抽样技术和CBES分类非平衡数据集

职为梅 郭华平 范 明
(郑州大学信息工程学院 郑州 450052)

摘 要 CBES是面向非平衡数据集分类的组合选择方法。相关的实验表明,CBES方法能大幅度提升基分类器的泛化能力。已有研究表明,抽样方法能有效提高分类器在非平衡数据集分类上的性能。因此,巧妙地将抽样技术应用到CBES方法中,进而提出基于抽样的CBES方法(SCBES),以期进一步提高CBES在稀有类上的性能。大量的实验表明,巧妙地使用抽样方法能进一步提高CBES方法在非平衡数据集分类上的性能。

关键词 非平衡数据集,组合分类器,组合选择,抽样技术
中图法分类号 TP181 **文献标识码** A

Sampling Techniques with CBES for Imbalanced Learning

ZHI Wei-mei GUO Hua-ping FAN Ming
(College of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract CBES is a method which can be used for classification of imbalanced datasets. Related experimental results show CBES can boost the generalization ability of the base classifier. Reported researches show sampling method can effectively improve the performance of rare data. In the paper, we skillfully used sampling methods into CBES, and then proposed a method, named sampling-based CBES (SCBES) to further improve the classification performance of rare data. The experimental results demonstrate SCBES can effectively improve the performance of classification for imbalanced datasets.

Keywords Imbalanced data sets, Ensemble, Ensemble selection, Sampling method

1 引言

非平衡数据集分类是一个既有挑战又有显著意义的问题,指的是在一个数据集中,一个类(多数类)的实例数目远远超过另一个类(少数类或稀有类)的实例数目,使得数据集中类比率失衡[1]。非平衡数据集分类有很多的应用,如欺诈性电话检测[2]、发现不可靠的通信客户[3]、文本分类[4]、金融交易[5]等等。在这些应用中,人们真正感兴趣的是稀有类实例,期望分类模型能更准确预测稀有类样本。然而传统的分类方法,如C4.5、Naviebayes、神经网络等是基于训练数据集构建模型拟合数据分布,这就使得模型很难对稀有类数据有效。因此,它们在非平衡数据集上效果很差[1]。

为了提高分类方法在稀有类数据上的分类性能,在文献[6]中我们将组合选择思想用于非平衡数据集分类,提出了基于实例的组合选择方法CBES(Case-Based Ensemble Selection),以进一步提高稀有类数据上的分类性能。实验结果表明,该方法可以进一步提高在稀有类数据上的分类性能[6]。

抽样方法也是解决非平衡数据集分类的一种有效方法[7]。本文中,我们巧妙地将抽样技术和基于实例的组合选择方法结合起来,设计并实现了SCBES(Sampling for CBES)方法。实验结果表明,巧妙地将抽样技术和组合选择思想结合起来应用于非平衡数据集分类能有效提升组合分类器在稀有类问题中的泛化性能。

本文第2节介绍基于实例的组合选择方法(CBES);第3节是介绍SCBES的基本思想;第4节是参数设置;第5节给出实验设计和相应的实验结果;最后是总结。

2 CBES方法

组合分类方法(集成分类方法)可以提高分类的性能[6],如,Bagging[8]、Boosting[9]、Random Forest[10]、Rotation Forest[11]、AdaBoost[12]等。在考察组合分类器的分类性能时,我们发现了一个有趣的现象,即大部分基分类器都具有很好的局部性,或者说每个基分类器都在某些局部区域具有更好的分类性能。产生这种现象的原因可能是自助抽样产生的随机扰动。这一现象表明:对于每个局部,总可以从基分类器库中找到一组更好的基分类器。基于这种观察,考虑稀有类数据的局部特征,我们将组合选择方法应用于非平衡数据集分类,设计并实现了基于实例的组合选择方法CBES(Case-Based Ensemble Selection)[6]。基于实例的组合选择方法的思想是:在训练阶段,CBES使用某种算法(实验时分别使用C4.5和Naviebayes作为分类算法)在数据集D上学习包含M个基分类器的分类器库H。训练分类器库的一个核心是保持每个

到稿日期:2013-05-21 返修日期:2013-07-16 本文受国家自然科学基金(60773048)资助。
职为梅(1977—),女,博士生,讲师,CCF会员,主要研究方向为数据挖掘,E-mail:iewmzhi@zzu.edu.cn;郭华平(1981—),男,博士,主要研究方向为数据挖掘、人工智能;范 明(1948—),男,教授,博士生导师,主要研究方向为数据挖掘、模式识别、人工智能。

分类器的差异性。论文使用 Bagging 为每个分类器构建有差异的分类器。在预测阶段，给定待分类样本 x_i，CBES 搜索 x_i 的 k 近邻用于构建选择集合 D_s，使用 D_s 从 H 选择若干个分类器构成子组合分类器 S，使用 S 预测 x_i 的类标号。在 20 个 UCI 数据集上的实验结果表明，将组合分类器选择思想合理地应用于非平衡数据集分类能有效提升组合分类器在稀有类问题中的泛化性能[6]。

3 基于实例的组合选择方法(SCBES)

抽样方法是解决非平衡数据分类问题的主要方法之一，也是非常有效的一种方法[12]。本文巧妙地将抽样方法和 CBES 方法结合起来，具体做法如下：学习得到组合分类器后，运用抽样技术对训练数据集 D 抽样后得到 D'，对于每一个待分类样本 x，在 D' 搜索 x 的 k 个近邻。抽样技术有很多，本文使用了常被用于非平衡数据集分类中的 4 种抽样技术：随机欠抽样[12](randomly undersampling the prevalent class)、随机过抽样[12](randomly oversampling the small class)、SMOTE[13](Synthetic Minority Over-sampling Technique)和基于 SMOTE 的边界抽样[14](BorderLineSMOTE sampling)。随机抽样具有代表性，是分类方法中常用的抽样方法，SMOTE 和 BorderLineSMOTE 方法是基于非平衡数据集分类而被提出来的抽样方法，常被用来提高稀有类分类的性能[13,14,16]。

3.1 随机过抽样和随机欠抽样

随机过抽样(以下简写为 OverSample)技术的思想是：抽样时只抽稀有类样本，改变抽样的比例使得训练集平衡。在二元分类中，假设训练集 D 由 D_{maj} 和 $D_{\min}$ 构成，其中 D_{maj} 是多数类实例集，$D_{\min}$ 是少数类(稀有类)实例集。随机过抽样技术从 $D_{\min}$ 中随机复制实例集 S，并将 S 添加到训练集 D 中。调整 $|S|$ 的大小使得 D 中类分布平衡。随机欠抽样(以下简写为 UnderSample)技术和随机过抽样相反，它的思想是：随机抽样时只抽多数类样本，也是通过抽样的比例使得训练集平衡。随机欠抽样技术从 D_{maj} 中随机选择实例集 S，并将 S 从训练集 D 中删除。通用调整 $|S|$ 的大小使得 D 的类分布平衡。随机抽样技术易于实现。

3.2 SMOTE

SMOTE 方法已被证明是一种非常有效的抽样方法[13]。SMOTE 的基本思想是考察稀有类实例的特征空间，向训练集中增加人工数据。具体方法如下：给定一个稀有类实例 $x \in D_{\min}$，对于特定的 k 值，计算 x 的 k 近邻，随机选择 k 近邻中一个按照式(1)产生的人工数据。

$$x_{new} = x + (x_i - x) \times \delta \tag{1}$$

式中，$x \in D_{\min}$，x_i 是 x 的一个近邻，$\delta \in [0,1]$ 是一个随机数，根据式(1) x_{new} 被增加到数据集中。SMOTE 方法通过一种特定的方式平衡了原始数据集，进而提高了非平衡数据集的分类性能。

3.3 BorderLineSMOTE

BorderLineSMOTE[14]是一种对 SMOTE 改进的抽样技术，与 SMOTE 相同，通过计算稀有类实例 x 的 k 近邻增加人工数据。与 SMOTE 不同，它只对边界上的 x 过采样。具体做法是：计算得到 x 的 k 近邻，并标记 k 近邻属于每个类的个数，如果 x 的近邻超过一半属于多数类，则 x 属于边界实例。确定边界实例 x 后，采用与 SMOTE 相同的方法对 x 过采样。

算法 1 SCBES 方法

训练阶段

输入：D 为训练数据集；M 为组合分类器库大小；sr 为采样比例

输出：组合分类器库 H

方法：

1. H=Bagging(D,M)
2. D′=Sample(D,sr)
3. return (H, D′)

预测阶段：

输入：H 为训练阶段学习的分类器库；D′为抽样后的样本集；x 为待分类样本；k 为待分类样本 x 的近邻数；disFunc 为样本间距离函数；pec 为子组合分类器 S 和原始组合分类器 H 大小的比率

输出：样本 x 的预测类标号

方法：

1. D_s=disFunc(D′,x,k) //在 D′中搜索 x 的前 k 个近邻
2. S=Φ
3. while |S|≤ pec * |H| do
4. 搜索能最大化给定指标(如 EBM[15])的 h∈H
5. S=S∪{h}
6. end while
7. y=S(x)
8. return y

3.4 SCBES 方法

本文提出的算法如算法 1 所示。在训练阶段，使用 Bagging 在训练集 D 上学习包含 M 个基分类器的分类器库 H，并返回抽样后的数据集 D'。在预测阶段，给定待分类样本 x，行 1 在 D' 上搜索 x 的 k 个近邻用于构建选择集合 D_s；行 3 到行 6 使用 D_s 从 H 选择 $pec * |H|$ 个分类器作为子组合分类器 S，这里我们使用能量指标 EBM[15]监督选择过程；行 7 使用 S 预测 x 的类标号，进而返回 x 的类标号(行 8)。

算法 1 中计算实例的距离(预测阶段行 1)使用了最常用的距离度量方法欧几里得距离。算法 1 中的另一个问题是组合选择的方法。本文采用贪心的组合选择方法选择最优或次优子组合分类器，具体过程参考文献[6]。算法 1 中有关参数取值的设置，将在第 4 节中详细给出。

论文选择上述抽样方法来提高非平衡数据分类。实验表明，巧妙地使用抽样技术后，基于实例的组合选择方法的分类性能有了明显提高，相关的实验结果在 5.3 节中给出。

4 参数设置

算法 1 中存在 3 个输入参数：k(近邻数)、pec(子组合分类器 S 和原始组合分类器 H 大小的比率)和采样比例 sr。k 和 pec 的值采用文献[6]中给出的值，分别为 7 和 0.125。

本节设计一组实验用于学习采样比例 sr 的大小。在此使用 2 个数据集(balloons 和 glass)作为代表数据集，关于数据集的具体描述见 5.1 节。论文使用 Bagging 建立大小为 200 的分类器库，使用 20 次交叉验证计算平均 f-measure 值，选择 C4.5 作为基分类器。相关结果见图 1—图 4。

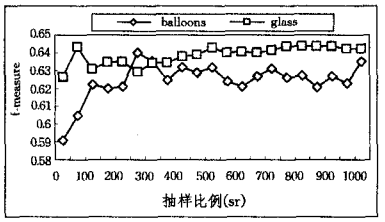

图 1　OverSample 下抽样比例 *sr* 对 CBES 性能 f-measure 的影响

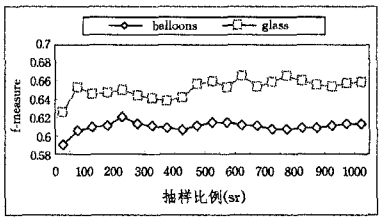

图 2　SMOTE 技术下抽样比例 *sr* 对 CBES 性能 f-measure 的影响

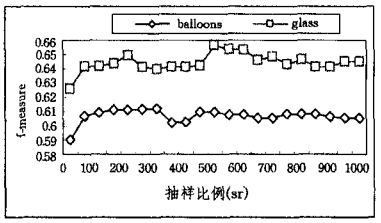

图 3　BorderLineSMOTE 下抽样比例 *sr* 对 CBES 性能 f-measure 的影响

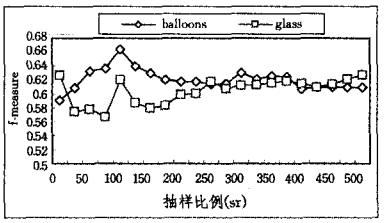

图 4　UnderSample 下抽样比例 *sr* 对 CBES 性能 f-measure 的影响

图 1—图 4 分别给出了基于 OverSample、SMOTE、BorderLineSMOTE 和 UnderSample 测试 *sr* 的结果。其中，OverSample、SMOTE 和 BorderLineSMOTE 中 *sr* 的值从 0% 递增到 1000%，表示抽样比例从 0(即不抽样)递增到 1000%。由图 1—图 3 可以发现当抽样比例选择 200%的时候，f-measure 的值趋于稳定，因此，实验时这 3 种抽样比例 *sr* 的值定为 200%。由于欠抽样的特殊性，在 UnderSample 中 *sr* 的值由 0%递增到 500%。由图 4 可以发现，当抽样比例为 100%时，f-measure 值趋于稳定，因此，UnderSample 的抽样比例 *sr* 定为 100%。从图 1—图 4 中可以看出，抽样后的 CBES 方法有助于提升算法在稀有类数据集上的泛化性能。

5　实验

5.1　实验数据

20 个数据集从 UCI 机器学习库中随机选取[17]。对于每个数据集，采用 5×2 折交叉验证分析算法的性能。为了保持数据集的不平衡性(少数类实例数/多数类实例数)稳定低于 0.3，我们保留最大实例数的类和最小实例数的类，并且采用随机欠抽样的方法删除某一个类的一些实例。处理后的数据集的详细情况如表 1 所列。

表 1　实验数据集信息

数据集	倾斜率	数据集	倾斜率
auto-mpg	0.15	kr-vs-kp	0.16
balance-scale	0.17	labor	0.30
balloons	0.20	letter-image	0.16
breast-cancer	0.29	pima	0.10
credit	0.11	promoters	0.17
german	0.10	sick	0.07
glass	0.22	sonar	0.09
hayes-roth	0.17	splice-junction	0.15
hepatitis	0.19	tic-tac-toe	0.18
ionosphere	0.17	vehicle	0.27

5.2　实验设置

为了评估 SCBES 的性能，我们对比了 OverSample-CBES、UnderSample-CBES、SMOTE-CBES、BorderLineSMOTE-CBES、CBES 和 Bagging(200)。算法 1 中的参数使用 $k=7$，$pec=0.125$，$sr=200$(OverSample、SMOTE 和 BorderLineSMOTE)或 $sr=100$(UnderSample)(见第 3 节)。实验使用 5×2 折交叉验证分析组合选择算法的性能。对于每次测试，使用 Bagging 建立包含 200 个分类器的分类器库，基分类器选择 C4.5。该部分同时给出了算法在 recall 和 f-measure 上的结果。

5.3　实验结果

表 2 展示了 OverSample-CBES、UnderSample-CBES、SMOTE-CBES、BorderLineSMOTE-CBES、CBES 和 Bagging(200)在所有数据上的 f-measure 值及标准差。

引入抽样技术以后，CBES 的性能有了进一步的提高，OverSample-CBES 相比于 CBES 在 8 个数据集上 f-measure 值有所提高。UnderSample-CBES、SMOTE、BorderLineSMOTE 相比于 CBES 分别在 11、13、11 个数据集上 f-measure 值有所提高。其中 UnderSample 的整体性能最好，把 CBES 的 f-measure 值平均提升了 3.08%，在 promoters 数据集上，UnderSample 将 CBES 的 f-measure 值提升了 27.5%。相比于 Bagging，OverSample-CBES、UnderSample-CBES、SMOTE、BorderLineSMOTE 分别在 18、18、18、18 个数据集上 f-measure 值有所提高。OverSample 把 Bagging 的 f-measure 值平均提升了 14.2%，而 UnderSample 把 Bagging 的 f-measure 值平均提升了 17.9%。这个结果也表明将抽样技术引入 CBES 后可以进一步提高 CBES 的泛化能力。由表 2 也可以发现 OverSample 的性能不是很好，这可能是因为数据集被处理后稀有类数据本身就很少，随机过抽样只是简单复制数据，没有提供更多有意义的数据，因此，效果不好。

图 5 展示了 OverSample-CBES、UnderSample-CBES、SMOTE-CBES、BorderLineSMOTE-CBES、CBES 和 Bagging(200)在所有数据上 recall 值的柱状图。图中，x 轴表示数据集在表 1 中的索引，y 轴表示算法在相应数据集上的 recall 值。由图 5 可以发现，OverSample-CBES 相比于 CBES 在 7 个数据集上 recall 值有所提高。UnderSample-CBES、SMOTE、BorderLineSMOTE 相比于 CBES 分别在 16、16、12 个数据集上 recall 值有所提高。其中 UnderSample 的整体性能最好，把 CBES 的 recall 值平均提升了 7.6%，在 glass 数据集上，UnderSample 将 CBES 的 recall 值提升了 17.2%。相比于 Bagging，OverSample-CBES、UnderSample-CBES、SMOTE、

BorderLineSMOTE 分别在 17、19、17、17 个数据集上 recall 值有所提高。OverSample 把 Bagging 的 f-measure 值平均提升了 14.1%，而 UnderSample 把 Bagging 的 f-measure 值平均提升了 23.7%。

表 2 各种抽样后的 CBES 和 CBES 以及 Bagging 在 f-measure 值上的比较

dataset	OverSample	UnderSample	SMOTE	BorderLineSMOTE	CBES	Bagging(200)
auto-mpg	0.5971(0.0949)	0.5932(0.0574)	0.6112(0.1148)	**0.6238(0.0717)**	0.5959(0.0919)	0.4212(0.1873) *
balance-scale	0.1727(0.0545)	**0.2813(0.0629)**	0.2452(0.0740)	0.2270(0.0928)	0.1799(0.0523)	0.0511(0.0469) *
balloons	0.6274(0.0781)	**0.6638(0.0430)**	0.6239(0.1053)	0.6097(0.0899)	0.5907(0.0782)	0.5388(0.1442)
breast-cancer	0.2457(0.0937)	0.2240(0.0649)	**0.2458(0.0784)**	0.2440(0.0946)	0.2248(0.0877)	0.0518(0.0598) *
credit	0.8193(0.0418)	0.8270(0.0378)	0.8133(0.0493)	0.8278(0.0484)	0.8175(0.0461)	**0.8336(0.0371)**
german	0.3569(0.0359)	0.3816(0.0374)	**0.4107(0.0294)**	0.3911(0.0244)	0.3551(0.0381)	0.3002(0.0513) *
glass	0.5939(0.2293)	0.6201(0.2082)	0.6018(0.2386)	0.5951(0.2195)	**0.6262(0.2240)**	0.5699(0.1910)
hayes-roth	0.7583(0.1045)	**0.8218(0.1700)**	0.8070(0.0897)	0.7505(0.1557)	0.7729(0.1492)	0.5788(0.1057) *
hepatitis	0.6770(0.0926)	0.6767(0.0704)	0.6888(0.0751)	**0.6995(0.0885)**	0.6832(0.0828)	0.6094(0.0955)
ionosphere	0.8548(0.0451)	0.8650(0.0455)	**0.8735(0.0488)**	0.8478(0.0493)	0.8562(0.0472)	0.8461(0.0435)
kr-vs-kp	0.9520(0.0141)	**0.9571(0.0109)**	0.9495(0.0167)	0.9495(0.0156)	0.9495(0.0156)	0.9271(0.0185) *
labor	**0.7628(0.1075)**	0.7182(0.1150)	0.6733(0.1013)	0.7504(0.0689)	0.7507(0.1045)	0.5949(0.1462) *
letter-image	0.9719(0.0100)	0.9656(0.0118)	**0.9734(0.0101)**	0.9731(0.0079)	0.9733(0.0101)	0.9617(0.0115) *
pima	0.1756(0.0837)	**0.2420(0.0638)**	0.2150(0.0666)	0.2076(0.0661)	0.1745(0.0887)	0.1250(0.0843) *
promoters	0.4177(0.1919)	**0.5360(0.1487)**	0.4323(0.2058)	0.4204(0.1966)	0.4204(0.1966)	0.2944(0.2211)
sick	0.8712(0.0270)	0.8572(0.0316)	0.8697(0.0282)	**0.8752(0.0279)**	0.8726(0.0275)	0.8483(0.0211) *
sonar	0.5406(0.0725)	**0.5784(0.0796)**	0.5546(0.1184)	0.5657(0.1005)	0.5532(0.0860)	0.4828(0.0424) *
splice-junction	0.9479(0.0100)	0.9469(0.0099)	0.9484(0.0105)	**0.9490(0.0115)**	0.9479(0.0100)	**0.9490(0.0115)**
tic-tac-toe	0.6223(0.0702)	0.6062(0.0564)	0.6067(0.0497)	0.6019(0.0568)	**0.6311(0.0607)**	0.3649(0.0926) *
vehicle	0.9440(0.0360)	**0.9584(0.0300)**	0.9455(0.0352)	0.9440(0.0360)	0.9455(0.0352)	0.9491(0.0334)
Avg	0.6455	0.6660	0.6545	0.6528	0.6461	0.5649

注：粗体字表示在相应的数据集上(行)，相应的算法(列)具有最高的 f-measure 值

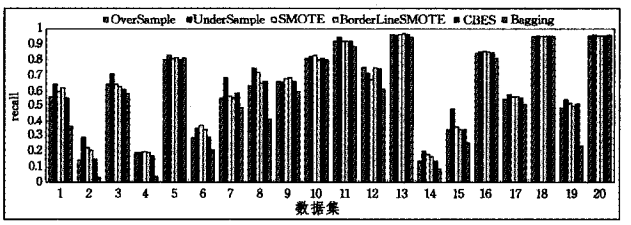

图 5 各种抽样后的 CBES 和 CBES 以及 Bagging 在 recall 值上的比较

由图 5 我们还可以发现，在大多数数据集上 SCBES 都取得了最高 recall 值。另外，在 recall 上，SCBES 表现出与表 2 类似的性质：(1)与 CBES 相比，SCBES 在大多数数据集上表现出更好或相当的泛化性能；(2)OverSample 的效果不太明显，而 UnderSample 的效果非常好。

表 3 给出了 SCBES 与 CBES 的对比结果，其中 win 表示 SCBES 比 CBES 的效果好；tie 表示 SCBES 和 CBES 的效果一样；loss 表示 SCBES 比 CBES 效果差。由表 3 可以发现，除 OverSample 外，其他 SCBES 的效果都好于 CBES，这说明抽样可以进一步提高 CBES 的泛化能力。

表 3 抽样后的 CBES 与 CBES 的对比

算法	win	tie	loss
OverSample-CBES	8	1	11
UnderSample-CBES	11	0	9
SMOTE-CBES	13	1	6
BorderLineSMOTE-CBES	11	2	7

以上结果表明，充分利用稀有类数据分布的局部特征能有效提高算法的分类性能。而抽样技术可以进一步提高组合选择方法在非平衡数据集分类上的性能。

结束语 事实上，本文的工作是基于实例的组合选择方法的后续工作。该方法将组合选择方法应用于非平衡数据集分类并考虑基分类器的局部性。本文进一步将抽样技术和组合选择方法巧妙地结合在一起，利用抽样技术改变数据的局部信息，从而找到对于待分实例性能最好的子组合分类器。相关实验结果表明，基于实例的组合选择方法和抽样技术相比于基于实例的组合选择能进一步提高分类器在稀有类数据集上的性能。这也充分说明了抽样技术对非平衡数据集分类有效。

参 考 文 献

[1] He Hai-bo, Garcia, Edwardo A. Learning from imbalanced Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[2] Fawcett T, Provost F. Combining Data Mining and Machine Learning for Effective User Profile[C]// Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. Portland, Oregon, USA, 1996: 8-13

[3] Ezawa K J, Singh M, Norton S W. Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management [C]// Proceedings of the International Conference on Machine Learning. Bari, Italy, 1996: 139-147

[4] Zheng Zhaohui, Wu Xiaoyun, Srihari Rohini. Feature Selection for Text Categorization on Imbalanced Data[J]. SIGKDD Explorations, 2004, 6(1): 80-89

[5] 黄浩，何钦铭，陈奇，等. 基于加权边界度的稀有类检测算法[J]. 软件学报，2012，23(5)：1195-1208

[6] 职为梅，郭华平，张银峰，等. 一种面向非平衡数据集分类问题的组合选择方法[J]. 小型微型计算机系统，2014，35

[7] 高嘉伟，梁吉业. 非平衡数据集分类研究问题进展[J]. 计算机科

学,2008,35(4):10-13

[8] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140

[9] Freund Y, Schapire R F. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1):119-139

[10] Breiman L. Random forests[J]. Machine learning, 2001, 45(1):5-32

[11] Rodriguez J J, Kuncheva L I, Alonso C J. Rotation forest: A new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10):1619-1630

[12] Sun Yan-min, Mobamed S K, Wong A K C. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12):3358-3378

[13] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-Sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16:321-357

[14] Han Hui, Wang Wen-yuan, Mao Bing-huan. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning [C]// Proceedings of International Conference on Intelligent Computing. Hefei, China, 2005:878-887

[15] Zhi Wei-mei, Guo Hua-ping, Fan Ming. Energy-Based Metric for Ensemble Selection[C]// Proceedings of 14th Asia-Pacific Web Conference. Kunming, China, 2012:306-317

[16] 曾志强,吴群,廖备水,等. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报, 2009, 37(11):2489-2495

[17] UCI repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/~mlearn/ MLRepository.html.

(上接第 40 页)

率,PLSA 分类方法低于 SVM,而兴奋、哀伤、放松 3 类的分类正确率高于 SVM。

结束语 本文将电影背景音乐分为兴奋、紧张、放松和哀伤 4 类。从近 60 部典型电影的原声音乐中截取出 703 个片段作为情感分类的数据集,创建了电影背景音乐情感数据集。提取了小节长度节奏模式特征、低音线模式特征及 MFCC 与音程特征。并运用 PLSA 方法对电影背景音乐进行情感分类。实验表明,PLSA 分类方法在电影背景音乐情感分类中,取得良好的分类效果,分类精度高于 SVM。

参 考 文 献

[1] Yang Y-H, Lin Yu-Ching, Cheng H-T, et al. Toward Multi-modal Music Emotion Classification[C]// Proceedings of the 9th Pacific Rim Conference on Multimedia. Berlin: Springer, 2008:70-79

[2] Lu Qi, Chen Xiao-ou, Yang D, et al. Boosting For Multi-Modal Music Emotion[C]// ACM 11th International Society for Music Information Retrieval Confe-rence. 2010:105-110

[3] Lin Yu-Ching, Yang Y-H, Chen H H, et al. Exploiting Genre for Music Emotion Classification[C]// IEEE International Conference on Multimedia & Expo, New York, 2009:618-621

[4] Lu Lie, Liu Dan, Zhang Hong-jiang. Automatic Mood Detection and Tracking of Music Audio Signals [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(1):5-18

[5] Liu Dan, Lu Lie, Zhang Hong-jiang. Automatic Music Mood Detection from Acoustic Music Data[C]// Proceeding of International Symposium on Music Information Retrieval. 2003:1-7

[6] Shi Yuan-yuan, Zhu Xuan, Kim H-G, et al. A tempo Feature via Modulation Spectrum Analysis and Its Application to Music Emotion Classification[C]// IEEE International Conference on Multimedia and Expo, Toronto Ont. 2006:1085-1088

[7] Yang Y-H, Liu Chia-Chu, Chen H H. Music Emotion Classification: A Fuzzy Approach[C]// ACM International Conference on Multimedia. New York, 2006:81-84

[8] Yang Y-H, Lin Yu-ching, Su Ya-fan, et al. Music Emotion Classification: A Regression Approach[C]// IEEE International Conference on Multimedia and Expo. 2007:208-211

[9] Schmidt E M, Trunbull D, Kim Y E. Feature Selection for Content-Based, Time-Varying Music Emotion Regression[C]// ACM Proceedings of the International Conference on Multimedia Information Retrieval. Mar. 2010:267-273

[10] 韩纪庆,张磊,郑铁然. 语音信号处理[M]. 北京:清华大学出版社, 2004:44-48

[11] Tzanetakis G, Essl G, Cook P. Audio Analysis Using the Discrete Wavelet Transform[C]// Proc. of World Student Environmental Summit. Sep. 2001:185-188

[12] Tsunoo E, One N, Sagayama S. Rhythm Map: Extraction of Unit Rhythmic Patterns and Analysis of Rhythmic Structure from Music Acoustic Signals[C]// IEEE International Conference on Audio, Speech and Signal Processing. March 2009:185-188

[13] Patterson R. Spiral Detection of Periodicity and the Spiral Form of Musical Scales[M]. Psychology of Music, 1986:44-61

[14] 韩纪庆,冯涛,郑贵滨,等. 音频信息处理技术[M]. 北京:清华大学出版社, 2007:41-46

[15] Zeng Zhi, Zhang Shu-wu, Li He-ping, et al. A Novel Approach to Musical Genre Classifition Using Probabilistic Latent Semantic Analysis Model[C]// IEEE International Conference on Multi-Media & Expo. 2009:486-489

[16] Robert L C, Jitendra V D, Bezdek J C. Efficient Implementation of the Fuzzy C-means Clustering Algorithms[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(2):248-255

[17] Thayer R E. The Biopsychology of Mood and Arousal [M]. 1989:10-15

[18] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines[M]. 2009:10-20