Cluster reduction using locality aware clustering

December 18, 2013

1 Definition

A Gaussian mixture model (GMM) is a parametric statistical model that assumes that the data originates from a weighted sum of several Gaussian sources. More formally, a GMM is given by $p(x|\Theta) = \sum_{j=1}^{k} \alpha_j p(x|\theta_j)$, where α_j denotes the weight of each Gaussian, θ_j its respective parameters and k denotes the number of Gaussian sources in GMM. Expectation maximization (algorithmi 1) is a widely used method for estimating parameter set of the model (Θ) using unlabeled data [1]. In the algorithms, we optimize GMM for dataset X, which has m examples, and each of the Gaussians is parametrized by mean μ, covariance matrix Σ and its prior ϕ.

TODO: definicija preisana iz enega clanka, povzemi po svoje

Algorithm 1 Standard EM GMM

  function EM_GMM(max_steps, X)
    $initialize(\mu, \Sigma, \phi)$
    for $step$ **in** $1..max\_step$ **do**
      for each $i \leftarrow 1..m$, $j \leftarrow 1..k$ **do**
        $w_j^{(i)} \leftarrow p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$
      end for

      $\phi_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$
      $\mu_j \leftarrow \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$
      $\Sigma_j \leftarrow \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$
    end for
  end function

---

2 Our method

Learning a mixture of Gaussian models on multidimensional vectors constructs a model that takes all features into account. Clusters fit the data well, but when displayed on a parallel coordinates display they overlap.

Learning a mixture of Gaussians for each feature independently leads to clusters that fit value distribution for each feature well, but now examples that belong to the same cluster on one feature connect to multiple clusters on the other feature which results in noise.

Our method (algorithm 2) optimizes each component of the parameters independently, but considers local neighborhood when calculating updates.

Algorithm 2 Modified EM GMM

---

    function OUR_EM_GMM($max\_steps, window\_size, X$)
        $initialize(\mu, \Sigma, \phi)$
        for $step$ **in** $1..max\_step$ **do**
            for f **in** $features$ **do**
                $XS \leftarrow select\_features(f, window\_size, X)$
                for each $i \leftarrow 1..m,\ j \leftarrow 1..k$ **do**
                    $w_j^{(i)} \leftarrow p(z^{(i)} = j | xs^{(i)}; \phi, \mu, \Sigma)$
                end for

$$\phi_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$
$$\mu_{j,f} \leftarrow \frac{\sum_{i=1}^{m} w_j^{(i)} xs^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$
$$\Sigma_{j,f} \leftarrow \frac{\sum_{i=1}^{m} w_j^{(i)} (xs^{(i)} - \mu_j)(xs^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

            end for
        end for
    end function

References

[1] A. P. Dempster, N.M. Laird, and D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* JRSSB, 39:1-38, 1997