用SOM网络从移动通信数据中识别相似个体

陈蔼祥
(广东商学院数学与计算科学学院 广州 510320)

摘 要 从移动电话通信数据中挖掘用户有用信息,被认为是富有研究价值的一项工作。利用从94个Nokia 6600用户中收集的实际通信记录,根据用户的通信统计情况,结合用户的调查数据,形成能被SOM网络接受的训练数据集。在此基础上,给出了基于SOM(Self-Organizing Map)的相似个体识别系统SSRS(Similary Subject Recognition System),该系统在训练好的SOM网络支持下,能够对群体中的个体进行有效分类,从而为预测个体将来行为奠定基础。

关键词 现实数据挖掘,移动通信,SOM

Using SOM to Recognize Similary Individuals from Mobile Phone Data

CHEN Ai-xiang
(School of Mathematics, Guangdong University of Business Studies, Guangzhou 510320, China)

Abstract Data collected from mobile phones have the potential to provide insight into the community's individual classification. Using the reality data collected from 94 Nokia 6600 users, a training dataset fed to SOM neural network was generated, the training dataset consists of user's correspondence statistics and survey data. A Similary Subject Recognition System(SSRS) building on SOM neural network was designed and realized. The system SSRS can carry on the effective classification to community's individual and lay the foundation of forecasting the individual future behavior.

Keywords Reality mining, Mobile phone, SOM

1 引言

移动通信技术是推动人类文明进程的最伟大技术之一。随着移动通信技术的快速发展,移动通信成本早已经降到了普通大众能够接受的程度,从而使移动电话能够普及到普罗大众。这一趋势为基于移动通信平台基础上的各种应用提供了可能。其中麻省理工的社会网络分析项目以及相应的移动社会软件[1,2]即是这类应用的典型代表。

对社会人群中的个体进行分类,识别个体之间的各种社会网络关系,是社会网络分析的核心任务。如果能够借用现代科技手段,识别特定人群的各种类别及其社会网络关系,这对于有针对性地开展社会公共事物的管理和组织、刑事案件侦破、劳资谈判、大型活动的动态管理、有效交友活动等都具有重要意义。早期对社会网络分析方面的研究,更多是基于调查问卷的形式获取数据[3],这使得研究只能局限于短期内小范围群体,从而极大地限制了这方面的研究的广度和深度以及研究结果的准确性。最近几年,出现了从个体在从事社会活动过程中产生的外部观察数据(比如e-mail[4,5]、通话日志[6,7]行为模式的观察数据)中进行数据挖掘的通信模式的研究趋势。Nathan Eagle博士及其项目组将手机作为社会传感器,通过用户手机以及手机上的蓝牙功能,产生大量日常通信数据以及用户的日常行为数据,然后在这些数据中用主分量分析、聚类等技术手段,挖掘并识别这些隐含在日常行为数据中的用户行为模式以及各种不同用户群体[8,9,12]。他们的研究结果表明,用他们的结果可以达到76%的预测精度以及96%的分类精度。Nathan Eagle通过收集不同社区用户手机的通信信息,然后根据手机通信信息分析不同社区的人际关系网络以及行为模式,发现城市社区和农村社区具有截然不同的人际关系网络结构,并且个体在新环境下会改变其行为模式,以适应新环境[10]。

对特定社会群体进行有效的识别和分类,是社会网络分析中的一个核心任务。如果能对某一群体中的人群进行有效分类,识别各类不同行为模式的人群,将对社区管理、大型活动组织、刑事案件侦破等都具有重要意义。对特定社会群体进行分类,首先需要选择合适的个体特征数据,例如个人兴趣爱好、行为模式等,以构成个体特征模式。我们只要选择合适的特征并获取特定群体中每个个体相关数据,就可以利用相关工具实现该群体的分类、识别任务。SOM就是其中一种可供我们选择和使用的一种神经网络。SOM网络是指自组织特征映射网络是由荷兰学者Teuvo Kohonen于1981年提出的仿生计算方法。SOM网络由于采用自组织竞争的工作模式,不需要任何导师信号,就能实现模式识别和分类功能。该网络的优点主要在于它能够自动地寻找学习样本中的内在规律和本质属性,并自组织、自适应地改变网络参数与结构,这使得SOM网络无需预先设定分类类别总数,就能够自动将差别很小的数据点归为一类,并且差别不大的类激发的神经

到稿日期:2010-09-13 返修日期:2010-12-31 本文受广东省自然科学基金(10451032001006140),广州市科技和信息化局项目(10C12140131),广东高校优秀青年创新人才培育项目(LYM10081),广东商学院校级科研项目(20084410592XM58)资助。

陈蔼祥(1978—),男,博士,副教授,主要研究方向为智能规划、模型诊断、自动推理、数据挖掘,E-mail:cax413@163.com。

元位置也是相邻的，因此网络具有良好的分类识别功能。

本文将在 Nathan Eagle 博士提供的 realitymining 数据的基础上构造个体特征数据，将其作为 SOM 神经网络的训练数据，创建并训练 SOM 网络，然后运用训练好的 SOM 网络实现个体类别的识别。本文第 2 节对本文所使用的现实数据集做一简要介绍；第 3 节给出个体特征集合，定义个体特征模式，并给出从现实数据集中构造个体特征模式的算法实现；第 4 节设计并实现了基于 SOM 网络的相似个体识别系统 SSRS，该系统运用算法得到的个体特征模式数据集创建并训练 SOM 神经网络，然后运用训练好的 SOM 网络，对群体进行分类并识别个体所属类别；第 5 节分析 SSRS 系统的功能和识别结果；最后进行了相应总结。

2 使用的数据

本文使用的数据来自麻省理工学院的社会网络分析项目。在一群参与项目的志愿者用户手机上预先安装特定的监控软件，这些监控软件以后台运行的形式收集手机的相关信息，包括手机位置（通过移动通信基站）、邻近的手机（通过蓝牙每 5min 扫描以自动发现 5～10m 范围内的手机）、通信信息（包括语音通话）、开启的应用程序（例如日历、游戏等）、手机的充电状态、手机的开关机状态等。参与项目的志愿者一共有 94 位，他们中 68 位是工作在校园内同一栋楼的同事（其中 90%为研究生、10%为职员），另外 26 位为麻省理工商学院学生。94 位志愿者手机中约有 30 位志愿者的手机监控软件会在晚间某个时段内与我们的服务器连接，上传当天的通信记录数据到我们的服务器，其他则直接将数据记录在手机的内存卡上，这些内存卡容量为 32M，可以存储约 4 个月的数据。当研究人员需要这些数据的时候，再从内存卡读取相关数据。数据搜集的时间为 2004 年 9 月份到次年的 6 月份，历时 10 个月。

收集得到的数据文件为 realitymining，整个数据文件大小有 56.6M，是一个非常庞大的数据。realitymiining 中，除了前面描述的有关通信信息之外，还包含了对每个手机用户所作的类似调查问卷形式的数据，记录在 s(n). surveydata 中。这些问题包括类似(1) have you travelled recently? (2) do you own a car? (3) how many miles to you live form MIT? 等共 25 个诸如此类的问题。有关数据集更为详细的介绍，请参考 Nathan Eagle 博士写的 the Reality Mining Data README 文件。感谢 Nathan Eagle 博士及其项目组提供的这些数据[11]，没有这些数据的支撑，本文的工作无从谈起。

3 分类数据的选择

首先给出关于个体的定义，然后定义个体的特征数据以及个体特征模式。

定义 1（个体及个体集） realitymining 数据集中每一个 subject 记录为一个体，记为 s，S 表示所有个体集合。

为了识别个体所属类别，对特定群体进行分类，需要选取合适的个体特征数据，并构造个体特征模式。为了使选取的特征能充分反映个体的爱好、兴趣、行为模式等关键特征，我们从 realitymining 数据文件中抽取每个个体的问卷调查数据 surveydata，作为个体的特征数据。surveydata 是麻省理工学院 Nathan Eagle 博士在其 reality mining 项目过程中对志愿者用户所作的一个在线调查结果。该问卷调查所涉及的问题包含类似你最近是否旅行、你的日常交通方式是什么、你习惯使用什么样的通信方式、你对本学期的学习是否满意等 25 个问题，每个问题都提供了各种不同的答案，并根据各种不同答案进行量化。有关该问卷调查数据的详细情况，请参考 the Reality Mining Data README 文件。然后根据 surveydata 中的数据，运用 SOM 网络对其进行分类，观察系统的分类能力和识别能力。

定义 2（问卷调查数据） surveydata 表示 realitymining 数据中的问卷调查数据，surveydata=〈trv1，car2，liv3，cmm4，evl5，sck6，trm7，phn8，msg9，wcf10，wce11，wcp12，wct13，pcf14，pce15，pcp16，pct17，exp18，sc19，ll20，c&r21，sup22，lsup23，qm24，ita25〉，其中

- trv1：问题 1-Have you travelled recently?
- car2：问题 2-Do you own a car?
- liv3：问题 3-How many miles to you live from MIT?
- cmm4：问题 4-How do you daily commute to MIT?
- evl5：问题 5-How much has your social network evolved since the start of Fall term?
- sck6：问题 6-Have you been sick recently?
- trm7：问题 7-How long into the term did it take for your social circle to become what it is today?
- phn8：问题 8-I use my phone(1) exclusively for work (2) school related matters…
- msg9：问题 9-How often do you send test message?
- wcf10：问题 10-The majority of my daily work communication is done through fact-face discussion?
- wce11：问题 11- The majority of my daily work communication is done through Email?
- wcp12：问题 12- The majority of my daily work communication is done through phone?
- wct13：问题 13- The majority of my daily work communication is done through text-message?
- pcf14：问题 14- The majority of my daily personal communication is done through face-face discussion?
- pce15：问题 15- The majority of my daily personal communication is done through Email?
- pcp16：问题 16- The majority of my daily personal communication is done through phone?
- pct17：问题 17- The majority of my daily personal communication is done through text-message?
- exp18：问题 18-I am satisfied with my experience at MIT thus far?
- sc19：问题 19-I am satisfied with my current social circle?
- ll20：问题 20-I feel I have learned a lot this semester?
- c&r21：问题 21-I am satisfied with the content and direction of my classes and research this semester?
- sup22：问题 22-I am satisfied with the support I received from mu circle of friends?
- lsup23：问题 23-I am satisfied with the level of support I have received from the other members in my Media Lab re-

search group/Sloan core team?

• qm24：问题 24-I am satisfied with the quality of our group meetings?

• ita25：问题 25-I am satisfied with how my research group interacts on a personal level?

除上述问卷调查数据外，我们选择用户使用手机而产生的一序列统计数据，作为用户的个体特征数据，这些数据称之为观察数据 obsdata。下面给出 obsdata 的形式定义。

定义 3(个体观察数据) obsdata 是关于用户使用手机的统计数据。obsdata=〈apps，my_minutes，timeon，comm_local，comm_sms，comm_voice，comm_data〉，其中

• apps：是用户手机上使用过的应用程序以及其使用频率的统计数据，这部分需要从 realitymining. mat 数据文件中用专门编写的子程序进行构造。

• min26：表示手机购买的月均通话时长

• tm27：手机的日均使用时长

• cl28：本地(基于波士顿)通信事件总数

• cs29：收发消息次数

• cv30：语音通话次数

• cd31：手机会话总次数(sessions initiated)

定义 3 中的 apps 字段是关于用户手机运行的应用程序的统计情况。该数据是通过安装在手机上的监控程序截获得到，并记录在 realitymining 文件中的相应字段。这些数据通过算法 2 从 realitymining 中抽取得到。下面给出 apps 的形式定义。

定义 4(应用程序使用频率统计) apps 字段是关于用户手机运行的应用程序的统计情况，apps=〈mnu1，ph2，log3，bu4，phb5，gs6，mce7，pb8，ss9，pf10，fe11，c12，cam13，a14，m15，f16，b17，w18，c19，pn20，vr21，mg22，np23，sa24，p25〉，其中

• mnu1：表示用户启动 menu 程序的频率

• ph2：表示用户启动 phone 程序的频率

• log3：表示用户启动 Logs 程序的频率

• bu4：表示用户启动 BtUi 程序的频率

• phb5：表示用户启动 Phonebook 程序的频率

• gs6：表示用户启动 gs 程序的频率

• mce7：表示用户启动 mce 程序的频率

• pb8：表示用户启动 Pinboard 程序的频率

• ss9：表示用户启动 ScreenSaver 程序的频率

• pf10：表示用户启动 profileapp 程序的频率

• fe11：表示用户启动 FExplorer 程序的频率

• c12：表示用户启动 Calendar 程序的频率

• cam13：表示用户启动 Camera 程序的频率

• a14：表示用户启动 Appmngr 程序的频率

• m15：表示用户启动 mmcapp 程序的频率

• f16：表示用户启动 FileManager 程序的频率

• b17：表示用户启动 Browser 程序的频率

• w18：表示用户启动 www 程序的频率

• c19：表示用户启动 ConnectionMonitorUi 程序的频率

• pn20：表示用户启动 PSLN 程序的频率

• vr21：表示用户启动 Voicerecorder 程序的频率

• mg22：表示用户启动 MediaGallery 程序的频率

• np23：表示用户启动 Notepad 程序的频率

• sa24：表示用户启动 SysAp 程序的频率

• p25：表示用户启动 Pop2 程序的频率

定义 5(个体特征模式) IFP=〈surveydata，obsdata〉

• surveydata：为前述问卷调查数据；

• obsdata：为定义 1 中的用户使用手机的统计数据。

我们对个体的 surveydata 和 obsdata 数据分别用 SSRS 系统进行分类，给出个体所属类别。详细实验结果见实验结果分析部分。

算法 1 构造个体特征模式算法

```
Gen_IFP(S,IFP)
  IFP←∅
  for i=1:|S|
      select an item s∈S
      ifp. surveydata=s. sruveydata;
      ifp. obs=data_analysis(ifp. obs,s);
      IFP←IFP∪ifp;
  end
end
```

算法 2 构造 obsdata 的算法

```
data_analysis (ifp. obs,s)
 ifp. obs. my_minutes=s. my_minutes;
 ifp. obs. timeon=s. timeon;
 ifp. obs. comm_sms=s. comm_sms;
 ifp. obs. comm_voice=s. comm_sms;
 ifp. obs. comm_data=s. comm_data;
   for i=1:|s. apps|
    switch s. apps{1,i}
     case 'camera'
       ifp. obs. apps. camera=ifp. obs. apps. camera+1;
     case 'Calendar'
       ifp. obs. apps. Calendar=ifp. obs. apps. Calendar+1;
     case 'Browser'
       ifp. obs. apps. Browser=ifp. obs. apps. Browser +1;
          ......
     end
end
```

4 基于 SOM 的相似个体识别系统 SSRS

算法 1 产生的 IFP 中记录的是个体特征模式。有了这些 IFP 记录后，即可用 SOM 网络对这些个体进行分类，并识别这些不同个体所属类别。为了更好地理解 SSRS 系统，先对 SOM 的工作原理作简要介绍，然后再描述 SSRS 系统。

4.1 SOM 网络基础知识

SOM 网络是一种自组织竞争神经网络，能够在不需要预先给定导师信号的情况下，通过对客观事物的反复观察、分析与比较，自动发现其内在规律，并对具有共同特征的事物进行正确归类。

SOM 网络的结构一般如图 1 所示。SOM 网络分成两层：输入层和特征映射层。输入层共有 R 个结点，能够接受 $R\times1$ 的列向量 P 作为网络的输入。特征映射层一共有 S 个神经元，可以将其按照特定的拓扑结构进行排列。一般会提供各种不同的拓扑结构函数，供网络初始化时调用。在 metlab 中就提供了六角结构、网格结构和随机结构函数，分别对

应 hextop()、gridtop()、randtop() 3 种函数，供调用者使用。特征映射层的每个神经元均与输入层每一神经元有连接权值，这些连接权值构成了 $S\times R$ 的连接矩阵 W。

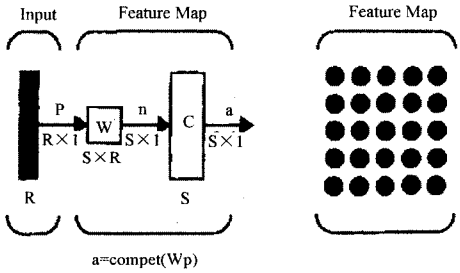


图 1　自组织特征影射

SOM 网络的最大特性是能够自动寻找样本中的内在规律和本质属性，自组织、自适应地改变网络的参数与结构，以实现网络的模式识别和分类的功能。学习时，每接到一个输入模式，SOM 即按照特定的距离函数 f，计算竞争层神经元权向量与输入模式之间的距离，距离最小的那个神经元为获胜神经元。

$$f(P,W_{i^*})=\min_{i\in\{1,\cdots,S\}} f(P,W_i)$$

式中，f 是距离函数，可选的距离函数有欧氏距离和夹角余弦距离。P 表示输入模式向量，W_i 表示竞争层第 i 个神经元与输入神经元的连接权向量。

一旦找到获胜神经元，SOM 采取胜者为王的策略，只对获胜神经元调整权值。也可以对所有位于获胜神经元邻域 $i\in N_{i^*}(d)$ 内的神经元的连接权向量 W_i 进行调整。权值调整公式如下：

$$W_i(q)=W_i(q-1)+\alpha(P(q)-W_i(q-1))$$

式中，q 表示迭代代数，邻域 $N_{i^*}(d)$ 含有在获胜神经元 i^*、半径 d 范围内的所有神经元结点索引，即 $N_i(d)=\{j,d_{ij}<d\}$。

可以看出，当某神经元获胜后，该神经元或其邻域内的神经元将获得调整权值的机会。调整的方向是朝输入模式 P 靠近，调整的幅度取决于学习率 α。而其他非获胜神经元，或者获胜神经元邻域外的那些神经元将受到抑制，得不到调整。经过多次迭代学习，不难想像相似的那些模式向量将逐渐聚集到一起，从而完成聚类和分类的目的。

前述的欧氏距离是描述两个模式向量距离相近程度的最常用的一个距离测度，其数学形式为

$$\|X-W_i\|=\sqrt{(X-W_i)^T(X-W_i)}$$

两模式向量的欧氏距离越小，两个向量越接近，因此认为这两个模式越相似。当两个模式完全相同时，其欧氏距离为零。如果对同一类内各模式向量设定阈值 D，将所有距离小于 D 的模式划分到同一类别，而距离超过 D 的向量模式划分到不同类别中，从而实现分类和模式识别的目的。

其他距离测度，比如余弦法等，限于篇幅，此处不赘述，感兴趣的读者请自行参考相关文献。有关 SOM 网络的详细介绍，请参考 Martin T. Hagan 等的相关著作[13]。

4.2　相似个体识别系统 SSRS

相似个体识别系统 SSRS 是以 metlab 神经网络工具箱中 SOM 网络工具函数为基础，在 GUIDE 辅助下设计并实现的视窗程序。程序界面如图 2 所示。

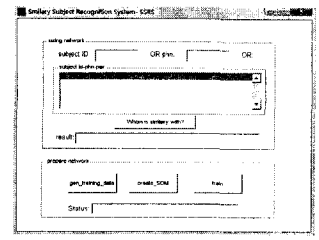


图 2　SSRS 关系识别系统

SSRS 系统可分成两个状态：网络准备状态和网络使用状态，分别对应图 2 中的下面和上面两部分。

在用 SSRS 进行相似个体识别之前，需要先准备 SOM 网络，图 2 中下半部分是用来准备网络的，共有 3 个按钮，分别用来产生训练数据、根据训练数据创建 SOM 神经网络、用训练数据对已创建的神经网络进行训练。prepare network 面板中的文本框主要用来显示系统所处状态。

当点击 gen_training_data 按钮时，系统将用算法 1 从 realitymining 文件中构造出个体特征模式集 IFP。当 IFP 准备好后，系统出现如图 3 所示的状态。

当点击 create_som 按钮时，系统首先对 IFP 执行归一化处理，将 IFP 中的数据全部转换成[0 1]区间中的数据，然后调用工具箱的 newsom 函数创建网络。一旦网络成功创建，即可点击 train 按钮对网络进行训练。当网络训练好后，系统会将个体的 id 号及其对应的电话号码(散列值)以列表的形式显示在 using network 面板的列表框中，辅助用户使用网络进行相似个体识别。图 4 是网络训练好后系统显示的状态。

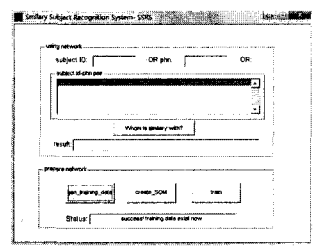


图 3　成功创建训练数据时的状态

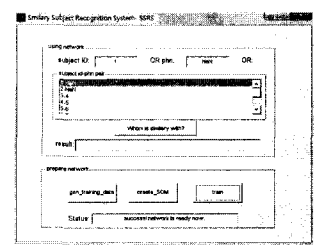


图 4　网络训练好后的状态

prepare network 面板中，必须按照 gen_training_data，create_som，train 的循序使用，方能正确运行。如果不遵循上述顺序，则系统会给出相应提示，使用户按照要求执行正确的动作。

当网络训练好后，系统会有提示，此时即可用训练好的网络进行识别。在系统的 using network 面板中，有一按钮 Whom is similary with?，只要在 subject ID 栏或 phn 栏文本框中，输入其中一个数值，则另一个文本框中会随即出现与用户输入的 ID/phn 号对应的 phn 号/ID，并且下面的列表框也会相应地定位到用户输入的 ID/phn 号对应的记录。如果用户输入的 ID/phn 号超出了系统能识别的号码(系统数据库)范围，则系统会有相应的提示。图 5 是在 subject ID 文本框输入数字 3(即第 3 个个体)后，再点击按钮 Whom is similary with? 对系统进行询问，系统将作出响应时的状态：系统在结果框中显示，个体 3 属于第 1 类，与之相似的个体有 4(5)、5(6)、6(7)、9(10)、10(11)(括号中的数字代表相应个体的电话号码散列值)。

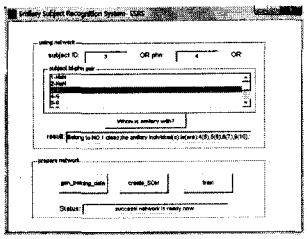


图 5　系统成功识别个体所属类别以及与该个体相似的其他个体

5　实验结果及分析

由于 realitymining 数据文件是. mat 格式的，因此出于方便考虑，本文实验是用 metlab7. 0 工具完成的，系统环境是 vista/intel(R) Core? 2Duo CPU T8100@2. 1GHz 2. 1GHz/2. 0G 内存。

5. 1　SSRS 对 surveydata 数据的分类实验结果

首先将 SOM 神经网络竞争层神经元设为 3，即希望网络将数据划分成 3 类，来看网络的分类情况。

为对 SSRS 系统的分类和识别能力能有一个直观的理解，我们记录了网络经过训练后的连接权矩阵，并将之绘制成二维平面图形，如图 6 所示。从图 6 可以看出，系统将数据分成 3 类，类别 1、2 与类别 3 的差异是比较明显的，但类别 1、2 的差别则较模糊。

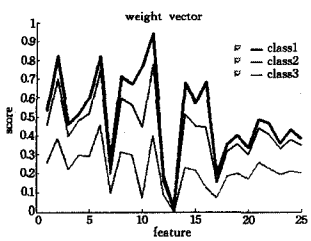


图 6　对 surveydata 分类时得到的代表类别中心的权向量

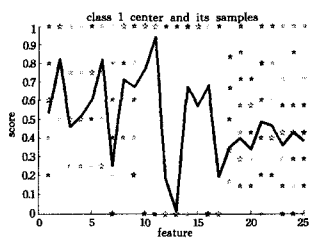


图 7　共有 54 个样本被划分到第 1 类

属于第 1 类的样本有：6，9，10，12，13，14，15，16，19，20，21，23，25，27，28，31，32，36，37，38，40，42，43，44，46，48，49，50，52，58，60，61，62，63，68，69，79，80，81，82，83，85，86，88，93，94，96，97，98，100，102，103，104，106，共 54 个样本，如图 7 所示。

属于第 2 类的样本有：4，5，7，11，17，22，26，30，33，35，41，53，55，56，57，65，66，67，70，71，72，73，74，75，76，77，78，84，87，89，90，91，95，101，共 34 个样本，如图 8 所示。

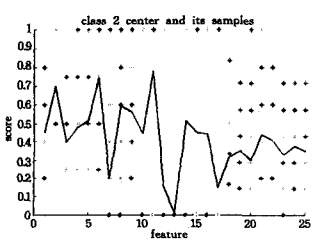


图 8　共有 34 个样本被划分到第 2 类

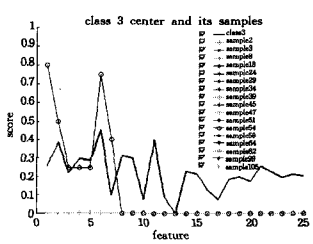
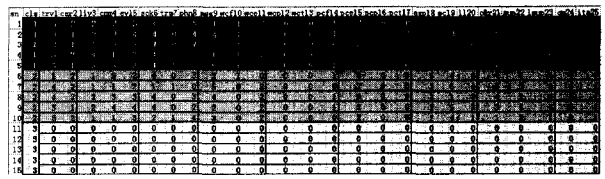


图 9　共有 18 个样本被划分为第 3 类

属于第 3 类的样本有：1，2，3，8，18，24，29，34，39，45，47，51，54，59，64，92，99，105，共 18 个样本，如图 9 所示。

另外，为了更好地理解各个类别间的差异情况，我们根据 SSRS 的分类情况，从每个类别中随机选取 5 个样本，3 个类别共 15 个样本，每个类别用不同颜色加以标示，形成表 1。表 1 中 sn 表示记录号，cls 是类别号，表示该条记录所属类别。其他列中，每列后面的数字分别对应 surveydata 中 25 个问题的序号，例如字段 trv1 中的数字“1”表示 surverdata 中 25 个问题中的第一个问题，即 have you travelled recently? 其余字段的含义可以类推得到。

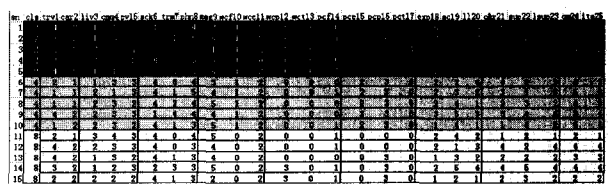
表 1　surveydata，神经元数为 3 时的分类情况(原始数据)



从表 1 可以看出，在将 SOM 神经网络竞争层神经元数设为 3 的情况下，SSRS 系统的分类结果可以使第 1、2 类和第 3 类之间具有明显的差别。而第 1、2 类之间，分类结果并不明显。第 3 类基本上取值为“0”，这是因为在进行数据处理时，把原始数据中所有取值为“NaN”的字段都用“0”代替。而原始数据中，存在比较多数据噪声，很多参与调查的用户没有回答问题，或者没有认真给出问题的答案。这些记录大多数都为“0”。SOM 轻易地将这样的一类数据与其他数据区分开来。这说明 SOM 神经网络虽具有一定的分类能力，但其分类结果还是比较粗糙的，这主要是因为竞争层神经元个数设得过少，类别总数过少的缘故。

我们将 SOM 竞争层神经元个数设为 10，然后再运行 SSRS 系统，并随机从中抽出 3 个类别，每个类别 5 个样本，形成表 2。

表 2　surveydata，神经元数为 10 时的分类情况(原始数据)



根据表 2 以及 car 字段的含义，可以比较肯定地说，第 2 类和第 8 类的群体是有车一族，而第 4 类的群体则是无车族。且根据 wcp(工作时是否常打电话)和 wct(工作时是否常发短信)的含义，2、4、8 这 3 类群体不太喜欢电话或短信进行沟通，而比较喜欢用 e-mail 沟通(wce 取值为 2 表示工作时用 e-mail 沟通)。此时类别的分类情况就得到了相对较好的效果。

5. 2　obsdata 的实验结果及其分析

对于 obsdata，我们同样先考察竞争层神经元数目只有 3 个的情况下权矩阵的情况。图 10 是权矩阵二维平面图。由图 10 可以看出，基本上 3 类向量是互相分离的，差异性比较良好。除少数几个特征指标相同外(由于噪声或数据缺失)，大部分指标都能得到较好的区分。

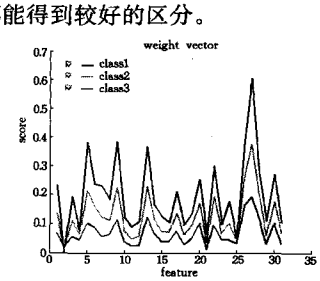


图 10　对 obsdata 分类时得到的代表类别中心的权向量

(下转第 283 页)

[2] Sezgin M, Sankur B. Survey over image thresholding techniques and quantitative performance evaluation [J]. Journal of Electronic Imaging, 2004, 13(1): 146-168

[3] Pun T. A new method for gray-level picture thresholding using the entropy of the histogram [J]. Signal Processing, 1980, 2(2): 223-237

[4] Pun T. Entropic thresholding: A new approach [J]. Computer Vision, Graphics, and Image Processing, 1981, 16(2): 210-239

[5] Kapur J N, Sajoo P K, Wong A K C. A new method for gray-level picture thresholding using the entropy of the histogram [J]. Computer Vision, Graphics, and Image Processing, 1985, 20(3): 273-285

[6] Abutaleb A S. Automatic thresholding of gray-level pictures using two-dimensional entropies [J]. Pattern Recognition, 1989, 47: 22-32

[7] Brink A D. Thresholding of digital images using two-dimensional entropies [J]. Pattern Recognition, 1992, 25(8): 803-808

[8] Chen W T, Wen C H, Yang C W. A fast two-dimensional entropic thresholding algorithm [J]. Pattern Recognition, 1994, 27(7): 885-893

[9] Gong J, Li L Y, Chen W N. Fast recursive algorithm for two-dimensional thresholding [J]. Pattern Recognition, 1998, 31(3): 295-300

[10] 吴成茂，田小平，谭铁牛. 二维熵阈值法的修改及其快速迭代算法[J]. 模式识别与人工智能，2010，23(1)：127-136

[11] 吴一全，潘喆，吴文怡. 二维直方图区域斜分的最大熵阈值分割算法[J]. 模式识别与人工智能，2009，22(1)：162-168

[12] 吴一全，张金矿. 二维直方图 θ-划分最大平均离差阈值分割算法[J]. 自动化学报，2010，36(5)：634-643

[13] 童小念，刘娜. 一种基于遗传算法的最优阈值图像分割算法[J]. 武汉理工大学学报：交通科学与工程版，2008，32(2)：301-304

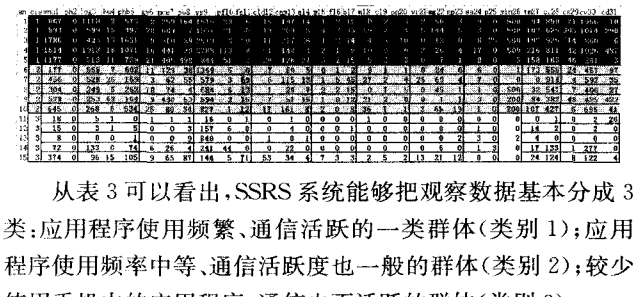
[14] 张云飞，张晔. 利用二维熵自动确定图像分割的阈值[J]. 哈尔滨工程大学学报，2006，27(3)：353-356

[15] 姚敏. 数字图像处理[M]. 北京：机械工业出版社，2008

（上接第 252 页）

同样，为了更好地理解各个类别间的差异情况，我们根据 SSRS 的分类情况，从每个类别中随机选取 5 个样本、3 个类别共 15 个样本，每个类别用不同颜色加以标示，形成表 3。表 3 中 sn 表示记录号，cls 是类别号，表示该条记录所属类别。其他列中，前 25 列中分别对应 apps 中相应字段，各字段含义见定义 3 和定义 4 相应部分。

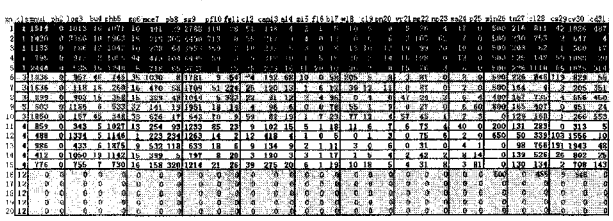
表 3　obsdata，神经元数为 3 时的分类情况（原始数据）



从表 3 可以看出，SSRS 系统能够把观察数据基本分成 3 类：应用程序使用频繁、通信活跃的一类群体（类别 1）；应用程序使用频率中等、通信活跃度也一般的群体（类别 2）；较少使用手机中的应用程序、通信也不活跃的群体（类别 3）。

当我们将 SOM 竞争层神经元数目设为 12 时，SSRS 对群体的分类将得到进一步细化。表 4 是将 SOM 竞争层神经元数目设为 12 时，根据 SSRS 系统的分类结果从中抽取 4 个类别（类别 1，3，4，12），每个类别随机抽取 5 个样本得到的数据。从表 4 中可以观察到类似表 3 的结论，而且比表 3 分类更明显、更加准确。

表 4　obsdata，神经元数为 12 时的分类情况（原始数据）



结束语　本文在 Nathan Eagle 博士的 reality mining 项目的基础上，用 Nathan Eagle 提供的数据，构造个体特征模式。用这些个体特征模式，创建并训练 SOM 神经网络，运用训练好的神经网络实现群体分类以及相似个体的识别。理论和实验结果表明，我们的个体特征模式的算法是有效的，用我们的方法进行相似个体识别能够达到良好的识别效果。

参 考 文 献

[1] Reality mining[EB/OL]. http://reality.media.mit.edu/

[2] Eagle N, Pentland A. Social Serendipity: Mobilizing Social Software[J]. IEEE Pervasive Computing (Special Issue: The Smart Phone), April-June 2005: 28-34

[3] Wasserman S, Faust K. Social Network Analysis: Methods and Applications[M]. New York: Cambridge Univ Press, 1994

[4] Kossinets G, Watts D. Empirical analysis of an evolving social netwok[J]. Science, 2006, 311: 88-90

[5] Ebel H, Mielsh L, Bornholdt S. Scale-free topology of e-mail networks[J]. Phys Reve, 2002, 66: 35103

[6] Aiello W, Chung F, Lu L. A random graph model for massive graphs[C] // Proceedings of the Thirty-sencond Annal ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2000: 171-180

[7] Onnela J, et al. Structure and tie strengths in mobile communication networks[J]. Proc Natl Acad Sci, 2007, 104: 7332-7336

[8] Eagle N. Mobile Phones as Social Sensors[M]. The Handbook of Emergent Technologies in Social Research. Oxford University Press, 2010

[9] Eagle N, Pentland A. Eigenbehaviors: Identifying Structure in Routine[J]. Behavioral Ecology and Sociobiology, 2009, 63(7): 1057-1066

[10] Eagle N, de Montjoye Y, Bettencourt L. Community Computing: Comparisons Between Rural and Urban Societies Using Mobile Phone Data[J]. IEEE Social Computing, 2009: 144-150

[11] Eagle N, Pentland A, Lazer D. Inferring Scoial Network Structure Using Mobile Phone Data[J]. Proceedings of the National Academy of Sciences (PNAS), 2009, 106(36): 15274-15278

[12] Eagle N, Pentland A, Lazer D. Inferring Social Network Structure Using Mobile Phone Data[J]. Proceedings of the National Academy of Sciences (PNAS), 2009, 106(36): 15274-15278

[13] Hagan M T, Howard B. Demuth. Neural network Design (Har/Dis edition)[M]. PWS Pub. Co., 1995