基于支持向量机的 MIDI 主旋律提取研究

崔建英[1] ,刘刚[2]
（1. 北京邮电大学自动化学院，北京 100876;
2. 北京邮电大学信息与通信工程学院，北京 100876）

摘要：哼唱检索是当前音乐信息检索研究和应用领域的一个热点，通常采用旋律信息作为音乐特征进行检索，MIDI 文件是构建音乐特征数据库的主要来源。如何从多音轨 MIDI 文件中提取出主旋律信息，是一个值得研究的问题。本文研究了一种基于支持向量机的 MIDI 主旋律提取方法，重点对主旋律提取中的特征提取、样本不平衡、结果可信度等问题进行了研究，实现 MIDI 文件主旋律的有效提取。实验证明，本文所述方法比已有算法具有更高的准确度。

关键词：哼唱检索；MIDI 主旋律；支持向量机；样本平衡；置信度
中图分类号：TP37

MIDI Melody Extraction Based-on Support Vector Machine

Cui Jianying[1], Liu Gang [2]
（1. Automation School, Beijing University of Posts and Telecommunications, Beijing 100876;
2. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: In this paper, a new method of MIDI melody extraction based on balanced data and Support Vector Machine (SVM) is introduced. The extracted melody could well construct the melody database for the hot investigation of Query-by-Humming (QBH) in Music Information Retrieval (MIR). First of all, extract the feature of each MIDI track as the data sample to be classified. Secondly, balance the data sample using a novel method called Random-SMOTE. Next, classify the balanced data to two categories (melody or accompaniment) and estimate the confidence of the results. Finally, obtain the extracted melody track according to the result of classification and confidence. Experimental results of our MIDI melody extraction show excellent accuracy outperforming the other existing approaches of MIDI melody extraction.

Keywords: query by humming; MIDI melody; Support Vector Machine; data sample balance; confidence

0 引言

在当前的音乐检索领域中，哼唱检索是一个研究热点。基于 MIDI 检索库的哼唱检索系统[1]（Query by Humming，QBH）是目前为止应用最为广泛的哼唱检索系统。MIDI 作为一种具有象征性音乐符号（symbolic music notation）的数字化文件，较容易从中提取出特征信息，进而能够用特定方法描述歌曲的主旋律。这种记录发音指令的文件格式，凭借其占用存储空间小、表达精确、通用性强等特点，成为目前音乐信息检索系统中十分重要的音乐格式，因此常被用来建立音乐检索的数据库。

当前的哼唱检索系统通常采用人工标注的方式得到 MIDI 主旋律。但随着音乐库的不断扩充，人工标注的工作量日益增加，很难满足现代海量的音乐信息检索需求。因此，能够自动提取 MIDI 主旋律对哼唱检索来说无疑是一个极其重要的贡献。

目前已有部分学者针对 MIDI 主旋律提取进行了研究。较早的 MIDI 主旋律提取算法是 Uitdenbogerd[2]和 Zobel 于 1999 年提出的一种基于简单规则的轮廓(Skyline)算法，提取出所有音轨中各时刻最高音的轮廓作为整首音乐的旋律曲线。因为大多数的音乐主旋律以最高音表现,所以 Skyline 算法提出的轮廓曲线能够在一定程度上描述主旋律，但是由于它将 MIDI

作者简介：崔建英，（1985-），女，硕士。E-mail: jianyingbupt@gmail.com
通信联系人：刘刚，（1973-），男，副教授，主要研究方向：语音识别。E-mail: liugang@bupt.edu.cn

中所有的音轨信息融合到一起，并且忽视了可能存在于较低声部的主旋律信息，对 MIDI 主旋律的描述并不准确。于是，Shan[3]于 2002 年提出了一种高音量音轨（High Volume Channel）算法。该算法通过一些措施删除掉乐器的通道，然后选取出具有最大音量的音轨作为主旋律。而 David（2006）[4]则采用随机森林作为分类器对音轨信息进行学习，并在此基础上建立了 MIDI 主旋律提取的系统。另外，Pedro J.详细描述了 MIDI 主旋律的概念，提出了一系列可以描述 MIDI 主旋律特征的统计描述符，并采用模糊遗传算法（Genetic Fuzzy Algorithm）实现主旋律的提取。其优点是具有较强的鲁棒性，准确率也较高，但是这种算法比较复杂，计算代价较大，需要寻找更为简单高效的方法。

鉴于上述的现有算法存在准确率低、算法复杂以及计算代价大等问题，本文将致力于寻找一种简单有效且准确率高的 MIDI 主旋律提取算法。在确定使用分类算法之后，首先需要考虑如何提取出能够有效区分主旋律和伴奏的音轨特征，以用来进行准确分类。由于 MIDI 本身的特点，还存在各类数据样本的比例相差悬殊的问题，需要考虑如何预先处理数据以免对分类产生不利影响。同时，在分类之后，还需要考虑分类结果的可信度。

论文结构如下: 第一节对 MIDI 主旋律提取做了系统描述。第二节讲述的是主旋律提取的理论和具体步骤。第三节讲述的是主旋律提取实验及结果。第四节是结论。

1 系统描述

标准的 MIDI 文件最多可含有 16 条音轨，其中第 10 条音轨为打击乐器，不包含主旋律信息，因此本文将不考虑第 10 条音轨。由于 MIDI 一般只含有一条主旋律音轨，而这条主旋律音轨基本包含了 MIDI 主旋律的全部信息，因此本文所研究的对 MIDI 主旋律的提取即为对 MIDI 主旋律音轨的提取。

提取流程图如下：

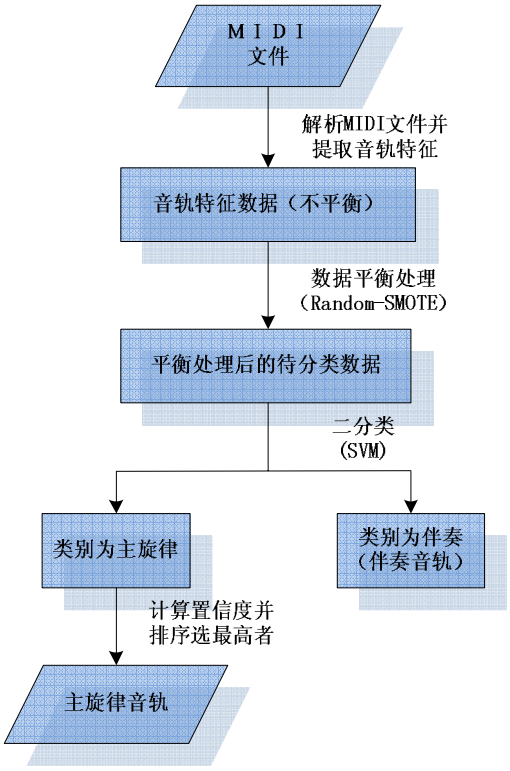

图 1　MIDI 主旋律提取流程

2 基于支持向量机的 MIDI 主旋律提取

2.1 音轨特征提取

MIDI 的构造方式[5]是以音乐元素的组织结构为基础的，它记录了音乐的全部乐谱信息和完整的演奏过程，音乐的多数基本特征都可以直接进行提取。图 2 所示的是 MIDI 文件的格式信息。

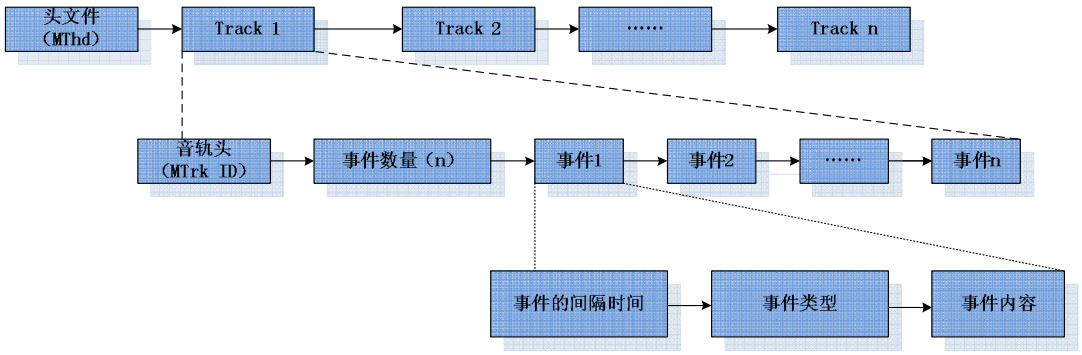

图 2 MIDI 文件的格式信息

本文提取的音轨特征包括音轨时长、音轨能量、音轨主音量、左右声道平衡、和弦在音轨中所占的比例、各音轨音程统计和十二平均律距离等。和弦指的是相同时刻有符合一定规律的两个或者两个以上音符同时发音，这种情况在伴奏音轨中出现较多。

其中, 音轨主音量、左右声道平衡都可以通过 MIDI 文件解析获得，而音轨时长、音轨能量、和弦在音轨中所占的比例、各音轨音程统计（反应音高波动情况）则可以从 MIDI 文件的事件信息中间接得到。但十二平均律距离，作为一个比较重要的音轨特征，有着相对复杂的提取过程。

假设 M 是一个由多条音轨（τ_i，1≤i≤16）组成的 MIDI，则可以表示为

$$M = \{\tau_1, \tau_2, \cdots, \tau_t\} \quad 1 \le i \le 16 \qquad \text{（式 2.1.1）}$$

假设 h_i 为音轨 τ_i 的音高直方图。因为根据音乐乐理中的十二平均律可知，音乐的一个八度音程是由 12 个半音组成的，所以 h_i 也可以用 12 个半音（音轨 τ_i 中的各个音高模 12 的等价值）来描述，由此可知 h_i 中的空间是 12 维的，具体表示如下：

$$h_i = \{h_{i1}, h_{i2}, \cdots, h_{i12}\} \qquad \text{（式2.1.2）}$$

因此，MIDI 中所有音轨的音高直方图集合 H 为

$$H = \{h_1, h_2, \cdots, h_i\} \quad 1 \le i \le 16 \qquad \text{（式2.1.3）}$$

假设 MIDI 中至少包含一个音符的音轨总数为 T，则第 i 维音高直方图的归一化定义如下：

$$\overline{h_i} = \frac{\sum_{k=1}^{16} h_{ki}}{T} \qquad \text{（式 2.1.4）}$$

于是，MIDI 中所有音轨的直方图归一化 h_A 为

$$h_A = \{\overline{h_1}, \overline{h_2}, \cdots, \overline{h_{12}}\} \qquad \text{（式 2.1.5）}$$

文献[6]中指出：MIDI 所有音轨的直方图归一化代表的是 MIDI 的平均旋律，它与 MIDI 的主旋律是相似的。而音高直方图 h_i 则反映了音轨 τ_i 的音高曲线特征。因此，音轨 τ_i 的音高曲线

与主旋律的距离就可以用 h_i 与直方图归一化 h_A 之间的距离来近似。

音高直方图 h_i 与直方图归一化 h_A 之间的距离定义如下：

$$d_i = D(h_i, h_A) \quad \text{（式2.1.6）}$$

其中，$D(h_i, h_a)$为计算 h_i 与 h_A 之间欧氏距离的函数。

假设 S 为音轨 τ_i 和主旋律之间的相似度。则距离 d_i 与相似度 S 成反比，即距离 d_i 越小，音轨 τ_i 和主旋律之间的相似度越大；反之亦然。由于距离 d_i 是根据音乐乐理中的十二平均律得到的，本文将其称为十二平均律距离。十二平均律距离能够体现各音轨与 MIDI 主旋律的相似程度，因此它是一个能够有效区分主旋律与伴奏的重要特征。

这样，通过一系列特征提取，就得到了原始的 MIDI 特征样本数据集

2.2 针对样本不平衡情况的预处理

MIDI 文件一般只含有一条主旋律音轨，音轨最多可有 16 条，这样就导致了在提取音轨特征作为分类的数据集时，数据集中的主旋律音轨数明显少于伴奏音轨数，二者比例相差可达十倍以上，这就构成了不平衡的分类数据集[7]。而其中数量上处于劣势的小类（即主旋律音轨这一类）的识别是分类的重点。小类样本的分布比较稀疏，且常被大类样本（伴奏音轨类）所包围，这样就为小类特征的学习带来极大的挑战。一般的分类算法对于不平衡数据集的分类效果并不好，往往将小类样本误分为大类。因此，为了提高分类效果，本文将对不平衡的特征样本数据集进行平衡处理。

对不平衡的特征样本数据集的平衡处理，本文采用的是一种基于 Random-SMOTE 的向上采样方法。SMOTE（Synthetic Minority Over-sampling Technique）是由 Chawla[8]等人提出的一种采样方法，目的是解决小类中样本数量过少的问题。但是它只在相邻近的小类样本之间线性插值，插值的结果是小类样本密集的地方依然相对密集，小类样本稀疏的地方依然相对稀疏，这就导致处于稀疏区域的样本不易识别、易被误分的问题。而 Random-SMOTE 方法在此基础上进一步优化，避免了此类问题的出现。这种方法主要是通过在小类的样本空间内随机生成新的小类样本，增加小类样本的数量。

Random-SMOTE 向上采样方法的基本思想是：对于每个小类样本 x，先从小类样本集中随机选择两个不同于 x 的样本 y_1、y_2，再以 x、y_1、y_2 为顶点构成一个三角形区域，然后根据向上采样倍率 N，在该三角形区域内随机生成 N 个新的小类样本。

如图 3 所示，生成新的小类样本的具体方法如下：

首先，在小类样本集中随机选择除 x 之外的两点 y_1、y_2，并在两点之间进行随机线性插值，生成 N 个临时样本 t_j（j=1,2,…,N）：

$$t_j = y_1 + rand(0,1) * (y_2 - y_1) \quad j = 1,2,\cdots N \quad \text{（式 2.2.1）}$$

其次，在样本点 t_j 与 x 之间进行随机线性插值，构造出 N 个新的小类样本 p_j（j=1,2,…,N）：

$$p_j = x + rand(0,1) * (t_j - x) \quad j = 1,2,\cdots N \quad \text{（式 2.2.2）}$$

其中，$rand(0,1)$ 表示区间(0,1)上的一个随机数。

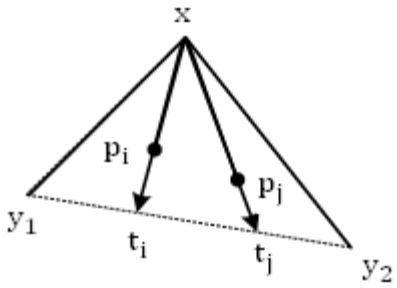

图 3 Random-SMOTE 中新样本的合成方法

这样，采用 Random-SMOTE 向上采样的方法，按照适当的向上采样倍率增加小类样本的数量, 就基本达到了两类分类样本的数据平衡。其中，增加倍率通过多次实验获得。

2.3 基于支持向量机的二分类

在对不平衡的特征样本数据集进行平衡处理之后，本文采用支持向量机作为分类器对处理好的数据进行二分类。

支持向量机（Support Vector Machine，SVM）是由 Vapnik 等[9]提出的一种基于统计学习理论的机器学习方法，是借助最优化理论解决分类问题的有力工具。它采用结构风险最小化原则（Structural Risk Minimization，SRM），综合考虑了经验风险和置信范围，从而具有良好的推广性能和较好的分类精确性。同时，它能够有效的解决过度拟合问题，在解决有限样本的学习问题时表现出优异的性能。因此本文选择支持向量机来进行二分类。

在分类过程中，本文采用的是台湾大学林智仁(Lin Chih-Jen)开发的 LibSVM 工具包，以径向基函数（Radial Basis Function, RBF）作为核函数，并通过交叉验证得到最佳 C 和 Gamma 参数。根据这些最佳参数，通过训练集样本的学习得到训练模型，然后再根据此模型对测试集数据进行分类。最终，测试集中的待分类样本经过二分类，得到两类结果并分别标记为主旋律音轨或者伴奏音轨。

2.4 分类结果的置信度估计

经过 SVM 的二分类，便可初步得到每个特征样本的所属类别：主旋律音轨或者伴奏音轨。为了在分类器的分类准确率尽可能高的同时，进一步估计出每一个测试样本分类结果的准确性，本文对分类结果作了置信度的估计，并最终根据置信度来确定每个特征样本的所属类别。

本文提出了一个自定义的置信度估计公式，用以评估待测样本分类结果的置信度。相关研究表明，待测样本 x_i 到 SVM 最优超平面的距离 $d(x_i)$和置信度之间呈正比[10]，即待测样本到分类面的距离越大，该样本的置信度相对越高；反之，待测样本到分类面的距离越小，该样本的置信度相对越低。因此，待测样本到最优超平面的距离 $d(x_i)$可以看作影响置信度的一个参量。除此之外，还有一个可能对置信度有影响的参量，即待测样本邻域内的训练样本与其属于同一个类别的概率 ρ_i。也就是说，若待测样本x_i为正样本，而与该待测样本距离最近的 j 个训练样本中，正样本的个数越多，表明x_i实际为正的可能性越大，正样本的个数越少，表明 x_i 实际为正的可能性越小。因此，待测样本 x_i 周围 j 个训练样本点与其同属一类的概率 ρ_i 的大小，可以相对反映出 SVM 对该待测样本的分类正确的可能性大小，即参数 ρ_i 与置信度之间成正比。

由此，推导出置信度估计的公式定义如下：

$$f(x_i) = exp(-\frac{1}{|d(x_i)| * \rho_i}) \quad \text{（式 2.4.1）}$$

式中，$d(x_i)$为待测样本 x_i 到 SVM 最优分类超平面的距离，ρ_i 为待测样本 x_i 周围 j 个训练样本点属于待测样本 SVM 分类结果这一类的概率，即

$$\rho_i = \frac{j_s}{j} \quad \text{（式 2.4.2）}$$

其中，j 为样本近邻点个数，j_s 为待测样本经过 SVM 分类得出的所属类别所包含的近邻点的个数。

在 SVM 测试阶段，用之前训练出的最优模型来预测各待分类样本，获取每个待分类样本 x_i 到 SVM 最优分类面的距离 $d(x_i)$，以及 SVM 对每个测试样本的分类预测结果。然后，在训练样本里找到 x_i 的 j 个近邻，计算出这 j 个近邻中和 x_i 的初始分类结果一样的概率 ρ_i。结合 $d(x_i)$和 ρ_i 这两个参数，根据置信度评估公式，即可得到该待测样本 x_i 的置信度 $f(x_i)$。最后，将属于同一首 MIDI 的分类结果的置信度排序，其中被分类标记为主旋律并且置信度最高的音轨就是所要提取的主旋律音轨。

3 实验结果

实验选取的数据库为 800 首中国流行音乐 MIDI 文件。随机选取其中 300 首作为训练集，其余 500 首作为测试集。

经过多次实验，获得平衡样本的最佳向上采样倍率为 5。实验证明，向上采样倍率不同，最终二分类的准确率也不同。同时这也充分证明了对待分类样本做平衡处理的重要性。实验对比见表 1。

表 1 根据不同向上采样倍率得到的分类的准确度

向上采样倍率	准确率
原始数据	0.391304
1	0.457665
2	0.665903
3	0.775743
4	0.880434
5	0.962264
6	0.873278

在二分类过程中，经过多次交叉验证得到训练模型的最佳 C 参数和 Gamma 参数分别为 2、8。经过二分类并对分类结果进行置信度估计之后，获得最终实验结果，参见表 2。表 2 中给出了提取出的主旋律音轨的准确率以及召回率。准确率 Prescision、召回率 Recall 计算如下：

$$\mathrm{Pr}esicion = \frac{TP}{TP + FP} \quad \text{（式 3.1）}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad \text{（式 3.2）}$$

其中，TP 是正确分类成主旋律的音轨数，FP 是将主旋律音轨错分成伴奏的音轨数，FN 是将伴奏错分成主旋律的音轨数。

表 2 不同主旋律提取算法的实验结果比较

算法	Recall	Precision
基于 SVM 的方法	0.816	0.962
Skyline	0.593	0.925
High Volume Channel	0.377	0.516
基于随机森林分类器的方法	0.585	0.930

实验结果表明，本文提出的基于支持向量机的 MIDI 主旋律提取算法，具有较高的准确率。而造成提取错误的主要原因，在于 MIDI 文件制作不规范、主旋律与伴奏极其相似造成误分类等。

4 结论

本文提出了一种基于支持向量机的 MIDI 主旋律提取算法，它在使用支持向量机进行二分类之前对数据比例相差较大的两类待分类样本做了平衡处理，并在二分类之后对分类结果做了自定义置信度的估计。通过这些处理，算法有效排除了待分类的两类数据间的不合理比例对支持向量机分类的影响，并提高了最终的主旋律提取结果的可信度。与目前现有的 MIDI 主旋律提取方法相比，本算法不仅简单易行，便于实现，而且进一步提高了主旋律提取的准确率。

本文仍然存在不足之处，例如在利用 Random-RMOTE 方法对待分类的 MIDI 主旋律特征数据（小类样本）进行向上采样预处理时，最佳向上采样倍率必须通过多次实验才能获得。因此，如何利用自适应方法自动获得最佳向上采样倍率将便是可以改进的地方。

[参考文献] (References)

[1] 郭敏，张卫强，刘加. 一种基于帧-音符方式的哼唱检索算法[J]. 清华大学学报(自然科学版)，2011，51（4）：561-565.
[2] A. Uitdenbogerd, J. Zobel. Manipulation of Music for Melody Matching[Z]. Bristol: ACM International Multimedia Conference, 1998.
[3] Isikhan C, Ozcan G. A survey of melody extraction techniques for music information retrieval[Z]. Thessaloniki: Conference on Interdisciplinary Musicology, 2008.
[4] David R., Pedro J., Antonio. Melody Track Identification in MIDI Files[Z]. Boston: American Association for Artificial Intelligence. 2006.
[5] 孙博文，张艳鹏，赵振国. 基于多音轨 MIDI 主旋律提取的音乐可视化表达[J]. 软件，2012，33（3）：64-66.
[6] Giyasettin OZCAN, Cihan ISIKHAN, Adil ALPKOCAK. Melodic Extraction on MIDI Music Files[Z]. CA: IEEE International Symposium on Multimedia. 2005.
[7] 叶志飞，文益民，吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报，2009，4（2）：148-156.
[8] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.. Smote: Synthetic minority over-sampling technique[J]. Artificial Intelligence Research, 2002, 16:321-357.
[9] Corinna Cortes, Vladimir Vapnik[J]. Support-vector networks. Machine Learning, 1995, 20:273-297.
[10] 赵行. SVM 分类器置信度的研究[D]. 北京：北京邮电大学，2010.