## 基于知识蒸馏和决策树的黑盒模型解释方法

1. Wang, J., Gou, L., Zhang, W., Yang, H., Shen, H.: Deep-vid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. IEEE Trans. Vis. Comput. Graph.25(6), 2168–2180 (2019).
2. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: Conitzer, V., Hadfield,G.K., Vallor, S. (eds.) Proceedings of the 2019 AAAI/ACM Conferenceon AI, Ethics, and Society, 131–138. ACM (2019).

## Deep-vid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation.

## (Ohio State University)

# Importance of the application

* 随着深度学习开始得到更广泛的应用，我们开始关注的不仅仅是模型预测的准确率和速度，更应该去理解为什么它会做出特定的预测。现有的分类器大多是较复杂的深度学习模型，通常将其作为黑盒模型来看待。之前的方法，通常将黑盒模型打开，通过梯度，层相关性反向传播以及类激活图，这些技术仅限于网络结构已知的情况，而在现实世界的应用中经常不知道具体的网络结构。因此，针对黑盒模型的解释尤为重要，通过这些理解能更好的对模型的参数进行调优以达到优异的性能，也为模型的设计拓宽思路。

# Problems cannot be solved with existing methods

* 大部分解释模型不适用于黑盒模型。
* 黑盒模型的解释方法，使用输入维扰动的方式为解释模型生成局部邻域数据。首先，在输入空间中对数据实例进行扰动是低效的，由于输入通常是高维的，并且扰动维的组合是巨大的。其次，通过直接扰动数据实例生成的邻居输入空间可能没有语义意义，导致在训练小模型时存在潜在的语义偏差。
* 现有的解释方法没有配套的可视化系统，来让用户了解解释模型的训练过程。

# Formal definition of the problem

一个分类问题，输入空间为（即图片），其对应的输出空间（即标签），给定训练集及相应的标签。采用一个神经网络作为预测模型：。训练目标是训练，利用知识蒸馏的方式，使得相同的输入情况下，f’(x)的预测和原始黑盒模型预测尽可能相似。而f’(x)是易于理解的简单线性模型。 训练高效的特征抽取器VAE，便于在高维空间对数据进行扰动，并能通过VAE中decoder给用户提供可视化的生成结果。

# What are the technical challenges with this approach?

* 如何训练得到易解释的线性模型？
* 如何生成高维扰动的邻域数据？
* 如何让用户了解解释模型的训练过程？

# Proposed Methods for these challenges

* 使用知识蒸馏的方式，使得线性模型抽取到黑盒模型的知识
* 预训练抽取高维特征的VAE，得到感兴趣数据样本点的高维特征。并通过均匀采样的方式，计算得到领域的样本点，以用于训练线性模型。通过展示线性模型的权值，可以得到，黑盒模型关注图片的关注点。
* 设计可视化系统，分析需求，主要从三个方面：对待解释模型的行为的可视化的理解， 灵活挖掘生成的邻居，用户能够通过系统交互式地训练和解释学生模型。针对三个需求分别为系统设计了三个模块。

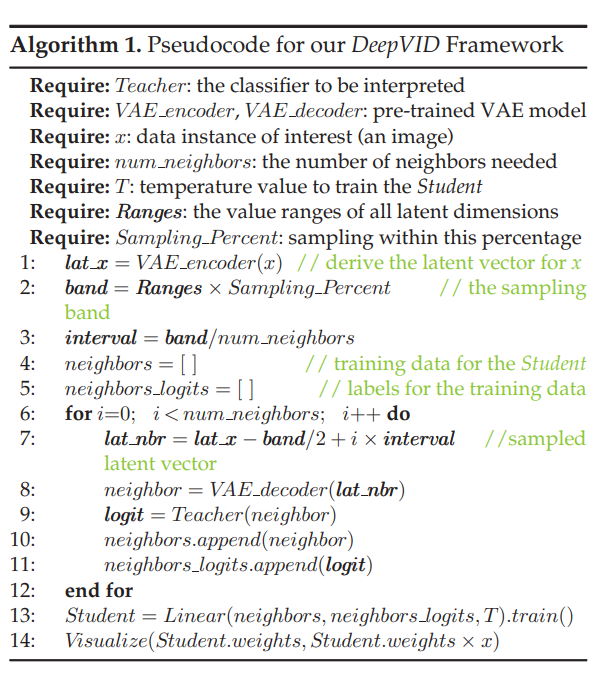


图1 deepvid 算法流程图

本文提出了新的模型解释框架，deepvid，并实现了系统，以帮助用户更好的理解解释模型的训练过程。首先利用数据集预训练VAE模型对应，上述算法图中第1步。根据感兴趣的样本点X，在2-12步中，通过扰动的方式，生成该样本点附近的邻域节点的数据。第13步，训练易理解的线性模型以得到黑盒模型的行为。第14步，以可视化系统的方式，展示权值，得到黑盒模型对于图像的关注点。

本文花了大量笔墨在系统的设计与实现上，通过与专家的交流与分析，将系统主要设计为3个层面：

（1）对待解释模型的行为的可视化的理解（TeacherView）如图2所示，具体来说，需要系统能够:可视化测试数据的特征空间进行展示，展示待解释模型对测试数据的整体表现，以及指定数据的概率。

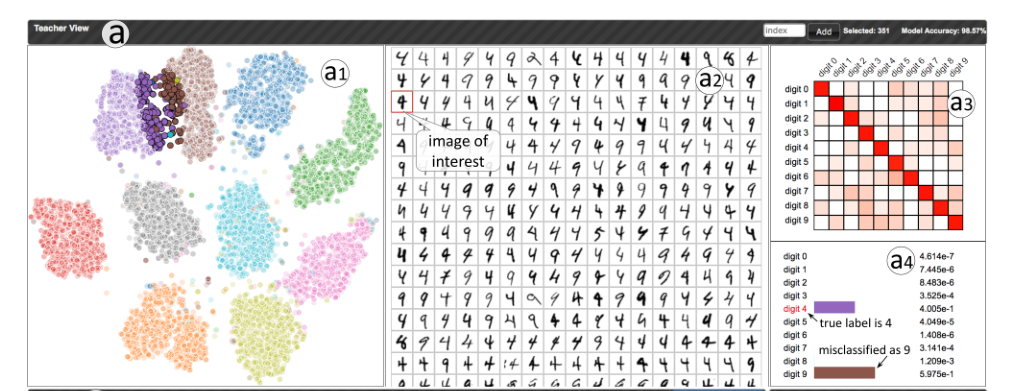


图2 Teacher View

1. 灵活挖掘生成的邻居，用户能够通过系统交互式地训练和解释学生模型（VAE View）如图3所示。具体来说，理解VAE中编码的语义，例如，对于MNIST手写数字，一些编码粗体，另一些来编码数字的斜体。系统能够交互式调整生成的邻居，使生成的邻居足够接近感兴趣的实例。在两个跨类边界的感兴趣的实例之间生成训练样本，例如，生成从用户感兴趣的数字4平滑变形为数字9的样本。

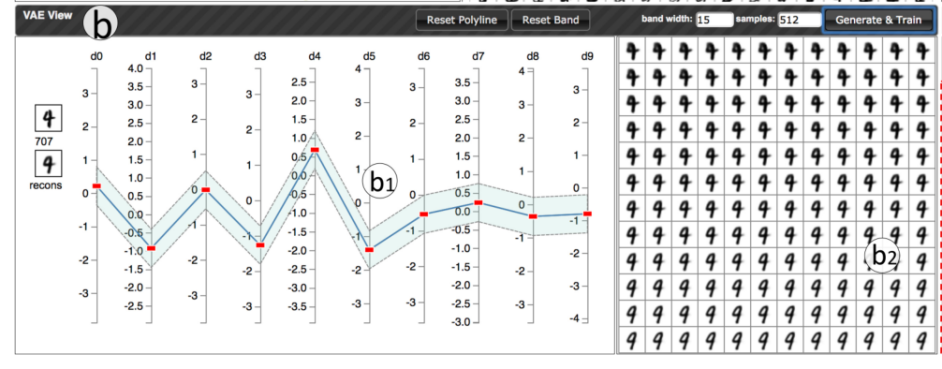


图3 VAE View

1. 用户能够通过系统交互式地训练和解释学生模型（Student View）如图4所示。帮助用户使用定制的超参数训练解释模型。在最顶部是一些超参数，用于去训练学生模型，学生模型是线性模型，以mnist数据集为例，输入是784个像素点，输出是10分类，待解释的模型是Lenet。红色框的部分是将学生模型的权重和带权的像素点可视化的结果，来展示每个像素点对于某个分类的贡献，越白代表贡献越大。右侧的另外一个子模块也展示了student对于这个图像的预测概率分布，应该尽可能和的教师模型的预测概率分布相近，帮助用户来判断小模型是否很好的蒸馏到大模型的知识。

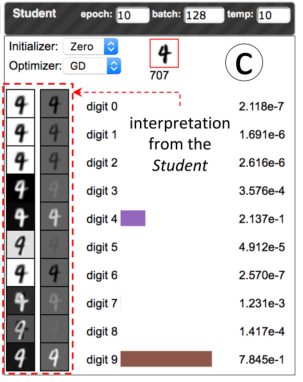


图4 Student View

**Experiments and Datasets**

* 数据集

论文采用了Mnist，CelebA两个数据集。 对于CelebA，对数据集进行预处理，resize成64大小的图片，选择三个二值的属性将类别分为8类，包括性别，是否是金发和是否戴眼镜。并选择了一个DFC-VAE来捕获语义信息。

* 对比方法

LIME，Sensitivity(基于梯度)，Gran-CAM（基于类激活图）。

* 实验
  + Case study

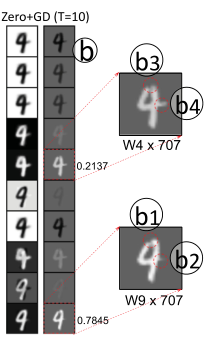


图5 Mnist case study

通过teacher view 发现这个4被误分类为9，因此就在该数据邻域，利用vae生成数据，来训练学生模型。训练得到的学生模型，将教师模型对这个数据分类到不同类中，分别进行解释，从上到下是0到9，可以看到误分类到9的主要原因是分类器高亮了这个数字的顶部以及忽略数字4的尾部。



图6 CelebA case study

在celeba数据集上case study。探索8170这张图片为什么被分类成class2(女生，金发和没有眼镜)。通过均匀采样从1682这张图片和它之间的样本点，得到该点的邻域数据。图6中左边的部分，展示了采样出重构的图片，即小模型的训练数据。从类别2中可以看到，分类器非常关注于头发部分的像素，说明分类器把头发特征作为分类的依据。

* + 领域专家的反馈

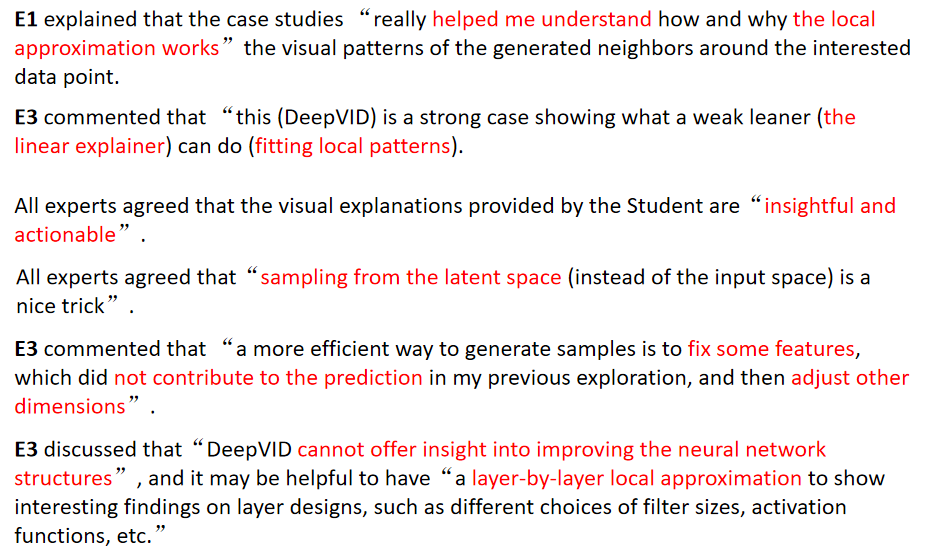


图7 领域专家的部分反馈

领域专家也对DeepVID给出了许多反馈，首先所有的专家都给出了积极的反馈，因为deepvid帮助他们解释和诊断模型。专家一认为，通过可视化生成的局部邻居，帮助它理解为什么局部近似可以work。专家三认为deepvid体现出了线性解释模型可以去拟合局部的模式。并且所有的专家都认为学生的可视化解释是直观且可行的关于vae视图是最受专家关注的，专家们都认为邻域生成的过程非常关系于最终的解释模型，并且认为从隐空间采样比输入空间采样是一个很好地选择。专家一还建议，固定一些在前面探索中，对预测无用的特征，然后只变换其他特征来进行扰动。专家三还认为，现有的deepvid无法通过解释之后对网络结构本身有更多的提升，建议利用例如层与层之间的局部近似，帮助用户去改进模型结构。那么实际这个建议，就是希望deepvid，能够结合黑盒解释和白盒解释。

* + 可视化与定量化评估

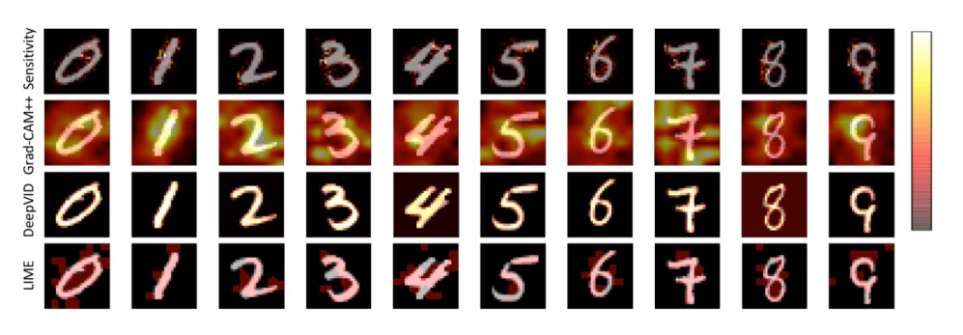


图8 可视化解释结果比较

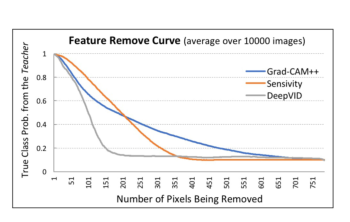


图9 插入删除曲线

通过和baseline的比较。图是将可视化的解释图和原本的图重叠起来所展示的，原始图片的颜色变成黑底灰字。 colormap 从黑到红到白，代表着像素点的重要性越来越高。从图中可以看到deepvid 更可以抓住灰色的像素点，证明了本文方法的优越性。 从插入删除曲线来看，也能显示本文方法的优越性。

# Differences of this Method compared to others

使用VAE模型来扰动感兴趣数据的隐向量，高效的得到重构的易解释的线性模型的训练数据。通过知识蒸馏将待解释的模型的知识传递给易解释的线性模型。通过可视化系统的方式，帮助用户理解待解释模型的行为，以及如何得到解释模型。

**Limitation of this work or your discovery**

* deepvid是非常依赖于预训练的VAE去提取语义的和进行重构的，而例如在cifar10这种数据集上，每个类别的数据比较少，vae效果会比较差，导致最终的解释也会很差。
* 方法创新性不是很强，将现有的方法进行组合，并设计了系统。

# **Faithful and Customizable Explanations of Black Box Models**

## (Harvard University)

# Importance of the application

* 在现实世界的决策中成功采用预测模型取决于决策者（如医生、法官）能够理解和信任他们的功能。只有决策者对预测模型的行为有一个清晰的理解，他们才能评估何时以及在多大程度上依赖于这些模型，检测其中的潜在偏差，并开发进一步细化模型的策略。
* 在本文中，关注一种新形式的解释，旨在帮助最终用户（例如，决策者如法官、医生）获得更深入的理解模型行为：不同的解释，描述了模型逻辑在不同的子空间感兴趣的忠实的和可解释的方式。医生可能热衷于了解模型是如何对不同的患者亚组进行预测的（见图1）。此外，她可能会对问一些问题感兴趣，比如“该模型如何预测与运动和吸烟的不同价值观相关的患者亚组？”。本文旨在可定制化的根据用户需求来生成解释。

# Problems cannot be solved with existing methods

* 无法在用户提出一些感兴趣的特征之后，定制化的生成解释模型。
* 无法定量的量化保真度、不模糊性和可解释性的概念

# Formal definition of the problem

* 设计一种两层决策集表示法，{(*q*1,*s*1,*c*1), (*q*2,*s*2,*c*2) · · · (*qM* ,*sM* ,*cM* )}， *qi* 和*si*是谓词的连词，*qi* 是用户感兴趣的子空间*，si*是内部决策的规则，*ci*是对应的分类。根据定义的保真度，不模糊性和可解释性的指标，转换为优化问题，求解最优的*qi* 和*si*。

# What are the technical challenges with this approach?

* 如何定义保真度，不模糊性和可解释性的指标？
* 如何求解该优化问题？

# Possible methods for this challenges

* 使用两层决策集的方式，来解释黑盒模型。外层决策集可用于自定义用户感兴趣的空间，内层决策集表示内部决策的规则。如下图10所示。定义具体保真度，不模糊性和可解释性的指标。

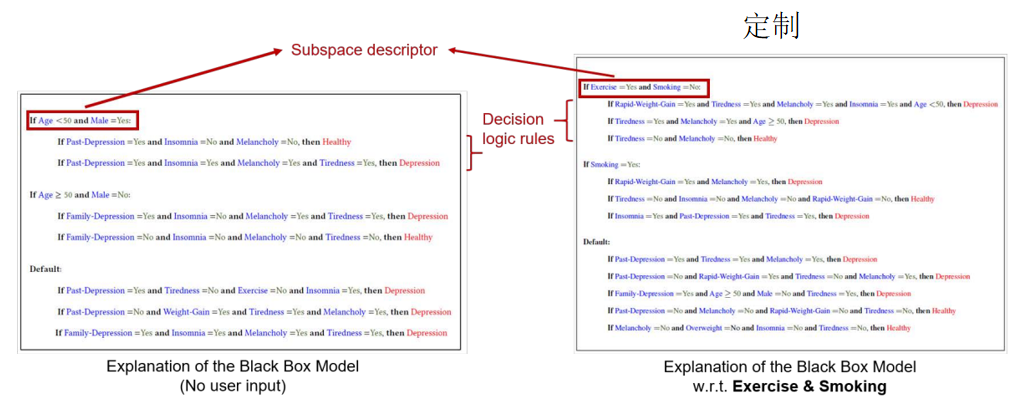


图10 两层决策集表示法

本文提出了一个能够根据用户感兴趣的子空间来给出相应解释的决策集解释模型。文中主要的创新点在于，使用两层决策集来解释原始模型的行为，定义了保真度，不模糊性和可解释性的指标，将其合并转换成一个优化问题，并进行总的优化，得到忠实且可定制化的解释模型。

本文对忠诚度（Fidelity）、不模糊性（Unambiguity）和可解释性（Interpretability）进行定量化的建模，如图11、12所示。第一个部分，忠诚度。它意味着得到的决策集应该和黑盒模型的预测尽可能相似。本文所采用的标准是决策集和黑盒模型之间预测的差值，希望这个值越小越好。那么第二个标准就是不模糊，无歧义性，其中就包含两个指标，第一个指标是 ruleoverlap，它的含义就是，规则重叠的部分，希望规则的重叠越少越好，代表着规则越清晰。第二个指标是cover，即每个规则下包含的数据，希望满足各个规则的数据越多越好。越多就代表着规则更具有概括性。最后一个标准就是可解释性size是决策集中三元组的数量，也就是有几条规则。可以把它理解为决策树的叶子节点的数量，maxwidth代表着q和s中最长的一条规则中有几个条件，此规则既包括用户感兴趣的子空间，也包括内部的决策规则。如果把内外ifthen理解成两颗决策树可以理解为两颗决策树的最大高度。numspreds指标代表着用户感兴趣的子空间，也包括内部的决策规则里面一共出现几种条件。numdsets 指标的含义是多个用户子空间中独特的条件的数量。此指标衡量的就是外部的子空间和内部的逻辑决策规则的条件的重叠数量，这个值应该重叠越小越好。

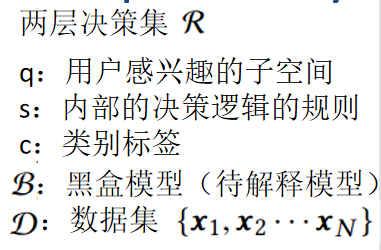


图11 符号定义

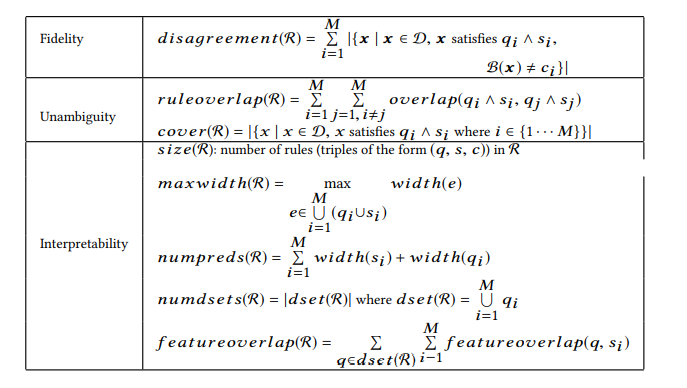


图12 忠诚度、不模糊性和可解释性的形式化定义

根据上述形式化的定义后，整合所有的目标后，形成以下的优化目标：

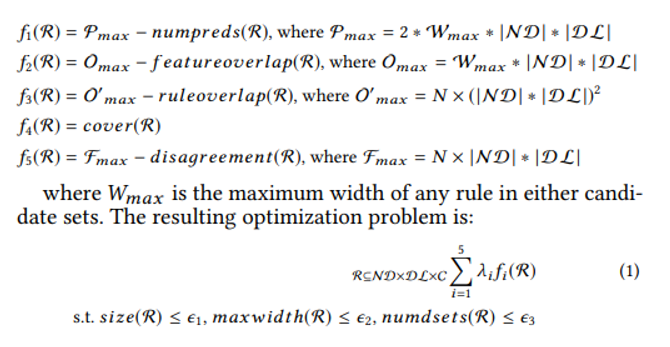


图13 优化函数

式中的ND、DL分别代表用户子空间和内部决策规则的条件的候选集合。在实践过程中，DL默认是由apriori算法挖掘得到的，而ND是通过用户的输入得到的，比如age>50和性别为女，因此完成了用户定制的需求。最终的优化函数选用了上述介绍的目标的加权和，权值是超参数需要自己来定，一部分指标作为人工设定的超参，使得解释模型忠实，不模糊和可解释。为了便于优化，对于所有越小越好的指标，都采用负数加最大值的方式，使得最终的目标函数是非负的，并且越大越好。该问题的优化算法直接采用近似局部搜索的优化方式求解最终的结果。

**Experiments and Datasets**

* 数据集

论文采用depression diagnosis 抑郁症数据集， 包括病史、症状、人口信息/ 是否健康的数据

* 对比方法

LIME-DS（变种：每一个局部模型都是一个决策集）

IDS（Interpretable Decision Sets）、BDL （Bayesian Decision Lists)。

* + 比较忠诚度和可解释性

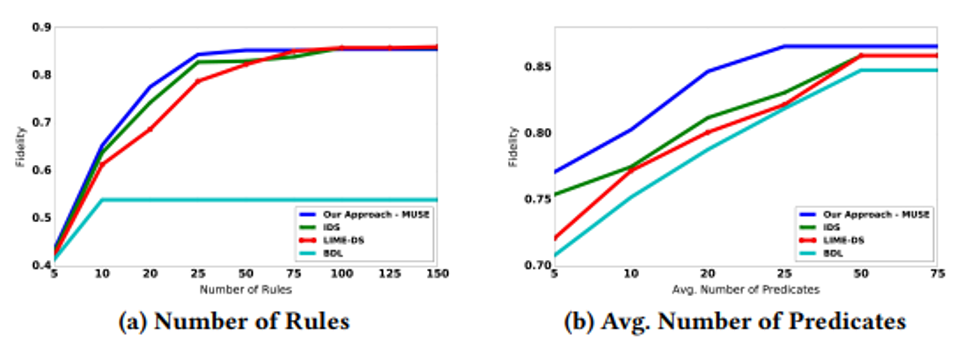


图14 不同方法忠诚度和可解释性的关系

为了了解不同的方法如何有效地在忠诚度和可解释性之间进行权衡，对于由本文的框架 在不需要用户输入感兴趣的特性和其他基线生成的解释，将忠诚度与各种可解释性指标进行对比。忠诚度的定义就是数据中由解释模型预测的标签与黑匣子模型预测的标签相同的数据的比例。图a和图b分别显示了使用本文方法与baseline对比的在抑郁症数据集和5层神经网络模型上的结果。a图的可解释性，用规则数量来表示，另一个用条件的平均数量来表示。从结果中可以看到，与其他最先进的基线相比，本文的方法产生的解释在忠实度与复杂性也就是可解释性之间提供了更好的权衡。

* + 用户实验

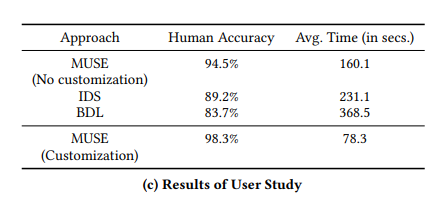


图15 用户实验结果

最后本文做了一个casestudy，希望评估用户使用本文的解释模型来理解和推理黑盒模型行为的便利性，实验结果如图15所示。作者设计了一个有33名参与者的在线用户研究，在这个研究中，每个参与者被随机呈现由本文方法和baseline方法产生的解释。参与者被问了5个问题，每一个问题都被设计用来测试他们对特征空间不同部分的模型行为(如解释所描述的)的理解。举一个例子来说，考虑一个65岁的女性患者。根据所提供的解释模型，你能绝对确定这个病人是健康的吗?如果不是，这个病人还需要具备哪些条件才能被标记为健康?这些问题与现实世界中的决策过程非常相似，在现实世界中，像医生、法官这样的决策者想要在特征空间的某些部分对模型行为进行推理。这些问题的答案可以根据黑盒模型来得到。在此基础上，计算了用户提供的答案的准确性，和回答每个问题所花费的时间，并以此来计算每个问题所花费的平均时间(以秒为单位)。右边的图显示了没有定制化情况下本文的方法和baseline的解释所获得的结果。可以看出，本文的解释模型，能够更好的帮助用户得到答案，准确度高于其他。并且，使用本文的解释模型用户的反应更快。

这些问题中有的问题还包括，问的是一个吸烟又不锻炼的病人的抑郁情况。作者对这些问题会向他们展示运动和吸烟在外层if-then的结果，模拟用户试图探索模型这些特征的效果。与不使用定制化的方法相比，准确度有了明显的提升，这说明允许用户从不同的角度去探索模型的行为，对于推理它在特征空间的不同部分的行为是非常有帮助的。

# Differences of this Method compared to others

在表示方法上使用了两层决策集，优势在于易于理解，并且可以自定义用户子空间。本文中量化了保真度、不模糊性、可解释性目标，通过优化问题求解得到最终的解释模型。

**Limitation of this work or your discovery**

* 该方法仅仅能适用于文本分类模型，当分类模型中分类较多，优化目标的时长将会较长。
* 使用决策集的方式进行解释，解释模型过于笼统，不能很好的表征原始黑盒模型的行为。