New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping

JEAN-CHARLES LAMIREL,\textsuperscript{a} CLAIRE FRANCOIS,\textsuperscript{b} SHADI AL SHEHABI,\textsuperscript{a} MARTIAL HOFFMANN\textsuperscript{b}

\textsuperscript{a} LORIA, Vandoeuvre-lès-Nancy (France)
\textsuperscript{b} URE/INIST-CNRS, Vandoeuvre-lès-Nancy (France)

The information analysis process includes a cluster analysis or classification step associated with an expert validation of the results. In this paper, we propose new measures of Recall/Precision for estimating the quality of cluster analysis. These measures derive both from the Galois lattice theory and from the Information Retrieval (IR) domain. As opposed to classical measures of inertia, they present the main advantages to be both independent of the classification method and of the difference between the intrinsic dimension of the data and those of the clusters.

We present two experiments on the basis of the MultiSOM model, which is an extension of Kohonen’s SOM model, as a cluster analysis method. Our first experiment on patent data shows how our measures can be used to compare viewpoint-oriented classification methods, such as MultiSOM, with global cluster analysis method, such as WebSOM. Our second experiment, which takes part in the ECSTES EEC project, is an original Webometrics experiment that combines content and links classification starting from a large non-homogeneous set of web pages. This experiment highlights the fact that break-even points between our different measures of Recall/Precision can be used to determine an optimal number of clusters for web data classification. The content of the clusters obtained when using different break-even points are compared for determining the quality of the resulting maps.

Introduction

In the procedure of information analysis, a general problem is the evaluation of the results of the data classification methods. The complexity of the studied topics, combined with the weaknesses of the most widespread objective classification quality estimators, such as inertia, may finally make it necessary to call on an expert of the studied domain for a subjective evaluation of the quality of the classification results. In this paper we propose new objective classification quality estimators for both evaluating and optimising the results of the classification and mapping methods, especially when they are applied to the domain of documentary databases. Our estimators have been experienced with two different methods. The first method consists in using the
estimators for comparing the efficiency of the viewpoint’s oriented data analysis methods with the efficiency of the global analysis methods on the same set of data, the latter being composed of a patent collection. The second method consists in using the estimators for optimising the results of an original Webometrics experiment that combines content and link classification, starting from a large non-homogeneous set of web pages.

A new set of measures for class quality evaluation

When comparing classification methods, one is faced with the problem of selecting reliable classification quality measures. The classical evaluation measures of the quality are based on the intra-class inertia and the inter-class inertia.\(^8,14\)

The intra-class inertia can be defined as:

\[
\frac{1}{|C|} \sum_{c \in C} \frac{1}{|C_c|} \sum_{e \in C_c} \| p_e - p_c \|^2
\]

(Eq. 1)

where \(C\) represents the set of classes associated with the classification, \(d\) represents a class member and \(p_e\) represents the profile vector (i.e. profile) associated with the element \(e\).

The inter-class inertia may be defined as:

\[
\frac{1}{|C|^2 - |C|} \sum_{c \in C} \sum_{c' \neq c} \| p_c - p_{c'} \|^2
\]

On the basis of these two measures, a classification is considered as satisfactory if it possesses low intra-class inertia as compared with its inter-class inertia. Nevertheless, these measures are often biased in several ways.

A first bias of these measures is related to the fact that the intrinsic dimensions of the class’s profiles (number of non-zero components in the profiles) are not of the same order of magnitude as the intrinsic dimensions of the data profiles. It is especially true in the documentary domain, where the average number of indexes in the documents is extremely low as compared with the dimension of their overall description space. This phenomenon causes an abnormal increase of the intra class inertia. Normalisation of the class profiles during the class construction phase, which is obtained with spherical classification models like Axial K-means,\(^9\) could indirectly help to solve this problem. Nevertheless an undesirable side effect of this solution is the averaging of the class profiles.

The intra-class inertia measures the distance of the class elements from the profile of a class to which they have been affected after the classification process. As a second bias, this latter measure might not be able to make a proper distinction between