**Integrated Mixture Model for Co-occurrence Data**

Loc Nguyen, Loc Nguyen’s Academic Network, Vietnam

Let *D* denote the co-occurrence data which is a set of co-occurrence events (*X*i, *Yi*) associated with values *V* such that *D* = {(*Xi*, *Yi*, *V*)} where *Xi* {1, 2,…, *M*}, *Yi* {1, 2,…, *N*}, and *V* is non-missing. Note that a co-occurrence event (*X*i, *Yi*) can occur many times, for example, there exists possibly three same triples in *D* as (1, 2, 1.0), (1, 2, 1.0), and (1, 2, 1.0). Hence, size of *D* which is denoted as |*D*| can be larger than *M*\**N*. Let *Z* = 1, 2,…, *K* be latent class of *X* and *Y*. Integrated mixture (IM) model *P*(*X*, *Y*, *V*) of co-occurrence data is defined as follows:

(Because we suppose that *V* only depends directly on latent class *Z*)

(Because we suppose that *X* and *Y* are mutually independent given *Z*)

(Because *PZ*(*V*) = *fZ*(*V*) is probabilistic density function when *V* is real variable)

Let

Note *αk*, *βk*(*X*), *γk*(*Y*), and *δk* are parameters of IM model. There are *K* parameters *αk*, *K*\**M* parameters *βk*(*X*), *K*\**N* parameters *γk*(*Y*), and *K* parameters *δk*. IM model is re-written as follows:

Let *x*(*V*) = *X* and *x*(*V*) = *X* be functions which map *V* to *X* and *V* to *Y*, respectively. For example, we have *x*(1, 2, 1.0) = 1 and *y*(1, 2, 1.0) = 2. According to EM algorithm, the conditional expectation *Q*(Θ | Θ(*t*)) is defined as follows:

Where,

Note Θ = (*αk*, *βk*(*X*), *γk*(*Y*), and *δk*)*T* be unified parameter of IM model. Suppose the distribution *fk*(*V* | *δk*) is normal distribution with mean *μk* and variance *σk*2 as follows:

Of course, we have *δk* = (*μk*, *σk*2)*T*. In E-step we calculate the probability *P*(*k* | *Vi*, Θ(*t*)). In M-step we maximize *Q*(Θ | Θ(*t*)) to estimate Θ(*t*+1) based on current Θ(*t*) and *P*(*k* | *Vi*, Θ(*t*)), which results out following estimation of Θ(*t*+1).

Where,

Estimation model is deduced from IM model as follows:

Given object *X* and feature *Y*, the value which is associated with *X* and *Y* is estimated as mean of *V* given the estimation model as follows: