**研究进展报告（2018.10.1）**

姓名：邓思宇

1. **本周计划**

**1.1具体任务**

1）完成基于randomwalk的主动学习实验部分

2）阅读文献《Personality-Based Active Learning for CF Recommender Systems》

3）学习TensorFlow—卷积神经网络

**1.2 任务完成情况**

（1）完成

（3）完成中

（4）完成中

1. **randomwalk—Active Learning**

**2. 1**决策信息系统

定义1．（决策信息系统[16]）决策信息系统定义成一个三元组：

 (1)

其中，*U*代表一个非空样本集合，也称论域；*C*代表一个非空条件属性集合；*d*指的是样本的决策属性．表1是一个决策信息系统．*U* = {*x*0, *x*1, *x*2, ..., *x*15}，C={*a*1, *a*2, *a*3, *a*4}．

**Table 1 Example of decision system**

1. **表1 决策信息系统**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *U* | *a*1 | *a*2 | *a*3 | *a*4 | *d* |
| *x*0 | 5.1 | 3.5 | 1.4 | 0.2 | is |
| *x*1 | 4.9 | 3.0 | 1.4 | 0.2 | is |
| *x*2 | 4.7 | 3.2 | 1.3 | 0.2 | is |
| *x*3 | 4.6 | 3.1 | 1.5 | 0.2 | is |
| *x*4 | 5.0 | 3.6 | 1.4 | 0.2 | is |
| *x*5 | 5.4 | 3.9 | 1.7 | 0.4 | is |
| *x*6 | 7.0 | 3.2 | 4.7 | 1.4 | iv |
| *x*7 | 6.4 | 3.2 | 4.5 | 1.5 | iv |
| *x*8 | 6.9 | 3.1 | 4.9 | 1.5 | iv |
| *x*9 | 5.5 | 2.3 | 4.0 | 1.3 | iv |
| *x*10 | 6.5 | 2.8 | 4.6 | 1.5 | iv |
| *x*11 | 6.3 | 3.3 | 6.0 | 2.5 | it |
| *x*12 | 5.8 | 2.7 | 5.1 | 1.9 | it |
| *x*13 | 7.1 | 3.0 | 5.9 | 2.1 | it |
| *x*14 | 6.5 | 3.0 | 5.8 | 2.2 | it |
| *x*15 | 7.6 | 3.0 | 6.6 | 2.1 | it |

定义2．（曼哈顿距离）向量*x =* [*x*1*, x*2*,…, xn*]与*y =* [*y*1, *y*2,…*, yn*]的曼哈顿距离为：

 (2)

1. 它表示在多维空间中两个点之间的距离．信息表的样本可以用向量表示．相应地，可以定义任意一组样本的相似度．

定义3．（相似度）给定一个决策信息系统*S* = ( *U*, *C*, *d* )，任意*x*, *y* ∈ *U*的相似度记为：

 (3)

根据式(2)(3)，可计算表1的决策信息系统中*sim*(*x*0, *x*6) = 0.13，*sim*(*x*3, *x*12) = 0.127．

定义4．（邻域）对于任意的样本*x* ∈ *U*，可以通过设置相似度阈值*θ*的方式确定其邻域，样本的邻域定义如下：

 (4)

相似度阈值*θ*越小，样本的邻域越大．根据表1所示的决策信息系统可以计算出*n*(*x*0, 0.5)={*x*1, *x*2, *x*3, *x*4}，*n*(*x*14, 0.2)={*x*6, *x*7, *x*8, *x*10, *x*11, *x*12, *x*13, *x*15}．

**2. 2随机游走**

无向加权图G=(V, E, w)．*V*是实例集合．设*n* = |*V*|．*E*是边集合，w是节点间的权重，用来衡量节点与节点间的相似度。节点S最多与k个邻居相连，邻居集合用表示。

定义5．（游走概率）给定加权图 G=(V, E, w)与起始点i，随机游走到i的邻居j的概率为：

|  |  |
| --- | --- |
| 算法一. | 权重优化算法 |
| Input: | 决策信息系统 . |
| Output: | 聚类信息块*bl*=[*bl*1, *bl*2, …, *blk*] . |
| 1. | **FOR** (each ) **DO** //找邻居,W*n*×*n*为通过相似度计算的权重矩阵 |
| 2. | kNearestNeighbors(i, k); |
| 3. | **ENDFOR** |
| 4. | Symmetrize(W); |
| 5. | **FOR** () **DO** |
| 6. | **FOR** (each ) **DO** // 计算转移概率矩阵 |
| 7. | **FOR** (each ) **DO** |
| 8. |  |
| 9. | **ENDFOR** |
| 10. | **ENDFOR** |
| 11. | **FOR** () **DO** //最多paraK步可达概率 |
| 12. | Pj=Pj-1P; |
| 13. | **ENDFOR** |
| 14. |  |
| 15. | **FOR** (each ) **DO** // 更新权重 |
| 16.  17.  18. | sij=  **IF**(sij > 6)**THEN**  wij = 0; |
| 19.  20. | **ELSE IF**  wij = exp(2k-wij)-1; |
| 21. | **END IF** |
| 22. | **ENDFOR** |
| 23. | **ENDFOR** |
| 24. | Return ; |

**2.3聚类方法**

得到权重矩阵W*n*×*n*后，将图G中连线划分为分割边或者非分割边．假设两节点权重足够大，可将该连边定义成非分割边，反之定义为分割边．这种边界划分方式基于一个阈值．采用深度优先遍历的方法对图G进行聚类。

**2.4主动学习**

主动学习阶段，利用上节中生成的信息块*bl*对代表样本进行标记和预测．

①如*bli*中存在未分类样本，则查询*bli*中部分样本的标签．

②如*bli*中已分类的样本数量足够大()且标签一致，则可预测该块中剩余样本的标签．

③增大阈值*threshold*，进行下一轮聚类、标记和预测．达到购买上限后，对不纯的块，采取投票的方式确定剩余未标记代表样本的标签．