LVQ 聚类算法在爆炸物 THz 光谱识别中的应用

赵晶晶[1],葛庆平[1],张存林[2]
ZHAO Jing-jing[1],GE Qing-ping[1],ZHANG Cun-lin[2]

1.首都师范大学 信息工程学院,北京 100037
2.首都师范大学 物理系,北京 100037
1.College of Information Engineering,Capital Normal University,Beijing 100037,China
2.Department of Physics,Capital Normal University,Beijing 100037,China
E-mail:jingjing518_510@163.com

ZHAO Jing-jing,GE Qing-ping,ZHANG Cun-lin.Application of LVQ clustering algorithm to identification of explosive by THz spectroscopy.Computer Engineering and Applications,2009,45(18):239-241.

Abstract: The terahertz(THz) technologies is one research hotspot in the domain of detecting explosive.Because the result gotten by clustered with original frequency-domain pattern is not satisfactory.This paper introduces second derivative curve that transformed from frequency-domain THz spectrum.Based on LVQ clustering algorithm and the new characteristic curve,an automatic detection system is designed and finished by VC++6.0.Appling LVQ to identification of explosive by THz spectroscopy.Experiment to the four kinds of explosive:RDX,DNT,TNT and HMX,trained with original frequency-domain pattern,the correct rate is 96%.After computing effective feature from transformed data,input to the same network,the correct rate up to 100%.The result shows that the system based on LVQ can be very capable of similar character clustering and has higher rate of identification.
Key words: The terahertz(THz);neural network;Learning Vector Quantization(LVQ) clustering algorithm

摘 要:运用 THz 光谱特性进行爆炸物的识别,是现代检测技术研究的一个热点。由于直接对原始数据进行聚类的识别率并不理想,首先对实验样本的 THz 频域光谱数据曲线进行二阶导数变换,得到了更能表现数据变化趋势和峰值的特征曲线,然后基于该特征曲线利用 LVQ 神经网络聚类算法,设计并用 VC++6.0 实现了 THz 光谱自动分类识别系统。分别对 RDX、DNT、TNT、HMX 四种爆炸物进行识别对比实验,运用原始数据训练出的分类器,识别率为 96%,运用变换过后的特征数据训练出的 LVQ 分类器,识别率可以达到 100%。实验证明,所设计的基于 LVQ 的神经网络分类器具有强大相似特征聚类功能和较高的识别率。
关键词:THz 技术;神经网络;学习矢量化网络(LVQ)聚类算法
DOI:10.3778/j.issn.1002-8331.2009.18.072 **文章编号**:1002-8331(2009)18-0239-03 **文献标识码**:A **中图分类号**:TP391

1 引言

在过去 10 年间,THz 技术受到了越来越多的关注,由于其光谱的特性, 使其非常具有吸引力并成为无损检测的一种工具。THz 波技术应用于安全检测方面,特别是对炸药及其相关材料检测研究优势主要有以下 4 点[1]:(1)不同炸药材料在 THz 波段具有不同的特征吸收, 可以用 THz 技术来进行炸药的鉴别;(2)THz 波可以穿透非金属和非极性材料,可以利用 THz 波来探测隐藏在这些包装材料中的炸药;(3)THz 波的能量比较低,不会导致生物组织电离,可以对人体和生物材料等进行无损检测;(4)相对于微波等光谱波段,由于 THz 具有较短的波长,材料在 THz 波段具有较高的空间的分辨率。

分析炸药及其相关材料的吸收光谱数据并进行处理,为建立 THz 爆炸物光谱自动识别检测系统奠定了基础。该系统的基本工作流程是:采集爆炸物的THz 数据,经过处理得到 THz 吸收系数, 提取吸收谱线的特征点的二阶导数值进行归一化,利用处理后的数据设计分类器,将待测光谱数据经过数据处理后以数组的形式输入到分类器中,输出光谱识别结果,即炸药所属类别。分类器可以采用矢量量化方法,或者利用人工神经网络也是很有效的方法[2-5]。矢量量化方法的识别效率高,但是模板的学习建立过程过于依赖主观判断,检测过程对数据质量的要求高,鲁棒性较差。

学习矢量量化网络(Learning Vector Quantization,LVQ)是一种混和网络,由输入层、竞争层和线性输出层组成。其在形式上类似于矢量量化(VQ)方法,但它又不同于矢量量化方法。传统的矢量量化方法可以通过矢量量化得到比原模板样本少得多的参考模板,并能直接基于这些参考模板对待识别样本进行判决或确认,但这种方法不具有自适应性。为此,Kohonen 提出了学习矢量量化方法。它的基本思想是在给定初始参考模板的基础上,使用有类别属性的模板样本,通过监督式自适应学习的方法来校正这些参考模板,经过若干次迭代后,所形成的

基金项目:北京市教委项目(No.KM200710028018)。
作者简介:赵晶晶(1984-),女,硕士研究生,主要从事机器视觉和图像处理研究;葛庆平(1951-),男,副教授,主要从事机器视觉和图像处理研究。
收稿日期:2008-04-22 **修回日期**:2008-07-11

参考模板就基本反映了模板样本的统计分布。该方法兼有矢量量化方法和神经网络方法的优点，将其应用于 THz 爆炸物光谱数据的自动识别系统中，建立神经网络模型，并利用实验样本进行验证。

2 LVQ 神经网络工作原理

2.1 LVQ 算法

学习矢量量化是 Kohonen 在 1989 年提出[6-7]，基本思想是在给定初始权值的基础上，使用有类别属性的训练样本，通过监督式自适应学习的方法来校正这些权值。经过若干次训练后，LVQ 网络中神经元的权值点就基本反应了训练样本点的统计分布。LVQ 的网络拓扑结构如图 1 所示，包括一个输入层，一个 Kohonen 层和一个输出层，其中 Kohonen 层也称为竞争层。

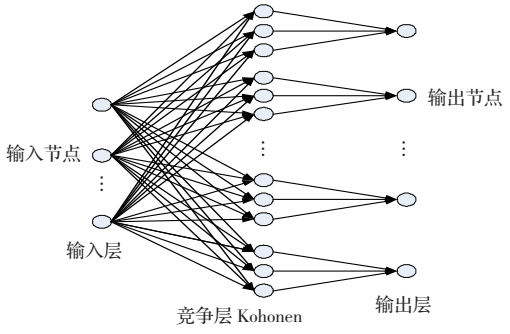

图 1 LVQ 神经网络的拓扑结构

LVQ 学习算法实质上是一种根据训练样本的特征，进行“奖-惩”的迭代学习算法。令 X 为样本集，W 为神经网络的神经元集，C 代表获胜神经元，L 代表输出类别集。LVQ 神经网络的基本实现过程是：首先确定各网络参数，选择学习样本，然后选择一些具有代表性的样本对权值进行初始化，这样可以大大地加快学习的过程。接着就可以寻求获胜神经元 C：

$$\|X-W_C\|=\|X-W_i\|,i=1,2,\cdots,M \tag{1}$$

用 Lw_C 代表与获胜神经元权值向量相关联的类，用 Lw_i 代表与输入向量相关联的类。

$$\begin{aligned} &W_C(n+l)=W_C(n)+\eta(n)[X-W_C(n)] \quad Lw_i=Lw_C \\ &W_C(n+1)=W_C(n)-\eta(n)[X-W_C(n)] \quad Lw_i\neq Lw_C \end{aligned} \tag{2}$$

其中，$\eta(n)$是学习速率。经过若干次训练后，所得到的权值矢量不再明显变化，说明网络已收敛。

2.2 LVQ 算法中学习速率的调整

LVQ 算法中学习速率 $\eta(t)$是个很重要的参数，它影响算法的稳定性和权值收敛的速度，是 LVQ 神经网络训练过程中需要重点考虑的参数。在定义学习速率的时候要贯彻快速稳定的原则，其通常有定学习速率、最优化学习速率和自适应学习速率三种定义方法[8]，其中最优化学习速率中所需的循环次数较少，但每次计算学习速率之前都要判断上一时刻的训练样本在该时刻神经网络中是否被正确分类。其定义如下：

$$\eta(t)=\frac{\eta(t-1)}{[(1+s(t)\eta(t-1)]} \tag{3}$$

若在 $t-1$ 时刻的训练样本被正确分类，在 t 时刻，$s(t)=+1$；

若在 $t-1$ 时刻的训练样本被错误分类，在 t 时刻，$s(t)=-1$。

对于具体问题还需要设定学习速率上限值，以防止学习过程中学习速率过大而导致学习过程不稳定，即当 $\eta(t)>\eta_0$ 时，$\eta(t)=\eta_0$。

3 THz 数据分析与处理

3.1 THz 吸收谱线

实验所用的数据由首都师范大学 THz 实验室提供，4 种炸药（RDX、HMX、DNT、TNT）的时域光谱经过傅里叶变换处理后得到 0.2~2 THz 的吸收系数，选取其中 57 个等频率间隔数据点进行归一化后拟合成谱线，通过谱线中吸收峰的位置和曲线趋势的一致性，研究人员已经可以对炸药种类进行人工识别。

3.2 THz 吸收谱线特征的提取

对于曲线特征的自动识别来说，必须通过输入数据的特征量来实现自动识别。特征量是对曲线自身性质的一种描述，好的模式特征应该具备以下几个条件[9]：（1）具有较好的类内一致性和类间区分度；（2）稳定性好，具有较好的抗噪能力；（3）具有较好的平移不变性、旋转不变性和尺度不变性等。

相对于原始数据点特征，提取数据曲线的二阶导数更能体现原始数据的变化趋势，并能对原数据吸收峰位置的数据进行加强，进而可以更有效地反映数据的信息。如图 2，提取原始 57 个数据点拟合曲线后中各点对应的二阶导数值，进行归一化后拟合得到的数据谱线，以第 2~56 个点的数值作为统计特征，组成一个 55 维向量。

4 LVQ 聚类算法的应用

本文的算法用 VC++6.0 实现，运用 LVQ 聚类算法对首都师范大学 THz 实验室所提供的人眼可识别的 299 组炸药数据进行自动识别的测试，其中包括 50 组 DNT 数据、50 组 TNT 数据、50 组 HMX 数据和 149 组 RDX 数据。

4.1 样本训练

4.1.1 基于最小学习误差增量的神经元自动生成算法

对于相同的实验环境，同类炸药的吸收谱线具有明显的相似性，但是，不同环境下，同类炸药的吸收谱线的细节可能具有多样性，用固定的 4 个输出类别制约了系统的兼容性和扩展性，针对这个问题，提出使用基于最小学习误差增量的神经元自动生成算法。

算法的实现过程如下：

（1）在训练前，将得到的数据进行数据特征提取，在 4 个类别中分别随机选取若干组特征向量作为训练样本，每组样本为 55 维的数据特征向量，将所有样本与权值（Kohonen 层特征向量）都进行归一化，使每一输入样本向量的模为 1，这样做的目的不但可以提高网络的准确性，而且可以大大加快整个训练的过程。

（2）预先设定主获胜神经元和次获胜神经元相应的学习误差上界 MAXERROR1 和 MAXERROR2。

（3）在训练的过程中，当产生与获胜神经元和次获胜神经元均不聚类（获胜样本和训练样本的向量的欧式距离超过误差上界）的样本时，系统新产生一个 Kohonen 层神经元组。

（4）如果 Kohonen 层产生了新的神经元，则同时更新权值数量和输出神经元数量，每一步都要对更新和产生的权值进行归一化处理。否则根据“奖-惩”迭代式对获胜神经元的权值进行更新。保证 Kohonen 层每一神经元组与一组输入特征结构对应，输出层根据现场数据的需求将不断变化。

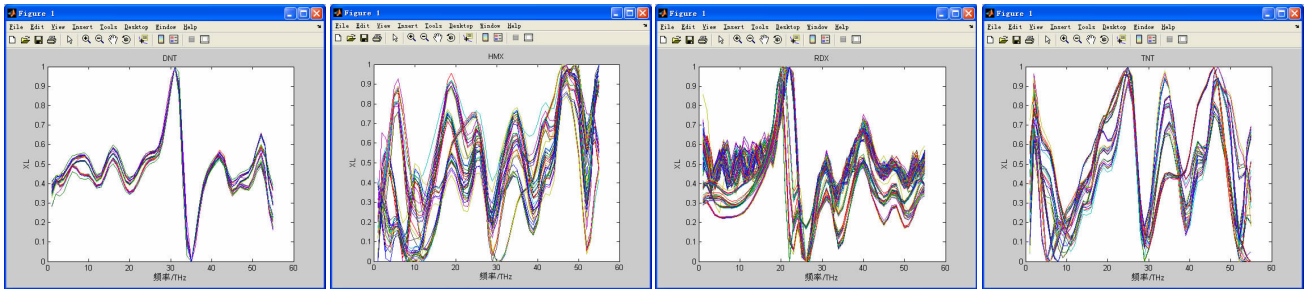

图 2　不同炸药的谱线特征提取

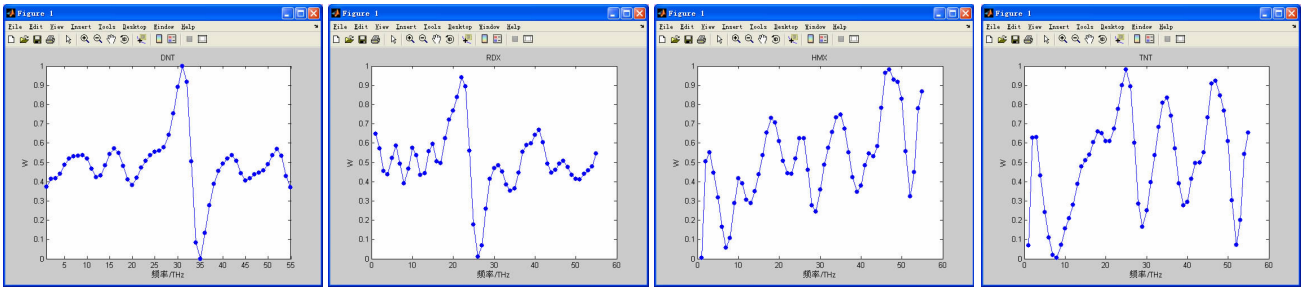

图 3　4 类炸药的分类特征权值谱线

4.1.2　样本训练参数的选择

神经网络的学习过程是离线进行的，识别系统只需要得到自动更新学习之后的权值数量和输出神经元的数量，综合比较了三类学习速率的定义方法，选择使用最优化学习速率，基于以下原因：(1)初始参数选择容易，学习过程简单；(2)较高的收敛稳定性；(3)较快的收敛速率。

具体的参数初始设置为：输入层神经元数为 47；竞争层神经元数为 4×47；初始输出层神经元数为 4，每个神经元对应一个单独的类别(RDX、DNT、HMX、TNT)。采用最优化学习速率：初始学习率为 $\eta(0)=0.02$；最大学习率为 $\eta_0=0.1$；相对学习率：$m=0.3$；窗口宽度：$\varepsilon=0.2$；迭代次数选择为 1 000，这些参数是在反复测试的基础上得出来的一组优化参数。如图 3 为训练后得到的 4 个类别特征分类权值曲线。

4.2　实验结果及分析

将待测的 299 组炸药待测样本数据分别进行二阶导数变换，提取出 55 维的数据特征向量，然后进行特征向量归一化，使每一输入样本向量的模为 1，最后将处理好的特征向量分别输入训练好的 LVQ 网络，利用通过训练得到的分类特征权值对这 299 组炸药待测样本进行识别分类。

表 1 显示了进行特征提取前和提取后的数据分别作为 LVQ 的输入向量所得到的识别率对比。

表 1　特征提取前后 LVQ 识别率对比

提取特征	前	后
输入特征向量维数	55	55
训练样本数目	47	47
识别率/(%)	96	100

结果显示，进行特征提取后进行识别的识别率从 96%提高到 100%。

表 2 显示了利用三种不同的方法进行炸药识别所得到的结果。

结果显示，运用矢量量化算法虽然可以自动识别，但选取模版的过程会消耗大量的模版空间，而且爆炸物检测依赖于有监督的学习过程，因此应用 LVQ 聚类算法作为机器识别的主要算法更加具有实用性和适应性，也更能满足用户需求。

表 2　LVQ 算法与以往模式识别方法的对比

主要识别方法	现有方法(人眼)	以往(矢量量化)	LVQ(本文)
自动识别	否	是	是
训练过程	人眼	无师训练	有师训练
样本识别率/(%)	100	80	100
客观	否	是	是

在众多需要安全检测的情况下，往往都需要对物品进行大批量地在线检测，此时人眼需要进行长时间的识别工作，很可能由于疲劳等主观因素造成检测失误，对有害物质的检测失误可能造成严重的后果，用机器代替人眼进行物品检测正是解决了人为因素所带来的诸多影响，可以很好地实现在线检测，具有客观性和可应用性。此外，对其它 THz 检测化学药品的光谱进行实验，发现识别率均在 90%以上，可见本文所设计的基于 LVQ 网络分类器在物质识别领域具有广阔的应用前景。

5　结论

本文对 THz 频域光谱的原始数据进行二阶导数变换，提取了更能反映不同物质曲线变化趋势的特征向量，并选取了 LVQ 神经网络聚类算法进行分类识别，相对于以往的模式识别方法，本文的算法更加适应于爆炸物的 THz 光谱识别[2-5]系统。基于本文算法运用 VC++6.0 实现了爆炸物 THz 光谱的自动识别分类系统，实现了机器代替人眼的识别过程。再者，已有报道表明，在 3THz 范围以内，有的炸药没有特征吸收峰[9-10]，实现无特征吸收峰炸药及其相关材料的检测，也是炸药检测研究中要面对的问题，本文所设计的基于数据二阶导数特征提取的 LVQ 网络分类器更加注重于数据的整体变化趋势，具有很好的应用前景。

参考文献：

[1] Federici J F，Schulkin B，Huang F，et al.Semicond[J].Sci Technol，2005，20：266.

(下转 244 页)