MALEENA – ИНТЕРАКТИВНАЯ СИСТЕМА МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ БИБЛИОТЕКИ OPENCV

К.А. Дробных, Н.А. Краснояров

Нижегородский госуниверситет им. Н.И. Лобачевского

Описывается интерактивная программная система MaLeEnA, реализующая графический интерфейс доступа к алгоритмам машинного обучения из библиотеки компьютерного зрения OpenCV. Система позволяет: загружать данные, настраивать параметры моделей, тренировать, загружать, сохранять и сравнивать модели, а также предсказывать значения на новых входных данных.

Введение

В настоящее время машинное обучение [1] применяется при решении многих задач компьютерного зрения, анализа речи, компьютерной лингвистики, медицинской диагностики, а также в процессе разработки интеллектуальных игр. Существует большое количество программных библиотек, содержащих реализацию алгоритмов машинного обучения. OpenCV [2] является одной из таких библиотек. Наличие интерактивной программной системы зачастую облегчает процедуру проведения экспериментов с использованием алгоритмов машинного обучения. Задача состоит в том, чтобы разработать графическое приложение, которое упрощает использование алгоритмов машинного обучения из библиотеки OpenCV, позволяет проводить эксперименты и сравнивать полученные результаты.

В настоящей работе описывается MaLeEnA – Machine Learning Environment A – интерактивная программная система для запуска алгоритмов машинного обучения из библиотеки компьютерного зрения OpenCV. MaLeEnA – это кроссплатформенное приложение, реализованное с использованием открытых библиотек OpenCV и Qt. Отличие данной системы от аналогов состоит в том, что она не требует установки дополнительного программного обеспечения и использует широко известную библиотеку компьютерного зрения OpenCV.

Существующие решения

Интерактивные приложения для работы с алгоритмами машинного обучения уже существуют на базе других библиотек, например, утилита rattle [3] для пакета R и приложение Weka [4]. Также необходимо отметить, что существуют и самостоятельные утилиты визуализации и анализа данных (Orange [5], RapidMiner [6] и др.).

Наше приложение имеет возможность автоматического определения формата записи данных (разделитель, имена переменных) в CSV-файле и позволяет задавать необходимое разбиение на обучающую и тестовую части. Поддержка работы с алгоритмом машинного обучения Gradient Boosting Trees является одной из отличительных особенностей разрабатываемой системы, т.к. такая возможность имеется только в системах Weka и RapidMiner.

Основные возможности

MaLeEnA позволяет выполнить полный цикл действий для решения задачи машинного обучения:

- загрузка данных (в известном формате CSV);
- настройка переменных (тип, подмножество используемых переменных);
- настройка параметров модели (реализована для алгоритмов Gradient Boosting Trees [7], Random Trees [8], Extremely Randomized Trees [9], Support Vector Machine [10], Classification And Regression Tree [11]);
- обучение модели, сохранение/загрузка в XML/YAML форматах, которые используются OpenCV;
- интерфейс для сравнения моделей, обученных с помощью предложенных алгоритмов на одних и тех же данных с различными параметрами;
- предсказание на новых данных.

Высокоуровневая архитектура

Система состоит из следующих компонент: Controller, Visualizator, Datafile, Models, Parameters + History.

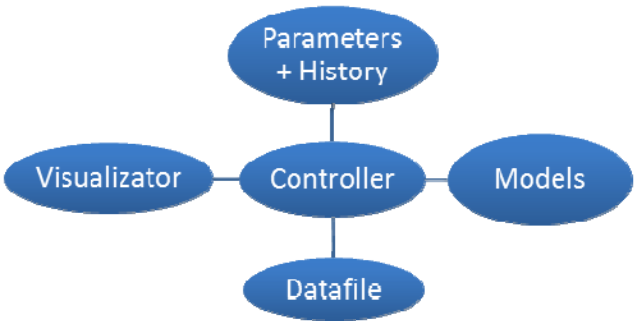

Рис. 1. Диаграмма взаимодействия компонент

Controller обеспечивает связь между компонентами (передача данных, вызовы).

Visualizator отвечает за отображение данных и взаимодействие с пользователем.

Компонент *Datafile* предоставляет возможности загрузки и настройки входных данных.

Models содержит функционал, обеспечивающий работу с моделями с использованием средств OpenCV (загрузка, сохранение, обучение, предсказание и т.п.).

Parameters + History реализует хранение параметров моделей и организует историю параметров обученных моделей (хранит параметры и ошибки обученных ранее моделей для дальнейшего сравнения их качества).

Иллюстрация основных возможностей

Рассмотрим решение задачи машинного обучения на данных Statlog (Heart) Data Set из коллекции UCI Machine Learning Repository [12].

- В процессе загрузки файла heart.dat система определяет формат записи данных (разделитель, имена переменных, признак отсутствующих значений).
- Диалог Variables позволяет выбрать response-переменную и сменить её тип (в данном примере это последняя переменная, categorical). При необходимости также можно изменить тип остальных переменных.
- Осуществляется выбор модели и настройка её параметров.
- Для запуска тренировки входные данные разбиваются на основную и тестовую части, с помощью одного из предложенных типов разбиения: можно указать пропорции разбиения, либо задать его вручную.

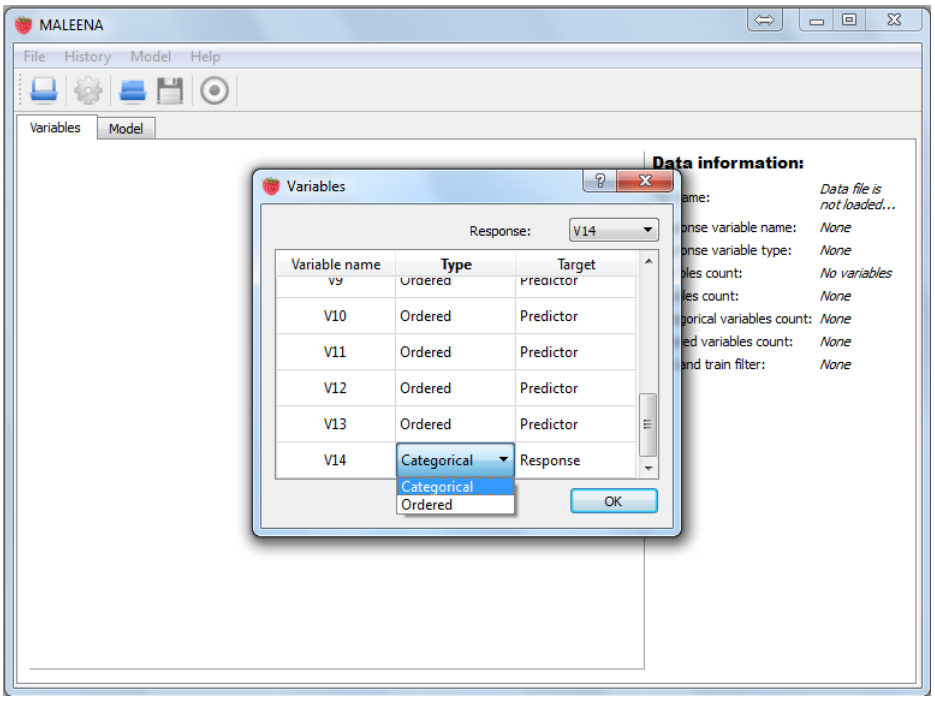

Рис. 2. Настройка переменных

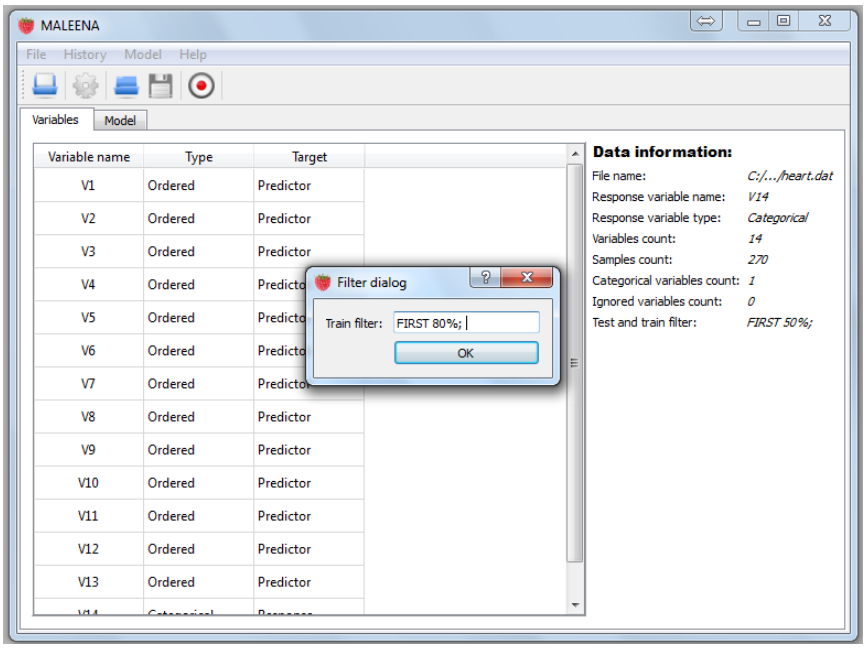

Рис. 3. Выбор разбиения данных на основную и тестовую части

- После тренировки нескольких моделей можно сравнить их и выбрать наиболее подходящую. На рис. 4 представлена диаграмма, используемая для сравнения моделей. Цветом на ней обозначается ошибка модели на тестовой части данных.
- Можно сохранить выбранную модель в файл или использовать её для предсказания значений на новых данных.

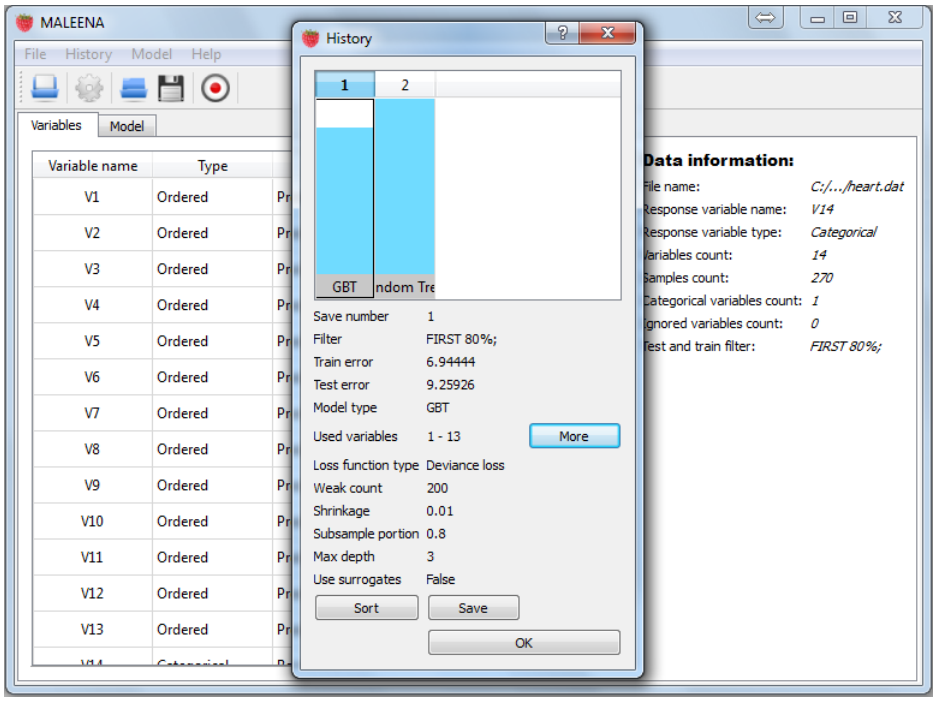

Рис. 4. Выбор разбиения данных на основную и тестовую части

Заключение

В работе описана система MaLeEnA – Machine Learning Environment A – интерактивная программная система, реализующая графический интерфейс доступа к алгоритмам машинного обучения из библиотеки компьютерного зрения OpenCV. Система позволяет загружать данные, настраивать параметры моделей, тренировать, загружать, сохранять и сравнивать модели, предсказывать значения на новых входных данных. Дальнейшее развитие системы подразумевает пользовательскую поддержку, исправление ошибок, доработку и расширение возможностей. Приложение будет выложено в открытый доступ в ближайшее время на сайт: http://ml.vmk.unn.ru/.

Работа выполнена в лаборатории «Информационные технологии» факультета ВМК ННГУ им. Н.И. Лобачевского.

Литература

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2008.
2. OpenCV Documentation – [http://docs.opencv.org/].
3. Rattle – [http://rattle.togaware.com/].
4. Weka – [http://www.cs.waikato.ac.nz/~ml/weka/index.html].
5. Orange – [http://orange.biolab.si/].
6. RapidMiner – [http://sourceforge.net/projects/rapidminer/].
7. Friedman J. Greedy function approximation: the gradient boosting machine // Annals of Statistics. 2001. V. 29, N. 5. P. 1189–1232.
8. Breiman L. Random Forests // Machine Learning. 2001. V. 45, N. 1. P. 5–32.
9. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees // Machine Learning. 2006. V. 63, N. 1. P. 342.
10. Burges C. A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining Knowledge Discovery. 1998. V. 2, N. 2. P. 121–167.

11. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and Regression Trees. Wadsworth & Brooks, 1984.
12. UCI Machine Learning Repository – [http://archive.ics.uci.edu/ml/].