позволяет организовывать отдельные коллекции документов одного типа или схожей тематики.

MongoDB, один из представителей документно-ориентированных СУБД, отлично подходящих для создания информационных приложений. Таким образом, можно сделать вывод что документно-ориентированные БД найдут своё применение в задачах, где требуется упорядоченное хранение информации, но нет множества связей между данными и не нужно постоянно собирать статистику по ним. Документы не требуют определения схемы – это значит, что каждый отдельный документ может состоять из любого количества уникальных полей – в отличие от реляционных баз данных, в которых при попытке хранить разнородные данные неизбежно появляются пустые поля.

И. В. Тимохин
Науч. рук. ***Н. Б. Осипенко,***
канд. физ.-мат. наук, доцент

ИСПОЛЬЗОВАНИЕ ПЕРЕКРЁСТНОЙ ПРОВЕРКИ ДЛЯ ОЦЕНКИ МОДЕЛИ ИДЕНТИФИКАЦИИ ЛЮДЕЙ ПО ФОТОГРАФИИ

При обучении модели для предсказания результата на каких-либо данных, необходимо оценивать качество предсказаний, предоставляемых моделью. Если для проверки модели использовать данные, на которых модель была обучена, то невозможно оценить способность модели к предсказаниям на новых данных, так как повышается вероятность переобучения.

Для решения этой проблемы могут применяться несколько способов. Простейшим является разбиение всей выборки исходных данных на две подвыборки (holdoutmethod): для обучения и для проверки. Размеры обоих подвыборок могут выбираться произвольно, но подвыборка для обучения обычно больше проверочной. В отличие от рассмотренного ниже метода перекрёстной проверки, такой метод требует меньше вычислительных ресурсов; однако оценка, даваемая методом, зависит от разбиения на подвыборки.

Одним из способов улучшить оценку является метод перекрёстной проверки (crossvalidator) [1]. Вся исходная выборка разбивается на k непересекающихся подвыборок. Затем модель обучают на $k - 1$ подвыборке, а для оценки используют одну неиспользованную для обучения подвыборку. Всего обучение проводится k раз, каждый раз меняя выбор подвыборок для обучения и для проверки. В качестве оценки всей модели используется среднее из k полученных оценок. Чем больше число k, тем менее отклонение полученной оценки, и тем больше вычислительных ресурсов требуется. Метод перекрёстной проверки позволяет оценивать модель при малой выборке исходных данных.

Различные методы оценки были использованы при создании модели для идентификации людей по фотографии их лих. Была взята выборка из 500 изображений, для каждого из которых было известно лицо какого человека находится на изображении. Модель обучалась определять, лицо какого человека находится на фото. При разбиении исходной выборки на обучающую и проверочную подвыборки, модель верно предсказывала результат в 63 % случаях, а при использовании метода перекрёстной проверки – в 71 % случаях, при $k = 500$.

Литература

1 Schneider, J. Cross Validation [Electronic resource] / Jeff Schneider – URL: https://www.cs.cmu.edu/~schneide/tut5/node42.html. – Date of access: 29.04.2017.