基于RSBoost算法的不平衡数据分类方法

李克文 杨 磊 刘文英 刘 璐 刘洪太
(中国石油大学(华东)计算机与通信工程学院 青岛 266580)

摘 要 不平衡数据的分类问题在多个应用领域中普遍存在,已成为数据挖掘和机器学习领域的研究热点。提出了一种新的不平衡数据分类方法RSBoost,以解决传统分类方法对于少数类识别率不高和分类效率低的问题。该方法采用SMOTE方法对少数类进行过采样处理,然后对整个数据集进行随机欠采样处理,以改善整个数据集的不平衡性,再将其与Boosting算法相结合来对数据进行分类。通过实验对比了5种方法在多个公共数据集上的分类效果和分类效率,结果表明该方法具有较高的分类识别率和分类效率。

关键词 不平衡数据,组合数据采样,Boosting,RSBoost
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.048

Classification Method of Imbalanced Data Based on RSBoost

LI Ke-wen YANG Lei LIU Wen-ying LIU Lu LIU Hong-tai
(College of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract The problem of class imbalance which is very common to many application domains becomes the research hotspot in data mining and machine learning. We presented a new classification method of imbalance data, called RSBoost, to increase the recognition rate of minority class and the classification efficiency. This approach uses SMOTE (synthetic minority over-sampling technique) and random under-sampling to balance the data sets, and then uses boosting method to optimize the classification performance. We conducted experiments using several public data sets to evaluate the performances of RSBoost and other four methods. The experimental results show that the approach proposed in this article can improve the classification performance and efficiency of imbalance data sets.

Keywords Imbalanced data, Mixed data sampling, Boosting, RSBoost

1 引言

不平衡数据的分类问题在多个应用领域普遍存在,已经成为数据挖掘和机器学习领域的研究热点。当其中一类数据的数量远多于其他类的数据时,传统的数据挖掘算法倾向于把样本划分到占统治地位的类(多数类)中。对于少数类而言,这样的算法往往是低效率甚至无用的[1,2],但实际上少数类样本才是真正的关注点,在实际应用中将少数类样本错分的代价也很大。因此提高少数类的分类精度,使我们可以有效地鉴别那些重要的但数量上较少的样本是十分迫切和关键的。

对原始数据集重采样、提出或改进优化分类算法是目前处理不平衡数据分类问题常见的方法。

对原数据集进行数据重采样是通过减少多数类的数量(欠采样)或增加少数类的数量(过采样)来平衡数据集的数据分布。欠采样和过采样技术都有优缺点。欠采样技术最明显的缺点是在删除多数类的过程中会造成原数据集中的部分数据信息丢失,其优点就是会明显减少训练数据的数据量,从而减少分类模型的训练时间。而过采样技术恰恰相反,原数据集中的所有数据都将保留到新的数据集中,因此数据集中的数据信息没有丢失,但其缺点就是新的数据集中数据量较大,分类模型的训练时间比较大。

除了随机欠采样和随机过采样,很多学者提出了改进的数据重抽样方法。Nitesh V. Chawla[3]等提出一种SMOTE数据过采样方法。该方法从每个少数类样本的 k 个(一般取5个)最近邻中随机选出一个近邻,在该样本和其被选的近邻之间随机线性地插入新的合成样本。

J. Laurikkala[4]提出了邻域清除法(Neighborhood Cleaning Rule,NCR),此方法是一种利用最近邻思想去除多数类样本的欠采样方法。其基本思想是提取数据集样本 x_i 并比较样本 x_i 与其3个近邻样本中多数样本的类别是否相同,若不相同,则 x_i 是多数类,在数据集中去除 x_i;若相同,则 x_i 是少数类,在数据集中去除3个近邻样本中的多数类样本。

目前还没有统一的结论来说明欠采样技术和过采样技术谁更具有优势,C. Drummond等人[5]提出欠采样技术建立的分类器优于过采样技术建立的分类器,Chris Seiffert[6]从模型训练复杂度和训练时间方面考虑,提出了相似的观点,但G. E. Batista等人[7]却认为过采样技术要优于欠采样技术,尤

到稿日期:2014-05-19 返修日期:2014-07-27 本文受山东省自然科学基金(ZR2013FL034)资助。
李克文(1969—),男,博士,副教授,硕士生导师,主要研究方向为软件工程、计算智能、大数据等,E-mail:likw@upc.edu.cn;杨 磊(1981—),男,硕士生,主要研究方向为软件质量与可靠性;刘文英(1968—),女,硕士,高级工程师,主要研究方向为嵌入式系统、软件工程;刘 璐(1990—),男,硕士生,主要研究方向为可信软件;刘洪太(1988—),男,硕士生,主要研究方向为软件质量与可靠性。

其是在数据集中存在类别重叠的问题时。

在提出或改进优化分类算法方面，研究人员也做了大量的相关工作。Chawla N. V 等人[8]提出了基于代价敏感的分类学习方法，代价敏感算法主要关注错分样例的代价，与不平衡数据学习之间存在很强的联系，可以用来解决不平衡数据的学习问题。Wang Chao-Xue 等人[9]提出一种新的基于权重的改进 KNN 算法用于对不平衡数据集进行分类。Joshi MV 等人[10]提出了改进的 Boosting 算法，此算法通过对多数类和少数类赋予不同的权值来提高少数类的分类精度。

将采样技术和集成学习算法相结合的方法也是一种解决不平衡数据分类问题的有效方法。目前，越来越多的不平衡数据分类方法将两种类型的方法进行结合。结合方法是通过数据采样技术对原始不平衡数据进行处理，以降低数据的不平衡性，然后采用集成学习的分类算法进行分类。Nitesh V. Chawla 等人[11]提出的 SMOTEBoost 算法就是一种 SMOTE 技术与 Boosting 技术相结合的方法。Li Xiong-Fei 等人[12]提出一种融合数据采样和 boosting 技术相结合的不平衡数据分类算法 PcBoost。但是，现有算法往往不能将两种方法的优点有效地结合在一起，比如：在通过数据采样技术对不平衡数据集进行处理之后，数据规模比较大，导致分类效率较低。除此之外，集成算法的选择往往影响着算法的分类精度。

针对现有算法存在着分类效率低或者分类精度不够高的问题，本文提出一种组合数据采样技术和 Boosting 技术相结合的 Resampling SMOTE Boost(RSBoost)算法来处理不平衡数据的分类问题。该方法是一种组合采样技术和 Boosting 相结合的方法，该方法首先使用 SMOTE 方法对少数类进行过采样处理以增加少数类的数量，生成新的数据集；其次对新的数据集在保持样品分布的情况下进行欠采样处理，从而生成训练数据集；再次使用 Boosting 算法对采样后的训练数据集进行分类学习，生成对应的训练模型；最后通过实验对比该方法与其它 4 种分类方法在多个公共数据集上的分类效果和分类效率。

2 基于 RSBoost 算法的不平衡数据分类方法

针对现有算法存在着分类效率低或者分类精度不够的问题，本文提出一种组合数据采样技术和 Boosting 技术相结合的 RSBoost 算法来处理不平衡数据的分类问题。

SMOTE 算法是一种比较好的过采样方法，可以有效增加少数类的数量从而平衡数据分布，但可能会因训练数据量过大而增加训练时间，从而降低模型的训练效率。因此，本文提出的组合数据采样技术是在增加少数类的数量后在保持数据分布的情况下对整数数据集进行随机欠采样，减少了训练数据集的数量，从而提高了算法效率。

Boosting 算法是一种集成分类算法，其在每次迭代过程中会增加错分样本的权重，从而使训练器在下一步训练中更多地关注错分样本。将多个弱分类器的组合提升为强分类器，从而提高对少数类和数据集整体的识别率。在不平衡数据集分类中，少数类更容易被错分，因此 Boosting 技术可以提高少数类的分类精度[13]。

RSBoost 算法在数据处理过程中，采用 SMOTE 算法增加少数类的数量后在保持数据分布的情况下对整数数据集进行随机欠采样，再与 Adaboost 算法相结合对数据进行分类。

与现有的不平衡数据分类方法相比，RSBoost 算法由于在采用 SMOTE 算法增加少数类之后，又对整个数据集进行随机欠采样，减少了数据集的规模，从而减少了模型训练时间，因此具有更高的分类效率；同时该算法与集成学习方法 Boosting 技术相结合，在每次迭代过程中使用经过处理的数据来训练弱分类器，并根据样本分类结果给样本赋予新的权值，经过多次迭代产生多个弱分类器，通过弱分类器权重投票得出最终输出结果。因此，该方法在提高分类效率的同时还能够增加少数类的分类精度。

算法具体流程如下：

给定训练集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，样本 $x_i\in X^d$ 是 d 维特征向量，类标记 $y_i\in\{P,N\}$，其中，P 对应少数类(正类，Positive)，N 对应多数类(负类，Negative)。

输入：训练集 S，基分类器 WL，过采样率 M，欠采样率 N

输出：分类模型 H(x)

Step1 初始化数据集中样本的权重：

$$D_1(i)=1/m \tag{1}$$

Step2 根据过采样率 M 对少数类进行 SMOTE 过采样处理后，在保持数据分布的情况下以欠采样率 N 对整个数据集进行随机欠采样处理，生成训练数据集 S_t'，其权重分布为 D_t'。

Step3 for t=1 to T

(1)根据训练数据集 S_t' 及其权重分布 D_t' 训练弱分类器 WL，并计算弱假设 $h_t:X\times Y\to[0,1]$

(2)计算 h_t 的伪损失

$$\varepsilon_t=\sum_{(i,y):y_i\neq y}D_t(i)(1-h_t(x_i,y_i)+h_t(x_i,y)) \tag{2}$$

(3)计算权重更新参数：

$$\beta_t=\frac{\varepsilon_t}{1-\varepsilon_t} \tag{3}$$

$$\omega_t=(1/2)\cdot(1+h_t(x_i,y_i)-h_t(x_i,y)) \tag{4}$$

(4)更新权重分布 D_t

$$D_{t+1}(i)=D_t(i)\beta_t^{\omega_t} \tag{5}$$

(5)归一化处理：

$$D_{t+1}(i)=\frac{D_{t+1}(i)}{\sum_i D_{t+1}(i)} \tag{6}$$

Step4 通过 T 个弱假设权重投票得到最终分类模型：

$$H(x)=\arg\max_{y\in Y}\sum_{t=1}^{T}h_t(x,y)\log\frac{1}{\beta_t} \tag{7}$$

本算法使用 SMOTE 过采样技术增加少数类数据数量来调节不平衡数据集的平衡度，从而平衡数据分布；对整个数据集在保持分布的情况下进行欠采样处理来减少用于训练的数据，降低数据集的规模，从而能够减少模型训练时间，提高算法的分类效率。同时，使用 AdaBoost 算法来提高分类器的分类精度。

3 实验与分析

3.1 不平衡数据分类的评价指标

在传统的分类学习中，由于其数据分布是相对平衡的，因此一般采用分类精度(即整体样本中正确分类的样本个数占总样本个数的百分比)来衡量一个分类器的性能。但是对于不平衡数据集而言，分类精度却因其更倾向于预测多数类而无法准确衡量分类器的性能[14]。例如，在一个二类不平衡问题中，假设数据集中多数类的样本数量占总数量的 99%，而少数类仅占总数量的 1%，那么分类器即使把所有样本都预

测为多数类，一样可到达99%的分类精度。但这样的分类器却没有实际应用的价值，因为在实际应用中我们更关注少数类。

不平衡数据集分类问题的评价指标也是机器学习和数据挖掘中的一个重要方面，目前常用的评价标准主要有：*F-value*、*G-mean*、*ROC* 曲线[15]、*AUC*[16]（*ROC* 曲线下面积）等。本文将采用 *F-value*、*G-mean*、*AUC* 来评价分类器的性能。同时，我们将时间 T 作为衡量分类器性能的参考，分类精度略微提高但时间代价极大的分类器显然也不是我们追求的目标。

以上评价标准是建立在二分问题混淆矩阵（见表1）的基础上。通常，在不平衡数据集中正类（Positive）和负类（Negative）分别代表少数类和多数类。其中，TP（True Positive）和TN（True Negative）分别表示正确分类的正类和负类样本的个数；FP（False Positive）表示实际为负类但被分为正类样本的个数；FN（False Negative）表示实际为正类但被分为负类样本的个数。

表1　二分问题混淆矩阵

实际类别 \ 分类结果	被分为正类	被分为负类
实际为正类	TP	FN
实际为负类	FP	TN

查准率：$Precision=\frac{TP}{TP+FP}$ (8)

查准率反应的是所有被分为正类的样本中实际为正类样品所占的比值。

查全率：$Recall=\frac{TP}{TP+FN}$ (9)

查全率反应的是所有实际上为正类的样本中被分为正类样品所占的比值。

$$F\text{-}value=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Recall+Precision} \quad (10)$$

其中，β 表示 *Recall* 与 *Precision* 的相对重要性，通常取 $\beta=1$。

Recall、*Precision* 和 *F-value* 都是针对正类（少数类）的评价标准，一般情况下，使用 *F-value* 作为不平衡数据集分类问题的评价准则。

$$G\text{-}mean=\sqrt{PositiveAccruacy\times NegativeAccuracy} \quad (11)$$

其中，

$$PositiveAccuracy=Recall=\frac{TP}{TP+FN} \quad (12)$$

$$NegativeAccuracy=\frac{TN}{TN+FP} \quad (13)$$

G-mean 是以少数类的分类正确率和多数类的分类正确率为基础的，通常作为衡量不平衡数据集整体分类性能的评价指标，较高的 *G-mean* 值说明分类器对于多数类和少数类样本都有不错的分类性能。

AUC：*ROC* 曲线能描述分类器在不同判别阈值时的分类性能，其横坐标为伪正确率：$FP_{rate}=\frac{FP}{FP+TN}$，纵坐标为真正确率：$TP_{rate}=Recall=\frac{TP}{TP+FN}$。在实际应用中，一般用 *ROC* 曲线与坐标轴围成的区域面积 *AUC* 值代替 *ROC* 曲线，来评价分类器的性能，*AUC* 的值越大，对应模型的预测性能越好。

3.2　实验结果与分析

为评价 RSBoost 算法对不平衡数据集分类问题的有效性，本文选择6个少数类和多数类样本比例不平衡的公共数据集进行实验，数据集的基本信息如表2所列。

表2　数据集信息

数据集名称	样例个数	属性个数	少数类个数	少数类比例(%)
CM1-n	344	38	42	12.2
JM1-n	9593	22	1759	18.3
hepatitis	155	20	32	20.6
credit-german	1000	21	300	30.0
breast-cancer	286	10	85	29.7
tic-tac-toe	958	10	332	34.7

实验在 weka 平台上使用部分公共数据集对比 C4.5、SMOTE、AdaBoost、SMOTE+Boost 和本文提出的 RSBoost 等方法的分类性能。其中，C4.5 决策树算法直接对不平衡数据集进行分类，SMOTE 算法中邻域 k 值设置为5，数据采样后少数类比例如表3所列，RSBoost 算法中欠采样后的样本数量为原总样本数的50%，基分类器算法使用 C4.5 算法，利用 weka 平台中 J48 分类器实现。

表3　数据采样后少数类比例

数据集名称	CM1-n	JM1-n	hepatitis	credit-g	breast-c	tic-tac-toe
少数类比例(%)	41.0	40.3	39.4	39.1	45.8	48.8

为增加实验数据的客观性，所有实验采用十折交叉验证得到结果。即将数据分为10份，9份用于训练，1份用于测试，最后用10次测试结果的平均值作为一次十折交叉验证的结果。

表4—表6分别列出5种方法在6个数据集上的 *F-value* 值、*G-mean* 值以及 *AUC* 值。

表4　5种方法在6个数据集上的 *F-value* 值对比

数据集	C4.5	SMOTE	AdaBoost	SMOTE+Boost	RSBoost
CM1-n	0.306	0.789	0.309	0.854	0.895
JM1-n	0.309	0.769	0.358	0.792	0.849
hepatitis	0.529	0.841	0.607	0.830	0.892
credit-german	0.442	0.655	0.482	0.657	0.755
breast-cancer	0.397	0.679	0.446	0.712	0.773
tic-tac-toe	0.775	0.854	0.938	0.972	0.955

表5　5种方法在6个数据集上的 *G-mean* 值对比

数据集	C4.5	SMOTE	AdaBoost	SMOTE+Boost	RSBoost
CM1-n	0.495	0.820	0.496	0.876	0.910
JM1-n	0.476	0.802	0.527	0.829	0.871
hepatitis	0.643	0.867	0.708	0.859	0.910
credit-german	0.572	0.715	0.612	0.716	0.788
breast-cancer	0.510	0.713	0.580	0.728	0.809
tic-tac-toe	0.825	0.855	0.949	0.972	0.954

表6　5种方法在6个数据集上的 *AUC* 值对比

数据集	C4.5	SMOTE	AdaBoost	SMOTE+Boost	RSBoost
CM1-n	0.594	0.841	0.741	0.949	0.960
JM1-n	0.649	0.851	0.684	0.892	0.926
hepatitis	0.708	0.895	0.812	0.949	0.961
credit-german	0.602	0.723	0.724	0.798	0.868
breast-cancer	0.584	0.739	0.631	0.787	0.869
tic-tac-toe	0.896	0.924	0.992	0.996	0.993

图1－图3示出了5个方法在6个数据集上的 *F-value* 值、*G-mean* 值以及 *AUC* 值的比较。在图1－图3中，我们使用 DS1、DS2、DS3、DS4、DS5、DS6 分别表示 CM1-*n*、JM1-*n*、hepatitis、credit-german、breast-cancer、tic-tac-toe 数据集。由图1－图3可知，C4.5算法对于不平衡度较高的数据集的结果较差，而 Adaboost 对本文中的所有数据集（不管是否是不平衡数据集）都能改进其分类性能，但对于不平衡数据集而言，虽能改进，但分类性能仍较差。通过改善数据集的不平衡性，发现数据集的少数类和数据集整体的分类性能都得到提高。本文提出的基于 RSBoost 算法的不平衡数据分类算法在6个数据集上都取得了较高的 *F-value* 值、*G-mean* 值和 *AUC* 值，只是在个别数据集中略差于 SMOTE＋Boost 方法。而 RSBoost 方法中训练数据集的训练数据个数仅为 SMOTE＋Boost 方法中训练数据个数的50%左右，因此 RSBoost 方法的分类效率要高于 SMOTE＋Boost 方法。以 JM1-n 数据集为例，其数据个数为9593，AdaBoost、RSBoost 和 SMOTE＋Boost 方法的训练建模时间分别为16.3s、5.69s、18.05s。显然对于数据量较大的数据集而言，RSBoost 方法具有的分类效率方面的优势同样值得我们关注。

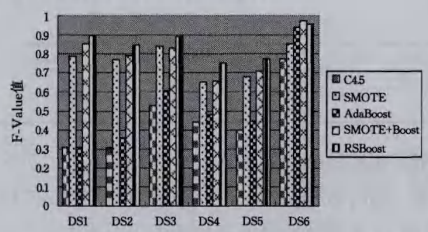

图1　5种方法在6个数据集上的 *F-value* 值对比柱形图

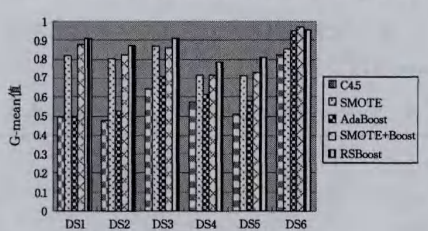

图2　5种方法在6个数据集上的 *G-mean* 值对比柱形图

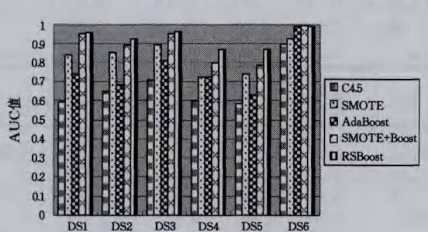

图3　5种方法在6个数据集上的 *AUC* 值对比柱形图

综上所述，本文提出的基于 RSBoost 算法的不平衡数据分类方法能有效地处理数据集的不平衡问题，对于少数类和数据集整体都具有较高的识别率。RSBoost 算法增加了少数类的样本的数量，因此，有效地平衡了数据集的不平衡性，从而提高了少数类的分类精度。同时，RSBoost 算法中总的训练数量不但少于 SMOTE 算法而且也少于原数集的样本数量。因而，本算法的分类效率也比较高。对于数据量较大的情况，本算法的分类效率将具有更大的优势。

结束语　在实际应用中很多领域都存在不平衡数据集，传统的分类算法往往倾向于处理多数类样本，而对少数类样本的识别率相对较低。SMOTE 算法虽然可以增加不平衡数据集中少数类的数量来平衡数据集，但是对于数据量较大的数据集，分类效率较低。本文提出的基于 RSBoost 算法的不平衡数据集的分类方法，增加少数类样本数量来平衡不平衡数据集同时可以减少训练样本总数量，并结合 Boosting 技术进一步提高分类器的性能。实验结果表明，此方法对于少数类和数据集整体都有较高的识别率，同时本方法的分类效率也有所提高。

参 考 文 献

[1] Batista G E A P A, Prati R C, Monard M C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29

[2] Gao Jia-wei, Liang Ji-ye. Research and Advancement of Classification Method of Imbalanced Data Sets [J]. Computer Science, 2008, 35(4): 10-13

[3] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-SamplingTechnique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[4] Laurikkala J. Improving Identification of Difficult Small Classes by Balancing Class Distribution[C] // Proceedings of the 8th Conference on AI in Medicine Europe: Artificial. 2001: 63-66

[5] Drummond C, Holte R C. C4.5, Class Imbalance and Cost Sensitivity: Why Under-Sampling beats Over-Sampling[C] // Proceedings of the ICML'03 Workshop on Learning from. 2003

[6] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance [J]. IEEE Transactions on System, MAN, and Cybernetics-PART A: Systems and Humans, 2010, 40(1): 185-197

[7] Batista G E, Prati R C, Monard M C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29

[8] Chawla N V, Cieslak D A, Hall L O, et al. Automatically Countering Imbalance and Its Empirical Relationship to Cost[J]. Data Mining and Knowledge Discovery, 2008, 17(2): 225-252

[9] Wang C X, Pan Z M, Ma C S, et al. Classification for Imbalanced Dataset of Improved Weighted KNN Algorithm[J]. Computer Engineering, 2012, 38(20): 160-163

[10] Joshi M V, Kumar V, Agarwal R. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements [C] // Proc of the 1st IEEE International Conference on Data Mining. San Jose, USA, 2001: 257-264

[11] Chawla N V, Lazarevic A, Hall L O, et al. Smoteboost: Improving Prediction of the Minority Class in Boosting[C] // Proc. of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Dubrovnik, Croatia, 2003: 107-119

[12] Li X F, Li J, Dong Y F, et al. A new learning algorithm for imbalanced data-PCBoost[J]. Chinese Journal of Computers, 2012, 35(2): 202-209

[13] Hothorn T, Buehlmann P, Kneib T, et al. mboost: Model-based boosting 2.0[J]. Journal of Machine Learning Research, 2010 (11): 2109-2113

（下转第267页）

达;最大等车时间为最大发车间隔 12min。

表 2 优化调度前后的相关指标

指标	优化前	优化后		
	全程车	全程车	区间车	大站快车
单程时长(min)	76	76	66.4	52
发车次数	136	84	29	20
发车间隔(min)	高峰期 5～9 平峰期 10～15	按客流需求产生,4/8/12		
乘客等待时间(min)	高峰期 5～9 平峰期 10～15	0.4/0.8/1.2/1.6/…/3.6/4/…/12		
总时间成本(元)	129786	119229		

与图 5 对应的组合公交调度模式产生的总费用为 119229 元,一天的总发车次数为 133。原先单一的全程车调度模式产生的时间成本费用为 129786 元,一天的总发车次数是 136。与单一调度相比,组合调度能有效减少乘客出行时间和系统成本,降低车辆拥挤程度,提高服务水平。

结束语 本文以公共交通服务为宗旨,研究了发车间隔不定的组合公交调度方法。首先通过站点客流数据建立了乘客时间成本模型,同时从运行时间角度出发,建立了公交公司的运营成本模型。充分考虑了不同模式车辆超车的问题,对到达站点的车辆次序进行重排。利用差分进化算法求解模型,对比调度前后的单线路运行时长、车辆运行次数、发车间隔以及总时间成本指标,由此说明发车间隔不定的公交组合调度模型的优越性。但公交实际出行情况复杂,乘客的出行规律有待进一步研究。

参 考 文 献

[1] Ceder A, Israeli Y. User and operator perspectives in transit network design[J]. Transportation Research Record: Journal of the Transportation Research Board, 1998, 1623(1): 3-7

[2] Ceder A, Hassold S, Dano B. Approaching even-load and even-headway transit timetables using different bus sizes[J]. Public Transport, 2013, 5(3): 193-217

[3] 李章维,郭冰冰,明洁,等. 基于居民出行行为分析的公交线路调度研究[J]. 计算机科学,2014,41(6A):94-97
Li Zhang-wei, Guo Bing-bing, Ming Jie, et al. Bus Line Scheduling Research Based on Residents' Travel Behavior Analysis [J]. Computer Science, 2014, 41(6A): 94-97

[4] 石琴,覃运梅,黄志鹏. 公交区域调度的最大同步换乘模型[J]. 中国公路学报,2007,20(6):90-94
Shi Qin, Qin Yun-mei, Huang Zhi-peng. Maximal synchronous transfer model of bus regional dispatching[J]. China Journal of Highway and Transport, 2007, 20(6): 90-94

[5] Guihaire V, Hao J K. Transit network design and scheduling: A global review[J]. Transportation Research Part A: Policy and Practice, 2008, 42(10): 1251-1273

[6] Schéele S. A supply model for public transit services[J]. Transportation Research Part B: Methodological, 1980, 14(1): 133-146

[7] Furth P G, Day F B. Transit Routing and Scheduling Strategies for Heavy-Demand Corridors (Abridgment) [M] // Advances in Bus Service Planning Practices. 1985: 23-26

[8] Ceder A. Designing Transit Short-Turn Trips with the Elimination of Imbalanced Loads[M] // Computer-Aided Transit Scheduling. Springer Berlin Heidelberg, 1988: 288-303

[9] Vijayaraghavan T A S, Anantharamaiah K M. Fleet assignment strategies in urban transportation using express and partial services[J]. Transportation Research Part A: Policy and Practice, 1995, 29(2): 157-171

[10] Eberlein X J, Wilson N H M, Bernstein D. Modeling real-time control strategies in public transit operations[M] // Computer-aided Transit Scheduling. Springer Berlin Heidelberg, 1999: 325-346

[11] Sun C, Zhou W, Wang Y. Scheduling combination and headway optimization of bus rapid transit[J]. Journal of Transportation Systems Engineering and Information Technology, 2008, 8(5): 61-67

[12] Codina E, Marín A, López F. A model for setting services on auxiliary bus lines under congestion[J]. Top, 2013, 21(1): 48-83

[13] 林培群,徐建闽. BRT 车站组的线路停靠组合优化模型[J]. 中国公路学报,2011,24(3):93-98
Lin Pei-qun, Xu Jian-min. Combinatorial optimization model of bus stop in BRT station-group[J]. China Journal of Highway and Transport, 2011, 24(3): 93-98

[14] Sun Chuan-jiao, Zhou Wei, Wang Yuan-qing. Scheduling Combination and Headway Optimization of Bus Rapid Transit [J]. Journal of Transportation Systems Engineering and Information Technology, 2008, 5(8): 61-67

[15] 裴玉龙,申翔浩,周侃. 高铁乘客换乘常规公交平均等候时间模型[J]. 交通运输工程学报,2013,13(6):76-82
Pei Yu-long, Shen Xiang-hao, Zhou Kan. Average waiting time model for passengers transfering from high-speed railway to bus [J]. Journal of Traffic and Transportation Engineering, 2013, 13(6): 76-82

[16] Omar J, Yasmin A. Synchronization of bus timetabling [J]. Transportation Research Part B: Methodological, 2012, 46(5): 599-614

[17] Storn R, Price K. Differential evolution－a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of Global Optimization, 1997, 11(4): 341-359

[18] Das S, Suganthan P N. Differential evolution: A survey of the state-of-the-art[J]. IEEE Transactions on Evolutionary Computation, 2011, 15(1): 4-31

(上接第 252 页)

[14] Ganganwar V. An overview of classification algorithms for imbalanced datasets[J]. International Journal of Emerging Technology and Advanced Engineering, 2012, 2(4): 42-47

[15] Gao S. An ensemble classifier learning approach to ROC optimization; Pattern Recognition[C] // 18th International Conference on ICPR. 2006: 679-682

[16] Hand D J, TillR J. A simple generalization of the area under the ROC curve for multiple[J]. Machine Learning, 2001, 45(2): 172-186