Résumé

Les méthodes à noyaux désignent une famille récente d'outils d'analyse de données, pouvant être utilisés dans une grande variété de tâches classiques comme la classification ou la régression. Ces outils s'appuient principalement sur le choix a priori d'une fonction de similarité entre paires d'objets traités, communément appelée "noyau" en apprentissage statistique et analyse fonctionnelle. Ces méthodes ont récemment gagné en popularité auprès des praticiens par leur simplicité d'utilisation et leur performance.

Le choix d'un noyau adapté à la tâche traitée demeure néanmoins un problème épineux dans la pratique, et nous proposons dans cette thèse plusieurs noyaux génériques pour manipuler des objets structurés, tels que les séquences, les graphes ou les images. L'essentiel de notre contribution repose sur la proposition et l'étude de différents noyaux pour nuages de points ou histogrammes, et plus généralement de noyaux sur mesures positives. Ces approches sont principalement axées sur l'utilisation de propriétés algébriques des ensembles contenant les objets considérés, et nous faisons ainsi appel pour une large part à la théorie des fonctions harmoniques sur semigroupes. Nous utilisons également la théorie des espaces de Hilbert à noyau reproduisant dans lesquels sont plongées ces mesures, des éléments d'analyse convexe ainsi que plusieurs descripteurs de ces mesures utilisés en statistiques ou en théorie de l'information, comme leur variance ou leur entropie. En considérant tout objet structuré comme un ensemble de composants, à l'image d'une séquence transformée en un ensemble de sous-séquences ou d'images en ensembles de pixels, nous utilisons ces noyaux sur des données issues principalement de la bioinformatique et de l'analyse d'images, en les couplant notamment avec des méthodes discriminantes comme les machines à vecteurs de support.

Nous terminons ce mémoire sur une extension de ce cadre, en considérons non plus chaque objet comme un seul nuage de point, mais plutôt comme une suite de nuages emboîtés selon un ensemble d'événements hiérarchisés, et aboutissons à travers cette approche à une famille de noyaux de multirésolution sur objets structurés.

Abstract

Kernel methods refer to a new family of data analysis tools which may be used in standardized learning contexts, such as classification or regression. Such tools are grounded on an *a priori* similarity measure between the objects to be handled, which have been named "kernels" in the statistical learning and functional analysis literature. The simplicity of kernel methods comes from the fact that, given a learning task, such methods only require the definition of a kernel to compare the objects to yield practical results.

The problem of selecting the right kernel for a task is nonetheless tricky, notably when the objects have complex structures. We propose in this work various families of generic kernels for composite objects, such as strings, graphs or images. The kernels that we obtain are tailored to compare clouds of points, histograms or more generally positive measures. Our approach is mainly motivated by algebraic considerations on the sets of interests, which is why we make frequent use of the theory of harmonic functions on semigroups in this work. The theoretical justification for such kernels is further grounded on the use of reproducing kernel Hilbert spaces, in which the measures are embedded, along with elements of convex analysis and descriptors of the measures used in statistics and information theory, such as variance and entropy. By mapping any structured object to a cloud of components, e.g., taking a string and turning it into a cloud or a histogram of substrings, we apply these kernels on composite objects coupled with discriminative methods, such as the support vector machine, to address classification problems encountered in bioinformatics or image analysis.

We extend this framework in the end of the thesis to propose a different viewpoint where objects are no longer seen as clouds of points but rather as nested clouds, where each cloud is labelled according to a set of events endowed with a hierarchy. We show how to benefit from such a description to apply a multiresolution comparison scheme between the objects.