用SOM网络从移动通信数据中识别朋友关系

陈蔼祥

(广东商学院数学与计算科学学院 广州 510320) (中山大学软件研究所 广州 510275)

摘 要 从移动电话通信数据中挖掘用户之间的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,设计并实现了一个行之有效的高效分析算法do&del,以构造手机号码对之间的通信模式,证明了do&del算法的可靠性,并给出了算法的理论复杂性和实验观察结果。然后在do&del算法的基础上,给出了基于SOM(Self-Organizing Map)的关系识别系统RRS(Relationship Recognition System),该系统能够在训练好的SOM网络支持下,有效识别双向朋友、单向/普通朋友、一般人等此类用户关系。

关键词 现实数据挖掘,移动通信,SOM,do&del算法

Using SOM to Recognize Friendship from Mobile Phone Data

CHEN Ai-xiang

(Department of Mathematics, Guangdong University of Business Studies, Guangzhou 510320, China)

(Software Research Institute, Sun Yat-Sen University, Guangzhou 510275, China)

Abstract Data collected from mobile phones have the potential to provide insight into the relational dynamics of individuals. In this paper, using the reality data collected from 94 Nokia 6600 users, an highly effective parsing algorithm named do&del was designed and realized to construct the communication pattern of a pair phone number. The do&del algorithm's correctness, theory time complexity and experimental observations were given. With the output of do&del as training data, a Relationship Recognition System building on SOM neural network was designed and realized. The system RSS can recoginze the reciprocal friends, nonreciprocal friends and reciprocal nonfriends.

Keywords Reality mining, Mobile phone, SOM, Do&del algorithm

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。

通过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,来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的用户群体[8,9]。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。仍然是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,他们的结果发现,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]。

社会个体之间的关系识别问题是社会网络分析中的一个核心任务,识别个体之间的内在关系,远比观察个体之间的外在关系要困难得多。个体关系的识别问题本质上可以看作是一个分类问题,只要我们能够获取个体之间的相关数据,我们就可以实现个体关系的识别。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

本文受广东省自然科学基金(10451032001006140),广州市科技和信息化局(10C12140131),广东高校优秀青年创新人才培育项目(LYM10081),广东省大学生创新实验项目(1059210097)资助。

陈蔼祥(1978-),博士,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。

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，差别不大的类激发的神经元位置也是相邻的，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，构造合适的通信模式数据，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 网络实现个体关系类型的识别。

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 节给出通信模式的相关定义，从现实数据集中产生通信模式的 do&del 算法，并给出了 do&del 算法停机的一个充分条件；第 4 节设计并实现了基于 SOM 网络的关系识别系统 RRS，该系统运用 do&del 算法得到的结果数据，创建并训练 SOM 神经网络，然后运用训练好的 SOM 网络，识别两个号码之间的朋友关系；第 5 节给出了 do&del 算法的运行效率方面的实验结果分析，并对 RRS 系统的功能和识别结果进行了分析；最后，我们进行了相应总结并给出了结论。

2 使用的数据[11]

本文使用的数据来自麻省理工学院的社会网络分析项目。这些数据是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 5 分钟扫描以自动发现 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%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 博士及其项目组所提供的这些数据[11]，没有这些数据的支撑，本文的工作无从谈起。

一般地，如此产生的数据一般来说都是巨大的，要在这些海量数据集中挖掘有用信息，除了挖掘对象本身的设计和考量外，算法的处理效率也是我们不得不关心的问题。将在下一节对此加以分析和讨论。

3 通信模式及其 do&del 构造算法

观察并识别两个体之间的内在关系(朋友关系或爱人关系)，远比观察两个体的外在关系(外貌、衣着、是否邻近等)要困难得多。来自认知科学领域的研究结果告诉我们，个体在社会活动中所呈现出来的时空模式可以揭示两个个体之间这种内在的关系类型[12,13]，例如，如果两个体在周末晚上呆在一起长达几个小时，这两个体极有可能关系非同一般。而如果两个体在某工作日下午呆在一起几个小时，就不能反映这两个体有亲密关系。借助认知科学的这一结果，Nathan Eagle 博士及其领导的团队提出了以下假定[14]：如果特定个体经常周末通信联系频繁，且常在一个地方出现，则这两个体很有可能是朋友关系，并将个体之间的关系进一步划分为双向朋友关系(互相认为对方是朋友)、单向/普通朋友关系(一方认为对方是朋友)、一般人关系(互不认为对方是朋友)。

上述假定为我们提供了一个从个体行为所产生的外部观察数据推断个体关系的基础。本文遵从上述假定，并在上述假定基础上，定义个体之间的通信模式，以便提供能为下一阶段 SOM 网络所使用的数据，达到通过通信模式这一观察数据来识别个体内在关系的目的。为此，本文首先给出通信事件的定义。

定义 1(通信事件及通信事件集) 事件是指两个体之间的一次通信，事件 $e=\langle$eventid, mphn, sphn, date, contact, description, direction, duration$\rangle$，其中：

- eventid：表示事件的序列号
- mphn：主号
- sphn：从号
- date：事件发生的时间
- description：通信类型，分为语音通话、短信等
- direction：表示事件发生的方向，取值分“outgoing”，“incoming”，分别对应 mphn 发起通话/发送短信和 mphn 接收电话/接收短信
- duration：表示通信持续时间。

集合 E 是由事件 e 组成的通信事件集。

根据 Nathan Eagle 的假定，两个体周末和工作日之间的通信情况以及通信方向，对于推断两个体内在关系，具有完全不同的意义。基于此，我们给出两个体间通信模式的定义如下。

定义 2(通信模式及通信模式集) 通信模式 cp 是两个体之间在一个时间段内通信统计情况，$cp=\langle$cpid, mphn, sphn, wnco, conp1, cod1, wnci, cinp1, cid1, wdco, conp2, cod2, wdci, cinp2, cid2, wnmo, wnmi, wdmo, wdmi$\rangle$，其中：

- cpid：表示通信模式序列号
- mphn，sphn 意义同定义 1
- wnco：表示主号 mphn 在周末呼叫从号 sphn 的次数
- conp1：表示主号 mphn 在周末呼叫从号 sphn 的未接的次数
- cod1：表示主号 mphn 在周末与从号 sphn 主叫通话时长
- wnci：表示主号 mphn 在周末被从号 sphn 呼叫的次数
- cinp1：表示主号 mphn 在周末被从号 sphn 呼叫未接的次数
- cid1：表示主号 mphn 在周末与从号 sphn 被叫通话的时长
- wdco：表示主号 mphn 在工作日呼叫从号 sphn 的次数
- conp2：表示主号 mphn 在工作日呼叫从号 sphn 未接

的次数

• cod2:表示主号 mphn 在工作日与从号 sphn 主叫通话的时长

• wdci:表示主号 mphn 在工作日被从号 sphn 呼叫的次数

• cinp2:表示主号 mphn 在工作日被从号 sphn 呼叫未接的次数

• cid2:表示主号 mphn 在工作日与从号 sphn 被叫通话的时长

• wnmo:表示主号 mphn 周末给从号 sphn 发短信次数

• wnmi:表示主号 mphn 周末接受从号 sphn 短信次数

• wdmo:表示主号 mphn 工作日给从号 sphn 发短信次数

• wdmi:表示主号 mphn 工作日接受从号 sphn 短信次数

集合 CP 为通信模式 cp 组成的通信模式集。

通信模式记录的是一对个体在某一时间段内的通信统计情况。约定 cp/E 表示基于事件集 E 的通信模式 cp。同理,CP/E 表示基于通信事件集 E 的通信模式集 CP。

一般地,实时监控数据得到的是一序列的通信事件,而通信模式则是某一时间段内特定号码之间的通信统计情况。由此产生了根据通信事件集 E 生成通信模式 cp 的问题。

定义 3 ($e \equiv p$) 对于事件 e 和通信模式 cp,如果 e. mphn $= cp$. mphn 且 e. sphn $= cp$. sphn,或者 e. mphn $= cp$. sphn 且 e. sphn $= cp$. mphn 则称 e 和 cp 匹配,记为 $e \equiv cp$。

定义 4(通信模式生成器) 通信模式生成器是函数 $f: M \times S \times E \to CP$,其中,$M$ 表示所有主号 mphn 的集合,S 表示所有从号 sphn 的集合,E 和 CP 的含义见前文相应部分。

显然,从计算的角度来看,上述生成函数 f 是可计算的,并且能在 $O(|M| \times |S| \times |E|)$ 时间内停机。因为只要遍历事件集 E,并根据 M 和 S 将 e 分类汇总到 CP 中做相应记录,算法即可成功终止。

对于函数 f 的计算算法的具体实现,可以有两种不同的方法。一种是以 E 为中心,遍历一次 E 生成所有 CP。另一种则是以 CP 为中心,遍历一次 E,生成一条 cp。两种方法,各有不同应用场合,方法一适合在 E 比较大,M 和 S 相对较小,即人群相对比较集中的情况,更多的是在构造 SOM 神经网络训练数据时使用,而方法二则可在 M 和 S 都比较大,即人群相对分散的情况下使用,该方法更多是在给定特定号码用 SOM 网络进行识别的时候使用。因此,应该根据不同应用场合选用不同方法。

对于实际通信数据而言,E 往往是巨大的。本文处理的实际数据集 realitymining 就是 94 个用户在历时 10 个月内产生的通信数据集,共有 56.6M。算法 1 是以 E 为中心的通信模式生成器 f 的具体实现算法。该算法是边生成边剪枝的前向处理算法。算法首先从 E 中提取一个记录 e,检查 CP 中是否有与 e 的主号 mphn 和从号 sphn 均匹配的记录 cp,如果存在,则调用 data_analysis 算法对记录 e 进行分类汇总,然后从 E 中删除事件 e;如果不存在该记录,则为这两个号码 e. mphn 和 e. sphn 构造新的 cp 记录,并添加到 CP 中。

大多数实际数据挖掘场合,要处理的数据量一般是巨大的。因此,算法 1 的性能和效率是非常值得关注的。本文将对算法 1 的可靠性和时间复杂度作相应的分析,分析的结果以下面两个定理的形式给出。

算法 1 通信模式生成函数 f 的 do&del 实现算法

```
do&del(E,CP)
  CP←∅
  for i=1:|E|
      select an item e∈E
        mphn←e.mphn;sphn←e.sphn;
      if there exist a cp∈CP such that cp≡e
          data_analysis(e,cp);
          E←E-e;
      else
          new cp;
          cp.mphn=mphn;cp.sphn=sphn;
          data_analysis(e,cp);
          CP←CP∪cp;
      end
  end
end
```

定理 1 设 E 是算法 1 的输入,E' 和 CP 是算法 1 的输出结果,算法 1 是可靠的,并且算法 1 停机后 E' 和 CP 的基数相等。

证明:对于 E,算法 1 将执行以下两种规则:如果 $\exists cp \in CP$,使 $e \equiv cp$,则删除一条记录(R1)。否则什么都不做(R2)。而对于 CP,算法 1 同样执行以下两种规则:$\exists cp \in CP$,使 $e \equiv cp$,则根据 e 更新 cp(R3),否则,建立新的与 e 对应的记录(R4)。当算法 1 遍历 E 后,任一对在原来 E 中出现的号码,由于规则 R1,结果 E' 中仅有一条记录与之对应。而任一对号码,由于规则 R4 的作用,在 CP 中有则仅有一条记录与之对应。因此,算法停机后,必有 E 和 CP 的基数相等。而由于遍历 E 的过程可确保 E 中记录无遗漏,因此,算法停机后得到的结果数据中,必有 CP/E 成立,即 CP 是基于 E 的,故算法 1 是可靠的。

算法 2 根据事件 e 计算通信模式 cp 的 data_analysis 算法

```
data_analysis (e,cp)
  switch e.description
    case 'voice call'
      switch e.date
        case {'Sunday','Saturday'}
          if e.direction=='outgoing'
            cp.weekend_call_out←cp.weekend_call_out+1 and
            cp.out_d1←cp.out_d1+e.duration or
            cp.o_not_pickup1←cp.o_not_pickup1+1
          else
            cp.weekend_call_in←cp.weekend_call_in+1 and
            cp.in_d1←cp.in_d1+e.duration or
            cp.in_not_pickup1←cp.in_not_pickup1+1
        otherwise
          if e.direction=='outgoing'
            cp.weekday_call_out←cp.weekday_call_out+1 and
            cp.out_d2←cp.out_d2+e.duration or
            cp.o_not_pickup2←cp.o_not_pickup2+1
          else
            cp.weekday_call_in←cp.weekday_call_in+1 and
            cp.in_d2←cp.in_d2+e.duration or
```

```
            cp. in_not_pickup2←cp. in_not_pickup2+1
        case 'short message'
          case {'Sunday','Saturday'}
            if e. direction=='outgoing'
              cp. weekend_msg_out←cp. weekend_msg_out+1
            else
              cp. weekend_msg_in←cp. weekend_msg_in+1
          otherwise
            if e. direction=='outgoing'
              cp. weekday_msg_out←cp. weekday_msg_out+1
            else
              cp. weekday_msg_out←cp. weekday_msg_out+1
      case 'packet data'
      otherwise
        error('this is impossible!')
end
```

定理 2 算法 1 的时间复杂度是 $O(|M|\times|S|\times|E|)$，其中 M 为主号空间，S 为从号空间，E 为事件空间。

证明：极端情况下，主号和从号的积空间 $M\times S$ 中的任一点，E 中至少有一条记录 e 与之对应。此时，必有 $|E|\geqslant|M\times S|=|M|\times|S|$。同时，根据定理 1 的可靠性，算法停机后，有 CP/E，故 $|CP|=|M\times S|=|M|\times|S|$。从而算法最终生成 CP 的时间复杂度为 $O(|M|\times|S|\times|E|)$。

值得一提的是，算法 1 中的剪枝策略是必要的，这可很大程度上避免海量数据处理时出现内存耗尽的情况，极大地降低对系统内存的要求。定理 1 和定理 2 的结果告诉我们，算法 1 总能在多项式时间内得到正确的结果。后文试验结果与分析部分，将给出算法 1 在不同规模数据下的实验对比结果。

4 基于 SOM 的关系识别系统 RRS

算法 1 产生的 CP 中记录的是各不同号码之间的通信模式，有了这些 CP 记录后，我们即可用 SOM 网络对这些通信模式进行分类，以识别这些不同通信模式的所属类别，从而达到揭示这些通信模式中隐含的不同关系的目的。为了更好地理解 RRS 系统，首先对 SOM 的工作原理作一简要介绍，然后再描述我们的 RRS 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，该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。

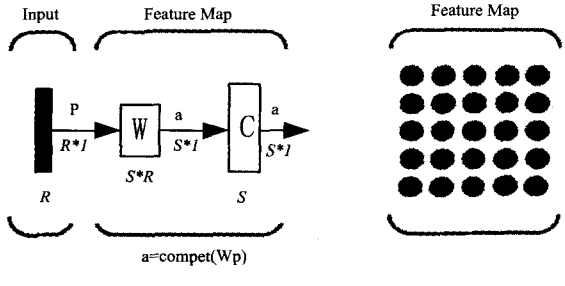

图 1 Self-organizing feature map

SOM 网络的结构如图 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 matlab 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()三种函数，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。

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。

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

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。权值调整公式按下式进行。

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

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。

前述提到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，其数学形式可用下式表示：

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

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，那么将所有距离小于 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，此处不作赘述，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[15]。

4.2 关系识别系统 RSS

关系识别系统 RSS 是以 matlab 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。

RSS 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 RSS 进行关系识别之前，需要先准备 SOM 网络，图 2 中下半部分是用来准备网络的，共有三个按钮，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。

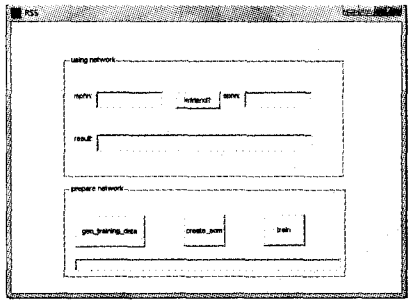

图 2　RSS 关系识别系统

当点击 gen_training_data 按钮时，系统将用算法 1 将 realitymining 文件中记录的通信事件，构造出号码之间的通信模式，从而产生通信模式集 *CP*。当 *CP* 准备好后，系统出现如图 3 所示的状态。

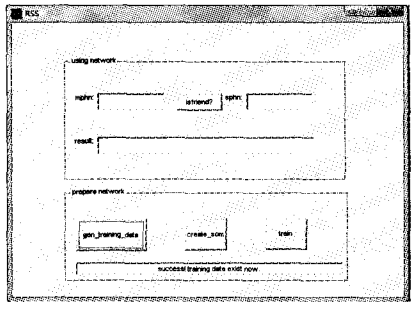

图 3　成功创建训练数据时的状态

当点击 create_som 按钮时，系统首先对 *CP* 执行归一化处理，将 *CP* 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，一旦网络成功创建后，即可点击 train 按钮对网络进行训练。

prepare network 面板中，必须按照 gen_training_data，create_som，train 的顺序使用，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，要用户按照要求执行正确的动作。

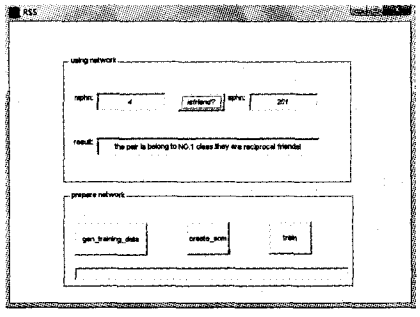

图 4　系统成功识别两号码之间的关系类别

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 isfriend?，该按钮两边分别有 mphn 和 sphn 两个文本框，要求你输入两个手机号码。按要求输入完两个手机号码后，点击 isfriend? 按钮，系统将识别这两个号码到底是属于哪一类别。例如，在 mphn 和 sphn 中分别输入 4 和 201(出于保护隐私的目的，这里的手机号都是将原始手机号通过单向散列函数处理后得到的手机号的散列值)。系统将认为这两个手机号属于第 1 类的关系，如图 4 所示。

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 matlab7. 0 工具完成的，系统环境是 vista/intel(R) Core™ 2Duo CPU T8100 @ 2. 1GHz 2. 1GHz/2. 0 G 内存。

5.1　算法 1 的实验结果

算法 1 是产生 RSS 系统训练数据的关键算法。为了测试算法的效率和性能。我们从 realitymining 中选取了不同规模的数据，以便对算法 1 进行测试。realitymining 共有 106 个 subject[1]，我们从中选择 10、20、30、40、50、70、90、106 个 subject 的通信记录，构造出 data10、data20、data30、data40、data50、data70、data90、data106 共 8 组不同规模的数据集。然后专门编写了测试算法 1 性能的测试程序 test. m，测试程序能够对 8 组数据分别调用算法 1 进行处理，然后记录算法在处理这些数据集时的各种表现，考虑的指标有处理事件总数、删除记录总数、算法运行的时间以及算法输出结果的文件大小。表 1 记录的是算法 1 在处理这 8 组数据时各个具体指标的详细情况。

表 1　算法 1 处理不同规模数据集时的性能表现

个体数	事件总数	删除记录数	算法运行时间(秒)	算法输出文件大小(kB)
data10	19896	18691	105. 14	11
data20	40310	37964	340. 97	24
data30	60180	56686	745. 38	35
data40	82101	77432	1365. 6	48
data50	94213	88641	1794. 5	56
data70	130690	123250	3186. 3	77
data90	159270	149330	4909. 4	99
data106	181270	169440	6597. 3	116

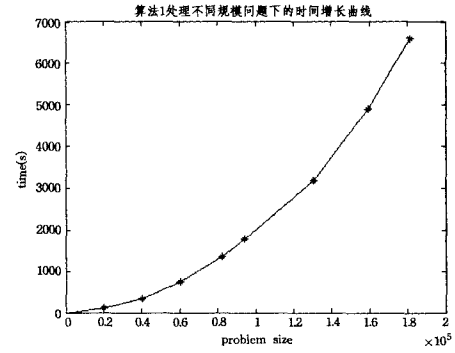

图 5　算法 1 时间复杂度实验观察结果

如果将通信事件总数看作是问题规模，根据表 1 中的数据，可以绘出算法 1 在处理各种不同规模问题的实际时间增长曲线(见图 5)。另外根据定理 2，可以绘出算法 1 的理论时间复杂度曲线(见图 6)。比较图 5 和图 6，我们可以发现，从实验观察到的时间增长曲线基本上与理论分析的结果是一致的。

[1]　此处的 106 个 subject 似与前文的 94 个 subject 矛盾。对此本文解释如下：前文的 94 个 subject 是来自 Nathan Eagle 博士写的 the Reality Mining Data README 文件，并且 Nathan Eagle 多篇论文中均是如此介绍其使用的数据集。但我们在实验中发现，realitymining 这个数据文件中，共有 106 个 subject。造成两者不一致的原因并不清楚，本文也未就此问题询问过 Nathan Eagle 博士。出于严谨真实的角度考虑，本文如实呈现这种不一致，并下注于此。

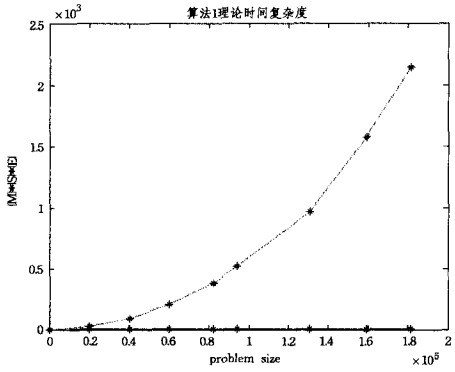

图 6　算法 1 理论时间复杂度曲线

5.2　RSS 的实验结果分析

为验证 RSS 的识别能力和识别准确性，对训练好的 SOM 网络(本文的实验数据是经过 100 代训练后得到的结果。实验中，训练代数设成 20 代得到的识别结果与训练代数为 100 时的结果差别不大)，本文选择 *CP* 中的若干条记录，将每条记录的 mphn 和 sphn 提交给 RSS，让 RSS 对这些号码之间的关系进行识别。为了方便我们分析 RSS 的识别结果，把 RSS 的识别结果以及对应号码之间的 *cp* 从 *CP* 中提取出来，放置到一个数据表中，同时，为方便对比，将 RSS 识别为同一类的 *cp* 记录排列放在一起，并根据不同类别标以不同颜色，以方便区分，形成图 7。

图 7 中 *sn* 表示记录的序列号，mphn 和 sphn 分别表示两个电话号码的散列值，*cls* 表示 RSS 识别的结果类别，*idx* 表示该 *cp* 在 *CP* 中的索引位置，其他字段的含义同定义 2。

sn	mphn	sphn	cls	idx	wnco	conp1	cod1	wnci	cinp1	cid1	wdco	conp2	cod2	wdci	cinp2	cid2	wnmo	wnmi	wdmo	wdmi
1	4	201	1	2	82	8	32479	0	0	0	224	10	49569	0	0	0	0	0	0	0
2	6	283	1	536	28	5	3182	13	1	5434	44	8	19105	16	3	8252	22	13	21	25
3	4	143	1	4	50	4	2553	51	5	1581	58	3	2148	45	2	1792	12	8	12	12
4	4	23	1	5	13	3	925	11	2	659	36	6	2113	15	5	1204	0	0	0	0
5	4	246	1	28	1	0	5	1	0	5	3	3	0	4	1	1436	96	47	189	113
6	5	491	2	303	4	1	247	5	0	742	7	1	282	4	1	475	0	0	0	0
7	4	230	2	23	0	0	0	0	0	0	5	3	741	3	1	105	0	0	2	2
8	6	757	2	537	0	0	0	0	0	0	1	0	354	0	0	0	11	13	6	7
9	4	224	2	30	5	0	1174	2	0	772	3	0	1166	0	0	0	0	0	0	0
10	4	223	2	34	3	0	55	1	1	0	2	1	43	4	3	313	0	0	0	0
11	5	340	3	305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
12	5	483	3	292	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	5	468	3	330	0	0	0	0	0	0	2	0	222	0	0	0	0	0	0	0
14	5	564	3	304	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	6	784	3	538	0	0	0	0	0	0	6	0	625	1	0	79	0	0	0	0
16	4	235	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1

图 7　RSS 的识别结果理解

从图 7 的结果来看，被 RSS 认为第 1 类的记录中，除第一条记录外，都具有较频繁的工作日和/或周末通话和/或短信联系，而且这种联系基本上是双向的，因此，被 RSS 划分为第 1 类，即双向朋友关系。对于第 1 条记录，看起来更像是属于第 2 类的单向朋友，因为该记录中的通信联系更多的是单向联系，但由于单向联系的每一指标值均较大，弥补了其他指标值为 0 而带来的距离的损失，从而干扰了 RSS 对其的识别。造成 RSS 识别错误的原因可能是由于原始数据带有噪声的缘故，因为在实际生活中记录 1 所对应的通信模式基本上不会出现，很难想像在周末 82 次主叫，工作日 224 次主叫的情况下，对方没有任何一次主动回叫，而且互相之间没有短信通信。我们有足够理由怀疑产生该记录的数据不是来自真实的通信数据。

相对于第 1 类关系，第 2 类关系的通信活跃程度就大为降低，整体通信频率远小于第 1 类。这些通信模式中，记录 7 和 8 少量的通信记录发生在工作日，而非周末，记录 8 虽互有短信联系，但不足以说明太多问题，因此这两条记录被 RSS 划分到第 2 类。6、9 和 10 这三条记录，看起来好像通信是双向的，但可以看出，这种双向通信并非对等的，更像是礼节性或例行公事式的对话，并且通信频率并不高，因此被 RSS 识别为第 2 类，即单向/普通朋友关系。

对于第 3 类，这些通信模式应该在 *CP* 记录中占绝大数，这是很好理解的。这类号码之间彼此通信并不活跃，相互之间较少甚至不通信。因此，RSS 将这一类号码识别为第 3 类，即一般人甚至陌生人关系。

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，构造个体之间的通信模式，并用这些通信模式数据，创建并训练 SOM 神经网络，再运用训练好的神经网络实现个体关系类型的识别。理论和实验结果表明，我们的生成通信模式的算法是有效的，用我们的方法进行关系识别能够达到良好的识别效果。

参 考 文 献

[1] Reality mining[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. Association for Computing Machinery, New York, 2000: 171-180

[7] Onnela J, et al. Structure and tie strengths in mobile communication networks[C]//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] Adelson R. Psychological status of the script concept[J]. Am Psychol, 1981, 36: 715-729

[13] Krackhardt D. Assessing the political landscape: structure, cognition, and power in organizations[J]. Admin Sci Q, 1990, 35: 342-369

[14] 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

[15] Hagan M T, Howard B. Demuth. Neural network Design[M]. PWS Pub. Co., Har/Dis edition. ISBN-13: 978-0534943325, 1995