还在辛辛苦苦提取特征？Embedding帮你自动来

原创 吕阅涵 信也科技拍黑米 2018-07-18

导语

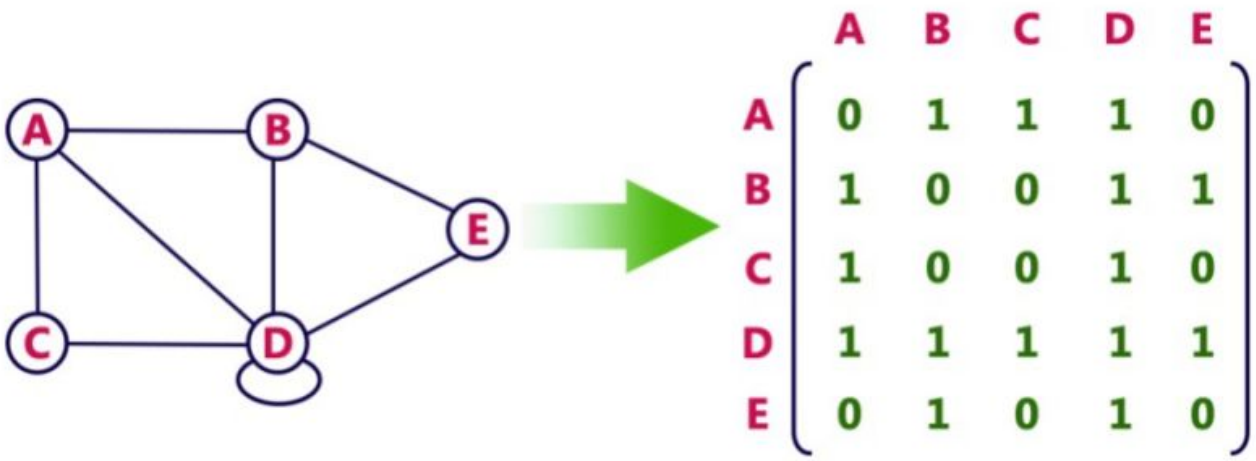
随着计算能力的提升，复杂网络表示学习(Network Representation)在社会学、通信网络、生命科学等多个领域都起着重要的作用，它也是拍拍贷与浙江大学通力合作的众多研究项目中的一员。

在拍拍贷的反欺诈系统中，利用表示学习里的Embedding方法，我们可以根据社交关系网络为用户自动生成特征。它已经在拍拍贷识别不良借入用户，提高模型效率，节约变量成本中逐渐起到了关键的作用。

面对用户的社交关系网络，如何为每个用户自动生成有效特征呢？本文将首先简要介绍几个强大的Embedding模型，再谈谈我们是如何在实际业务场景中利用这些模型的。

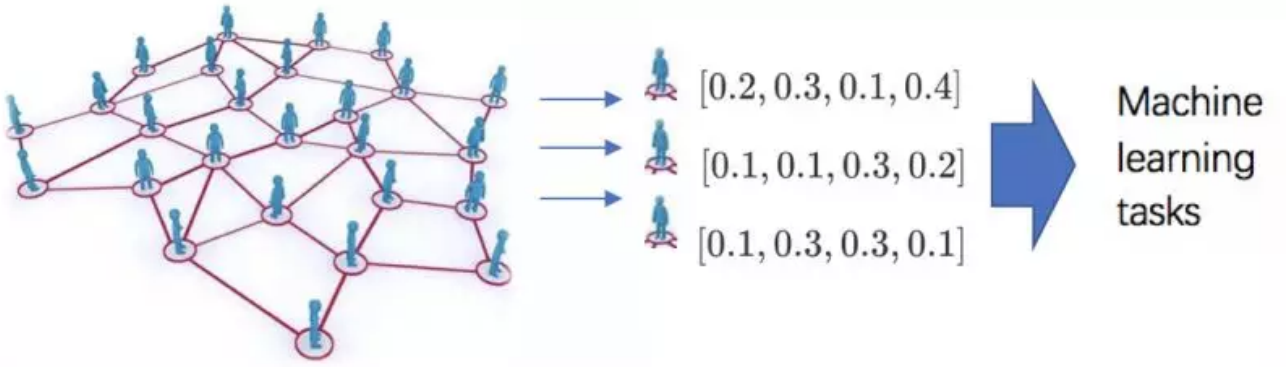
如何描述一个网络

面对多个对象，我们首先根据对象之间普遍存在的某种关系将它们连接起来，从而以网络的形式获得整个对象群体。我们知道在数据结构中，最早使用的描述网络的语言是邻接矩阵，它同义于one-hot编码。



尽管邻接矩阵可以捕捉到网络的全部信息，但信息的抽象程度很低，数据过于稀疏也使得邻接矩阵的维度太高（=节点数量的平方），节点对应的one-hot向量不适合作为节点的特征输入模型。

Embedding致力于将庞大的节点网络（十万级以上）作为输入，最终为每个节点学习出可以有效代表该节点的低维（百维左右）的向量，最终将向量或向量生成的变量分数作为其他机器学习任务的输入变量。



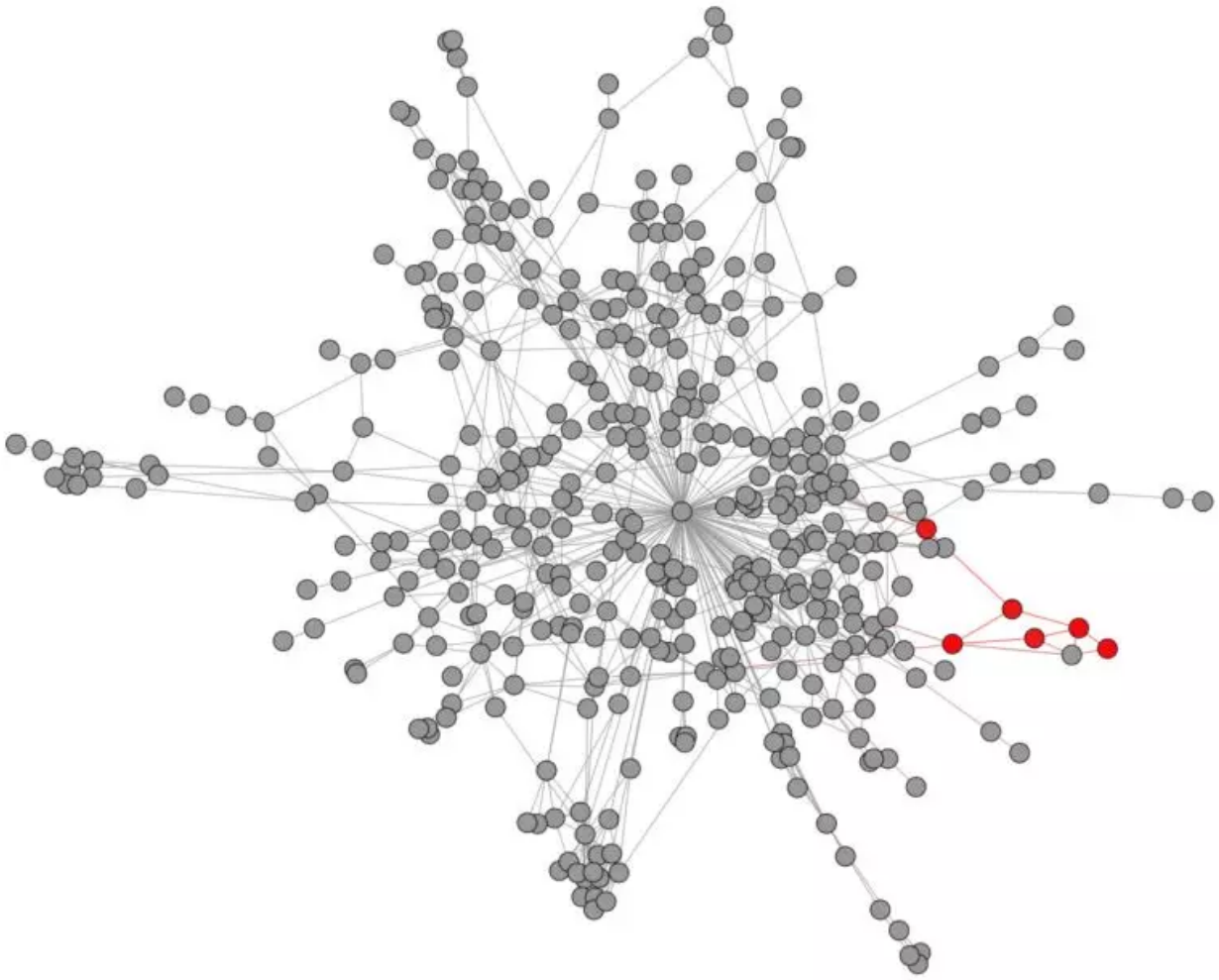
在以用户行为为目标变量的机器学习问题里，标准的解法是挖掘出用户的其他维度特征信息，作为模型的输入，再利用已有的各种回归、分类模型完成预测问题。在目前日益注重用户隐私保护，减少变量输入成本，提高建模效率的期待下，我们希望可以根据较少而公开的用户信息形成对用户较为有效的描述。Embedding仅利用用户所提供的某种社交网络信息，进行无监督或半监督的网络表示学习，从而自动生成有限维度的用户特征，将其作为输入提供给其他预测问题。

几个重要的 Embedding模型

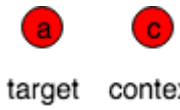
复杂网络Embedding发展至今，大致可以分为两种基本类型：一类重在挖掘节点临近关系，以Deepwalk[1]、node2vec[2]、LINE[3]为代表，被称为基于临近关系的表示学习，其二是重在挖掘节点结构特征的，以struc2vec[4]为代表，被称为基于网络结构的表示学习。

基于临近关系的表示学习

在自然语言处理领域，随着Word2vec算法的提出，我们通过考察单词与背景文本共同出现的概率，可以将单词映射为有限维欧式空间中的向量，从而实现对文本的数值化分析。沿着这个思路，Deepwalk提出使用类似的方法。它考察节点与临近节点共同出现的概率，



上图是一个用户网络，通过从每个节点出发的多次随机游走，我们获得了类似于语料库的长句子集合，每个句子等同于一处连通的微型用户关系网络（如图中的红色部分）。



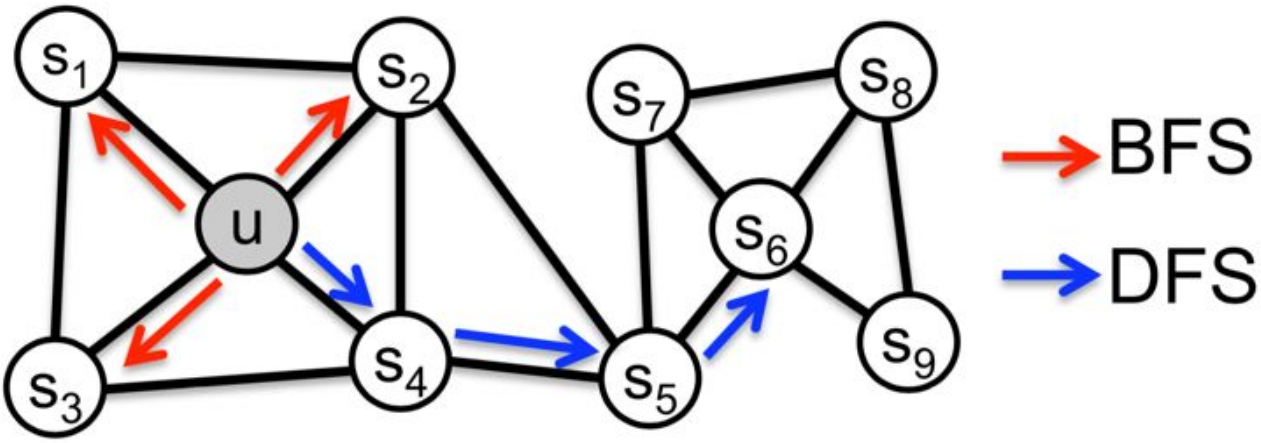
对每个句子利用skipGram算法：滑动一个短的窗口提取出(单词，背景)对，以背景节点出现的后验概率生成损失函数，

$$Loss = -\log P(c|embedding(a))$$

再利用梯度下降来更新每个节点的embedding向量。

$$embedding(a) = embedding(a) - learningrate * \frac{\partial Loss}{\partial embedding(a)}$$

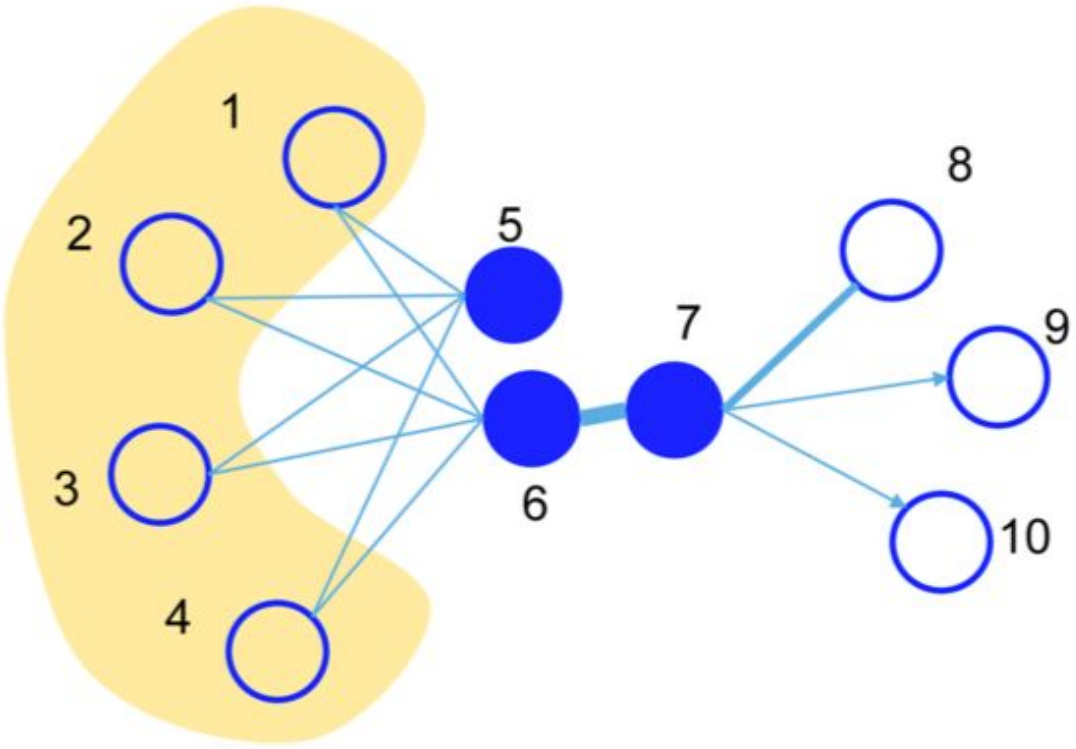
与Deepwalk类似的还有Node2vec算法，它与前者最大的不同之处在于：在第一步生成用户“语料库”的过程中，随机游走算法可以是深度优先(DFS)或广度优先(BFS)的。



上图是对Node2vec这一特征的刻画。 在此图中，BFS方式会给出类似于(u, s1, s2, s3)的随机游走结果，而DFS会给出类似于(u,s4,s5,s6)的随机游走结果，这可以通过简单的调整概率转移函数来实现。

可想而知，BFS出的长句子，skipGram处理之后，在同一社群(cluster)中的节点会具有相似的表示向量了；而DFS所给出的长句子，在skipGram处理过后，具有类似角色(role)的节点，比如u和s6则会具有更加接近的表示向量。

而LINE是另外一种刻画局部网络特征的方法。在LINE中我们定义一阶和二阶临近相似度。一阶相似度刻画相邻节点，而二阶相似度刻画拥有相似邻居的两个节点。



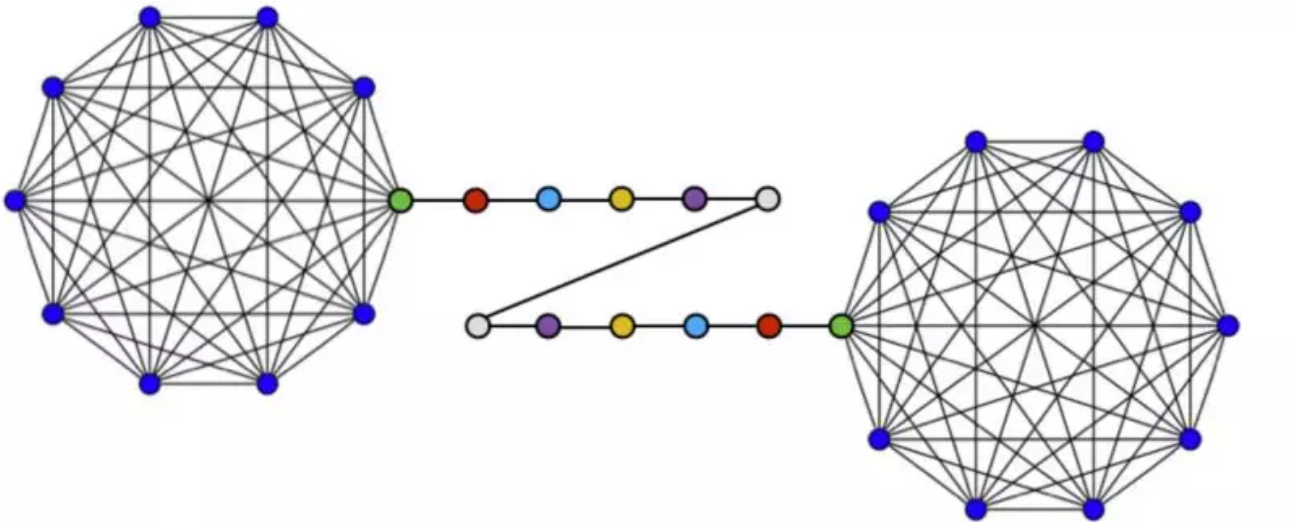
如图所示，通过LINE的方法对该网络做embedding， 6号节点和7号节点应当具有相似的一阶临近相似度，而5号与6号节点应当具有相似的二阶临近相似度。在学习迭代的过程中，同样的，两个节点的一阶和二阶相似度都首先被映射到损失函数域，再利用梯度下降来更新节点的embedding向量。

基于网络结构的表示学习

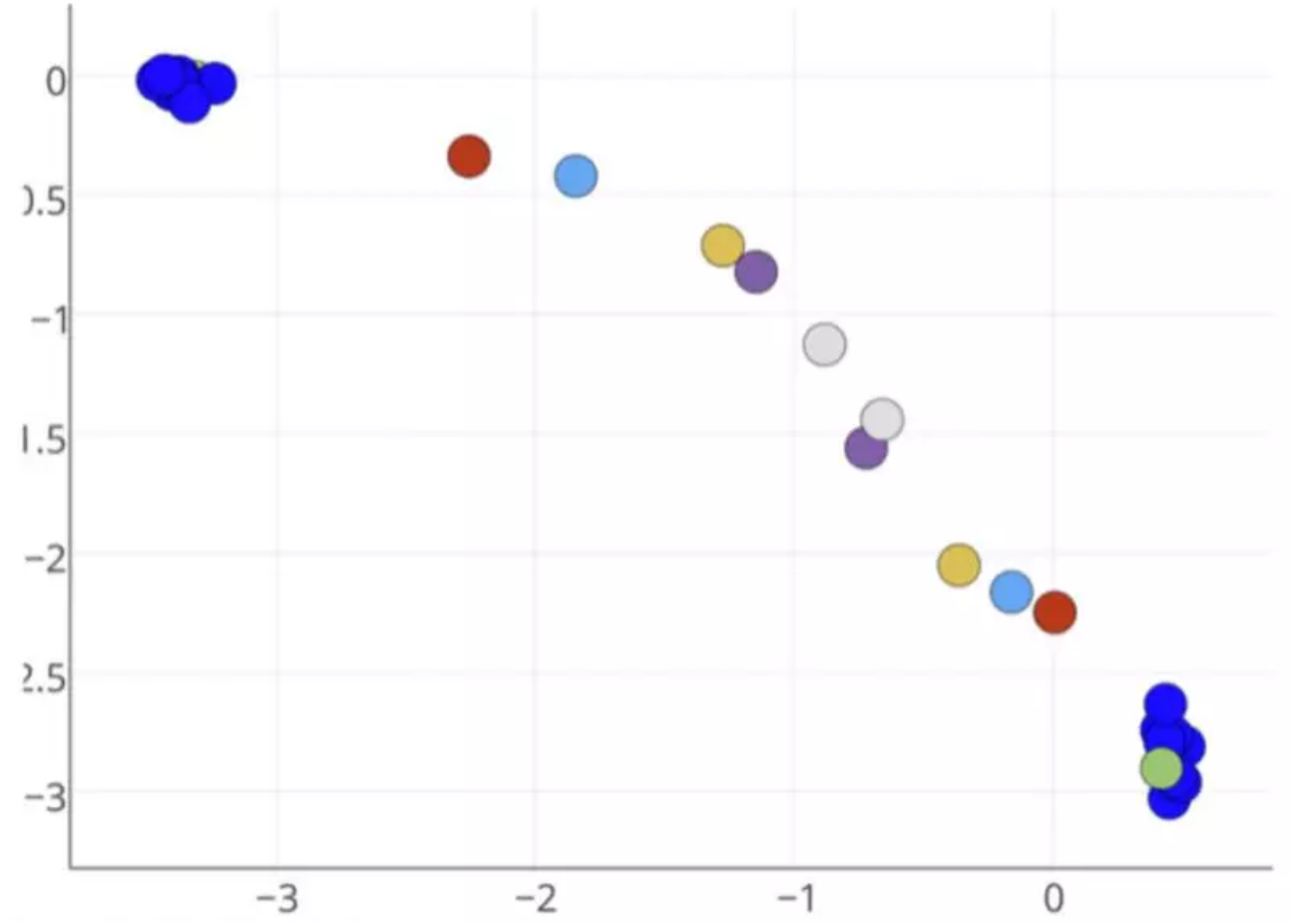
聪明的小读者可以发现，上面提到的三种方法都要求在训练的过程中，参与比较的两个节点之间须有路径相连，或者说，越临近的节点的embedding向量越趋于相似。然而事实上对于大部分的业务场景，网络并不是全连通的，甚至被分成很多很小的cluster。另外，并不是临近的节点就应当是相似的，这些embedding方法忽略了节点结构上的相似性。

struc2vec的作者们提出了新的方法。对于网络中的任意一个节点，首先计算它的k阶邻居们的度数序列。其次，两个节点的相似性则由它们各自k阶邻居度数的有序序列来定义。这里距离采用Dynamic Time Warping (DTW) 来定义。

举一个论文中提到的例子。将一个barbell图视为一个小型的网络，如下所示

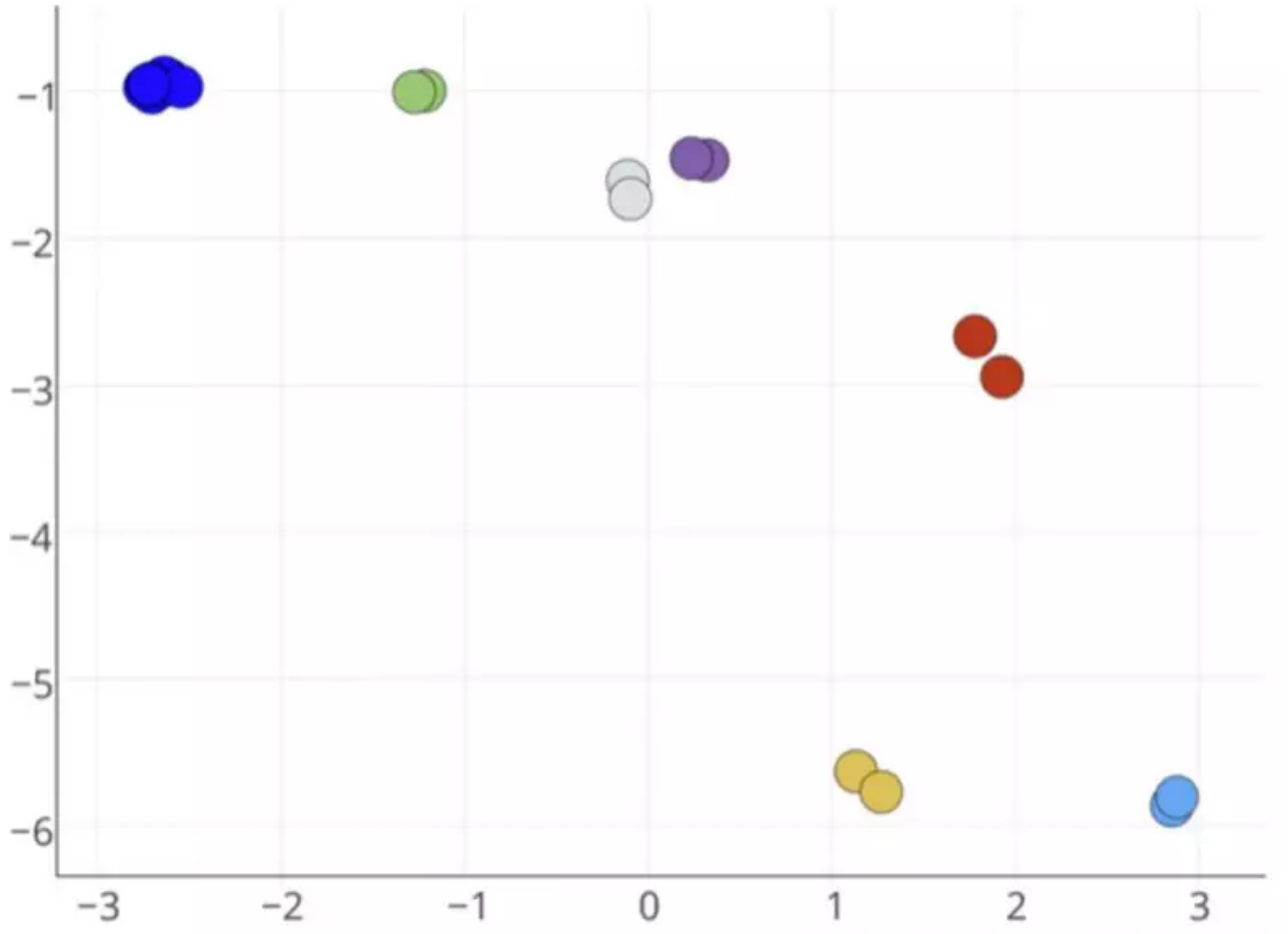


很显然，相同颜色的节点在网络中是完全对称的，应该具有非常接近的embedding向量。如果利用Node2vec将节点映射到2维空间，其结果却并不如我们所料：



蓝色的节点们被映射到了距离最远的两部分。实际上，在原图中离的越近的节点，在embedding二维空间中离得越近。

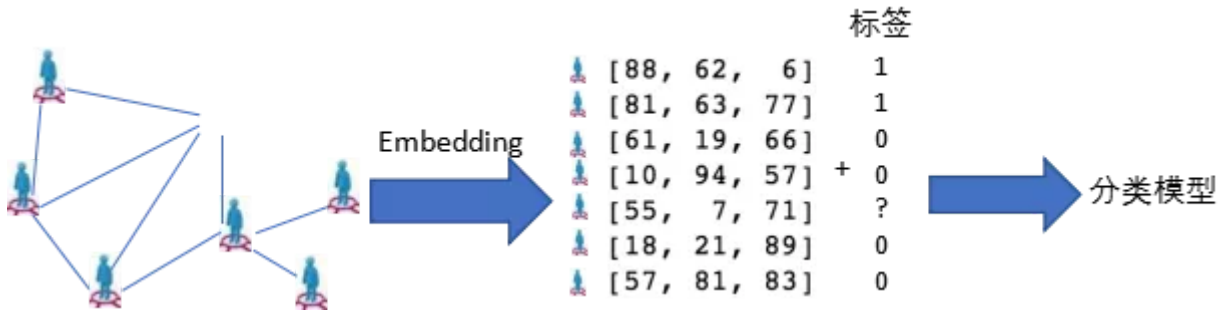
如果利用抓结构特征的struc2vec算法来embedding，就会得到我们想要的结果：



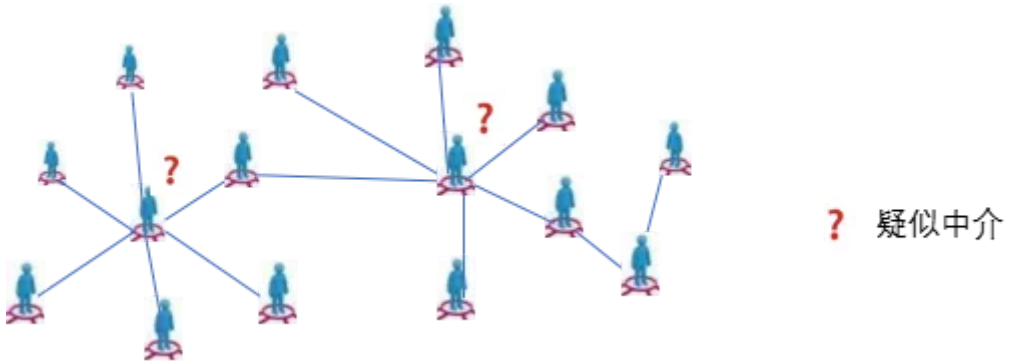
在网络中完全对称的节点，在embedding空间中也最为接近，这正是我们想要的结果。

应用场景举例

在预测逾期用户还款时，将所有借入用户根据他们所提供的联系人信息组成一个联系网络。对联系网络使用基于临近关系的embedding方法后，相关联的用户便具有了相近的表示向量。基于用户周围其他用户的逾期行为会影响他本身的假设，我们可以合理的将表示向量作为分类模型的输入，预测用户是否逾期。



不良中介会对业务带来极大的欺诈风险。我们基于中介在借出用户网络中具有特殊的结构，利用struc2vec模型，我们将结构相似的用户映射到低维空间中相临近的区域中。不良中介由于具有相似的结构特征，就会被映射到临近的区域。如果已知其中部分中介的标签，我们就可以精准的抓出另外一部分的不良中介。



模型延伸

半监督学习

聪明的小读者会再次注意到，上面提到的不论是基于临近关系还是基于网络结构的表示学习方法，在embedding的过程中都没有引入节点的标签信息，实际上我们可以这么做。

在训练一对儿节点的过程中，我们会得到一个(单词|背景)对，如(a, c)。在计算好这一对节点对应的embedding向量的损失函数之后，我们可以给它再乘以一项两个节点标签的f函数

$$Loss = -\log P(c|embedding(a)) \cdot f(label(c), label(a))$$

根据不同业务场景的需求，可以良定义这个f函数，使得不同标签的节点在embedding空间之间的距离被放大或者缩小。

模型的混合

我们还可以通过计算多个损失函数，将不同的模型混合起来，使得embedding向量既捕捉到局部网络的临近性特征，又可以很好的描述网络的结构特征。

后记

复杂网络的表示学习是拍拍贷与浙江大学通力合作的众多人工智能项目中的一员，目前还处于前期蓄力阶段，部分阶段性成果已经在反欺诈、催收、借出等场景中展开实验，我们将继续对算法的钻研和对落地场景的持续探索。

参考文献

[1] Perozzi, B., Al-Rfou', R. & Skiena, S. (2014). DeepWalk: Online Learning of Social Representations.. *CoRR*, abs/1403.6652.

[2] Grover, A. & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks.. *KDD*, .