Data Driven Research

Editor:

1. methods

In this part we explain the methods we used to analyse the correlations between MeSH terms and of repositories. Basically, we consider all the trends throughout time as a *continues-time random process*, which is simply an infinite dimensional random variable. For example, let $\mathbf{y}(t)$ is a random process, infinite dimensional random variable, for repository trend which for each repository is different and it changes over time. For the sake of simplicity, We discretize the time for the random process $\mathbf{y}(t)$ by choosing to take samples every year. Thus we can express the random process of the trends of GEO repository as a limited sequence of random variables: $\mathbf{y}_{\mathrm{G}}(t) = [\mathbf{y}_{\mathrm{G}}(t_1), \cdots, \mathbf{y}_{\mathrm{G}}(t_T)]$. In the same way we define random process $\mathbf{y}_{\mathrm{K}}(t)$ for the trend of GenBank Repository, and random preocesses $\mathbf{x}_{\mathrm{M}i}(t)$ for i^{th} MeSH term in the dataset. Thus the goal is to find correlations, though local, between MeSH terms, $\mathbf{x}_{\mathrm{M}i}(t)$, and repositories, $\mathbf{y}_{\mathrm{G}}(t)$ and $\mathbf{y}_{\mathrm{K}}(t)$.

In our study we set $t_1 = 2004$ and $t_T = 2014$ and evaluate variables at the end of each year.

For the repository trends, we assigned the value of to the instantiation of each random variable by computing the normalized number of citations for each year, for example

$$\mathbf{y}_{\mathrm{G}}(t_j) = \frac{\mathbf{c}_{\mathrm{G}}(t_j)}{\sum_{i=1}^{T} \mathbf{c}_{\mathrm{G}}(t_i)} \tag{1}$$

where $\mathbf{c}_{\mathrm{G}}(t_i)$ is the number of times that datasets of GEO repository has been cited between t_{i-1} and t_i. The same quantity can be computed readily for GenBank repository.

For the MeSH term trends we need to evaluate the influence of each Mesh term in each year. In other words, in each year, we are interested in evaluating how much a MeSH term is a predictor of repository. For evaluating values of $\mathbf{x}_{\mathrm{M}i}(t)$, for each t we train a linear Support Vector Machine to predict(SVM) repository, based on MeSH terms that has been attributed to the datasets. So, we build the dataset $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ where n is the number of datasets $x_i \in \mathbb{N}^d$ is the MeSH term representation of the i^{th} dataset and $y_i \in \{GEO, GenBank\}$ is the label of that dataset. Here d is the number of mesh terms in our MeSH vocabulary and $x_i j$ is an integer value implying how many times the j^{th} MeSH term has appeared in the all papers that used the i^{th} dataset. We partitioned the dataset to disjoint set $\mathcal{D}^{(t)}$, which each-year-dataset $\mathcal{D}^{(t)}$ only contain datasets that has been used in the papers of that year. Then, for each year we trained an linear SVM, find $\mathbf{w}^{(t)}$ for all t, for each of the datasets and chose the hyperparameters so that the cross validation accuracy is maximized.

The linear SVM classifier trains a linear decision function $f(x_i) = \mathbf{w}^T x_i$ where x is the feature vector and $\mathbf{w} \in \mathbb{R}^d$ is the weight vector which is optimized in the SVM's training process. In addition to the model for predicting class labels, weights can be considered as a measure of how much each feature is predict of the class labels. For instance, the js for which $\mathbf{w}_j = 0$ (or close to zero) we can infer that corresponding features have zero (or negligible) influence on the decision value, and we can ignore them.

In this paper, we use the weights $\mathbf{w}_j^{(t)}$ to quantitatively measure the influence of j^{th} MeSH term in t^{th} year, and set $\mathbf{x}_{\mathrm{M}j}(t) = [\mathbf{x}_{\mathrm{M}j}(t_1), \cdots, \mathbf{x}_{\mathrm{M}j}(t_T)] = [\mathbf{w}_j(t_1), \cdots, \mathbf{w}_j(t_T)]$.

Having specified all the trends, we can now analyze the (local and global) correlations between each MeSH trend $\mathbf{x}_{\mathrm{M}j}(t)$ repository trend $\mathbf{y}_{\mathrm{G}}(t)$ and $\mathbf{y}_{\mathrm{K}}(t)$.