

РАСПРЕДЕЛЁННАЯ ВЫЧИСЛИТЕЛЬНАЯ СИСТЕМА
«ПОЛИГОН АЛГОРИТМОВ КЛАССИФИКАЦИИ»

Воронцов К.В., Инякин А.С., Лисица А.В., Романов М.Ю.,

Стрижов В.В., Хачай М.Ю.*, Чехович Ю.В.

Вычислительный центр РАН, Россия

**Уральский государственный университет, Россия*

ВЦ РАН, 119333, Москва, Вавилова 40, voron@ccas.ru, inyakin@ccas.ru

«Полигон» предназначен для массового тестирования алгоритмов классификации на реальных задачах и представления результатов тестирования через web-интерфейс.

Задача классификации представляется в виде матрицы признаковых описаний объектов, в которой выделен целевой признак. Дополнительно могут задаваться типы признаков и матрица штрафов. База задач хранится централизованно на главном сервере «Полигона». Начальный набор задач формируется из репозитория UCI, кроме того, пользователи имеют возможность загружать свои задачи.

Алгоритм классификации принимает на входе матрицу обучающей выборки, матрицу тестовой выборки, набор управляющих параметров, и на выходе выдаёт матрицы классификаций обучающей и тестовой выборки. Алгоритм реализуется в виде вычислительного модуля (*плагина*, *plug-in*), в соответствии с внутренними стандартами системы «Полигон», и размещается на вычислительном сервере.

Вычислительным сервером может быть любой компьютер, подключённый к Интернету, на котором установлен «Менеджер вычислительного сервера», осуществляющий запуск плагинов и обмен данными между плагинами и центральным сервером. На одном вычислительном сервере может работать несколько плагинов.

Пользователям и группам пользователей предоставляется возможность устанавливать права доступа к своим задачам и алгоритмам для других пользователей и групп.

Методика тестирования основана на стандартной процедуре $T \times N$ -кратного скользящего контроля ($T \times N$ -fold cross-validation), в результате которой все объекты оказываются классифицированными как контрольные ровно по T раз, и как обучающие по $T(N-1)$ раз. Эти результаты используются для вычисления набора показателей качества, существенно расширенного по сравнению с общепринятой методикой. *Основной показатель качества* — средняя частота ошибок на контроле, с доверительным интервалом. Кроме того, вычисляются: эмпирические распределения основного показателя и величины переобученности (разности частоты ошибок на обучении и контроле); ROC-кривые на

обучении и контроле; разложение средней ошибки на вариацию и смещение (bias-variance); распределение вариации и смещения по объектам; эмпирическое распределение отступов (margin) на обучении и контроле для вещественно-значных классификаторов; профиль представительности объектов, позволяющий идентифицировать выбросы, и ряд других показателей и графиков [1].

Формируя запрос на получение отчёта, пользователь задаёт список задач, список алгоритмов из числа доступных системе, управляющие параметры каждого алгоритма и параметры методики тестирования. В ответ на запрос строится отчёт в виде таблицы «задачи × алгоритмы» и детальные отчёты по всем запускам. Один раз вычисленные результаты тестирования сохраняются в базе данных «Полигона» и при повторном запросе выдаются без обращения к алгоритмам.

Преимущества «Полигона» в сравнении с WEKA, R, Matlab и другими системами, применяемыми для тестирования алгоритмов.

1. Стандартные системы устанавливаются локально на компьютере пользователя, оставляя ему возможность модифицировать методику тестирования, задачи и алгоритмы. Поэтому эмпирические результаты, представляемые в разных статьях, оказываются несопоставимыми, даже если тестирование проводилось на одних и тех задачах из UCI.

2. В открытых системах WEKA, R невозможно использовать коммерческие алгоритмы, хотя публикация результатов тестирования выгодна их владельцам. «Полигон» решает возникающие технические и юридические проблемы, так как алгоритмы остаются «на территории» владельца, и он имеет возможность ограничивать доступ к ним.

3. В «Полигоне» нет ограничений на язык программирования, на котором реализуются алгоритмы. WEKA допускает только Java, что снижает скорость выполнения алгоритмов и затрудняет перенос в систему готовых алгоритмов, написанных на других языках.

4. К настоящему времени сложилась де-факто стандартная методика тестирования, основанная на $T \times N$ -кратном скользящем контроле. «Полигон» существенно расширяет эту методику, вычисляя набор показателей, с разных сторон характеризующих качество алгоритмов.

Разработка системы осуществляется при поддержке РФФИ, проекты № 07-07-00372, 07-07-00181.

Литература

1. Воронцов К. В., Инякин А. С., Лисица А. В. Система эмпирического измерения качества алгоритмов классификации // Всеросс. конф. ММРО-13. М.: МАКС Пресс, 2007. С. 577–581.
2. Полигон алгоритмов // www.MachineLearning.ru.