Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems

Author:-

Chris Biemann

Deepanshu Vijay 201302093
Arushi Dogra 201302084

Introduction

- Clustering is defined as the task of grouping together objects in such a way that those objects that are similar to each other comes in the same group.
- The Chinese Whispers algorithm provides a basic yet very effective way to partition the nodes of a graph.
- It is a randomised clustering algorithm and is time linear in the number of edges and is capable of partitioning very large graphs in comparatively short time.

Relevant Work

- MCL : A clustering algorithm for graphs. http://micans.org/mcl/.
- Important concepts from Graph Theory (cf. Bollobás 1998).
- https://en.wikipedia.org/wiki/Cluster_analysis

Algorithm

- The algorithm works as follows in a bottom-up fashion.
- Initially all the nodes are assigned to different classes.
- Then the nodes are processed for a small number of iterations.
- In each iteration the nodes inherit the strongest class in the local neighborhood. This is the class whose sum of edge weight is maximum to the current node. If there are multiples Strongest class available, one of them is randomly chosen.
- One important thing to note here is that the classes are updated immediately. This means a node can obtain classes that were introduced in the same iteration.

Algorithm

```
initialize:
 forall v_i in V: class(v_i)=i;

while changes:
 forall v in V, randomized order:
   class(v)=highest  ranked  class
            in neighborhood of v;
```

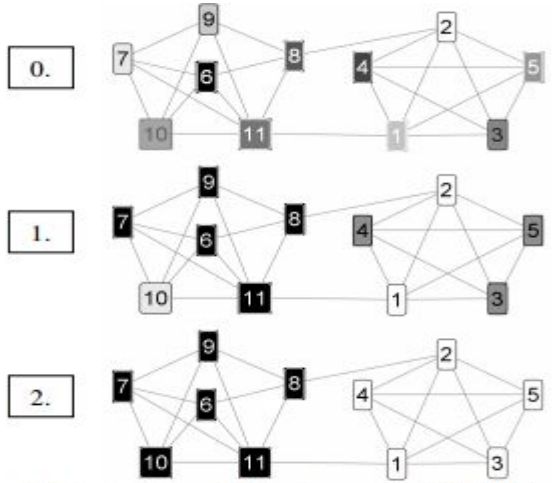

Figure 2: Clustering an 11-nodes graph with CW in two iterations

- Formally, the algorithm does not converge. But usually the class do not change after few iterations. The number of iterations depends on the diameter of the graph. Larger the diameter more number of iterations will be required.

Chinese Whispers as matrix operation

- The Chinese Whispers is a special case of Markov-Chain-Clustering(MCL). MCL is the parallel simulation of all possible random walks up to a finite length on a graph G.
- The idea is that random walkers are more likely to end up in the same cluster where they started than walking across clusters.
- MCL simulates flow on a graph by repeatedly updating transition probabilities between all nodes, eventually converging to a transition matrix after k steps that can be interpreted as a clustering of G. This is achieved by alternating an expansion step and an inflation step.
- The expansion step is a matrix multiplication of MG with the current transition matrix.
- The inflation step is a column-wise non-linear operator that increases the contrast between small and large transition probabilities and normalizes the column- wise sums to 1. The k matrix multiplications of the expansion step of MCL lead to its time-complexity of O(k*n2).

Chinese Whispers as matrix operation

- Let maxrow(.) be an operator that operates row-wise on a matrix and sets all entries of a row to zero except the largest entry, which is set to 1.
- By applying maxrow(.), Dt-1 has exactly n non-zero entries. This causes the time-complexity to be dependent on the number of edges, namely O (k·|E|). In the worst case of a fully connected graph, this equals the time-complexity of MCL.

$D^0 = I_n$
for t=1 to iterations
$D^{t-1} = \text{maxrow}(D^{t-1})$
$D^t = D^{t-1}A_G$

Figure 4: Matrix Chinese Whispers process. t is time step, I_n is the identity matrix of size n×n, A_G is the adjacency matrix of graph G.

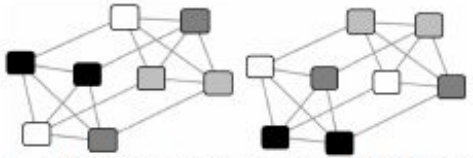

Figure 5: oscillating states in matrix CW for an unweighted graph

Thank You