基于多段间隔监督度量学习的病人相似度算法①

李世强[1,2], 倪嘉志[1,2], 刘　杰[1], 叶　丹[1]

[1](中国科学院软件研究所 软件工程技术研发中心, 北京 100190)
[2](中国科学院大学, 北京 100190)

摘　要: 伴随着医疗卫生服务的信息化进程推进, 病人相似度成为了医疗电子健康数据的二次利用中的重要问题. 在已有医疗专家对病人健康数据的评估信息下, 可以将病人相似度问题转化为有监督的距离度量学习问题. 通常的做法是对病人的医疗健康数据打标签来作为监督信息. 在现有的病人相似度计算工作中, 对监督信息的利用是很局限的; 多是比较两个不同病人的标签是否完全相等来判断病人相似与否; 在实际中, 病人的标签往往是多个维度, 这种比较忽略了标签本身的相似性. 本文将病人的诊断数据作为监督信息, 在度量学习中, 根据标签的相似程度将目标病人的邻居区分开来, 形成多段间隔, 更充分地利用监督信息. 在基于多标签的 KNN 分类评估实验中, 该算法学习出的相似度度量在 Hamming Loss 和 *a*-Accuracy 两种指标下性能有很大提升.

关键词: 电子健康记录; 病人相似度; 监督距离度量学习; 多标签分类

Patient Similarity Based on Supervised Metric Learning of Multi-Margin

LI Shi-Qiang[1,2], NI Jia-Zhi[1,2], LIU Jie[1], YE Dan[1]

[1](Software Engineering Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: With the development of medical and health services informatization, patient similarity becomes an important task in reuse of Electronic Health Records (EHR). By using the physician feedback on EHR data, patient similarity problem can be transformed to supervised distance metric learning problem, the supervised information usually comes from the tags we make on one patient's EHR data. In the existing work of Patient similarity Computing, the utilization of supervised is pretty circumscribed, the similarity of two different patients is often depended on their EHR data tags' completely equality. But in fact, the patient's tags contains many dimensions, that methods ignores tags' own similarity. In this work, we use the patient's diagnose data as the supervised information and divide the target patient's neighbor area into many margins based on their similarity using metric learning. The supervised information is also more fully used in this algorithm. Finally, in the multi-label KNN classification evaluation experiment, the similarity metric learned from this algorithm performs better than other algorithms in Hamming Loss and *a*-Accuracy.

Key words: electric health record; patient similarity; supervised metric learning; multi-label classification

近年来, 伴随着医疗卫生服务的信息化进程推进, 产生了大量的医疗电子健康数据. 数据内容主要包括来自医院的电子病历、区域卫生信息平台采集的居民健康档案等, 其中大量充斥着非结构化/半结构化的数据, 包括图像, office 文档, 以及 XML 结构文档等. 如何合理高效地二次利用这些数据具有重要的现实意义. 这些数据的一个常见使用场景是: 医疗专家通过检索引擎访问病人的健康数据来获得病人的健康信息概览或者查询特定的细节问题. 然而, 基于关键字的精确或模糊匹配查询的有效性依赖于用户提交的查询短语, 且聚焦在单个病人的数据上, 更适用于结构化数据; 没有能够利用复杂电子健康数据中潜在的大量医疗知

① 基金项目:国家自然科学基金(U1435220);军队后勤科技项目(AWS4R013)
收稿时间:2016-03-07;收到修改稿时间:2016-04-08　[doi:10.15888/j.cnki.csa.005444]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

识[1,2]. 而基于病人相似度的案例查询就成了重要的技术补充.

案例查询中的核心部分便是病人相似度度量. 病人相似度度量的目标是根据病人的关键医疗特征来捕捉和量化病人间的相似性. 除了案例检索, 病人相似度还可以应用在病人群体识别, 治疗方案的比较研究, 以及病人风险分级等诸多领域.

除了利用已有的距离度量来直接计算病人相似度[3,4], 还可以将病人相似度问题转化成有监督的距离度量学习问题, 从而利用医疗专家的专业反馈信息. Wang Fei, Jimeng Sun 等人将 LMNN[5]算法的思想应用到病人相似度上, 提出了局部监督的病人相似度度量学习算法(LSML[6]). 该算法利用医疗专家对每一个病人所打的标签作为监督信息, 通过比较标签是否相等, 来识别出以某个病人为中心的一定范围内的同构邻居(homogeneous neighborhood)和异构邻居(heterogeneous neighborhood); 再通过拉近(pull)同构邻居, 推远(push)异构邻居来学习出一种泛化的 Mahalanobis 距离.

然而, 该算法简单地比较标签是否相同, 忽略了不同标签间也具有相似性这一重要信息; 此外, 对于已有的医疗健康数据, 医疗专家再次人工地评估病人间的相似性或者为病人打标签将耗费巨大的工作量, 如何从已有的医疗数据中获取监督信息也是一个关键且有实际意义的问题.

基于此, 本文提出了一种改进的病人相似度度量学习算法. 该算法从已有的医疗健康数据中获取监督信息来构造病人标签, 并将不同标签间的相似性考虑进来, 根据不同的标签相似性来产生多个间隔, 从而学习出一种泛化的 Mahalanobis 距离. 实验证明, 该算法学习出的相似度度量能有效地提升 KNN 分类器的准确率. 并且易于扩展应用到不同的场景.

1 数据与病人表示模型

医疗健康数据的特点是数据结构复杂、数据量大、且很多指征是随着时间变化的. 单个病人的健康信息淹没在这数据海洋之中, 由于数据量以及数据结构的不同, 无法直接在原始数据上进行运算. 因此进行度量学习之前, 需要采取一种合适的模型来对病人进行表示, 以期保留足够的医疗信息并且能够适应于不同的相似度算法.

对病人的表示通常可以采用基于向量(vector)的方法, 或者基于矩阵的(matrix)方法. 后者实际上是前者在时间序列上的细化, 矩阵的行方向是时间序列, 矩阵的每一列都对应一个向量. 以住院数据为例, 基于向量的表示将病人的多次住院记录合并, 对同一指征在多次住院记录中的值进行统计; 而基于矩阵的表示则将多次住院记录的值分别保留. 显然基于矩阵的表示能够更多地保留信息, 并且能够反映病人信息在时间纵向上的变化趋势. 但是该模型的复杂性也意味着其需要更复杂的相似度算法, 比如序列相似度算法, 以及更高的计算代价. 在实际应用中, 基于向量的表示在可扩展性上有着巨大的优势, 可以很容易地应用于各种机器学习算法或相似度算法上, 如 KNN, SVM 等. 综合利弊, 本文采取基于向量的方法来从病人的基本信息, 用药信息, 化验信息, 诊断信息等信息中提取特征来表示病人.

因此, 每一个病人即是一个特征向量. 病人的原始医疗健康信息中的各种指征往往无法直接应用于特征向量, 如血细胞是化验数据中的一个原始指征, 但是血细胞出现在不同的化验样本中, 包含有不同的医疗信息且度量标准也往往不同: 在血液样本中, 需要检测血细胞浓度; 而在尿液样本中, 只需检测血细胞出现与否. 因此, 原始指征血细胞无法作为一个单独的特征, 需要与不同的样本结合起来. 预处理工作中的一个很大挑战就是从原始指征中提取和转化特征.

针对不同特征的数据类型, 采取不同的方式来记录其值. 对于静态特征, 如年龄, 性别, 种族等, 使用固定的值或编码来表示; 对于时序的数值特征, 如血细胞在血液样本中的浓度, 采用其统计值(平均值, 方差, 中位数)来表示; 对于时序的离散特征, 比如血细胞在尿液样本中出现与否, 对其进行统计计数. 通过上述方法获得的病人特征向量往往具有超高的维数, 通常还需要通过特征选取方法来进行降维.

2 监督病人相似度学习算法

病人相似度的目标是选取合适的度量来刻画病人在特定医疗场景下的相似性. 有诸多的距离度量可以用来计算用以表示病人的特征向量间的相似度, 如欧式距离, Mahalanobis 距离等. 然而, 病人的健康数据复杂且包含有专业信息, 直接应用这些度量并不能很好地体现出病人健康数据在医疗背景下的意义, 一种直观的想法是引入专家对病人相似度的人工判断, 来

© 中国科学院软件研究所 http://www.c-s-a.org.cn

帮助计算相似性. 因此, 病人相似度计算问题可以转化成有监督的距离度量学习(Supervised Distance Metric Learning)问题. 在该问题背景下, 需要解决两个重点问题:

① 如何获取病人相似度的监督信息?

② 如何利用监督信息学习出来一种新的相似度度量?

2.1 病人标签相似度

通常来讲, 监督信息需要专家人工给出. 特定到病人相似度问题, 对于一份包含有个病人的训练数据, 医疗专家需要对个病人对之间的相似度进行评估, 这种人工开销是巨大的, 且评估效果是不稳定的. Wang Fei[6], Jimeng Sun[6]等人采取了打标签的方式, 如此, 医疗专家只需要对 N 个病人分别进行标签分类, 在后续计算中, 通过判断不同病人间的标签是否相等来断定其是否相似. 这样虽然可以大大减少工作量, 但是病人的医疗健康数据的复杂性不是一个简单的标签就可以概括的, 且简单地比较标签的相等性, 只能判断病人在该标签下是否严格相等, 丢失了病人的相似程度这一重要信息.

在典型的医疗场景中, 病人的基本信息, 各种化验数据, 监测数据, 用药数据等, 都是医疗专家做出诊断的依据. 也就是说在病人的医疗健康数据中, 诊断信息即是医疗专家对病人的整体情况所做的人工评估. 因此, 可以将病人的诊断信息转化为病人标签, 从而作为度量学习中的监督信息.

在一份典型的病人医疗健康数据中, 诊断信息往往出现多次, 即一个病人对应多种诊断. 按照规范, 诊断信息通常使用 ICD-10 进行编码. 令 C 表示 ICD-10 编码全集. y_a 和 y_b 分别表示病人 a 和 b 的标签, 则一种可能的情况是 $y_a=\{c_1,c_2,...,c_m\}$, 其中 $c_i\in C$ 且 $1\le i\le m$; 同理, $y_b=\{c_1,c_2,...,c_n\}$, 其中 $c_j\in C$ 且 $1\le j\le n$.

在 Wang Fei[6], Jimeng Sun[6]等人的工作中, 由于病人只有一个单值标签, 只能通过比较标签是否相等来作为相似度学习的监督信息, 而这并不符合病人相似度的实际情况, 因为病人的信息不能仅通过一个标签进行概括, 也无法简单的通过比较标签的相等性来作为相似与否的先验条件. 而在有了更复杂的标签信息后, 可以计算标签的相似程度来作为相似度学习的监督信息. 在上述表示形式下的标签, 其相似度可采用 Jacacard 系数来计算:

$$J(y_a,y_b)=\frac{|y_a\cap y_b|}{|y_a\cup y_b|} \tag{1}$$

其中, y_a 和 y_b 分别表示病人 a 和 b 的标签信息, $0\le J(y_a,y_b)\le 1$.

2.2 病人相似度学习

在获得相似度的监督信息后, 病人相似度问题可以转化为有监督的距离度量学习问题. 首先对距离度量的基本概念进行定义.

定义 1. 假设 χ 是数据点的集合, 对任意的向量 $\forall \vec{x}_i,\vec{x}_j,\vec{x}_k\in\chi$, 将满足以下 4 个条件的映射 $D:x\times x\to R$ 称作为距离度量(Distance Metric):

a) 非负性: $\mathrm{D}(\vec{x}_i,\vec{x}_j)\ge 0$;

b) 一致性: $\mathrm{D}(\vec{x}_i,\vec{x}_j)=0\Leftrightarrow \vec{x}_i=\vec{x}_j$;

c) 对称性: $\mathrm{D}(\vec{x}_i,\vec{x}_j)=\mathrm{D}(\vec{x}_j,\vec{x}_i)$;

d) 次可加性: $\mathrm{D}(\vec{x}_i,\vec{x}_j)+\mathrm{D}(\vec{x}_j,\vec{x}_k)\ge \mathrm{D}(\vec{x}_i,\vec{x}_k)$;

当放宽一致性条件使得 $\vec{x}_i=\vec{x}_j\Rightarrow \mathrm{D}(\vec{x}_i,\vec{x}_j)=0$, 这时称 D 为伪度量(Pseudo Metric). 为了简化后续讨论和工作, 文中使用伪度量来作为距离度量, 只在必要的时候指出其与严格度量之间的区别.

在对原始数据做一个线性映射 $\vec{x}'=L\vec{x}$ 后, 原始数据被映射到一个新的空间, 在该空间上计算欧式距离, 可以得到一种新的距离度量, 该度量可以表示为:

$$\mathrm{D}_L(\vec{x}_i,\vec{x}_j)=\left\|L(\vec{x}_i-\vec{x}_j)\right\|_2 \tag{2}$$

该线性转化是由公式中的对角矩阵 L 来定义, 很容易对公式(2)进行变形. 定义如下矩阵:

$$M=LL^T \tag{3}$$

显然, M 是一个半正定矩阵, 根据 M 可以重写公式(2):

$$\mathrm{D}_M(\vec{x}_i,\vec{x}_j)=\sqrt{(\vec{x}_i-\vec{x}_j)^T M(\vec{x}_i-\vec{x}_j)} \tag{4}$$

这种形式的伪度量即是泛化的 Mahalanobis 度量. 令 $X=\left[\vec{x}_1,\mathrm{L},\vec{x}_j\right]\in R^n$ 来表示病人集合的特征矩阵, $Y=[\vec{y}_1,...,\vec{y}_n]^T$ 来表示相应的标签向量, 其中 $y_i\subseteq\{c_1,c_2,...,c_m\}$ 表示病人 $\vec{x}_i$ 的标签, 需要强调的是 y_i 本身也是一个集合. 我们的目标即是学习出如公式(4)的泛化 Mahalanobis 度量, 在 Kilian Q. Weiberger[5]以及 Wang Fei[6], Jimeng Sun[6]等人的工作中, 由于监督信息仅是单个标签, 因此其思路是在训练过程中将标签相同的邻居拉近, 将标签不同的邻居推远, 从而学习

© 中国科学院软件研究所 http://www.c-s-a.org.cn

出一个间隔. 该过程的示意如图 1 所示. 而在标签本身也具有相似性的情况下, 这种无区别地将所有标签不同的邻居同等对待, 是不合理的. 一个启发式的方法是根据标签的相似性程度, 学习出多种间隔. 为简单起见, 将标签的相似度粗略的分成三个层次: 1)完全相同; 2)部分相同; 3)完全不同, 从而可以学习出两段间隔, 该过程如下图 2 所示; 这种简化是为了更聚焦在算法核心部分. 可以很容易的证明, 该算法可以扩展以支持更细粒度的标签相似度划分.

根据上述标签相似度的等级, 可以将某个病人实例 $\vec{x}_i$ 的 k 个邻居分成三类. 为简单起见, 借用论文[6]中的部分术语, 将这三类邻居分别定义如下.

定义 2. 同构(homogeneous)邻居: 记为 N_i^o, 是与 $\vec{x}_i$ 具有相同标签的、最近的 N_i^o 个邻居.

定义 3. 相似(similar)邻居: 记为 N_i^s, 是与 $\vec{x}_i$ 具有部分相同标签的、最近的 N_i^s 个邻居.

定义 4. 异构(heterogeneous)邻居: 记为 N_i^e, 是与 $\vec{x}_i$ 不具有任何相同标签的、最近的 N_i^e 个邻居.

在上述定义中, $|\bullet|$ 表示集合元素的个数. 将 $\vec{x}_i$ 的邻居分好类后, 分别计算各个类别到 $\vec{x}_i$ 的距离的平方和, 分别如下:

$$O_i = \sum_{j:\vec{x}_j \in \Psi_i^o} \mathrm{D}_M\left(\vec{x}_i, \vec{x}_j\right)^2 \tag{5}$$

$$S_i = \sum_{k:\vec{x}_k \in \Psi_i^s} \mathrm{D}_M\left(\vec{x}_i, \vec{x}_k\right)^2 \tag{6}$$

$$E_i = \sum_{l:\vec{x}_l \in \Psi_i^e} \mathrm{D}_M\left(\vec{x}_i, \vec{x}_l\right)^2 \tag{7}$$

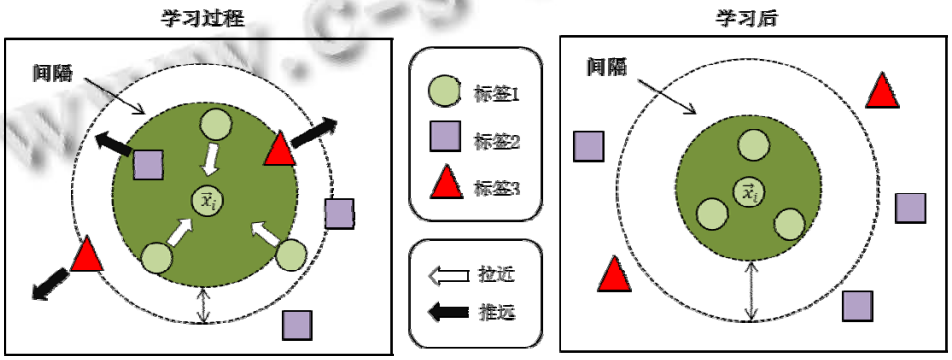

图 1 不考虑标签相似度的学习过程, 标签 2 与标签 1 具有更高的相似度, 但在该学习过程中, 由于只计算标签相等性, 标签 2 与标签 3 被同等对待, 从而学习出一个间隔.

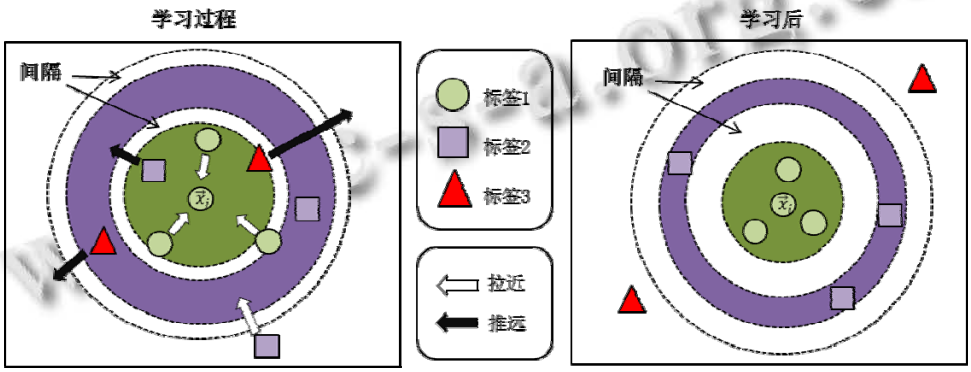

图 2 考虑标签相似度的学习过程, 标签 2 与标签 1 具有更高的相似度, 因此在该学习过程中, 标签 2 与标签 3 被区别对待, 带有标签 3 的样本被推出地更远, 从而学习出两个间隔.

当前的目标是最大化两段间隔, 这两段间隔可分别表示为:

$$G_1 = \sum_{i=1}^{n}\left(S_i - O_i\right) \tag{8}$$

$$G_2 = \sum_{i=1}^{n}\left(E_i - S_i\right) \tag{9}$$

因此, 优化目标可转化为最小化如下公式:

© 中国科学院软件研究所 http://www.c-s-a.org.cn

$$\begin{aligned} J &= -\mu G_1 - (1-\mu) G_2 \\ &= \sum_{i=1}^{n} \mu (O_i - S_i) + (1-\mu)(S_i - E_i) \end{aligned} \tag{10}$$

其中, $\mu \in [0,1]$ 是权重参数, 该参数的值可以通过交叉验证来进行调优. 接下来对公式(10)进行展开, 需要用到如下矩阵迹运算的性质:

(1) $tr(x \cdot x^T) = x^T \cdot x$

(2) $tr(A+B) = tr(A) + tr(B)$

(3) $k \cdot tr(A) = tr(k \cdot A)$

其中 x 是列向量. 根据公式(3), 公式(4)和公式(5), 可对 O_i 做如下变换:

$$\begin{aligned} O_i &= \sum_{j:\vec{x}_j \in ¥_i^o} \mathrm{D}_{\boldsymbol{M}} (\vec{x}_i, \vec{x}_j)^2 \\ &= \sum_{j:\vec{x}_j \in ¥_i^o} (\vec{x}_i - \vec{x}_j)^T L L^T (\vec{x}_i - \vec{x}_j) \\ &= \sum_{j:\vec{x}_j \in ¥_i^o} tr\left(L^T (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T L \right) \\ &= tr\left(L^T \cdot \sum_{j:\vec{x}_j \in ¥_i^o} (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T \cdot L \right) \end{aligned} \tag{11}$$

同理, 对 S_i, E_i 做同样的变换可以得到:

$$S_i = tr\left(L^T \cdot \sum_{k:\vec{x}_k \in ¥_i^o} (\vec{x}_i - \vec{x}_k)(\vec{x}_i - \vec{x}_k)^T \cdot L \right) \tag{12}$$

$$E_i = tr\left(L^T \cdot \sum_{l:\vec{x}_l \in ¥_i^o} (\vec{x}_i - \vec{x}_l)(\vec{x}_i - \vec{x}_l)^T \cdot L \right) \tag{13}$$

将公式(11), (12), (13)代入公式(10)化简后可得:

$$J = tr\left(L^T \cdot \left(\mu \Sigma_o + (1-2\mu)\Sigma_s + (\mu - 1)\Sigma_e \right) \cdot L \right) \tag{14}$$

其中,

$$\Sigma_o = \sum_{i=1}^{n} \sum_{j:\vec{x}_j \in ¥_i^o} (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T \tag{15}$$

$$\Sigma_s = \sum_{i=1}^{n} \sum_{k:\vec{x}_k \in ¥_i^o} (\vec{x}_i - \vec{x}_k)(\vec{x}_i - \vec{x}_k)^T \tag{16}$$

$$\Sigma_e = \sum_{i=1}^{n} \sum_{l:\vec{x}_l \in ¥_i^o} (\vec{x}_i - \vec{x}_l)(\vec{x}_i - \vec{x}_l)^T \tag{17}$$

因此, 该距离度量学习问题可以形式化为给定权重参数 $0 \le \mu \le 1$, 求:

$$\min_{L:L^T L = I} J \tag{18}$$

其中, J 是公式(14), 正交性限制 $L^T L = I$ 是为了减少 L 的信息冗余. 为了简化计算, 根据论文[6], 定义以下三个对称方阵.

定义 5. 同构邻接矩阵 H^o 是一个 $n \times n$ 的方阵, 其中方阵中的元素

$$h_{ij}^o = \begin{cases} 1, & x_j \in N_i^o 或 x_i \in N_j^o \\ 0, & 其他 \end{cases} \tag{19}$$

定义 6. 相似邻接矩阵 H^s 是一个 $n \times n$ 的方阵, 其中方阵中的元素

$$h_{ij}^s = \begin{cases} 1, & x_j \in N_i^s 或 x_i \in N_j^s \\ 0, & 其他 \end{cases} \tag{20}$$

定义 7. 相似邻接矩阵 H^e 是一个 $n \times n$ 的方阵, 其中方阵中的元素

$$h_{ij}^e = \begin{cases} 1, & x_j \in N_i^e 或 x_i \in N_j^e \\ 0, & 其他 \end{cases} \tag{21}$$

接下来定义 $g_i^o = \sum_j h_{ij}^o$, 定义 $G^o = diag(g_1^o, g_2^o, ..., g_n^o)$; 同理定义 $g_i^s = \sum_j h_{ij}^s$, $G^s = diag(g_1^s, g_2^s, ..., g_n^s)$; 定义 $g_i^e = \sum_j h_{ij}^e$, $G^e = diag(g_1^e, g_2^e, ..., g_n^e)$. 定义 Laplacian 矩阵 1^o, 1^s, 1^e 分别如下:

$$1^o = G^o - H^o \tag{22}$$

$$1^s = G^s - H^s \tag{23}$$

$$1^e = G^e - H^e \tag{24}$$

根据上述定义, 可将公式(5), (6), (7)进行转换; 具体推导过程参见文献[6].

$$O = \sum_{i=1}^{n} O_i = 2tr\left(L^T X 1^o X^T L \right) \tag{25}$$

$$S = \sum_{i=1}^{n} S_i = 2tr\left(L^T X 1^s X^T L \right) \tag{26}$$

$$E = \sum_{i=1}^{n} E_i = 2tr\left(L^T X 1^e X^T L \right) \tag{27}$$

根据上述定义, 公式(18)可以转换为:

$$\min_{L^T L = I} tr\left(L^T \cdot X \cdot \left(\mu 1^o + (1-2\mu) 1^s + (\mu - 1) 1^e \right) \cdot X^T \cdot L \right) \tag{28}$$

根据 Ky Fand 定理[7], 上述公式有解析解, 其解的形式为 $L^* = [l_1, l_2, ..., l_d]$, 其中 $l_1, l_2, ..., l_d$ 是矩阵 $X \cdot \left(\mu 1^o + (1-2\mu) 1^s + (\mu - 1) 1^e \right) \cdot X^T$ 的负特征值所对应的特征向量.

2.3 算法步骤描述

在上述两段间隔的设定下, 该多间隔监督距离度量学习算法的完整过程如下所示.

输入: 数据特征矩阵 X, 监督信息 Y, 同构邻居数量 N_i^o, 相似邻居数量 N_i^s, 异构邻居数量 N_i^e, 间隔权重参数 μ.

步骤: 1.根据公式(19), (20), (21)分别构造邻接矩阵 H^o, H^s, H^e;

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.根据公式(22), (23), (24)分别构造 Laplacian 矩阵 1^o, 1^s, 1^e;

3.计算出矩阵 $X\cdot\left(\mu 1^o+(1-2\mu)1^s+(\mu-1)1^e\right)\cdot X^T$ 的所有 d 个负特征值, 按大小排序后, 所对应的特征向量;

4.将 L 设置为上述特征向量构成的 d 维矩阵.

3 实验与评估

相似度的人工评估成本高, 开销大, 且评估结果不稳定. 本实验采用间接评估的方式, 通过评估学习出来的相似度度量的应用来间接评估该相似度度量. 具体来说, 在特定的医疗背景下, 病人相似度度量可应用于多标签的KNN分类器, 来预测病人在该医疗背景下, 患某些疾病的概率, 并通过相应的指标来评估预测效果.

3.1 实验数据

本实验采用的数据来自合作医院的 2009-2014 年 5 年间的住院数据, 在经过去隐私化, 清理过滤后, 共提取出 4669 个病人案例. 从这些病人的基本数据、化验数据、用药数据、手术数据中提取特征, 共提取出 3068 个特征, 在这些特征上, 我们进行了最基本的特征提取, 具体做法是过滤掉方差近似为 0 的特征, 这种特征所含信息量极少. 由于数据在高维空间下的稀疏性, 有较多特征的方差近似为 0, 最后共保留下 1036 个特征.

3.2 监督信息

理论上相似度度量的监督信息应该是病人相似度的度量矩阵, 该矩阵记录了由专家人工给出的两两病人间的相似度的值. 然而在实现中, 对用于训练的病人语料进行两两相似度的人工评估是一项开销巨大的工作, 另外为了保证监督信息的正确性, 往往需要综合多位专家的结果. 在 Wang Fei[6], Jimeng Sun[6]等人的工作中, 采取了一种简化的方式, 只人工地对每个病人打标签, 在算法学习过程中, 通过比较标签是否相等来作为监督信息. 这种方式虽然避免了人工进行两两病人相似性比较所带来的巨大开销, 但是病人的总体情况往往是无法通过一个单维度的标签就能够概括地, 且病人是否相似也并不能通过比较单个标签来判断, 这就使得监督信息缺乏足够的信息量. 另外, 即使只进行人工打标签, 其工作量也不容忽视.

然而, 在病人健康数据中, 诊断信息是专家根据病人当前情况所做出的阶段性结论. 从这个角度来看, 诊断信息即是病人某一时期的标签, 病人的所有诊断信息构成了病人总体情况的总结. 因此, 一个启发式的思路是在假设医疗专家对病人的诊断正确的情况下, 该诊断即是专家人工所做的标签. 无需再另外的人工进行打标签. 这样, 每个病人的多个诊断形成了对该病人多维度的总结, 这种总结比单个标签更具有信息量.

本文选取冠心病作为特定的医疗背景, 根据国际 ICD-10 编码规范, 将冠心病(缺血性心脏病)分为 6 个类目: 心绞痛(I20)、急性心梗(I21)、随后性心梗(I22)、心梗的某些近期并发症(I23)、其他急性缺血性心脏病(I24)和慢性缺血性心脏病(I25). 文中将非冠心病的其他疾病统归为一类. 这样, 可将病人的诊断数据中的疾病编码共分为 7 类, 即 7 种标签. 每个病人可以有多次诊断, 即有多种疾病标签. 对每个病人的疾病标签集合进行编码, 即可形成病人相似度学习的监督信息.

3.3 评估指标

本实验通过评估多标签分类的准确性来间接反映相似度度量的有效性. 目前已有多种多标签分类问题的评估标准, 较主要的有 Hamming Loss[8], One-Error[9], Coverage[9], 和 α -Accuracy[10]. 设测试样本集为 $S=\{(x_1,Y_1),(x_2,Y_2),...,(x_m,Y_m)\}$, 其中 Y_i 是 x_i 的标签集合, $1\le i\le m$, 且 $Y_i\subseteq Y=\{c_1,c_2,...,c_n\}$; $H:x\to y$ 表示假设函数, 令 $\hat{Y}_i$ 为 x_i 的预测的标签集合; $f:x\times y\to R$, $f(x_i,c)$ 表示将 x_i 的标签预测为 c 的概率, $rank_f(x,c)$ 是对的标签 x 预测 c 的排序, 当 $f(x,c_1)>f(x,c_2)$ 时, $rank_f(x,c_1)<rank_f(x,c_2)$.

Hamming Loss: 衡量预测所得标签与样本实际标签之间的不一致程度, 即样本具有某标签 c_i 但未被识别出, 或者不具有标签 c_i 缺被误判的可能性. 该值越小表明算法性能越好.

$$hloss_s(H)=\frac{1}{m}\sum_{i=1}^{m}\frac{xor\left(Y_i,\hat{Y}_i\right)}{|Y|} \tag{29}$$

One-Error: 该指标衡量了预测所得的可能性最高标签为伪正例的可能性. 该值越小表明算法性能越好.

$$1-error_s(H)=\frac{1}{m}\sum_{i=1}^{m}\left\{\left[\arg\max_{c\in Y}f(x_i,c)\right]\notin Y_i\right\} \tag{30}$$

Coverage: 该指标衡量了在对样本的预测标签的排序队列中, 从隶属度最高的标签开始, 平均需要跨域多少标记才能覆盖样本所拥有的全部标记. 该值越

© 中国科学院软件研究所 http://www.c-s-a.org.cn

小表明算法性能越好.

$$\text{coverage}_s(H)=\frac{1}{m}\sum_{i=1}^{m}\max_{c\in Y_i} rank_f(x_i,c)-1 \quad (31)$$

a-Accuracy: 该指标衡量了预测标签的平均准确度. 该值越大表明算法性能越好.

$$\alpha\text{ - Accuracy}(H)=\frac{1}{m}\sum_{i=1}^{m}\left(1-\frac{\left|\beta\left(Y_i-\hat{Y}_i\right)+\gamma\left(\hat{Y}_i-Y_i\right)\right|}{\left|\hat{Y}_i\cup Y_i\right|}\right)^{a} \quad (32)$$

其中, $\alpha\ge 0$, $0\le\beta$, $\gamma\le 1$, $\beta=1|\gamma=1$.

3.4 实验结果与分析

本实验设置 3 个对照组, 除本文所提的多段间隔度量学习(MMSML)计算出的度量外, 另外 3 个对照组分别为(1)文献[3]中算法(LSML)计算出的度量; (2)标准 Mahalanobis 距离; (3)欧式距离. .

在横向对比前, 首先需要确定 MMSML 算法中的 μ 的取值. 在保持标签的分布的前提下, 随机抽取 1000 份用作训练数据, 从余下数据中抽取另外 1000 条用作测试数据. 令 $N_i^o=N_i^s=N_i^e=5$, KNN 算法中的,k 值设置为 5, 设置 μ 的值从 0.1 开始, 以 0.1 步进递增, 直至 0.9. 实验结果如下表所示.

表 1　μ 的不同取值对算法性能的影响

μ	hloss	1-error	coverage	aA
0.1	0.10129	0.02049	0.04098	0.78985
0.2	0.10129	0.02049	0.04098	0.78985
0.3	0.10129	0.02049	0.04098	0.78985
0.4	0.09953	0.02049	0.04098	0.79345
0.5	0.09251	0.02049	0.03689	0.80665
0.6	0.07845	0.02049	0.03689	0.83546
0.7	0.07553	0.02049	0.03279	0.84026
0.8	0.07260	0.02049	0.03689	0.84747
0.9	0.07377	0.02049	0.03689	0.84507

从上表中可看出, MMSML 算法总体性能随着 μ 的增大而提高, μ 值的增大意味着同构邻居与相似邻居间的间隔比重增大. 在该测试集上, one-error 指标保持不变, 即说明算法对最优标签的决策是稳定的. 在 Hamming Loss 和 α -Accuracy 指标下, μ 取 0.8 时取得最优值, 综合考虑, 设置 μ 值为 0.8.

分别设置不同的 k 值, 评估欧氏距离, Mahalanobis 距离, LSML 度量, MMSML 度量四种不同度量的性能. k 的取值从 3 开始, 以 2 为步长递增, 直至 17. 其中, 部分实验的详细结果如下表 2 所示. Hamming Loss 指标, One-error 指标, Coverage 指标以及 α -Accuracy 指标的变化趋势如图 3, 图 4, 图 5 所示.

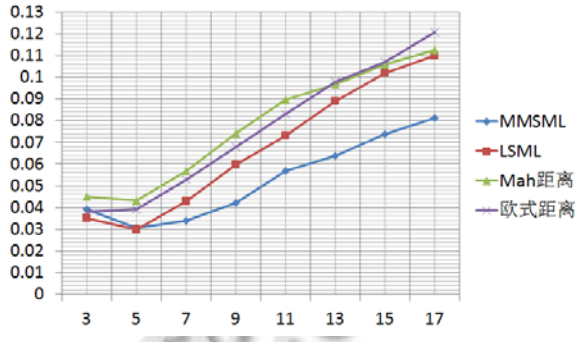

图 3　Hamming Loss 变化趋势

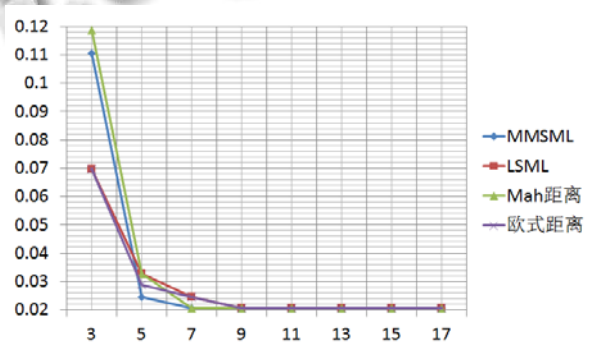

图 4　1-error 变化趋势

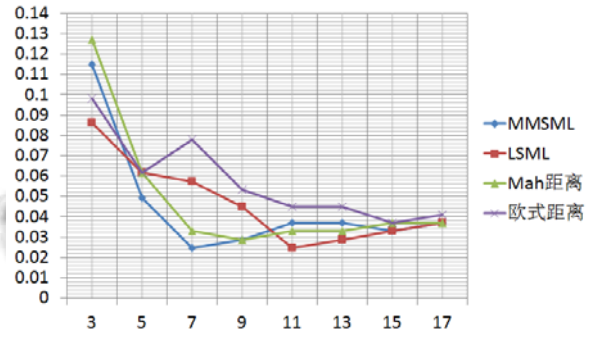

图 5　Coverage 变化趋势

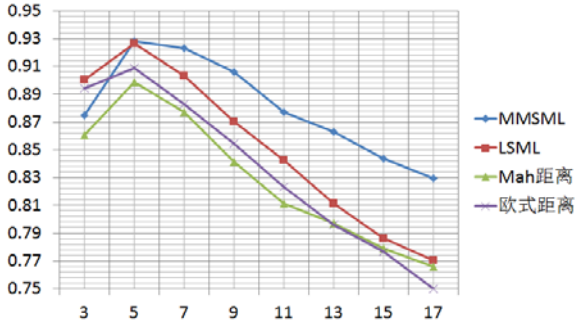

图 6　α -Accuracy 变化趋势

© 中国科学院软件研究所 http://www.c-s-a.org.cn

从实验结果中可知, MMSML 算法学习出的相似度度量在 Hamming loss 和 α -Accuracy 两种指标下具有明显优势, 这两种指标衡量了算法的准确度. 并且随着 k 的增大, 优势进一步扩大. 在图 3, 图 6 中, 两种指标表示的性能随着 k 增大而衰减, 但是 MMSML 的衰减速率是最低的. 分析原因, 是因为 MMSML 算法能够在纵向距离上形成层次化的区分, 能够利用监督信息, 按照监督信息的相似程度将目标样本的邻居不同程度地推远和拉近. 而其他度量则不具备或者具备有限的这种能力. LSML 只能区分出同构和异构邻居, 当 k 值增大, 需要检查更大范围时, 相似邻居和异构邻居混杂在一起. 而 MMSML 算法能够将异构邻居进一步从相似邻居的临界范围内过滤出去, 从而提高准确率.

为进一步验证算法的稳定性. 将 k 值固定为 15. 在不同训练集和测试集上进行试验. 试验结果如下表 3 所示. 在下述结果中, MMSML 算法在不同数据集上均表现出优势, 表明算法具有良好的稳定性.

表 2 四组对照试验在不同 k 取值下的部分实验结果

	hloss			1-error			coverage			*aA*		
k	5	9	15	5	9	15	5	9	15	5	9	15
欧式距离	**0.039**	**0.068**	**0.107**	0.0246	0.0205	**0.0205**	**0.0615**	**0.0533**	**0.0369**	**0.909**	**0.855**	**0.776**
Mah 距离	**0.043**	**0.074**	**0.106**	**0.0328**	**0.0205**	**0.0205**	0.0615	0.0287	**0.0369**	**0.899**	**0.841**	**0.779**
LSML	0.030	**0.060**	**0.102**	**0.0328**	**0.0205**	**0.0205**	0.0615	**0.0451**	0.0328	**0.926**	**0.870**	**0.786**
MMSML	0.030	0.042	0.074	**0.0287**	**0.0205**	**0.0205**	0.0492	0.0287	0.0328	0.929	0.906	0.844

表 3 四组对照试验在不同数据集下的实验结果(测试集大小等于训练集)

	hloss			1-error			coverage			*aA*		
训练(测试)	800	1200	2400	800	1200	2400	800	1200	2400	800	1200	2400
欧式距离	**0.127**	**0.107**	**0.149**	**0.0894**	**0.0205**	**0.2012**	**0.1789**	**0.0369**	**0.3527**	**0.729**	**0.776**	**0.638**
Mah 距离	**0.118**	**0.106**	**0.138**	**0.0894**	**0.0205**	**0.2012**	0.1707	**0.0369**	**0.3278**	**0.743**	**0.779**	**0.650**
LSML	**0.128**	**0.102**	**0.140**	**0.0894**	**0.0205**	**0.2012**	**0.1789**	0.0328	**0.3568**	**0.727**	**0.786**	**0.660**
MMSML	0.101	0.074	0.129	**0.0894**	**0.0205**	**0.2012**	0.1707	0.0328	0.3320	0.780	0.844	0.678

4 结语

本文将病人相似度问题转化为有监督的度量学习问题. 针对人工评估来产生监督信息所存在的开销大等困难, 提出利用诊断信息作为距离度量学习的监督信息. 同时, 在LSML算法上进行改进, 将监督信息本身的相似程度作为学习的重要标准, 更充分地利用监督信息. 可以很容易地证明, 该算法可以扩展应用于不同的监督信息相似程度, 从而学习出多段间隔. 另外本文从真实的住院数据中提取特征构造实验数据, 通过多标签 KNN 分类来间接地评估所学习出的度量. 实验结果表明, 本文所提出的相似度度量在不同大小的数据集上, 在 Hamming Loss 和-Accuracy 指标下均有明显优势, 表现出较好的准确性和稳定性.

参考文献

1 Natarajan K. Analysis of Search on Clinical Narrative within the EHR[Thesis]. Columbia University, 2012.

2 Partners BS. Secondary uses of electronic health record (EHR) data Applications for the life science and healthcare industry. 2012.

3 Chan L, Chan T, Cheng L, Mak W. Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). 2010. 467–470.

4 Gottlieb A, Stein GY, et al. A method for inferring medical diagnoses from patient similarities. BMC Medicine, 2013, 11(1): 194.

5 Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 2009, 10: 207–244.

6 Sun J, Wang F, et al. Supervised patient similarity measure of heterogeneous patient records. ACM SIGKDD Explorations Newsletter, 2012, 14(1): 16–24.

7 Zha H, He X, Ding C, Simon H. Spectral Relaxation for K-means Clustering. MIT Press, 2001: 1057–1064.

8 Elisseeff A, Weston J. A kernel method for multi-labelled classification. Advances in Neural Information Processing Systems, 2001: 681–687.

9 Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000, 39: 135–168.

10 Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757–1771.

© 中国科学院软件研究所 http://www.c-s-a.org.cn