Приложение

R: Справочная карта по Data Mining

Приведена по Yanchang Zhao, yanchang@rdatamining.com (с незначительными изменениями и дополнениями). Последнюю версию см. на http://www.RDataMining.com. Имена пакетов представлены в круглых скобках с подчеркиванием. Рекомендуемые пакеты и функции показаны жирным шрифтом.

СТАТИСТИКА (STATISTICS)

Описательная статистика (Summarization)

summary()	обобщение результатов
describe()	краткие описательные статистики данных (Hmisc)
boxplot.stats()	диаграммы размахов (box plot) и сопряженные с ними статистики

Дисперсионный анализ (Analysis of Variance)

aov()	оценивание и вывод таблицы дисперсионного анализа
anova()	расчет таблицы анализа дисперсий (или девианса) для одной или более моделей

Статистические тесты (Statistical Tests)

chisq.test()	тест хи-квадрат для таблиц сопряженности и оценки качества подгонки моделей
ks.test()	тест Колмогорова-Смирнова
t.test()	тест по *t* -критерию Стьюдента
prop.test()	тест на значимость заданной пропорции или их эквивалентность
binom.test()	точный биномиальный тест

Регрессионный анализ (Regression Functions)

lm()	линейные модели
glm()	обобщенные линейные модели
gbm()	обобщенные регрессионные бустинг-модели (gbm)
predict()	метод для получения предсказанных значений
residuals()	остатки модели, разности наблюдаемых и предсказанных значений
nls()	нелинейная регрессия
gls(), gnls()	построение линейных и нелинейных моделей методом обобщенных наименьших квадратов (nlme)

Модели со смешанными эффектами (nlme) (Mixed Effects Models)

lme(), nlme()	линейные и нелинейные модели со смешанными эффектами

Principal Components and Factor Analysis

princomp(), prcomp()	анализ главных компонент и факторный анализ

Обнаружение выбросов (Outlier Detection)

boxplot.stats()$out	перечень наблюдений, выходящих за пределы интервала,

	который ограничен "усами" диаграммы размахов
lofactor()	расчет фактора локальных выбросов по LOF-алгоритму (DMwR или dprep)
lof()	параллельная реализация LOF-алгоритма (Rlof)

Прочие функции

var(), cov(), cor()	дисперсия, ковариация, корреляция
density()	вычисление оценки ядерной плотности
cmdscale()	многомерное шкалирование (MDS)

Пакеты

gbm	обобщенные регрессионные бустинг-модели
nlme	линейные и нелинейные модели со смешанными эффектами
Rlof	параллельная реализация LOF-алгоритма
extremevalues	поиск экстремальных значений в одномерных данных
outliers	использование нескольких общих методов обнаружения выбросов
mvoutlier	поиск многомерных выбросов с использованием робастных методов

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ (TIME SERIES ANALYSIS)

Преобразование и отображение (Construction & Plot)

ts()	создание объектов класса "временной ряд"
plot.ts()	метод для визуализации объектов класса "временной ряд"
smoothts()	сглаживание временных рядов (ast)
sfilter()	удаление сезонных флуктуаций с использованием скользящего среднего (ast)

Декомпозиция (Decomposition)

decomp()	декомпозиция временного ряда по фильтру квадратного корня (timsac)
decompose()	классическая сезонная декомпозиция по скользящему среднему
stl()	сезонная декомпозиция временного ряда по локальной регрессии
tsr()	декомпозиция временного ряда (ast)
ardec()	декомпозиция временного ряда по авторегрессии (ArDec)

Прогнозирование (Forecasting)

arima()	построение моделей ARIMA для одномерного временного ряда
predict.Arima()	прогнозируемые значения по моделям ARIMA
auto.arima()	подгонка оптимальной модели ARIMA для одномерного ряда (forecast)
forecast.stl(), forecast.ets(), forecast.Arima()	прогнозирование ряда с использованием STL, ETS и ARIMA моделей (forecast)

Корреляция и ковариация (Correlation and Covariance)

acf()	автокорреляция и автоковариация временного ряда
ccf()	кросс-ковариация и кросс-корреляция двух одномерных временных рядов

Пакеты

forecast	анализ и визуализация прогнозов одномерного временного ряда
hts	анализ и прогнозирование иерархических и сгруппированных рядов
TSclust	функции для кластеризации временных рядов
dtw	динамическая трансформация шкалы времени (Dynamic Time Warping, DTW)
timsac	функции для анализа и преобразования временных рядов
ast	анализ временных рядов
ArDec	декомпозиция временного ряда, основанная на авторегрессии
dse	средства для создания многомерных, линейных и инвариантных моделей временных рядов

ФУНКЦИИ ДЛЯ ВЫДЕЛЕНИЯ АССОЦИАТИВНЫХ ПРАВИЛ И ПАТТЕРНОВ ПОСЛЕДОВАТЕЛЬНОСТЕЙ (ASSOCIATION RULES AND SEQUENTIAL PATTERNS FUNCTIONS)

apriori()	поиск ассоциаций по алгоритму APRIORI, который последовательно подсчитывает частоты одновременно происходящих событий, априори отсекая маловероятные (arules)
eclat()	поиск частых наборов событий по алгоритму Eclat, который перебирает классы эквивалентности и их пересечения (arules)
cspade()	поиск частых фрагментов последовательностей по алгоритму cSPADE (arulesSequences)
seqefsub()	поиск частых подпоследовательностей (TraMineR)

Пакеты

arules	поиск частых, максимальных или закрытых наборов событий и ассоциативных правил с использованием двух алгоритмов: Apriori и Eclat.
arulesViz	визуализация ассоциативных правил
arulesSequences	дополнение к arules для обработки и выделения частых последовательностей
TraMineR	поиск, описание и визуализация последовательностей объектов или событий

КЛАССИФИКАЦИЯ И ПРЕДСКАЗАНИЕ (CLASSIFICATION & PREDICTION)

Деревья решений (Decision Trees)

ctree()	деревья условного вывода, рекурсивное разбиение для непрерывных, цензурированных, упорядоченных, категориальных и многомерных откликов в рамках условного вывода (party)
rpart()	деревья рекурсивного разбиения и регрессии (rpart)
mob()	рекурсивное разбиение, приводящее к созданию деревьев, конечные узлы которого содержат статистические модели для соответствующих поднаборов данных (party)
varimp()	важность предикторов (party)

Случайный лес (Random Forest)

cforest()	ансамбли моделей, создаваемые с использованием алгоритмов "случайный лес" и "бэггинг" (party)

randomForest()	случайный лес (randomForest)
importance()	важность предикторов (randomForest)

Нейронные сети (Neural Networks)

nnet()	построение нейронной сети с одним скрытым слоем (nnet)
neuralnet()	обучение нейронных сетей (neuralnet)
mlp(), dlvq(), rbf(), rbfDDA(), elman(), jordan(), som(), art1(), art2(), artmap(), assoz()	различные типы нейронных сетей (RSNNS)

Машины опорных векторов (Support Vector Machine - SVM)

svm()	обучение машины опорных векторов для регрессии, классификации или оценки плотности вероятности (e1071)
ksvm()	машины опорных векторов (kernlab)

Байесовские классификаторы (Bayes Classifiers)

naiveBayes()	наивный баейсовский классификатор (e1071)

Оценка эффективности моделей (Performance Evaluation)

performance()	рассчитывает различные меры качества предсказательных моделей (ROCR)
PRcurve()	кривые чувствительности и специфичности классификатора (DMwR)
CRchart()	графики кумулятивной чувствительности классификатора (DMwR)
roc()	построение ROC-кривых (pROC)
auc()	вычисление площади под ROC-кривой (pROC)
ROC()	визуализация ROC-кривой (DiagnosisMed)

Пакеты

party	рекурсивное разбиение
rpart	деревья рекурсивного разбиения и регрессии
rpartOrdinal	деревья классификации для откликов с упорядоченными категориями
rpart.plot	метод для визуализации rpart-моделей
randomForest	классификация и регрессия на основе случайного леса
caret	модели классификации и регрессии
nnet	нейронные сети встречного распространения и лог-линейные модели для откликов с несколькими классами
RSNNS	реализация Штутгартского Симулятора Нейронных Сетей в R (SNNS)
neuralnet	обучение нейронных сетей обратного распространения и устойчивого обратного распространения с учетом или без учета весовых коэффициентов
e1071	функции для анализа латентных классов, Фурье-преобразования коротких временных рядов, нечеткой кластеризации, обучения машин опорных векторов, вычисления кратчайшего пути, бэггинг-кластеризации, построения наивного байесовского классификатора и др.
ROCR	визуализация результатов оценки качества классификаторов
pROC	визуализация и анализ ROC-кривых

КЛАСТЕРИЗАЦИЯ (CLUSTERING)

Кластеризация, основанная на разбиение (Partitioning based Clustering)

Разбиение данных на *k* групп с последующей попыткой улучшить качество кластеризации путем перемещения объектов из одной группы в другую

kmeans()	выполняет кластеризацию по методу *k* средних для некоторой матрицы с данными
kmeansruns()	вызывает функцию kmeans() для выполнения кластеризации по методу *k* средних и выполняет нахождение оптимального числа кластеров с использованием нескольких начальных расположений их центров (fpc)
pam()	реализация метода "разделение вокруг медоидов" (PAM) (cluster)
pamk()	реализация метода "разделение вокруг медоидов" (PAM) с одновременным нахождением оптимального числа кластеров (fpc)
kmeansCBI()	интерфейс для взаимодействия с функциями из пакета kmeans (fpc)
cluster.optimal()	поиск оптимального числа кластеров в некотором наборе данных (bayesclust)
clara()	кластеризация для больших наборов данных (cluster)
fanny(x,k,…)	выполнение нечеткой кластеризации с *k* кластерами (cluster)
kcca()	кластеризация по *k* центроидам (flexclust)
ccfkms()	кластеризация с использованием конъюгатных выпуклых функций (cba)
apcluster()	кластеризация по методу "передачи сообщений" (affinity propagation) на основе входной матрицы сходств (apcluster)
apclusterK()	кластеризация по методу "передачи сообщений" для нахождения *k* кластеров (apcluster)
cclust()	выпуклая кластеризация, включая метод *k* средних и два других метода (cclust)
KMeansSparseCluster()	кластеризация по методу *k* средних с одновременным нахождением информативных переменных (sparcl)
tclust(x, k, alpha,…)	кластеризация по методу *k* усеченных средних (часть наблюдений удаляется из рассмотрения) (tclust)

Иерархическая кластеризация (Hierarchical Clustering)

Иерархическая декомпозиция данных либо снизу вверх (агломерация), либо сверху вниз (разделение)

hclust()	иерархический кластерный анализ по матрице расстояний
birch()	алгоритм BIRCH для кластеризации очень больших объемов данных с использованием CF-деревьев (birch)
pvclust()	иерархическая кластеризация с одновременным вычислением *p*-значений путем извлечения бутстреп-выборок разного размера (pvclust)
agnes()	агломеративный иерархический кластерный анализ (cluster)
diana()	иерархический кластерный анализ на основе разделения (cluster)

`mona()`	иерархический кластерный анализ на основе разделения для данных, представленных только бинарными переменными (`cluster`)
`rockCluster()`	кластеризация с использованием алгоритма Rock (`cba`)
`proximus()`	кластеризация на основе алгоритма Proximus для данных, представленных только бинарными переменными (`cba`)
`isopam()`	алгоритм кластеризации Isopam (`isopam`)
`flashClust()`	оптимальная иерархическая кластеризация (`flashClust`)
`fastcluster()`	быстрая иерархическая кластеризация (`fastcluster`
`cutreeDynamic(), cutreeHybrid()`	выделение кластеров на дендрограммах (`dynamicTreeCut`)
`Hierarchical SparseCluster()`	иерархическая кластеризация с одновременным нахождением информативных переменных (`sparcl`)

Модели, основанные на кластерах (Model based Clustering)

`Mclust()`	кластеризация на основе статистических моделей (`mclust`)
`HDDC()`	кластеризация на основе статистических моделей для данных большой размерности (`HDclassif `)
`fixmahal()`	кластеризация с использованием фиксированных точек и расстояния Махаланобиса (`fpc`)
`fixreg()`	кластеризация на основе регрессии по фиксированным точкам (`fpc`)
`mergenormals()`	кластеризация, основанная на слиянии компонент смешанного гауссова распределения (`fpc`)

Кластеризация, основанная на плотности (Density based Clustering)

Формирование кластеров путем объединения участков с плотным расположением точек

`dbscan(data, eps, MinPts,…)`	нахождение кластеров произвольной формы с использованием параметров `eps` (радиус, в пределах которого лежат точки-соседи) и `MinPts` (пороговое значение плотности расположения точек) (`fpc`)
`pdfCluster()`	кластеризация на основе ядерной плотности вероятности (`pdfCluster`)

Другие техники кластеризации (Other Clustering Techniques)

`mixer()`	кластеризация на основе случайных графов (`mixer`)
`nncluster()`	быстрая кластериазация на основе алгоритма "restarted minimum spanning tree" (`nnclust`)
`orclus()`	кластеризация на основе алгоритма ORCLUS (`orclus`)

Отображение результатов кластеризации (Plotting Clustering Solutions)

`plotcluster()`	визуализация групп данных (`fpc`)
`bannerplot()`	горизонтально ориентированный "биплот", изображающий результат иерархической кластеризации (`cluster`)

Оценка качества кластеризации (Cluster Validation)

`silhouette()`	вычисление и извлечение информации по "силуэтам" кластеров (`cluster`)
`cluster.stats()`	вычисление нескольких статистик качества кластеризации на основе матрицы расстояний (`fpc`)

`clValid()`	вычисление статистик качества для нескольких алгоритмов кластеризации и числа кластеров (`clValid`)
`clustIndex()`	вычисление нескольких индексов кластеризации, которые можно использовать для нахождения оптимального числа кластеров (`cclust`)
`NbClust()`	позволяет вычислить 30 индексов для оценки качества кластеризации и нахождения оптимального числа кластеров (`NbClust`)

Пакеты

`cluster`	кластерный анализ
`fpc`	различные методы кластеризации и валидации полученных решений
`mclust`	кластеризация, основанная на статистических моделях
`birch`	кластеризация очень больших наборов данных с использованием алгоритма BIRCH
`pvclust`	иерархическая кластеризация с расчетом *p*-значений
`apcluster`	кластеризация на основе метода "передачи сообщений"
`cclust`	методы выпуклой кластеризации, включая алгоритм *k* средних, алгоритм обновления на новых данных, алгоритм "нейронного газа", а также вычисление индексов для нахождения оптимального числа кластеров
`cba`	кластеризация для решения бизнес-задач, включая такие алгоритмы, как Proximus и Rock
`bclust`	байесовская кластеризация на основе иерархической модели, подходящая для нахождения групп в данных большой размерности
`biclust`	алгоритмы для нахождения двумерных кластеров
`clue`	ансамбли кластерных решений
`clues`	метод кластеризации, основанный на локальной регуляризации
`clValid`	валидация кластерных решений
`clv`	методы валидации кластерных решений, включая популярные методы внутренней и внешней валидации
`cluster.bayesclust`	статистические тесты для валидации кластеров, полученных на основе геномных данных
`clustsig`	анализ статистической значимости кластеров, а также разницы между кластерами
`clusterSim`	поиск оптимальной процедуры кластеризации для имеющихся данных
`clusterGeneration`	имитация кластеров
`gcExplorer`	инструмент для графического анализа результатов кластеризации
`hybridHclust`	гибридный иерархический кластерный анализ на основе "совместных кластеров"
`Modalclust`	иерархический кластерный анализ на основе мод
`iCluster`	кластеризация на основе разнородных генетических данных
`EMCC`	эволюционные методы Монте-Карло (EMC) для кластеризации
`rEMM`	расширяемая марковская модель (EMM) для кластеризации потоковых данных

ОБРАБОТКА ТЕКСТОВ (TEXT MINING)

Импорт, очистка и подготовка текстов (Importing Text, Cleaning and Preparation)

`readPDF()`	извлечение текста и метаданных из документов формата PDF (`tm`)
`Corpus()`	построение корпуса, т.е. коллекции из нескольких документов (`tm`)
`tm_map()`	преобразование текстовых документов, т.е. стемминг, удаление стоп-слов и т.д. (`tm`)
`tm_filter()`	отфильтровывание документов из корпуса (tm)
`TermDocumentMatrix()`, `DocumentTermMatrix()`	создание терм-документных и документ-термных матриц (`tm`)
`Dictionary()`	создание словаря их текстового вектора или терм-документной матрицы (`tm`)
`stemDocument()`	стемминг слов в документе (`tm`)
`stemCompletion()`	восстановление полной формы слов после стемминга (`tm`)
`SnowballStemmer()`	стемминг по алгоритму Snowball (`Snowball`)
`stopwords(language)`	возвращает стоп-слова из разных языков (`tm`)
`removeNumbers()`, `removePunctuation()`, `removeWords()`	удаление чисел, знаков пунктуации или некоторого набора слов из документа (`tm`)
`removeSparseTerms()`	удаление редких термов из терм-документной матрицы (`tm`)

Обнаружение часто встречающихся термов и ассоциаций (Frequent Terms and Association)

`findAssocs()`	находит связи между термами в терм-документных матрицах (`tm`)
`findFreqTerms()`	находит часто встречающиеся термы в терм-документных матрицах (`tm`)
`termFreq()`	формирование вектора с частотами термов для заданного документа (`tm`)

Тематическое моделирование (Topic Modelling)

`LDA()`	подгонка LDA-модели (Latent Dirichlet Allocation) (`topicmodels`)
`CTM()`	подгонка CTM-модели (Correlated Topics Model) (`topicmodels`)
`terms()`	извлечение наиболее вероятных термов для заданной темы (`topicmodels`)
`topics()`	извлечение наиболее вероятных тем для заданного документа (`topicmodels`)
`polarity()`	индекс полярности (в анализе тональности текстов) (`qdap`)
`textcat()`	классификация текстов на основе *n*-грам (`textcat`)

Визуализация текста (Text Visualization)

`wordcloud()`	создание "облака слов" (`wordcloud`)
`comparison.cloud()`	создание "облака слов" для сравнения частоты встречаемости этих слов в разных документах (`wordcloud`)

commonality.cloud()	"облако слов", общих для нескольких документов (wordcloud)

Пакеты

tm	набор утилит для анализа текстов
topicmodels	подгонка LDA- и CTM-моделей
wordcloud	создание "облака слов"
lda	подгонка LDA-моделей
wordnet R	интерфейс к лексической базе данных WordNet
RTextTools	автоматическая классификация документов путем обучения с учителем
qdap	набор утилит для анализа естественного языка и документов
sentiment140	анализ тональности текстов с использованием бесплатного сервиса sentiment140
tm.plugin.dc	дополнительный модуль для пакета tm, позволяющий организовать распределенные вычисления при анализе текстов
tm.plugin.mail	дополнительный модуль для пакета tm, облегчающий работу с текстами электронной почты
textir	набор утилит для выполнения анализа тональности текста и оценивания статистических различий между документами
tau	набор утилит для анализа текстов

АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ И ФУНКЦИИ ДЛЯ РАБОТЫ С ГРАФАМИ (SOCIAL NETWORK ANALYSIS AND GRAPH MINING FUNCTIONS)

graph(), graph.edgelist(), graph.adjacency(), graph.incidence()	создание графов на основе структур данных таких типов, как "ребра" (edges), "список ребер" (edge list), "матрица расстояний" (adjacency matrix) и "матрица встречаемости" (incidence matrix) соответственно (igraph)
plot(), tkplot(), rglplot()	статичные, интерактивные и трехмерные изображения графов (igraph)
gplot(), gplot3d()	визуализация графов (sna)
vcount(), ecount()	подсчет числа ребер и узлов (igraph)
V(), E()	доступ к узлам и ребрам графа (igraph)
is.directed()	является ли граф направленным? (igraph)
are.connected()	связаны ли два узла графа? (igraph)
degree(), betweenness(), closeness(), transitivity(), evcent()	различные меры центральности графа (igraph, sna)
edge.density()	плотность графа (igraph)
add.edges(), add.vertices(), delete.edges(), delete.vertices()	добавление узлов и ребер (igraph)
neighborhood()	нахождение соседей заданного узла графа (igraph, sna)
get.adjlist()	получение списков соседних узлов или ребер (igraph)
nei(), adj(), from(), to()	индексирование узлов и ребер графа (igraph)

`cliques()`, `largest.cliques()`, `maximal.cliques()`, `clique.number()`	обнаружение клик, т.е. полных подграфов (igraph)
`clusters()`, `no.clusters()`	нахождение максимально связанных элементов графа и их количества (igraph)
`fastgreedy.community()`, `spinglass.community()`	алгоритмы обнаружения сообществ в графах (igraph)
`cohesive.blocks()`	нахождение "сцепленных блоков" (кластеров) в графах (igraph)
`induced.subgraph()`	извлечение части графа (igraph)
`mst()`	алгоритм минимального остовного дерева (igraph)
`components()`	нахождение максимально связанных компонентов графа (igraph)
`shortest_paths()`	нахождение кратчайшего пути между узлами (igraph)
`%->%, %<-%, %--%`	операторы индексирования ребер графа (igraph)
`get.edgelist()`	возвращает список ребер в виде матрицы с двумя столбцами (igraph)
`read.graph()`, `write.graph()`	импорт и экспорт графов в виде файлов разных форматов (igraph)

Пакеты

igraph	анализ и визуализация графов
sna	анализ социальных сетей
d3Network, networkD3	R-интерфейс к JavaScript-библиотеке D3 для построения графов, деревьев, дендрограмм и диаграмм Санки
RNeo4j	взаимодействие с базами данных Neo4j из среды R
statnet	набор инструментов для описания, визуализации, анализа и имитации графов
egonet	меры центральности для анализа социальных сетей
network	инструменты для создания и модификации графов
bipartite	визуализация двураздельных графов и вычисление некоторых описательных статистик
blockmodeling	обобщенное и классическое моделирование блоков в размеченных графах
diagram	визуализация простых графов (сетей), построение потоковых диаграмм
NetCluster	кластеризация элементов графа
NetData	наборы данных к лабораторным работам по анализу социальных сетей с помощью R от McFarland et al.
NetIndices	расчет различных индексов, включая индексы для описания структуры пищевых сетей
NetworkAnalysis	оценка статистических различий между взвешенными или невзвешенными графами
tnet	анализ взвешенных, бимодальных и динамических графов

ФУНКЦИИ ДЛЯ РАБОТЫ С ПРОСТРАНСТВЕННЫМИ ДАННЫМИ (SPATIAL DATA ANALYSIS FUNCTIONS)

`geocode()`	геокодирование с использованием сервиса Google Maps (ggmap)

plotGoogleMaps()	визуализация пространственных данных на картах Google Maps (plot-GoogleMaps)
qmap()	быстрое построение карт (ggmap)
get_map()	запросы к сервисам Google Maps, OpenStreetMap, или Stamen Maps для построения карт (ggmap)
gvisGeoChart(), gvisGeoMap(), gvisIntensityMap(), gvisMap()	гео-диаграммы и карты от Google (googleVis)
GetMap()	загрузка статичной карты с сервера Google (RgoogleMaps)
ColorMap()	цветовое кодирование и изображение уровней некоторой переменной на карте (RgoogleMaps)
PlotOnStaticMap()	совмещение графиков с географическими картами (RgoogleMaps)
TextOnStaticMap()	нанесение текстовых меток на карты (RgoogleMaps)

Пакеты

plotGoogleMaps	визуализация пространственных данных на картах Google Maps и сохранение результатов в виде HTML-виджетов
RgoogleMaps	работа с картами Google Maps в R
ggmap	визуализация пространственных данных с использованием сервисов Google Maps и OpenStreetMap
plotKML	визуализация пространственных и пространственно-временных объектов с использованием сервиса Google Earth
SGCS	кластеризация геоинформационных данных с использованием пространственных графов
spdep	инструменты для поиска пространственных зависимостей

ГРАФИЧЕСКИЕ ФУНКЦИИ (GRAPHICS FUNCTIONS)

plot()	функция общего назначения для визуализации данных
barplot(), pie(), hist()	столбиковые диаграммы, круговые диаграммы и гистограммы
boxplot()	диаграммы размахов
stripchart()	одномерные диаграммы рассеяния
dotchart()	точечная диаграмма Кливленда
qqnorm(), qqplot(), qqline()	графики квантиль-квантиль
coplot()	категоризованные графики
splom()	категоризованные матрицы диаграмм рассеяния (lattice)
pairs()	матрицы диаграмм рассеяния
cpairs()	улучшенные матрицы диаграмм рассеяния (gclus)
parcoord()	диаграммы параллельных координат (MASS)
cparcoord()	улучшенные диаграммы параллельных координат (gclus)
parallelplot()	диаграммы параллельных координат (lattice)
densityplot()	график ядерной функции плотности (lattice)
contour(), filled.contour()	контурные диаграммы
levelplot(), contourplot()	контурные диаграммы (lattice)
mosaicplot()	мозаичная диаграмма
assocplot()	диаграмма ассоциаций

smoothScatter()	диаграммы рассеяния с цветным представлением плотности вероятности; позволяют визуализировать большие массивы данных
sunflowerplot()	диаграмма рассеяния типа "подсолнух"
matplot()	изображение столбцов одной матрицы в зависимости от столбцов другой матрицы
fourfoldplot()	визуализация таблиц сопряженности размером 2×2×*k*
persp()	трехмерные диаграммы поверхностей
cloud(), wireframe()	трехмерные диаграммы рассеяния и диаграммы поверхностей (lattice)
interaction.plot()	график взаимодействий между двумя переменными
iplot(), ihist(), ibar(), ipcp()	интерактивные диаграммы рассеяния, гистограммы, столбиковые диаграммы и диаграммы параллельных координат (iplots)
pdf(), postscript(), win.metafile(), jpeg(), bmp(), png(), tiff()	сохранение графиков в файлах разных форматов
gvisAnnotatedTimeLine(), gvisAreaChart(), gvisBarChart(), gvisBubbleChart(), gvisCandlestickChart(), gvisColumnChart(), gvisComboChart(), gvisGauge(), gvisGeoChart(), gvisGeoMap(), gvisIntensityMap(), gvisLineChart(), gvisMap(), gvisMerge(), gvisMotionChart(), gvisOrgChart(), gvisPieChart(), gvisScatterChart(), gvisSteppedAreaChart(), gvisTable(), gvisTreeMap()	различные интерактивные диаграммы, созданные с использованием Google Visualisation API (googleVis)

Пакеты

ggplot2	реализация принципов "грамматики графических элементов"
ggvis	интерактивный вариант реализации принципов "грамматики графических элементов"
googleVis	интерфейс между R и Google Visualisation API для создания интерактивных диаграмм
d3Network, networkD3	R-интерфейс к JavaScript-библиотеке D3 для построения графов, деревьев, дендрограмм и диаграмм Санки
rCharts	создание интерактивных диаграмм с использованием различных JavaScript-библиотек
lattice	продвинутая высокоуровневая система для визуализации данных с упором на многомерные данные
vcd	визуализация категориальных данных
iplots	интерактивные диаграммы

ПРЕОБРАЗОВАНИЕ ДАННЫХ (DATA MANIPULATION)

transform()	преобразование таблицы данных
scale()	центрирование и нормирование столбцов матриц и таблиц
t()	транспонирование матриц
aperm()	транспонирование массивов
table(), tabulate(), xtabs()	формирование сводных таблиц
stack(), unstack()	объединение и декомпозиция векторов в соответствии с уровнями некоторого фактора

split(), unsplit()	разбиение данных на группы в соответствии с уровнями некоторого фактора(-ов) и обратная этому операция
reshape()	конвертирование таблицы данных в "широкий" или "длинный" формат
merge()	слияние двух таблиц данных (подобно join-операциям в базах данных)
aggregate()	вычисление сводных статистик для отдельных групп данных
by()	применение произвольной функции к таблице данных, разбитой на группы в соответствии с уровнями некоторого фактора(-ов)
melt(), cast()	разбиение таблицы данных на составляющие элементы с последующим формированием новой таблицы с измененной формой и/или содержимым (reshape)
sample()	формирование случайных выборок
complete.cases()	обнаружение записей в таблице данных, которые не содержат пропущенных значений
na.fail, na.omit, na.exclude, na.pass	обработка пропущенных значений

Пакеты

dplyr	высокоэффективный набор утилит со стандартизованным синтаксисом для работы с таблицами данных
reshape	гибкий инструмент для изменения формы таблиц данных и их агрегирования
reshape2	усовершенствованная версия пакета reshape
tidyr	результат дальнейшего усовершенствования пакета reshape2; позволяет легко изменять форму таблиц данных при помощи функций spread() и gather()
data.table	набор утилит для высокоэффективной работы с таблицами данных (индексирование, join-операции с сохранением порядка, присваивание значений, группирование и т.д.)
gdata	различные утилиты для манипуляций с данными
lubridate	набор функций для работы с датами и временем
stringr	набор функций для работы с символьными данными

ФУНКЦИИ ДОСТУПА К ДАННЫМ (DATA ACCESS FUNCTIONS)

save(), load()	сохранение и загрузка объектов типа RData
read.csv(), write.csv()	импорт и экспорт файлов формата .csv
read.table(), write.table(), scan(), write()	импорт и экспорт данных
read.xlsx(), write.xlsx()	импорт и экспорт Excel-файлов (xlsx)
read.fwf()	ипорт данных, которые хранятся в виде файлов с фиксированной шириной по лей
write.matrix()	экспорт матрицы или таблицы данных (MASS)
readLines(), writeLines()	запись и чтение текстовых строк из файла
sqlQuery()	выполенение SQL-запросов к базе данных ODBC (RODBC)

`sqlFetch()`	чтение таблицы из базы данных ODBC (RODBC)
`sqlSave()`, `sqlUpdate()`	сохранение и обновление таблиц в базе данных ODBC (RODBC)
`sqlColumns()`	выяснение структуры таблиц в базе данных (RODBC)
`sqlTables()`	получение списка таблиц, имеющихся в базе данных (RODBC)
`odbcConnect()`, `odbcClose()`, `odbcCloseAll()`	открытие и закрытие соединения с базой данных ODBC (RODBC)
`dbSendQuery`	выполенение SQL-запроса к базе данных (DBI)
`dbConnect()`, `dbDisconnect()`	открытие и закрытие соединения с системой управления базами данных (DBI)

Пакеты

RODBC	доступ к базам данных ODBC
foreign	чтение и запись данных в сторонних форматах, таких как Minitab, S, SAS, SPSS, Stata, Systat и др.
sqldf	выполнение SQL-подобных SELECT-запросов к таблицам R
DBI	DBI-интерфейс между R и реляционными DBMS
RMySQL	драйвер для соединения с базами данных MySQL
RJDBC	доступ к базам данным через интерфейс JDBC
RSQLite	драйвер для соединения с базами данных RSQLite
ROracle	DBI-драйвер для соединения с базами данных Oracle
RpgSQL	DBI/RJDBC-интерфейс для работы с базами данных PostgreSQL
RODM	интерфейс для работы с Oracle Data Mining
xlsx	чтение и запись файлов в форматах Excel 97/2000/XP/2003/2007
xlsReadWrite	чтение и запись Excel-файлов
WriteXLS	создание файлов формата Excel 2003 (xls) из таблиц данных R
SPARQL	SPARQL-драйвер для выполнения запросов SELECT и UPDATE

ФУНКЦИИ ДОСТУПА К ДАННЫМ ЧЕРЕЗ ВЕБ-ИНТЕРФЕЙС (WEB DATA ACCESS FUNCTIONS)

`download.file()`	загрузка файлов из Интернета
`xmlParse()`, `htmlParse()`	разбор файлов XML и HTML (XML)
`userTimeline()`, `homeTimeline()`, `mentions()`, `retweetsOfMe()`	извлечение разнотипных данных из сети Twitter (twitteR)
`searchTwitter()`	поиск в сети Twitter по поисковой фразе (twitteR)
`getUser()`, `lookupUsers()`	получение информации о пользователе сети Twitter (twitteR)
`getFollowers()`, `getFollowerIDs()`, `getFriends()`, `getFriendIDs()`	получение списка фоловеров и друзей (или их идентификаторов) того или иного пользователя сети Twitter (twitteR)
`twListToDF()`	конвертирование списков twitteR в стандартные таблицы данных (twitteR)

Пакеты

twitteR	набор утилит для работы с Twitter API
RCurl	R-клиент для выполнения запросов по стандартным сетевымпротоколам (HTTP/FTP/...)

XML	чтение и создание документов в форматах XML и HTML
httr	набор утилит для работы с URL и HTTP (построен на основе RCur, но проще в использовании)

ФУНКЦИИ ДЛЯ ВЗАИМОДЕЙСТВИЯ С ИНСТРУМЕНТАМ ОБРАБОТКИ "БОЛЬШИХ ДАННЫХ" MAPREDUCE, HADOOP И SPARK

`mapreduce()`	спецификация и выполнение задач MapReduce (rmr2)
`keyval()`	создание объектов типа "ключ – значение" (rmr2)
`from.dfs(), to.dfs()`	чтение и запись объектов R при работе со сторонними файловыми системами (rmr2)
`hb.get(), hb.scan(), hb.get.data.frame()`	чтение таблиц HBase (rhbase)
`hb.insert(), hb.insert.data.frame()`	запись таблиц HBase (rhbase)
`hb.delete()`	удаление записей из таблиц HBase (rhbase)

Пакеты

rmr2	анализ данных в среде R в стиле MapReduce на Hadoop-кластере
rhdfs	соединение с Hadoop Distributed File System (HDFS)
rhbase	соединение с NoSQL базой данных HBase
Rhipe	инструменты для работы с Hadoop из среды R
SparkR	тонкий R-клиент для работы с Apache Spark
RHive	распределенные вычисления на основе запросов к HIVE
Segue	выполнение параллельных вычислений с использованием облачного сервиса Amazon Elastic Map Reduce (EMR)
HadoopStreaming	утилиты для использования R-скриптов при обработке потоковых данных на Hadoop-кластере
hive	распределенные вычисление, основанные на парадигме MapReduce
rHadoopClient	R-клиент для работы Hadoop

БОЛЬШИЕ МАССИВЫ ДАННЫХ (LARGE DATA)

`as.ffdf()`	преобразование таблицы данных в формат ffdf (ff)
`read.table.ffdf(), read.csv.ffdf()`	чтение данных из текстового файла и сохранение в виде ffdf-объекта (ff)
`write.table.ffdf(), write.csv.ffdf()`	сохранение ffdf-объектов в виде текстовых файлов (ff)
`ffdfappend()`	добавление обычной таблицы данных или таблицы ffdf к существующей таблице ffdf (ff)
`big.matrix()`	создание стандартной "большой матрицы" (объект типа `big.matrix`), размер которой ограничен доступным объемом RAM (bigmemory)
`read.big.matrix()`	создание "большой матрицы" путем чтения из ASCII-файла (bigmemory)
`write.big.matrix()`	запись "большой матрицы" в файл (bigmemory)
`filebacked.big.matrix()`	создание "большой матрицы" в виде файла хранящегося на диске (размер такой матрицы может превышать доступный объем памяти) (bigmemory)
`mwhich()`	усовершенствованные "which"-подобные команды для работы с большими матрицами (bigmemory)

Пакеты

ff	хранение больших массивов данных на диске с эффективным использованием памяти, а также набор функций для быстрого доступа к таким данным
ffbase	стандартные статистические функции для пакета ff
filehash	простая база данных типа "ключ – значение" для работы с большими массивами данных
g.data	создание и поддержка пакетов для работы с данными типа "delayed data"
BufferedMatrix	объекты для хранения матриц с данными во временных файлах
biglm	регрессионный анализ для данных, которые не помещаются в памяти компьютера
bigmemory	набор утилит для работы с матрицами данных очень большого объема
biganalytics	расширение пакета bigmemory, содержащее дополнительный набор аналитических функций
bigtabulate	table-, tapply-, и split-подобные функции для работы с объектами классов matrix и big.matrix

ПАРАЛЛЕЛЬНЫЕ ВЫЧИСЛЕНИЯ

Функции для организации параллельных вычислений

sfInit(), sfStop()	запуск и завершение работы вычислительного кластера (snowfall)
sfLapply(), sfSapply(), sfApply()	параллельные версии функций lapply(), sapply(), apply() (snowfall)
foreach(...) %dopar%	параллельное выполнение циклов (foreach)
registerDoSEQ(), registerDoSNOW(), registerDoMC()	регистрация последовательного, SNOW и многопоточного бэк-энда для выполнения параллельных вычислений с помощью пакета foreach (foreach, doSNOW, doMC)

Пакеты

snowfall	"обертка" на основе функционала пакета snow, предназначенная для более эффективной разработки программ для параллельных вычислений в среде R
snow	организация параллельных вычислений в R
multicore	параллельное исполнение R-кода на машинах с несколькими процессорами
snowFT	расширение пакета snow для разработки робастных и воспроизводимых приложений, и для удобной организации параллельных вычислений
Rmpi	интерфейс для работы с MPI (Message-Passing Interface)
rpvm	R-интерфейс для работы с PVM (Parallel Virtual Machine)
nws	набор утилит для координации параллельных вычислений
foreach	конструктов foreach-циклов для R
doMC	адаптор к пакету multicore для выполнения параллельных foreach-вычислений
doSNOW	адаптор к пакету snow для выполнения параллельных foreach-вычислений
doMPI	адаптор к пакету Rmpi для выполнения параллельных foreach-

	вычислений
doParallel	адаптор к пакету multicore для выполнения параллельных foreach-вычислений
doRNG	генератор случайных чисел, позволяющий выполнять воспроизводимые параллельные вычисления на основе foreach-циклов
GridR	исполнение R-кода на удаленных машинах и кластерах
fork	набор функций для одновременной работы с несколькими процессами R

ИНТЕРФЕЙС К WEKA И ДРУГИМ ЯЗЫКАМ ПРОГРАММИРОВАНИЯ (INTERFACE TO WEKA AND OTHER PROGRAMMING LANGUAGES FUNCTIONS)

Пакет RWeka – это R-интерфейс к Weka, который позволяет работать с функциями Weka из среды R:

- Ассоциативные правила: `Apriori()`, `Tertius()`
- Регрессия и классификация: `LinearRegression()`, `Logistic()`, `SMO()`
- "Ленивые" классификаторы: `IBk()`, `LBR()`
- Мета-классификаторы: `AdaBoostM1()`, `Bagging()`, `LogitBoost()`, `MultiBoostAB()`, `Stacking()`, `CostSensitiveClassifier()`
- Классификаторы на основе правил: `JRip()`, `M5Rules()`, `OneR()`, `PART()`
- Деревья классификации и регрессии: `J48()`, `LMT()`, `M5P()`, `DecisionStump()`
- Кластеризация: `Cobweb()`, `FarthestFirst()`, `SimpleKMeans()`, `XMeans()`, `DBScan()`
- Фильтры: `Normalize()`, `Discretize()`
- Стемминг слов: `IteratedLovinsStemmer()`, `LovinsStemmer()`
- Токенайзеры: `AlphabeticTokenizer()`, `NGramTokenizer()`, `WordTokenizer()`

Другие языки

`.jcall()`	вызов метода Java (rJava)
`.jnew()`	создание нового объекта Java (rJava)
`.jinit()`	инициализация Java Virtual Machine (JVM) (rJava)
`.jaddClassPath()`	добавляет JAR-файлы к пути класса (rJava)

Пакеты

rJava	низкоуровневый интерфейс между R и Java
rPython	вызов функций Python из R

ФУНКЦИИ ДЛЯ ГЕНЕРАЦИИ ДОКУМЕНТОВ И ОТЧЕТОВ (GENERATING DOCUMENTS AND REPORTS FUNCTIONS)

`Sweave()`	сочетание текста с R или S-кодом для автоматического формирования отчетов
`xtable()`	экспорт таблиц в форматах LaTeX или HTML (xtable)

Пакеты

knitr	пакет общего назначения для формирования динамических отчетов в среде R
xtable	экспорт таблиц в форматах LaTeX или HTML

R2HTML	создание HTML-отчетов
R2PPT	формирование презентаций Microsoft PowerPoint
RPMG	графический интерфейс пользователя (GUI) для интерактивных R-сессий
Red-R	графический интерфейс пользователя с открытым кодом для визуального программирования на языке R
rattle	графический интерфейс пользователя для Data Mining на языке R
latticist	графический интерфейс пользователя для выполнения визуального разведочного анализа данных

Создание графических интерфейсов пользователя и веб-приложений

shiny	фреймворк для разработки веб-приложений в R
svDialogs	создание диалоговых окон
gWidgets	платформо-независимый набор инструментов для разработки графических интерфейсов пользователя
R AnalyticFlow	программа для выполнения анализа данных путем рисования блок-схем, определяющих последовательность аналитических операций

Редакторы для разработки кода на R

RStudio	бесплатная интегрированная среда разработки (IDE) для R
Tinn-R	бесплатный графический интерфейс пользователя для R
Rpad	веб-интерфейс для R в виде рабочих книг

ССЫЛКИ НА ОБУЧАЮЩИЕ РЕСУРСЫ В ИНТЕРНЕТЕ

Веб-сайт RDataMining:	http://www.rdatamining.com http://www2.rdatamining.com
RDataMining группа в LinkedIn (20,000+ подписчиков):	http://group.rdatamining.com или https://www.linkedin.com/groups/4066593
RDataMining в сети Twitter (2,500+ фоловеров):	@RDataMining
Проект RDataMining на сайте R-Forge:	http://www.rdatamining.com/package http://package.rdatamining.com