Rtkpp: Un package pour faire l'interface entre R et la bibliothèque STK++

S. Iovleff [a]

[a]Université Lille 1 (UMR 8524), Inria Lille Nord Europe (Modal Team)
59650 Villeneuve d'Ascq - France
serge.iovleff@inria.fr

Mots clefs : C++, Big data, parallélisme .

Introduction

STK++ (The Statistical ToolKit, `http://www.stkpp.org`) est une librairie écrite en C++ qui est développée de manière continue depuis dix ans. Elle a connu un essor particulièrement important ces trois dernières années avec sa participation à plusieurs activité de valorisation et de transfert, en particulier par la création de packages R s'appuyant sur STK++.
La dernière étape significative a consisté à la mettre à disposition de la communauté R au travers du package rtkpp que nous détaillerons dans cette présentation.

Description de STK++

La librairie STK++ est une librairie écrite en C++ qui est divisée en plusieurs projets dont les principaux sont :

1. le projet *Arrays* propose un ensemble classe de tableaux dont les principales fonctionnalités sont
   - la gestion native des données manquantes,
   - ajouté, enlevé, permuté, lignes et colonnes de tableaux facilement,
   - utilisation de techniques de "template programming" pour optimiser les parcours ds tableaux,
   - utilisation des "template expression" pour évaluer sans coûts additionnels des expressions complexes,
2. le projet *STatistiK* fournit des classes qui englobent les principales lois de probabilités et les traitements statistiques usuels,
3. le projet *Algebra* fournit les méthodes d'algèbre linéaire usuelles et une interface avec lapack,
4. les projets t *Reduct* et *Regress* fournissent des classes pour faire de la réduction de dimension, effectuer des régressions linéaires et des régressions splines,...
5. le projet *Clustering* est dédié à l'apprentissage non-supervisé.

Plusieurs de ces projets sont parallélisés et permettent de répondre à certains défis posés par le Big data.

Présentation de rtkpp

Le but du package rtkpp est de faire le lien entre l'environnement R et la librairie STK++. Il utilise le (incoutournable) package Rcpp [1] pour réaliser ce lien. Le package Rcpp propose de spécialiser deux fonctions pour réaliser les transferts de données entre me R et le C++

```
Rcpp::wrap<>() // conversion from C++ to R
Rcpp::as<>() // conversion from R to C++
```

Ces deux fonctions ont été étendu de manière à pouvoir utiliser directement des tableaux R au sein de STK++. Toutefois rtkpp propose un support direct de certaines structures de donnÃ©es R. Voici un exemple

```
RcppExport SEXP countNA( SEXP r_matrix)
{
  BEGIN_RCPP
  // wrap R matrix in a STK matrix without data copy
  STK::RMatrix<double> m_data(r_matrix);
  // use STK::wrap for a direct evaluation without intermediaries arrays
  return Rcpp::List::create(
     Rcpp::Named("rows") = STK::wrap(STK::countByRow(m_data.isNA())),
     Rcpp::Named("cols") = STK::wrap(STK::count(m_data.isNA()))
                          );
  END_RCPP
}
```

Utilisation de rtkpp par d'autres packages

rtkpp est actuellement utilisé par deux packages :

• HDPenReg [2] qui effectue des régressions en grande dimension avec des pénalités lasso, fusion et fused-lasso.

• MixAll qui permet d'estimer des modèles de mélange gaussien, gamma, poisson en présence de données manquantes.

De nombreux autres packages devraient suivre dans les mois qui suivent.

Références

[1] Dirk Eddelbuettel, Romain François (2011). Rcpp: Seamless R and C++ Integration *Journal of Statistical Software*, **40**(8), 1-18

[2] Quentin Grimonprez (2015). HDPenReg: High-Dimensional Penalized Regression. *http://cran.r-project.org/web/packages/HDPenReg/*