Journal of Machine Learning Research xx (2017) xx-xx Submitted 09/17; Published x/xx

1

Introduction to Statistical Computing in Scala - an Implementation of the K-Nearest Neighbors Classifier

Roxana Tesileanu ROXANA.TE@WEB.DE
Environmental Statistics and Bioinformatics
National Institute of Forest Research and Management (INCDS)
Brasov, Romania

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

Statistical computing in ecology evolves at a high speed, mainly because researchers have recognized the advantage of being able to design their algorithms according to their needs. The present paper introduces the implementation in Scala language of the k-Nearest Neighbors (kNN) classifier, which can be applied also on small datasets, a situation commonly encountered in ecological research, and discusses its possible role within a more complex learning pipeline for classification tasks.

Keywords: learning pipeline for classification, machine learning in ecology, diffusion processes, k-Nearest Neighbors classification, Scala language, Simple Build Tool (SBT), multivariate analyses

1. Introduction

One of the drivers of the machine-learning progress is the great amount of data born within and gathered by networked and mobile computing systems which necessitates further processing in order to gain insights into the specific fields from which it originates (**?**). The "Big Data phenomenon" is real but, unfortunately, it can not be generalized to all research areas, especially some areas of ecological research which investigate systems which are by nature data-poor and will probably remain as such unless cost-intensive data collection projects are being proposed and financed. This by no means implies that such areas cannot take advantage of the progress of machine-learning and take the best out of the existing datasets. On the contrary, machine-learning, being a study field which "sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with automatic improvement over time, and inference and decision-making under uncertainty" (**?**), is welcomed in real-world environmental decision-making facing the need of proposing courses of action under the cloud of uncertainties and regarding issues characterized by multiple attributes (**?**). Moreover, machine-learning is used not only in analyzing data from observational studies (which might benefit from larger time series datasets), but also from experimental studies (which can't produce large amounts of data but benefit from an experimental design, delivering data which provides stronger inference about effects and causal relationships (**?**)) (**?**). Indeed, a combination of machine-learning classifiers with the benefits of a MANOVA experimental design used for collecting training data, might produce useful research results, in ecological research. Thus, in order to be able to use the power of machine-learning algorithms adapted to small ecological datasets, a statistical package written in Scala is currently under development at INCDS.

The idea of adapting the use of machine-learning algorithms to real-world data requirements is not new. Hastie and Tibshirani (**?**), describe the Discriminant Adaptive Nearest Neighbor (DANN) procedure as a locally adaptive form of nearest neighbors classification using a modified linear dis-

criminant analysis (LDA) procedure to estimate an effective metric for computing neighborhoods. The main idea of combining distance-based approaches with approaches based on the association between variables is driven by the fact that a multivariate dataset can be analysed in both ways (**?**): calculating scores for the derived variables (components) for each object, and calculating dissimilarity measures (distances) for every possible pair of objects.

The present paper introduces the Scala functions used to run the kNN classification algorithm and then it delineates its role within a real-world classification pipeline for reducing error.

2. Running the kNN algorithm

In order to be able to run the kNN algorithm written using the Scala language, one should consider which tools to choose for the different components of a development system - which are according to Rehman and Paul (**?**): the hardware platform, the operating system, editors, compilers and assemblers, debuggers, version control system and bug tracking. For implementing Scala projects, working with the Scala Build Tool (SBT) (http://www.scala-sbt.org/) condenses the list of the components of a development system but also requires a default structure of the project directory with the source files of user-defined packages located in the src subdirectory (**?**) following a standard pattern. This is the reason for which the source files of the scalaML package containing the kNN classifier are placed at the end of a chain of subdirectories within the project directory (see appendix A through E). Another advantage of using SBT for the statistical computing workflow, beside condensing the list of components of a development system, is the interactive use of the Scala REPL, a tool for evaluating expressions in Scala (https://docs.scala-lang.org/overviews/repl/overview.html), within a SBT session (for further details and references (**?**, see)). In order to access the functions for the kNN classification, the package needs to be imported into the current REPL session.

The kNN classifier reads CSV and text files for classification tasks, uses basic vector operations for handling variables and objects, runs the kNN algorithm on the data and computes the error rate of the kNN classifier.

2.1 Basic vector operations

Because most statistical algorithms manipulate datasets which are actually collections of vectors, manipulating vectors is an essential task. The scalaML package offers this functionality, using the Scala main types while developing machine-learning algorithms. This makes the process of synchronizing types in further development stages easier. The vector operations are found in the BasicVectorOP.scala source file (see appendix A) and include: vector addition, vector subtraction, elementwise multiplication and dot product for any two vectors of type Array[Double], and matrix multiplication for any two matrices of type Array[Array[Double]], where each Array[Double] represents an observation (object), and all elements with the same index of inside arrays (i.e. all elements with the same index inside of all Array[Double] within the Array[Array[Double]]) represent a column (i.e. a variable) of the multivariate dataset (**?**), (**?**), (**?**), (**?**), (**?**).

The documentation of the kNN classifier offeres worked examples of functions' application in REPL ((**?**, see)).

2.2 Reading files for classification tasks

The ReadFile.scala source file (appendix B) of the scalaML package contains the function used to read CSV and text files for classification tasks. It returns a tuple with the information needed for the kNN algorithm (a data matrix of type Vector[Array[Double]], the data labels in the form of a Vector[Int], and the used classes in the form of a Range[Int]) (**?**), (**?**), (**?**), (**?**) ((**?**, see) for details).

2.3 KNN classification function

The kNN classification function (see appendix C) implements in Scala the pseudocode of the kNN algorithm (**?**), (**?**), (**?**), (**?**). It starts by calculating the Euclidean distance between a new object (which is going to be classified by the algorithm) and every object of the dataset, it sorts the distances in increasing order, it then takes k objects with the lowest distances and finds the majority class among those k objects, returning it as the prediction for the class of the new object (**?**). The documentation of the kNN classifier gives a thorough example of its use ((**?**, see)). The variables can also be scaled from 0 to 1 before using the kNN classifier, if a normalization is required. For this purpose a normalizing function can be used (see appendix D) **?**, **?**, **?**, **?**.

2.4 Testing framework

In order to test the accuracy of the kNN classifier, an error rate can be computed using the testing function provided in the TestFrameWorkClassif.scala source file (see appendix E) (**?**), (**?**), (**?**). The error rate is given by the total number of errors (misclassified observations) divided by the total number of tested observations ((**?**, see) for details on its use).

3. The learning pipeline for reducing error

The machine-learning workflow for classification tasks for high-dimensional datasets (i.e. more than three variables recorded for each object) generally consists of the following steps: data collection, reading the data in R, Scala, or other programming language used for statistical computation, analyzing the data using multivariate analyses like (MANOVA, PCA, cluster analysis, etc.), applying and testing the classification algorithm (kNN, Bayes classifier, etc.), deciding on and argumenting the choice of the most appropriate model. Real-world workflows are itterative, requiring a dynamic configuration of the used algorithms and eventually the application of different algorithms to compare the results. Such workflows are called in machine-learning practice "learning pipelines" (**?**).

The present paper proposes a combination of the MANOVA experimental design and analysis, with a further data classification step using the kNN classifier. The background idea is that, if two or more groups differ in their group centroids, they are, globally speaking, suitable for classification tasks. If, not, then we probably miss the variables which make a difference, or, we don't really have two different populations, and further classifications are possible but miss their point. In the first case, we further investigate (add new variables or remove current variables), and if the MANOVA test indicates group differences, we are able to find a first decision boundary using the linear combination with the highest eigenvalue.

The decision boundary ideally separates the different groups. Though, variables overlap in the decision space, forming a "transition zone". This is where the residual variance comes from. KNN can then be further applied to investigate the situation occurring in the neighborhood of the objects located in the transition zone, and the best techniques to do this logically appear to be distance-based (i.e. based on dissimilarity measures between objects). Now, the question is how we delineate the transition zone? We can check the objects which produce the MANOVA residual (analysis of the residual) (**?**). For high residuals we choose a higher k number, and for low residuals we choose a lower k number. This approach could represent a helpful investigation pipeline for diffusion processes, possibly requiring the use of alternative dissimilarity measures identifying the correct direction for neighbor selection, and certainly needs further investigation. Nevertheless, it promises a high practical relevance for environmental data analysis.

4. Conclusion

Scala has already been adopted for developing machine-learning algorithms and is one of the languages in which Spark's MLlib (Spark's library for machine learning functions) is written. Its main functionality is to offer parallel algorithms which run well on data clusters, so it is used for analysing large data sets. Some classic algorithms are not included because they were not designed for parallel platforms (**?**). For ecological research purposes, being able to run machine-learning algorithms on small datasets is essential. As a result, a package for statistical computing written in Scala is currently under development at INCDS. In addition, this offers the opportunity of exploring new learning pipelines advancing the field of machine-learning research. One such example is combining the MANOVA experimental design and analysis with the use of the kNN classifier, for adjusting the decision boundary in classification tasks involving multivariate datasets.

Acknowledgment

The author would like to thank Jeff Druce and Mike Reposa from CRA for very useful learning tips and Georgeta Ionescu from INCDS for giving me the financial support needed.

Note

This paper is part of the project "Experimenting with Scala and R for multivariate analyses", which is stored on my GitHub profile under the following link: https://github.com/RoxanaTesileanu/multivariate_analyses. Further documentation on using the kNN classifier **?**, and on using Linux as a development platform for Scala projects **?** can be found under the following link: https://www.researchgate.net/profile/Roxana_Tesileanu/public

Appendix A: Source file - BasicVectorOP.scala

Please follow the link to the source file:
https://github.com/RoxanaTesileanu/multivariate_analyses/blob/
master/DeepLearning/src/main/scala/com/mai/scalaML/
BasicVectorOP.scala

Appendix B: Source file - ReadFile.scala

Please follow the link to the source file:
https://github.com/RoxanaTesileanu/multivariate_analyses/blob/
master/DeepLearning/src/main/scala/com/mai/scalaML/
ReadFile.scala

Appendix C: Source file - kNN.scala

Please follow the link to the source file:
https://github.com/RoxanaTesileanu/multivariate_analyses/blob/
master/DeepLearning/src/main/scala/com/mai/scalaML/
kNN.scala

Appendix D: Source file - AutoNorm.scala

Please follow the link to the source file:
https://github.com/RoxanaTesileanu/multivariate_analyses/blob/

master/DeepLearning/src/main/scala/com/mai/scalaML/
AutoNorm.scala

Appendix E: Source file - TestFrameWorkClassif.scala

Please follow the link to the source file:
https://github.com/RoxanaTesileanu/multivariate_analyses/blob/
master/DeepLearning/src/main/scala/com/mai/scalaML/
TestFrameWorkClassif.scala