生能否通过考试、提前退课、顺利毕业等，这实质上是一个二分类问题的研究。然而，学生学习成效的多分类预测比二分类更有利于学生个性化教学的实施，但多分类预测难度较大，目前国内外有关慕课数据的多分类预测的研究很少。本文提出基于数据复杂度的纠错输出编码（error correct output codes，ECOC）多分类算法，对慕课数据进行挖掘，以期实现对学生成绩的多分类预测，即预测学生成绩的四个等级（优、良、合格、不合格），为个性化和差异化的教学干预提供理论基础与技术条件。

1　基于数据复杂度的 ECOC 分类算法

本文提出的基于数据复杂度的 ECOC 分类算法，首先利用数据复杂度降低两个分类的复杂度，然后，利用 ECOC 算法实现多个二分类预测，从而实现多分类预测。该算法主要分为二个步骤：基于数据复杂度的二分类调整方法；基于 ECOC 的多分类预测。

1.1　基于数据复杂度的调整方法

数据复杂度是通过分析数据特征来衡量数据样本分类的难易程度[13-14]。数据复杂度越高的分类问题，算法越难以实现正确的分类预测。因此，本文采用了多种数据复杂度方法降低二分类问题的分类复杂程度，它们分别是费希尔判别率（F_1）、交叉重叠体积（F_2）、重叠区域数据点数（F_3）、同类实例与异类实例距离比（N_2）、基于最近邻分类器的错误率（N_3）、INN 分类器的误差率（N_4）、非线性分类器的非线性特点（L_3），以及基于质心匹配的方法（C_1）[13]。

1.2　ECOC 多分类算法

ECOC 多分类算法最早由通信领域为解决信号传输问题而提出，其处理问题的关键在于选择高效优秀的编码策略和解码策略[15]。ECOC 多分类算法的主要策略是将多类问题转化为多个二分类问题进行求解。ECOC 多分类算法包括三个基本步骤：编码、训练、解码。

编码策略的设计是为了得到实现多类分解和基分类器集成的编码矩阵。编码矩阵是由二元（-1，+1）或者三元（-1，+1，0）组成，其中 -1 代表负类，0 代表对应的类在分类时被忽略，+1 代表正类；行向量表示一个类别，列向量表示一个基二分类器。此外，编码策略分数据无关和数据相关两种类型。数据无关主要是指编码时不依赖数据样本，数据相关是指编码时样本特征的分布紧

密联系着编码矩阵。数据无关算法包括一对一（One Vs. One，OVO）、一对多（One Vs. All，OVA）、稀疏随机（sparse random）和密集随机（dense random）策略。数据相关算法包括 DECOC、Forest-ECOC、ECOC－ONE 等。编码矩阵行列应尽可能不同，减少基二分类器和码字之间的相关性。训练步骤就是通过训练数据对 ECOC 及其相应的基分类器进行训练，从而调整 ECOC 及其基分类器的参数。

解码策略是指在测试某个样本时，每个基二分类器都会产生一个相同长度的向量作为输出，然后计算出输出向量与编码矩阵中的码字之间的距离，选择距离最小的码字作为最终结果，并将相应的类别标签赋予该样本。解码策略包括 AED、ED、ELB、ELW、HD、LAP、LLB 等。

1.3 算法框架

首先，基于数据复杂度的 ECOC 分类算法需将所有的类别随机分成数量尽可能相同的两组，然后对每个类别进行相同的数据复杂度评估，再交换两个组中具有最高复杂度的两个类，从而降低两个组组内整体的数据复杂度，提高对应基二分类器的泛化能力。通过降低类别之间的数据复杂度，使得分类算法更容易区分不同类别，从而达到提高预测准确率的目标。在追求降低组内整体复杂度时，F1 和 F3 两种数据复杂度评价指标是相反的，应尽可能提高组内 F1 或 F3 指标的值才能降低复杂性。不断调整以上类别分组过程以达到最优分配，即组内数据复杂度达到最低，形成类别树形分布。将节点编码为 +1 或 －1（没有参与分类的节点编码为 0），形成编码矩阵的一列，得到编码矩阵后，再用训练分类器对样本进行预测，算法整体框架如图 1 所示。

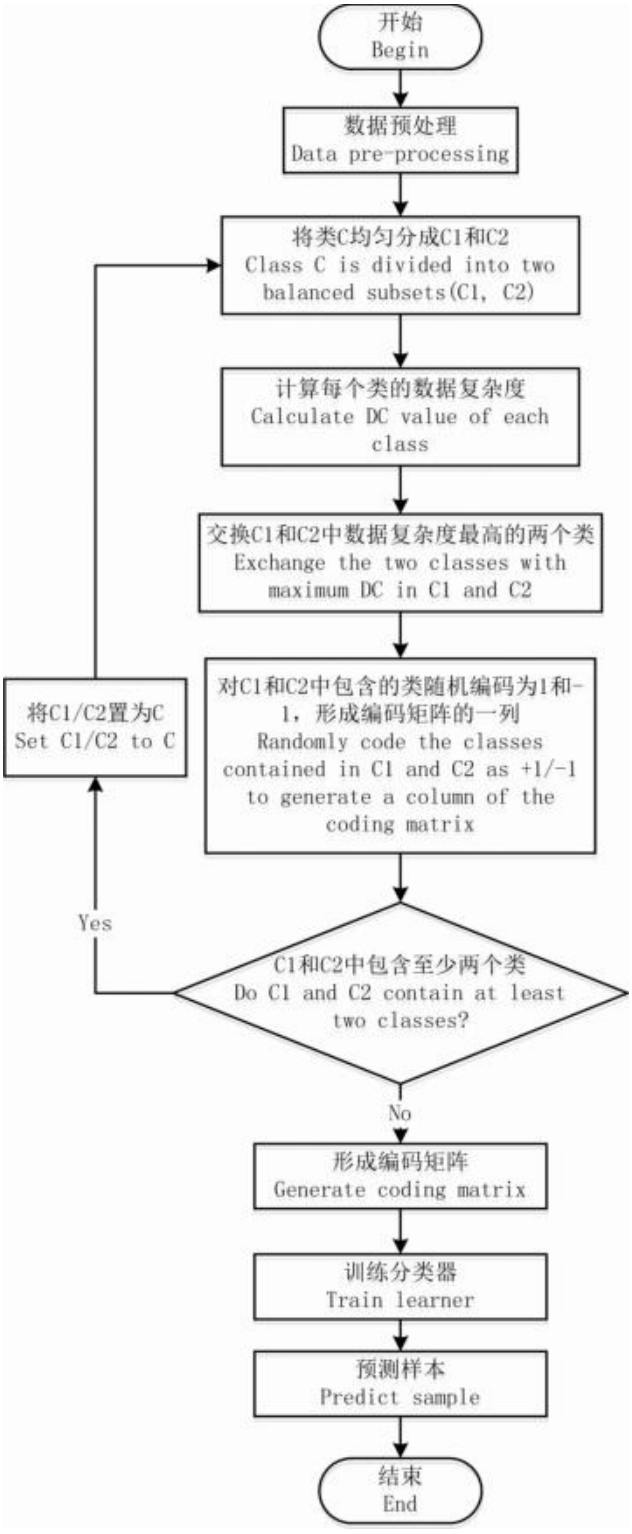

图 1 基于数据复杂度的 ECOC 分类算法框架

Fig.1 The framework of ECOC method based on data complexity

2 实验结果与分析

2.1 数据集与实验参数设置

收集并整合三所高校的计算机基础课程的慕课数据，预处理后的类别数量、特征数量、样本数量如表 1 所示。根据学生的期末成绩，把学生样本划分成四类：分数位于 0（含 0）到 60 之间判为第 4 类，分数位于 60（含 60）到 75 之间判为第 3 类，分数位于 75（含 75）到 85 之间判为第 2 类，分数

位于 85（含 85）到 100（含）之间判为第 1 类，亦即为不合格、合格、良与优四个等级。该课程的知识点包括硬件系统组成、操作系统与常用软件、音频和图像处理、数据库技术、算法、无线通信技术、加密解密算法，以及网络安全等。将相应的知识点做成教学视频放于慕课平台供学生学习，且每章均有课后习题，另有阶段性小测，以巩固学生对知识点的掌握。本研究收集了学生在慕课平台观看教学视频的行为数据（包括访问时间、访问时长、所用电子设备）、教学视频中的答题数据、章节作业数据、小测数据，以及线下期末成绩等大量细粒度的学习行为数据。

表 1　学生慕课学习行为数据集

Tab. 1　Data sets of students' learning behavior in MOOC

数据集 Data set	类别数量 Number of classes	特征数量 Number of features	样本数量 Number of samples
学校 A School A	4	32	2430
学校 B School B	4	80	3107
学校 C School C	4	151	3611

算法在 Matlab 上实现，并调用 Sklearn 工具包。为充分验证算法的有效性，实验均采用十折交叉法进行测试，取平均值作为评价标准；均采用了两种基二分类器，即 SVM 和朴素贝叶斯（NB）；所有算法的解码策略均采用 ELW（即指数损失加权解码）方式。五种传统 ECOC 算法，即编码策略为 OVO、OVA、DECOC、ECOC – ONE、Forest – ECOC 的均进行了相应的实验，以对比本文提出的基于数据复杂度的 ECOC 与传统 ECOC 的算法性能。

2.2　实验结果分析

如表 2 和表 3 所示，准确率最高的实验结果采用下划线并加粗标识。从表 2 基于 SVM 的 ECOC 算法的预测结果可以发现：在学校 A 数据集上，基于数据复杂度的 ECOC 算法和其他 ECOC 算法性能相近，准确率在 81% 左右；在学校 B 数据集上，基于质心匹配的数据复杂度 C1 的 ECOC 算法预测准确率最高，达到 73.32%，与基于数据复杂度的 ECOC 算法性能相近，但比传统 ECOC 算法准确率平均提升约 0.6%；在学校 C 数据集上，基于质心匹配的数据复杂度 C1 的 ECOC 算法获得最高准确率，可达 73.55%，基于数据复杂度的 ECOC 算法相比其他 ECOC 算法准确率平均提升约 0.51%；在所有编码方式中，基于质心匹配的数据复杂度 C1 的 ECOC 算法获得最优平均准确率为 75.95%。综上分析可得，基于数据复杂度的 ECOC 算法相比于传统 ECOC 算法具有更优的性能。

表 2　基于 SVM 的不同 ECOC 算法的准确率对比

Tab. 2　Accuracy of different ECOC methods based on SVM

基分类器 Base learner	编码方式 Encoding method	学校 A School A	学校 B School B	学校 C School C	平均准确率 Average accuracy
SVM	ECOC – F_1	0.8097	0.7287	0.7327	0.7570
	ECOC – F_2	0.8099	0.7264	0.7308	0.7557
	ECOC – F_3	0.8101	0.7308	0.7317	0.7575
	ECOC – N_2	0.8105	0.7282	0.7320	0.7569
	ECOC – N_3	0.8101	0.7288	0.7285	0.7558
	ECOC – N_4	0.8105	0.7269	0.7288	0.7554
	ECOC – C_1	0.8099	0.7332	0.7355	0.7595
	ECOC – L_3	0.8107	0.7282	0.7324	0.7571
	OVO	0.8101	0.7290	0.7262	0.7551
	OVA	0.8109	0.7166	0.7166	0.7480
	DECOC	0.8105	0.7218	0.7312	0.7545
	ECOCONE	0.8107	0.7171	0.7288	0.7522
	Forest – ECOC	0.8103	0.7280	0.7292	0.7558

基于 NB 的 ECOC 算法的结果如表 3 所示。在所有算法中，基于最近邻分类器的错误率数据复杂

度 N3 的 ECOC 算法在三个数据集上平均分类准确率最高，达到 68.95%。在学校 A 数据集上，OVA 算法取得最好的准确率，即 71.56%，其余 ECOC 算法的准确率相近；在学校 B 数据集上，DECOC 算法取得最好的准确率，即 70.26%；在学校 C 数据集上，基于最近邻分类器的错误率数据复杂度 N3 的 ECOC 算法取得最好的准确率，即 67.63%。从平均准确率上看，基于数据复杂度的 ECOC 算法和传统 ECOC 算法并没有明显的差距。

由表 2 和表 3 的结果可得，在学校 A、B、C 三个数据集上，基于 SVM 的 ECOC 算法的性能明显优于基于 NB 的 ECOC 算法，平均预测准确率分别提高了约 11%、4%、6%。由图 2 可以看出，基于 SVM 的 ECOC 算法预测的平均准确率（三个数据集的准确率平均值）明显优于基于 NB 的 ECOC 算法。进一步可以发现，在采用 SVM 为基分类器情况下，ECOC 算法结果总体波动较小，其中基于数据复杂度的 ECOC 算法性能更加稳定，且优于传统 ECOC。相反地，在基于 NB 分类器情况下，ECOC 算法准确率波动较大，其中基于数据复杂度的 ECOC 算法效果较好。

表 3 基于 NB 的不同 ECOC 算法的准确率对比

Tab. 3 Accuracy of different ECOC methods based on NB

基分类器 Base learner	编码方式 Encoding method	学校 A School A	学校 B School B	学校 C School C	平均准确率 Average accuracy
NB	ECOC – F_1	0.6872	0.6775	0.6660	0.6769
	ECOC – F_2	0.6918	0.6789	0.6685	0.6797
	ECOC – F_3	0.6844	0.6780	0.6679	0.6768
	ECOC – N_2	0.6829	0.6770	0.6673	0.6757
	ECOC – N_3	0.6913	0.7010	0.6763	0.6895
	ECOC – N_4	0.6868	0.6900	0.6679	0.6816
	ECOC – C_1	0.6858	0.6783	0.6669	0.6770
	ECOC – L_3	0.6922	0.7002	0.6693	0.6872
	OVO	0.6924	0.6751	0.6687	0.6787
	OVA	0.7156	0.6637	0.6485	0.6759
	DECOC	0.6926	0.7026	0.6665	0.6872
	ECOCONE	0.7024	0.6749	0.6613	0.6795
	Forest – ECOC	0.6833	0.6885	0.6672	0.6797

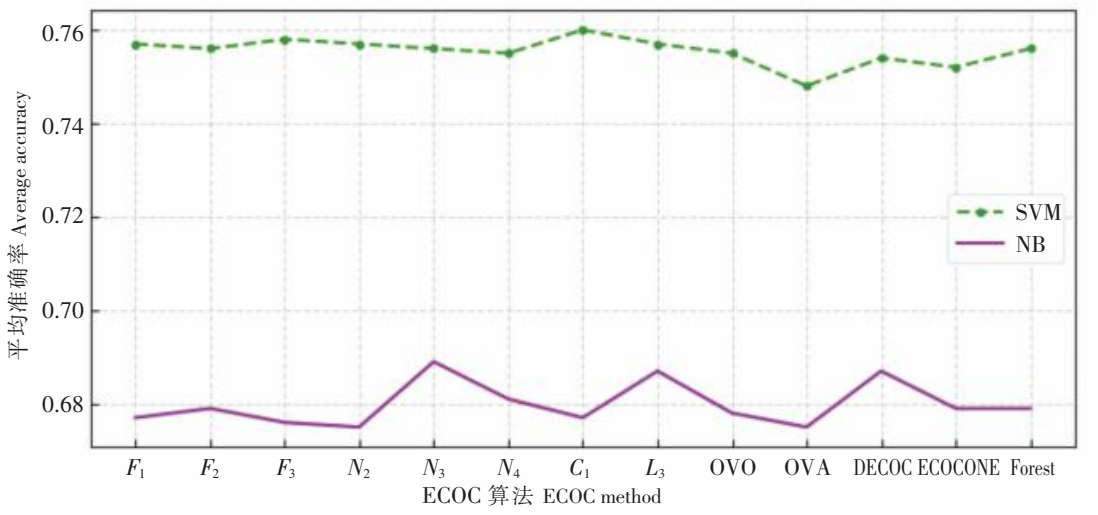

图 2 基于 SVM 和 NB 的 ECOC 算法的三所高校平均准确率对比

Fig.2 Average accuracy of different ECOC methods based on SVM and NB in three school data sets

3　结论

本文收集并整合三个高校学生的慕课学习行为数据，设计了包括多种基于数据复杂度的 ECOC 算法，并与传统 ECOC 算法在内的十多种基于 ECOC 的多分类算法进行比较。实验结果表明，相比于传统的 ECOC 算法，基于数据复杂度的 ECOC 多分类算法预测精度更高、更稳健，其对学生成绩进行四个等级的预测，平均准确率可达 75% 以上，为提前教学干预提供了参考。

未来可考虑用更多的数据复杂度算法来计算三所高校数据的复杂度，衡量数据内部分布情况。此外，算法中基分类器的多样性影响着 ECOC 算法的性能，因此，丰富基分类器的多样性也是值得深入研究的方向。

[参 考 文 献]

[1] SEATON D T, BERGNER Y, CHUANG I, et al. Who does what in a massive open online course? [J]. Communications of the ACM, 2014, 57(4): 58-65.

[2] KAY J, REIMANN P, DIEBOLD E, et al. MOOCs: so many learners, so much potential. [J]. IEEE Intelligent Systems, 2013, 28(3): 70-77.

[3] GARDNER J, BROOKS C. Student success prediction in MOOCs [J]. User Modeling and User-Adapted Interaction, 2018, 28(2): 127-203.

[4] 蒋卓轩, 张岩, 李晓明. 基于 MOOC 数据的学习行为分析与预测 [J]. 计算机研究与发展, 2015, 52(3): 614-628.

[5] QIU J, TANG J, LIU T X, et al. Modeling and predicting learning behavior in MOOCs [C] //Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco: ACM, 2016: 2835842, 93-102.

[6] XU B, YANG D. Motivation classification and grade prediction for MOOCs learners [J]. Computation Intelligence and Neuroscience, 2016: 4. DOI:10.1155/2016/2174613.

[7] ZHANG M, ZHU J, WANG Z, et al. Providing personalized learning guidance in MOOCs by multi-source data analysis [J]. World Wide Web, 2019, 22(3): 1189-1219.

[8] YU C, WU J, LIU A, et al. Predicting learning outcomes with MOOC clickstreams [J]. Education Sciences, 2019, 9(2): 104. DOI:10.3390/educsci9020104.

[9] MEIER Y, XU J, ATAN O, et al. Predicting grades [J]. IEEE Transactions on Signal Processing, 2016, 64(4): 959-972.

[10] XU J, MOON K H, Van Der SCHAAR M. A machine learning approach for tracking and predicting student performance in degree programs [J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(5): 742-753.

[11] ULLOACAZAREZ R L, LOPEZMARTIN C, ABRAN A, et al. Prediction of online students performance by means of genetic programming [J]. Applied Artificial Intelligence, 2018, 32(9/10): 858-881.

[12] HUSSAIN M, ZHU W, ZHANG W, et al. Using machine learning to predict student difficulties from learning session data [J]. Artificial Intelligence Review, 2019, 52(1): 381-407.

[13] SUN M, LIU K, WU Q, et al. A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis [J]. Pattern Recognition, 2019, 90: 346-362.

[14] CANO J R. Analysis of data complexity measures for classification [J]. Expert Systems with Applications, 2013, 40(12): 4820-4831.

[15] DIETTERICH T G, BAKIRI G. Solving multiclass learning problems via ECOC [J]. Journal of Artificial Intelligence Research, 1994, 2(1): 263-286.

[16] ZHOU L, WANG Q, FUJITA H. One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies [J]. Information Fusion, 2017, 36: 80-89.

(责任编辑　朱雪莲　英文审校　黄振坤)