VALIDATION AND EXTENSION OF THE CLASSIFICATION MODEL

Update and validation of classification models can never be regarded as finished. Each application to a new area or new entities require a – at least in part – update and validation of the underlying model. As new data sets become available at our centre, these can be seen as additional opportunities for the application but also for the validation of the model.

1.1 Validation based on other input variables

One of the crucial steps in a cluster analysis is the choice of input variables. Milligan and Cooper (1987) describe this choice as the second phase in their seven-step structure of organization of the clustering process. The analysis and model presented in this thesis are based on research profiles using the subject classification scheme developed by Glänzel and Schubert (2003). This scheme consists of 16 major fields in the natural and life sciences (13 fields), social sciences (2 fields) and humanities (1 field). Of course, this is not the only scheme or subject classification system available. Other subject classification schemes exist or are being developed and thus can be used for the validation of the classification model.

The WoS, the data of which are used in this thesis, provides the “ISI subject Categories” as classification scheme. This system assigns journals to about 250 different subject categories across fields in natural and life sciences, social sciences and arts & humanities. Using this very fine grained system would create vectors in a 250 dimensional space. Such a high dimensionality can easily become problematic for clustering exercises especially when vectors contain zero’s on many dimensions. This will be the case for specialized institutions which make up two thirds of the database. Furthermore, many journals are assigned to two or more different subject categories so that the “discriminative power” of this scheme is rather low. So prior to clustering this data, an aggregation is needed to reduce the dimensionality in the data and to avoid all too frequent multiple assignment of journals. However, this is exactly what has been done by Glänzel and Schubert resulting in classification system hierarchically on top of the ISI classification. Other statistical techniques like PCA or factor analysis could be used to statistically aggregate these 250 categories but this would only be a statistical exercises with an aggregate not that different from the one used already and thus not being a real validation of the clustering and classification model.

Other subject classification schemes available at or developed by our centre can serve as real alternatives.
1.1.1 Scopus

With the release of Elsevier’s bibliographic citation database Scopus a different system became available. In this database document or papers are assigned to 335 different classes using Elsevier’s ASJC system. Also here the above mentioned problem of a high dimensionality of the field space come into play.

Two different approaches can be taken to tackle this problem.

A mapping can be created between Scopus ASJC system (All Science Journals Classification) and the 16 fields in the Glänzel and Schubert scheme used for the calculation of the research profiles. The advantage is that this approach allows the direct comparison of the research profiles from both systems. The main disadvantage lies in the mapping itself as it can unintentionally induce the similarity with the WoS classification system.

A new hierarchical system can be developed to aggregate the 335 different classes of the ASJC system to a higher level. A possibility for this aggregation is the Scopus’ Source Subject Classification system where 4 subject areas are divided in 30 categories. Another option is a statistical aggregation using PCA or factor analysis or other techniques. Both options take the full peculiarities and specialties of the Scopus system into account but undermines the direct comparison of research profiles with the ones used and described in the thesis. The WoS profile is a vector based on 16 fields while the Scopus subject area vector would have 30 fields or categories. A real comparison is only possible after the development of a complete new clustering and classification model for the institutions into 8 groups and then compare membership between both classification models.

1.1.2 Hybrid clustering

The subject classification scheme used in this thesis is based on journal assignments. Each journal is assigned to at least one science field and all papers published within one journal get the same subject classification. For the main part of papers in the database this is a correct classification but sometimes it does not hold. For instance, one of the papers on bibliometric indicators and relational charts (Glänzel et al, 2008) I had co-authored has appeared in the journal ‘Archivum Immunologiae et Therapiae Experimentalis’. According to our subject classification this paper would be assigned to “Clinical and Experimental Medicine I (General and Internal Medicine)” as a paper published in a journal in the field of Immunology. However, the above mentioned paper is a pure bibliometric one.

Jointly with the Electrical Engineering Department at our university a novel subject delineation strategy was developed (see Glänzel et al., 2009; for bioinformatics). This strategy combines the analysis and mining of textual components like keywords, titles and abstracts with the bibliometrical, citation-based techniques. This hybrid approach enables a subject classification at the
paper level by using both formal links between papers in citations and informal links induced by text similarity. Results of this approach can also be used to create vectors to locate institutions in a field space.

1.2 Validation based on other entities or institutions

The first phase in the structure of a clustering process described by Milligan and Cooper (1987) is the selection of entities or elements to be clustered. I have used European research institutions as entities for my clustering. However, as described in the next chapter, also data on US research institutions became available at our centre. This new dataset creates opportunities for a replication analysis with two independent datasets. A replication analysis involves a rerun of the clustering analysis with different choices made at some point in the process. In this case a shift from European data to US institutions would involve a different choice in the first phase of the process. Afterwards the agreement or consistency between the two results is measured and evaluated.

McIntyre and Blashfield (1980) proposed an approach suitable for replication analysis in cluster analysis for different datasets. After the clustering of the first dataset the centroids for each cluster are calculated. A k-means clustering is done on the second dataset using the centroids from the first analysis as input points. The second dataset is also directly cluster-analyzed using the same distance measure, aggregation method and number of clusters as the clustering of the first dataset. In a last step a measure of consistency can be measured between the two different clusterings of the second dataset. McIntyre and Blashfield (1980) suggest the kappa statistic but Milligan and Cooper (1987) suggest the corrected Rand index proposed by Hubert and Arabie (1985).

The results of this replication analysis can enhance the external validity of the classification model.
References


