通过移动端传感器数据个性化理解血糖动态

Sam Royston 计算机科
学系 纽约大学，2016
sfoxroyston@gmail.com

arXiv:2302.01400v1 [cs.HC] 2 Feb 2023

摘要

连续血糖监测仪彻底改变了糖尿病患者管理血糖的能力，并为人工胰腺系统铺平了道路。在本文中，我们将连续血糖监测数据与通过智能手机收集的传感器输入相结合，并利用这些数据为患者和临床医生提供分析工具。在为期9个月的时间里，我们通过一款定制的iOS应用程序，从一名自由生活的1型糖尿病患者处收集了GPS数据、活动分类数据和血糖数据。该数据集的新颖之处在于其规模、包含GPS数据，以及数据是在不干扰患者日常生活的情况下从自由生活的患者处收集的。我们介绍了一种基于GPS和活动数据测量生活方式事件发生情况的方法，并展示了这些事件能够捕捉食物摄入的实例，因此与血糖变化相关。最后，我们将这些事件表示整合到系统中，以创建有用的可视化和通知，帮助患者管理糖尿病。

1 引言

1型糖尿病是一种自身免疫性疾病，其特征是无法产生胰岛素，因此也无法调节血糖。糖尿病患者必须手动注射胰岛素，以维持血糖水平稳定；血糖控制不佳则与多种危险副作用有关。有效调节血糖既困难又耗时；我们的系统将部分数据收集、调节和预测工作负载转移到普适计算硬件上。此外，可穿戴技术的当前趋势表明，连续血糖监测技术的侵入性将逐渐降低，并且最终可能集成到消费级健身设备中。血糖是一项信息丰富的变量，有助于描述人体新陈代谢对膳食或未进食状态的反应；如果连续血糖监测方法的侵入性能够降低，这些信息或许对各类非糖尿病患者以及依赖这项技术的糖尿病患者都有用。

我们认为，随着移动设备生成的传感器“轨迹”日益普及，以下目标将成为面向糖尿病及其他领域的个人健康分析平台的重要基础原则。因此，我们所介绍的集成系统正是基于这些原则设计的：

优先考虑个人层面的有效性 在某些情况下，健康数据的敏感性可能会限制大规模汇总分析，因此我们应当关注适用于 $n = 1$ 的方法。基于此，我们重点研究从一位日常生活环境中的 1 型糖尿病患者收集的大型数据集。

持续且不易察觉的数据收集。 所有数据均由手机自动收集，随后在远程服务器上进行汇总、清理和分析。数据收集期间对受试者日常生活的影响微乎其微，参与者只需打开移动应用程序，使其在 iPhone 上作为后台进程运行即可。

利用无处不在的情境数据 将身体活动估计、GPS位置和食品购买交易等形式的数据与连续血糖监测仪（CGM）的数据相结合，从而构建易于理解且可付诸行动的表征和机器学习模型。

健康应用开发者首先且最重要的是要满足患者的需求和偏好。例如，如果一种采用多种外部变量进行血糖预测的先进模型，需要患者主动输入相当一部分数据，那么对患者而言，这种模型就没什么用。简而言之，患者的时间是最宝贵的资源，我们希望尽可能少地占用它。

我们将糖尿病护理行动 (DCA)定义为患者采取的一种行动，涉及食物摄入、胰岛素剂量或未摄入食物或胰岛素。糖尿病护理行动通常包括检查血糖（如果有相关工具可用）。为说明数据收集与患者效用之间的权衡，我们来看一个糖尿病护理行动：患者查看连续血糖监测、胰岛素泵历史记录，回想自己吃过什么，然后决定采取行动；行动可以是纠正性的，也可以是预防性的。典型的糖尿病护理行动可能包括胰岛素“推注”剂量、纠正性糖摄入，或调整泵基础率

允许为个人或课堂使用而免费制作本作品的全部或部分数字版或纸质副本，前提是不得为营利或商业利益制作或分发副本，且副本首页须附有本声明和完整引文。以其他方式复制、再版、发布到服务器或分发至列表，均须事先获得特定许可和/或支付费用。

ACM

。糖尿病患者一天中可能会采取许多糖尿病护理行动，根据行动的质量，每次可能耗费患者 15 秒到 5 分钟（质量较差或“精简”的行动可能会忽略某些考量，例如检查连续血糖监测或胰岛素泵历史记录）。如果我们开发一款分析应用，每天需要花 20 分钟录入数据，那么就必须证明，相比于像上述那样成功完成糖尿病护理行动，把这段时间用于数据录入更值得。我们预计，由于患者对自身经历和过往代谢反应有独到的了解，这种“值得花费时间”的权衡几乎不可能证明数据录入更有优势；因此，我们的目标是专注于几乎不占用患者时间的数据收集方法。

用于理解不同类型数据的技术构成了一个系统，旨在实时提升患者的认知水平和糖尿病管理能力。在本文中，我们的贡献包括

- 介绍一个大规模的单人数据集，该数据集以大规模形式提供了真实世界预测因子的全新组合
- 一种从非结构化智能手机传感器数据中构建离散“事件”特征的方法，这些特征与血糖变化相关

2 动机与相关工作

此前关于血糖数据分析与预测的研究涵盖了多种方法。该领域的机器学习研究在实际操作上面临一些挑战，原因包括：

连续血糖监测基准数据集并未公开提供

由于医疗数据具有敏感性，连续血糖监测研究的数据集通常不会公开。这会影响研究的可复现性，也使不同算法之间无法直接比较。例如，血糖控制良好的糖尿病患者产生的连续血糖监测数据波动较小，可能更容易通过多种回归技术进行预测。本文使用静态预测（即最后一个已知值）的误差，来帮助衡量数据的“易预测”程度。

连续血糖监测仪是一个黑箱 连续血糖监测仪使用专有技术，通过组织间液葡萄糖水平估计血糖水平。这些系统可能采用了卡尔曼滤波器等技术，但美敦力和 Dexcom 等公司几乎没有提供明确的相关文档。

我们可以将预测问题分为两种子类型：**低血糖预测**和**精确血糖预测**。低血糖预测将问题设定为分类任务，旨在预测一段测量值是否会在某个时间跨度后导致低血糖（血糖偏低）。预测低血糖尤其重要，因为它对患者的健康构成最直接的威胁；在特别严重的情况下，若不予治疗，可能导致意识丧失、癫痫发作，甚至在极少数情况下导致死亡。Dassau 等人采用投票系统，结合各种预测器，以高准确率预测低血糖事件 [3]。

相比之下，精确的血糖预测属于回归问题，通常通过测量均方根误差（RMSE）或实际值与预测值之间的R^2误差来评估。已有研究采用人工神经网络 [2], , 以及支持向量机和卡尔曼滤波器 [1], [12]。

在 [1], 中，研究者使用卡尔曼滤波器将纳入胰岛素水平和食物摄入的代谢模型拟合到数据，并将该模型定义的代谢变量用作支持向量机的特征，以进行预测。 [insert] 此前曾尝试将原始活动数据和剂量数据直接添加到时间序列血糖特征中，以训练支持向量回归机（SVR），但未能成功；这一尝试凸显了特征工程在该问题中的重要性。Marling 等人在此基础上，采用代谢模型，并结合患者记录的活动和胰岛素剂量，构建了更完善的特征，用于支持向量回归器。Marling 等人报告称，与不包含活动数据的模型版本相比，包含活动数据的版本取得了虽小但具有统计显著性的提升。Rollins 等人[4] 从信号处理的角度研究该问题，并记录了一位身穿生物传感服的 2 型糖尿病患者连续 3 周的数据。

许多连续血糖监测预测建模研究的目标之一，是助力人工胰腺（AP）的开发。用药理学术语来说，胰岛素吸收动力学 [11] 和碳水化合物代谢的速度都很慢，因此经典的控制理论框架不足以有效控制血糖水平。基于这一原因，人们开始关注解决预测任务，以服务于未来的人工胰腺控制系统。

3 数据发现与预测

在本系统中，我们重点关注利用由位置和活动数据构建的“事件”特征，以提高预测血糖变化的能力。

3.1 数据集

我们介绍一个数据集，其中包含一名25岁男性1型糖尿病患者的间歇性位置、活动和血糖测量值，以及其他元数据。

活动类别（*181,635*条测量值）即使应用程序未运行，活动类别也会在后台持续分类，因此是收集频率最高的数据类型。这是一个分类变量，取值为’步行’, ’跑步’, ’骑行’, ’静止’, ’乘车’, ’静止乘车’, 或’未知’

位置（*61,817* 次测量值）应用在后台运行时，通过 iPhone 上的全球定位系统主动测量位置，并将数据直接发送到服务器。全球定位系统的轮询策略由移动端操作系统管理；设备移动时生成的数据多于静止时，

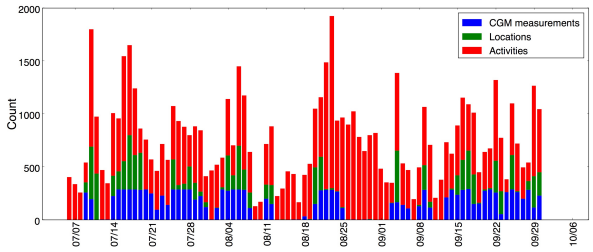


图 1. 每日测量频率

如果设备无法与 GPS卫星建立连接，则完全不会生成任何数据

食品购买 服务器会查询各种银行API，以编目带时间戳的食品购买记录，其中可能包含其他元数据，例如购买地点和成本

血糖（*50,478* 次测量）外部 Dexcom 连续血糖监测设备通过皮下传感器每 5 分钟测量一次血糖水平（ mg/dl），并通过蓝牙与用户的手机通信。

由于采样频率范围较广，我们对数据进行了预处理，将其转换为由 280 多万个特征向量组成的数组，每个特征向量代表一个 15 秒的时间窗口。虽然连续血糖监测测试间隔为每 5 分钟一次，但通过每 15 秒截取一次快照，我们得以捕获更多活动数据，其采样频率有时高达每分钟 4 次以上。

3.2 识别相关事件

利用位置和活动数据的重要第一步，是确定餐食及其他类型事件发生的候选时间。原始位置和活动数据本身并不能构成可行的特征，也无助于提高预测准确率 [1], 因此，我们构建了专门的特征，用以表征对不同地点的访问。*3.2.1* 活动数据插补

如图 1 所示，尽管连续血糖监测采样具有一定规律性，但位置和活动更新的间隔并不规则，而是由 iOS 内部机制决定，并受电池续航、相对运动强度等因素影响。因此，需要稳健的插补方法来填补不完整数据。本节将介绍用于插补活动测量值的方法。

最近邻插补活动数据最简单、最有效的方法之一，是选取待插补时间段的最近邻活动值。然而，当活动模式较为复杂时（例如活动测量值经常交替变化），这种方法就会受到干扰。

滚动模式 该方法会用出现频率最高的活动，填补样本中所有缺失的活动预测

在该样本中

通过使用**逻辑回归** 进行填补，我们可以通过预测每种可能活动的概率，了解活动之间更复杂的关系；这些概率是样本时间窗口内各项活动的线性组合。具体而言，我们对时间t 的活动预测为

$$\underset{a\in A}{\operatorname{argmax}}\left(\mathbf{x}\cdot\mathbf{w}_a\right)$$

此外，我们可以方便地计算给定活动a 的似然，具体如下：

$$\frac{\exp(\mathbf{x}\cdot\mathbf{w}_a)}{\sum_{i\in A}\exp(\mathbf{x})\cdot\mathbf{w}_a}$$

其中，$\mathbf{x}$ 是在以t为中心、宽度为 $2k+1$ 的时间窗口内对活动进行的独热编码，而$\mathbf{w}_a$ 是与活动a对应的 $|A|\times 2k+1$ 权重矩阵。因此，权重的全集$\mathbf{W}$ 可解释为一个大小为$|A|\times|A|\times 2k+1$的张量。通过反向传播算法 [8], 对$\mathbf{W}$ 进行学习，使其最小化 [9] 中定义的负对数似然损失。该方法的一个重要特点是，最终得到的是标量似然值，因此可以对其应用阈值函数。

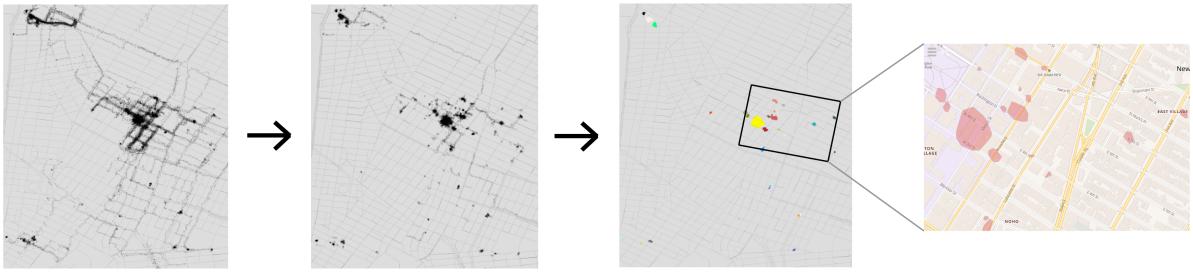
3.2.2 位置筛选与聚类

使用上述活动插补方法之一，我们可以根据位置测量发生时插补出的活动来筛选大多数位置测量值（尽管有少数测量值可能不在任何已观察到的活动范围k 内，此时便无法进行插补）。对于逻辑回归插补方法，我们可以根据某一时间活动对应的可能性来筛选位置。

本节仅筛选**静止** 状态下的位置测量值，因为这些测量值很可能是在用户进食时采集的。当然，这一规则也有例外，但我们将这些情况（即在移动途中进食）视为未观测事件。

为了识别一个人一天中的重要事件，我们必须通过聚类将经纬度空间离散化。由于*K-Means* 聚类要求簇必须是凸形且填满空间，因此不适用于这种情形。此类限制存在问题，因为它们不符合非凸的人类移动模式，还会给数据稀少或没有数据的区域分配标签，从而导致泛化能力较差。专为GPS数据设计的时间序列方法（例如 [6] 中所述的方法）或许也适用于聚类，但在实践中，对于我们涉及多次访问同一位置的场景，我们难以找到合适的参数。基于密度的聚类，即*DBScan*，是一种具有两个参数——*eps* 和*MinPts*[5]——的凝聚式聚类方法。DBScan以确定性方式构建簇$\mathbf{C}$，使其满足以下条件：$\forall p,q\in C_i,\forall C_i\in\mathbf{C}$

1. 若 $\mathrm{p}\in\mathrm{C}_i$，且存在一条由 n 个点组成的路径 p, q，其中 $\mathrm{p}_i,\ldots\mathrm{p}_n$, $\mathrm{p}_i=\mathrm{p}$, $\mathrm{p}_n=\mathrm{q}$，并且对于 $\mathrm{i}<\mathrm{n}$ς$(\mathrm{p},\mathrm{q})=|\mathrm{N}_E\mathrm{ps}(\mathrm{p}_i)|\geq\mathrm{MinPts}\wedge\mathrm{p}_{i+1}\in\mathrm{N}_E\mathrm{ps}(\mathrm{p}_i)$ 均为真，则 $\mathrm{q}\in\mathrm{C}_i$



(a) Raw GPS traces (b) After filtering by activity (c) After running DBScan (d) Integration of convex hulls into web-based map interface

图 2. 数据流图

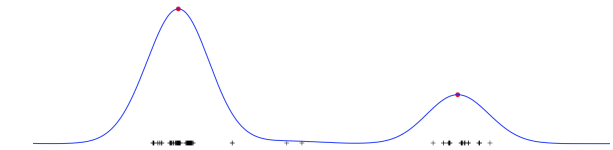


图 3. 通过核密度估计对 3 小时内的事件进行离散化

2. 注意到一般而言 $\varsigma(p, q) \neq \varsigma(q, p)$，如果 $p, q \in C_i$，那么$\exists o \in C_i$，使得 $\varsigma(o, p) \wedge \varsigma(o, q)$ 成立。

换句话说，上述条件确保$p, q \in C_i$中的所有点之间都存在一条路径，并且该路径符合由*Eps* 和*MinPts*设定的密度约束。*DBScan* 利用空间哈希算法提升效率，非常适合我们的用途，因为它会将低密度区域中的点舍弃为“噪声”，并且能够识别非凸聚类。需要注意的是，如果不先根据推断出的活动进行筛选，由于存在虚假模式（例如路径交叉处，用户经常沿路径行走，导致路径交叉点被识别为聚类），使用*DBScan*识别关键位置会变得困难。

通过*DBScan*确定簇后，接下来的步骤是按照第 3.2.3 节所述，进一步处理这些聚类位置的时间戳；或者计算各簇对应的凸包，以用于地图界面，并可能用于定义地理围栏。

从几何形态上看，该方法识别出的簇与实际的餐馆及其他常去位置高度吻合。然而，最终用于定义这些簇的点只是采集到的全部 GPS 轨迹中的一小部分。

3.2.3 精确确定事件发生的时间

算法识别出的特定“热点”在空间上由一组经纬度对界定。不过，它也对应着在该目的地发生的一次或多次离散访问，每次访问都会贡献一部分经纬度，这些经纬度共同构成整个聚类。识别用户一天中的关键事件的最后一步，是为事件的某次发生关联一个具体时间：这项任务类似于沿时间轴进行一维聚类。首先，为了生成观测脉冲的平滑版本，我们使用带宽为 2 小时的h 进行高斯核密度估计。高斯核密度估计根据测量值时间$\mathbf{X}$计算概率密度函数 $\rho(y)$

$$\rho(y) = \sum_{x \in \mathbf{X}} \exp\left(\frac{(y-x)^2}{h^2}\right) \tag{1}$$

函数 $\rho(y)$ 是平滑的，因此很容易找到某个时间间隔内的局部最大值集合。这些最大值用于将不同就诊期间产生的血糖时间序列对齐到同一位置。在我们的系统中，我们对血糖读数的 2 小时窗口进行对齐，从而能够计算与特定位置类型相关的血糖汇总统计数据。

事件识别框架中有多个超参数，例如 *eps* 和 *MinPts* 参数，以及 *DBScan* 算法的参数，还有密度估计步骤中的核宽度参数。这些超参数经过选择，用于识别以餐食频率发生的事件，而不是如上厕所等更频繁发生的事件。

3.3 理解连续血糖监测动态

本节讨论分析连续血糖监测设备生成的数据这一问题。我们发现，在适当的条件下，线性回归及其变体表现出色，优于其他方法。

3.3.1 用于预测的线性模型

一个用于预测单个变量（未来t 分钟后的血糖）的简单线性回归模型，会为一组特征选择权重w，以便最小化凸代价函数

$$f(\mathbf{X}, \mathbf{y}) = \left|\left|\mathbf{w}^\top \mathbf{X} - \mathbf{y}\right|\right|_2^2 \tag{2}$$

其中，$\mathbf{X}$ 是由$n \times m$ 构成的矩阵，表示n 个样本和m 个特征；$\mathbf{y}$ 是n 个响应组成的向量，$\mathbf{w}$ 是由m个权重组成的向量。当我们将过去k 个血糖读数作为特征时，权重$\mathbf{w}$ 的大小可以解释为每个过去测量值在估计值中的重要性。

估计值。式（1）的变体，如套索回归和岭回归，也分别通过$\mathbf{w}$ 的L_1 范数和L_2范数对权重进行惩罚，以避免权重学习随机噪声：

套索回归

$$f(\mathbf{X},\mathbf{y}) = \left|\left|\mathbf{w}^\top\mathbf{X} - \mathbf{y}\right|\right|_2^2 + \lambda||\mathbf{w}||_1 \quad (3)$$

L_1 惩罚项可提高$\mathbf{w}$的稀疏性。这有助于防止无关权重学习虚假噪声（即对应于很久以前的值的权重）。

岭回归

$$f(\mathbf{X},\mathbf{y}) = \left|\left|\mathbf{w}^\top\mathbf{X} - \mathbf{y}\right|\right|_2^2 + \lambda||\mathbf{w}||_2^2 \quad (4)$$

L_2 惩罚项限制权重$\mathbf{w}$的大小。

弹性网络

$$f(\mathbf{X},\mathbf{y}) = \left|\left|\mathbf{w}^\top\mathbf{X} - \mathbf{y}\right|\right|_2^2 + \alpha\left(||\mathbf{w}||_2 + \lambda||\mathbf{w}||_1\right) \quad (5)$$

弹性网络采用加权组合，将套索回归和岭回归中使用的损失函数结合起来。

全变差正则化

$$f(\mathbf{X},\mathbf{y}) = \left|\left|\mathbf{w}^\top\mathbf{X} - \mathbf{y}\right|\right|_2^2 + \lambda||\nabla(\mathbf{w})||_1\big) \quad (6)$$

总变差正则化会使权重 ∇（$\mathbf{w}$）的有限差分算子具有稀疏性。直观而言，对于此类时间序列应用，采用变差正则化先验是合理的，因为我们可以预期，测量值之间的时间间隔越短，它们就越相似。

上述模型各有优缺点。每个模型都定义了一个关于 $\mathbf{w}$、可高效求解的凸优化问题，但其缺点在于假设每个样本的观测值数量固定。

3.3.2 非参数核回归

为了处理不均匀采样的数据，我们采用 [10] 中所述的 Nadaraya 和 Watson [13] 核回归公式。我们希望在任意时间点计算缺失变量的期望值，而不对观测值的间隔作任何假设。

$$f(\mathbf{X},\mathbf{y}) = \sum\left(\frac{\mathbf{w}^\top\mathbf{x}_i}{\mathbf{w}^\top\delta_i} - y_i\right)^2 \quad (7)$$

其中，δ_i 与样本向量 $\mathbf{x}_i$ 大小相同；对于每个 $d_j \in \delta_i$ 和 $x_j \in \delta_i$

$$d_j = \begin{cases} 1 & x_j \text{ contains an observation} \\ 0 & \text{otherwise} \end{cases}$$

简而言之，δ_i 仅用于标记哪些时间槽包含测量值，而不论该测量值是什么。我们可以将式（7）的分母视为一个归一化项，它使我们能够从观测值数量可变的数据中有效学习。通过用 $\mathbf{w}^\top\cdot\delta_\mathbf{x}$ 进行归一化，我们可以确保 $\mathbf{y}$ 的估计值不会随观测值数量而变化。请注意，当 $\mathbf{w}$ 与向量 c δ_{obs}正交时，损失函数（6）会趋于无穷大，这意味着式（6）不是凸函数。$\mathbf{w}^\top\cdot\delta_\mathbf{x} = 0$ 可能发生的情形，与不同的 δ_i 数量一样多。因此，使用随机梯度下降和反向传播来学习$\mathbf{w}$是自然的选择，因为这些技术常用于非凸情形。近年来，反向传播已成功用于训练深度神经网络进行图像分类，但它也可用于通过任何可微目标函数学习权重。此外，这种方法的优势在于，随机梯度下降是一种在线算法，因此无需假设信号是平稳的。事实上，我们可以通过反向传播在线学习第 3.3.1 节中的线性回归模型的权重，不过收敛时间可能更长。这种方法的主要缺点是，它不具备*3.3.1*中所讨论的凸代价函数所提供的保证。

3.3.3 纳入外生特征

通过第3.2节中的技术得到的滤波标签，可以转换为独热编码，用于描述用户何时静止于经常访问的位置。接下来需要确定哪些位置对血糖预测具有重要意义。在本研究中，我们手动选取了部分属于饮食地点的簇，并将其视为“重要”地点（即熟食店、泰国菜、披萨），而忽略其余簇。这种方法存在一个问题：所有观测值都等于 1（即 True），因此式（6）中的独热编码归一化项与分子相同。为了将独热编码重新表示为二值观测值向量，我们将任何位置测量或购买都计为一次观测，即使该位置不被视为“相关事件”也是如此。一个值得进一步研究的方向是缺失值比例与式（5）相较于式（1）的性能提升之间的关系。

$$f(\mathbf{X},\mathbf{y}) = \sum\left(\frac{\mathbf{w_{bg}}^\top\mathbf{x_{bg}}_i + \mathbf{w_{exog}}^\top\mathbf{x_{exog}}_i}{\mathbf{w_{bg}}^\top\delta_{\mathbf{bg}_i} + \mathbf{w_{exog}}^\top\delta_{\mathbf{exog}_i}} - y_i\right)^2\Gamma \quad (8)$$

4 系统

该系统由网页服务器、移动端应用程序和 Dexcom 制造的外部连续血糖监测设备组成，负责数据收集、分析，并向用户报告信息。为了能够更自由地选择机器学习工具，所有分析工作负载都在远程而非用户设备上执行。

我们的系统能够向用户提供实时分析和预测是一项重要优先事项，但要实现这一点仍面临一些障碍。

- 有些数据（例如信用卡消费）不会实时更新

- 设备制造商对

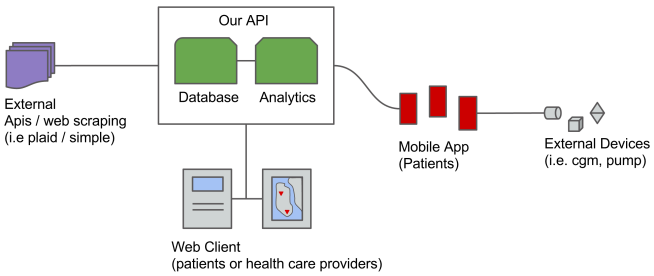


图 4. 数据流图

后台应用的使用有所限制，这是出于电池消耗和隐私方面的考虑

目前，所有软件都部署在两个位置之一：移动应用程序或网页服务器后端。

4.1 移动应用程序

大部分相关数据收集工作都是通过移动应用程序完成的。该应用程序运行于 iPhone 上，使用 Objective-C 编写。移动应用程序通过 Apple 的 HealthKit 框架访问连续血糖监测记录的数据，这些数据会通过蓝牙传输到手机。CoreLocation 框架用于接收频繁且高精度的位置更新，同时设置地理围栏以进行被动跟踪。

有趣的是，Apple 的硬件加速活动分类系统会自动将推断出的活动存储并记录在设备的本地数据库中，记录范围可追溯至过去 7 天。这意味着，应用程序需要在后台运行才能记录位置数据，而活动数据则会独立记录。这一特性在一定程度上解释了活动测量值相对于其他数据源为何更加频繁。应用程序启动后，会同步并上传自上次运行以来收集的活动数据。

4.2 位置跟踪

迄今为止，移动端数据收集系统面临的最大问题，是持续进行位置跟踪对手机电池续航的影响。尽管 Apple 已对位置数据的收集进行了大量优化，这也是位置数据并非均匀采样的原因之一，但启用跟踪时，电池续航时间不会超过 24 小时，有时甚至更短。GPS 坐标的采集频率取决于用户的移动速度，这很可能是因为对于大多数应用程序而言，静止状态下的位置更新并无用处。移动用户的位置更新频率约为每 10–15 秒一次。如果应用程序对用户设备的电池续航产生不利影响，整个系统的实用性就会降低，因此我们采用了更节能的地理围栏策略。不过，这意味着移动应用程序采用两种不同的数据收集模式：

主动跟踪 此模式会收集原始 GPS 轨迹，以形成新的簇或“热点”
被动跟踪，应用程序仅对进入和离开先前学习到的热点作出响应。

在 iOS 中实现被动跟踪相对直接，因为 CoreLocation 框架提供了地理围栏应用程序接口。Apple 将每个应用程序可激活围栏的数量限制为 20 个；根据我们的经验，这不足以跟踪所有已学习的簇。为解决这一问题，我们采用在设备本地实现的嵌套地理围栏策略，根据用户当前所在的较大超级围栏主动更新被跟踪的区域。

4.3 服务器

服务器会定期收集数据，例如在部分探索性分析中使用的借记卡食品消费。来自不同来源的数据意味着，必须投入大量工作来整理数据格式，以便将其导入机器学习组件、进行长期存储，并转换为适当格式，以便快速下载并填充前端可视化。服务器应用程序需要兼顾结构化数据库的灵活性，同时存储经过处理的数据副本，以便快速发送至客户端应用程序。每个测量值都会编入索引并存储在 MongoDB 实例中，以便在需要时执行复杂查询。然而，对于用户向服务器发出的多数数据请求，服务器返回的都是预处理过的压缩文件，其中已去除无关元数据。

4.4 数据处理

我们关注的许多数据并非按固定间隔报告，且来自不同来源；这意味着，有时只有通过匹配时间戳，才能匹配同时发生的测量值。依赖时间戳可能会引发问题，例如某个组件的系统时钟与其他组件不同步。我们的系统曾几次遇到这个问题。减少此类故障的一种可行方法，是要求由单个（软件）组件负责与服务器通信，而不是由设备上的多个组分别更新远程数据库。这样，数据到达数据库时就已按同时发生的条目分组，从而至少可以部分消除数据交叉引用这一故障点。

4.5 界面

应用程序包含两个用户界面：网页和移动端。移动应用程序主要用作数据收集平台，而非数据可视化工具包，但仍包含地图界面。网页应用前端是数据可视化的主要界面，并利用 MapBox 对 WebGL 的支持来为地图添加注释。借助这些工具，可以在瓦片地图界面上绘制 50,000+ 条线，而不会显著降低性能。通过展示用户记录路径的每一段，我们希望让体验更加有趣，并帮助唤起用户的记忆。我们利用基于浏览器的筛选功能，确定 GPS 信息的显示范围。

通过第3.2节所述方法确定的相关簇以凸包的形式绘制在地图界面上，并可响应用户交互。用户选择某个簇（即“热点”）后，即可查看该位置所有访问对应的汇总CGM统计数据。



图5. 地图界面显示在已识别的“热点”访问期间及之后的血糖均值和标准差

5 结果

我们使用第3.1节所述的数据集评估了一系列预测方法，并解读由此学习得到的模型。为了比较不同的预测方法，我们将测试集限定为仅包含完整观测值，以便对所有方法进行比较。对于核回归，训练集除了包含其他方法所使用的完整样本外，还包含不完整样本。我们表明，核回归能够利用不完整数据来提升预测性能。

5.1 活动预测

我们使用逻辑分类器，以近期活动窗口作为输入来插补活动。我们将各类别的 F 分数（1 对其余）与滚动均值和最近邻基线的分数进行了比较。我们还计算了每种插补器的总体准确率。总体而言，逻辑分类器优于两种基线，这主要是因为它能够学习乘车与静止乘车测量值频繁交替的规律。对于步行和骑行等较少见的测量值，最近邻的表现最佳。逻辑分类器系统的主要优势在于，它能够为给定时间的某项活动生成概率，这在第3.3.2节所述的筛选步骤中很有用。

5.2 事件识别

在第 3.2 节中，我们概述了一种利用普遍部署的传感器数据识别个人日常生活中离散事件的方法。为了证明这些事件与糖尿病护理和分析相关，我们计算了特定类型的生活方式事件发生与某一时间间隔内血糖升高超过 30毫克/分升这一事件之间的皮尔逊相关系数。我们比较了不同时间间隔和不同事件类型的相关性。不出所料，相关性最高的一组事件是通过人工参照聚类边界与已知和食物摄入相关的位置而选出的那组事件。虽然尚不清楚为什么一般的生活方式事件发生与加速度计检测到的“静止”行为实例相比，相关性更高，但可能的解释包括

倾向于重复事件 受试者可能更倾向于在曾经去过的地方吃东西或摄入糖分，而不是在新的位置。第3.2节所述的生活方式事件筛选方法基于密度，因此不太可能识别用户首次访问新位置的情况。如果用户确实习惯于重复日常行为，那么所述方法更适合用于检测其摄入食物的时间。

应用程序激活偏差 只有同时具备活动估计和位置数据，才能测量生活方式事件。无论应用程序是否正在运行，都会进行活动估计；但要追踪位置，应用程序必须在后台运行。用户可能会倾向于在摄入食物时确保应用程序处于活动状态，因为这很可能是研究中我们试图追踪的最重要事件，因此位置测量是否存在与血糖升高之间可能因此出现轻微相关性。

此外，从图8中可以看到，最大相关性出现在检测到事件约60分钟后。这是合理的，因为食物摄入导致的血糖升高至少需要10–15分钟才能被连续血糖监测检测到，而且某些类型的食物可能会使血糖在之后数小时内逐渐升高。

5.3 血糖预测

对数据进行了预处理，使其具有零均值和单位方差。根据我们的经验，线性回归模型的预测表现出乎意料地稳健；无论是将滞后值用作特征，还是在支持向量机中使用斜率和曲率等特征，其表现都优于支持向量机。此外，在各类线性模型中，全变差正则化略占优势。非参数核回归表现良好，但仅使用过去血糖数据的模型（7）与使用外生信息的模型（8）在性能上没有差异。总体而言，在预测任务中使用外生数据源颇具挑战，且往往会导致性能下降。因此，我们建议将第3.2节所述特征用于帮助用户理解其生活方式、饮食习惯与血糖动态之间的关系。图8表明，这种关系确实显著，尤其是在被标记为潜在食物地点的热点中。

6 结论

我们介绍了一种系统，可在采集血糖测量值的同时收集情境数据，并介绍了一种从这些数据中提取有用特征的方法。此外，我们

活动	样本模式	最近邻	逻辑分类器	
总体准确率	0.78	0.79	**0.87**	
静止	0.89	**0.96**	0.95	
乘车	0.23	0.07	**0.83**	
静止乘车	0.12	0.02	**0.46**	
步行	0.15	**0.44**	0.35	
骑行	0.13	**0.53**	0.34	

图6. 各插补器在各活动类别上的 F分数（不包括跑步，因为测试集中的案例数量太少）

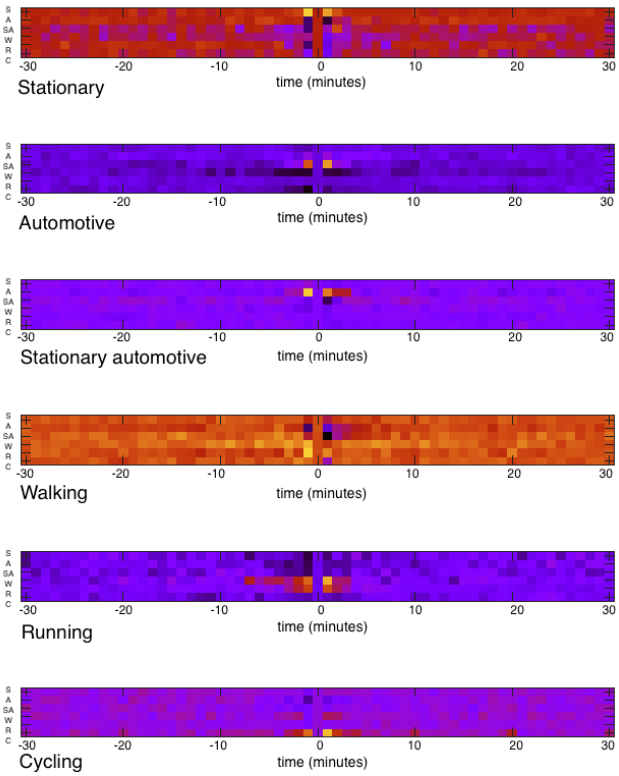


图7. 用于插补活动的权重 W。请注意，乘车信号会过度赋予静止乘车信号作为预测变量的权重，反之亦然。这是因为它们经常交替出现。

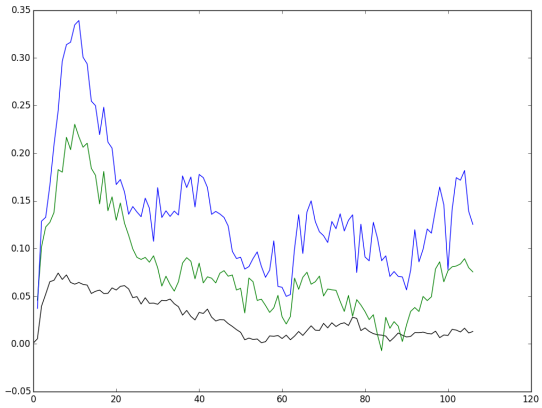


图8. 不同非血糖信号随时间与血糖的相关性："静止"测量值（黑色）、生活方式事件发生（蓝色），以及根据聚类是否对应食物消费地点进行筛选后的生活方式事件发生（绿色）。

时间跨度	静态	线性回归	套索回归 $\lambda = 1.0$	岭回归
15	14.3	10.42	10.43	10.42
30	25.86	20.5	20.49	20.49
60	43.52	37.64	37.63	37.64

弹性网络 $\alpha = 1.0, \lambda = 0.5$	TV 正则化	非参数核回归
10.43	-	**10.42**
20.49	20.4	**18.04**
37.64	-	**31.89**

图9. 血糖预测的交叉验证结果

还介绍了使用线性模型在不同时间范围内进行血糖预测的基准结果。这项工作的主要目的是展示同时收集情境数据和血糖数据所带来的可能性。"事件"特征能够将血糖统计数据与常去地点具体关联起来。我们希望，随着数据量增加，本文所述的此类情境特征能够为预测与控制系统作出贡献。

7 参考文献

[1] R. Bunescu、N. Struble 和 C. Marling。利用生理模型和支持向量回归预测血糖水平。国际机器学习与应用会议，2013年。

[2] P.-G. C、F. A、S. G、C. C、G. EJ、R. M、de Leiva A 和 H. ME。用于根据连续血糖监测进行在线血糖预测的人工神经网络算法。《糖尿病科技与治疗》，2010年。[3] D. E、C. F、L. H、B. BW、Z. H、J. L、C. HP、W. DM、B. BA 和 D. F. 3rd。利用连续血糖监测的实时低血糖预测套件：人工胰腺的安全网。《糖尿病护理》，2010年。[4] M. Eren-Oruklu、A. Cinara、D. K. Rollins 和 L. Quinn。用于估计未来葡萄糖浓度和低血糖警报的自适应系统辨识。《自动化》，2012年。[5] M. Ester、H.-P. Kriegel、J. Sander 和 X. Xu。一种基于密度的 al-

在含噪声的大型空间数据库中发现簇的算法。知识发现与数据挖掘，1996年。[6] J. H. Kang、W. Welbourne、B. Stewart 和 G. Borriello。从位置轨迹中提取地点。第2届美国计算机协会 *WLAN*热点上的无线移动应用与服务国际研讨会论文集。[7] T. Kanungo、D. M. Mount、N. S. Netanyahu、C. D. Piatko、R. Silverman 和 A. Y. Wu。一种高效的k均值聚类算法：分析与实现。电气电子工程师学会模式分析与机器智能汇刊，2002年。[8] Y. LeCun、L. Bottou、G. B. Orr 和 K.-R. Muller。高效反向传播。技术报告，AT&T实验室图像处理研究部。[9] Y. LeCun、S. Chopra、R. Hadsell、M. Ranzato 和 F. J. Huang。基于能量的学习教程。即将发表于《结构化数据预测》，2006年。[10] N. Razavian 和 D. Sontag。用于根据实验室检验结果进行诊断的时间卷积神经网络。工作论文，2015年。[11] R. S. Sherwin、K. J. Kramer、J. D. Tobin、P. A. Insel、J. E. Liljenquist、M. Berman 和 R. Andres。人体胰岛素动力学模型。《临床研究杂志》，1974年。[12] N. W. Struble。使用机器学习回归模型测量血糖变异性并预测血糖水平。硕士论文，俄亥俄大学，2013年。[13] G. S. Watson。平滑回归分析。《印度统计学杂志》，1964年。