ИСПОЛЬЗОВАНИЕ МЕТРИЧЕСКИХ ПРИЗНАКОВ В РЕШАЮЩИХ ДЕРЕВЬЯХ НА ПРИМЕРЕ ЗАДАЧИ КЛАССИФИКАЦИИ ТИПОВ ЛЕСНЫХ МАССИВОВ

Китов Виктор Владимирович
кандидат физико-математических наук, ведущий научный сотрудник лаборатории облачных технологий и аналитики больших данных РЭУ им. Г. В. Плеханова.
Адрес: ФГБОУ ВПО «Российский экономический университет имени Г. В. Плеханова», 117997, Москва, Стремянный пер., д. 36.
E-mail: v.v.kitov@yandex.ru

Методы классификации по характеру принятия решения делятся на методы, использующие глобальную оптимизацию (все наблюдения обучающей выборки), и локальную оптимизацию (наблюдения только в малой окрестности исследуемого объекта). Перспективным направлением исследований является совмещение преимуществ каждого подхода в одном объединенном классификаторе. В статье предложен метод объединения этих подходов за счет встраивания локальных метрических признаков в подход, использующий глобальную оптимизацию. Данный подход продемонстрирован для случая, когда в качестве классификатора, использующего глобальную оптимизацию, применяются методы случайного леса (random forest) и особо случайных деревьев (extra random trees). Предложены различные варианты метрических признаков. Перспективность указанного подхода проиллюстрирована на примере решения задачи классификации типа лесных массивов, в которой добавление предложенных метрических признаков существенно улучшило точность классификации.
Ключевые слова: классификация, решающие деревья, метрические признаки, тип лесного покрова.

THE USAGE OF METRIC FEATURES IN PREDICTION WITH DECISION TREES DEMONSTRATED ON THE TASK OF FOREST COVER TYPE CLASSIFICATION

Kitov, Victor V.
PhD, Leading Researcher of the Laboratory of Cloud Technologies and Analysis of Big Data of of the PRUE.
Address: Plekhanov Russian University of Economics, 36 Stremyanny Lane, Moscow, 117997, Russian Federation.
E-mail: v.v.kitov@yandex.ru

Methods of classification by nature of decision-making divide on methods using global optimization (all training samples are used), and local optimization (only samples in the neighbourhood of the studied object are used). The perspective direction of research is combination of advantages of each approach in one integrated classifier. In article the method of combination of these approaches by embedding of local metric features into the approach using global optimization is proposed. This approach is shown for a case when the classifier using global optimization is random forest and extra random trees. Various variants of metric features are evaluated. Performance of the proposed approach is illustrated on the forest cover type prediction task, where it leads to significant improvement in classification accuracy.
Keywords: classification, decisive trees, metric signs, type of a forest cover.

Один из самых распространенных методов машинного обучения для задач регрессии и классификации – решающие деревья, а также композиции решающих деревьев, такие как случайный лес (random forest), особо случайные деревья (extra random trees) и бустинг решающих деревьев. Это связано с тем, что решающие деревья имеют встроенную функциональность отбора признаков, инвариантны к монотонным преобразованиям признаков, позволяют обучаться одновременно на разных типах признаков (непрерывных, порядковых и номинальных), а также с тем, что они быстро обучаются и быстро вычисляют прогнозы и могут прогнозировать не только сами классы, но и их вероятности [4].

Широкое распространение решающих деревьев требует от исследователя понимания алгоритма обучения и свойственных ему ограничений. Как известно, решающее дерево при обучении использует глобальную оптимизацию, настраиваясь на всех данных для выбора корня и используя все более и более ограниченную информацию для выбора дочерних узлов. Глобальная оптимизация производится не полностью, а «жадным» алгоритмом, который оптимизирует только выбор каждого отдельного узла дерева. В связи с двумя перечисленными свойствами решающее дерево находит некоторую аппроксимацию зависимости в данных в целом по всей выборке. В противоположность этому подходу другой известный метод машинного обучения – метод ближайших соседей – осуществляет прогнозирование, основываясь только на локальной окрестности исследуемого объекта, используя данные о ближайших к нему объектах обучающей выборки.

Поскольку представленные алгоритмы используют принципиально разные подходы к выбору прогноза, то естественно ожидать, что некоторая их комбинация позволит получить результат, превосходящий по точности результат каждого метода в отдельности. Концепция объединения разнородных прогнозирующих алгоритмов, компенсирующих ошибки друг друга, известна в англоязычной литературе как stacking или ensemble learning и основана на том, что прогнозы каждого из прогнозирующих алгоритмов подаются на вход некоторому объединяющему предсказателю, который выдает окончательный прогноз.

Альтернативным подходом к объединению разнородных алгоритмов является добавление новых признаков, которые используются другими алгоритмами. Обычно производится кластеризация с числом кластеров, равным числу классов, и признаковое пространство расширяется за счет расстояний между рассматриваемым объектом и центрами кластеров. Например, ряд авторов к признакам относят площади треугольников, образованных объектом и центрами всевозможных кластеров [1], другие – добавляют к ним расстояние до ближайшего кластера и сумму расстояний до всех остальных кластеров [2], третьи – считают признаком сумму

расстояний до ближайшего кластера и ближайшего соседнего объекта [3].

Нами предлагается метод уточнения решающих деревьев за счет добавления к ним метрических признаков, характеризующих расстояния до ближайших объектов обучающей выборки, а также ряда производных признаков. Подход демонстрируется на примере решения задачи классификации типов лесных массивов, представленной на соревновании kaggle.com[1], и дает существенное улучшение в точности классификации для этой задачи.

Постановка задачи

Рассматривается задача автоматического определения типа лесного массива в различных точках карты по геологическим данным. Рассматривается территория четырех заповедников в США. Всего существует 7 допустимых типов леса для рассматриваемой территории: ель обыкновенная, скрученная широкохвойная сосна, желтая сосна, тополь, осина, ель Дугласа и криволесье. Обучающая выборка состоит из 15 120 объектов, а контрольная – из 565 892 объектов. Признаки, позволяющие предсказать тип лесного покрова, состоят из непрерывных и дискретных. Непрерывные признаки включают высоту над уровнем моря, азимут, угол наклона земли, горизонтальное и вертикальное расстояние до ближайшей реки/озера, горизонтальное расстояние до ближайшей дороги, горизонтальное расстояние до лесных пожаров и уровень освещенности в 9, 12 и 15 часов. Дискретные признаки включают тип почвы и название природного заповедника.

Предлагаемое решение

Задача решалась с использованием библиотеки машинного обучения scikit-learn. Поскольку часть признаков – непрерывные, а часть – дискретные, то напрашивалось использование ансамблей решающих деревьев. В качестве таких ансамблей использовались методы случайного леса (random forest) и особо случайных деревьев (extra random trees). Поскольку пропорции классов в обучающей выборке равные, а в тестовой – сильно неравномерные, то при обучении и оценивании моделей объекты обучающей выборки учитывались с весами, соответствующими априорным вероятностям классов в тестовой выборке. В качестве критерия неравномерности моделей использовался критерий Джини. При настройке узлов деревьев выборка была постоянной (сэмплирование осуществлялось только по признакам), так как случайность выборки не давала улучшения в качестве. Параметры, наиболее существенно влияющие на качество, настраивались по сетке значений, используя кроссвалидацию: доля случайно отобранных признаков принимала значения {0,05; 0,1; … 0,5}, а минимальное число наблюдений в листовых узлах – {1; 5; 10; 20; 40}.

Для повышения точности производились следующие преобразования признаков:

1. Использовались все известные признаки. Дискретные признаки представлялись векторами индикаторных функций для каждого возможного значения (one-hot encoding).

2. Признаки, соответствующие освещенности в 9, 12 и 15 часов, заменялись на единственный признак, измеряющий максимальную освещенность в течение дня. Биологическая интерпретация заключается в том, что прорастание ростка зависит от максимальной освещенности, при этом не важно, когда именно эта освещенность достигается.

3. Поскольку всего было 40 типов почвы (2 не использовались, поскольку по ним не было наблюдений в обучающей выборке), то для уменьшения количества признаков изменялось представление типа почвы – вектор индикаторных функций был заменен вектором условных вероятностей каждого класса при заданном типе почвы.

4. Добавление метрических признаков:

[1] URL: https://www.kaggle.com/c/forest-cover-type-prediction

– расстояний до объектов каждого класса обучающей выборки;

– относительных расстояний до объектов каждого класса по отношению к минимальному расстоянию;

– минимального расстояния до объекта какого-либо класса в обучающей выборке.

5. Удаление признаков, характеризующих заповедник, поскольку метрические признаки дают более точную информацию об окружении рассматриваемого объекта.

6. Добавление индикатора, что типы почвы для текущего объекта и для ближайшего объекта обучающей выборки совпадают.

7. Добавление индикатора того, что три ближайших соседа объекта принадлежат одному классу и объект принадлежит треугольнику с вершинами, образованными его тремя ближайшими соседями.

8. Добавлен периметр треугольника, полученного на шаге 7 (в случае если три ближайших соседа принадлежат разным классам, периметр полагался равным бесконечности).

Особенность расчета метрических признаков состоит в том, что среди признаков не было координат точек в явном виде. Из географических признаков были лишь ближайшее горизонтальное расстояние до дороги, до лесных пожаров, до воды, а также ближайшее вертикальное расстояние до воды и высота над уровнем моря. Поэтому расстояние вычислялось как взвешенная сумма этих признаков, где веса подбирались по точности классификации объектов обучающей выборки методом ближайшего соседа по скользящему контролю на отдельных объектах (leave-one-out). При этом расстояние между объектами, принадлежащими разным заповедникам, полагалось равным бесконечности.

В табл. 1 приведены изменения в точности для каждого изменения состава признаков для методов random forest и extra random trees. Точность измерялась по кросс-валидации на обучающей выборке, а также на контрольной выборке, используя модель, настроенную по всей обучающей выборке.

Таблица 1

Точность классификации методов на обучающей и контрольной выборке

Шаг	Extra random trees		Random forest	
	Обучение (*CV*)	Контроль	Обучение (*CV*)	Контроль
1	0,7842	0,7952	0,7855	0,7936
2	0,7848	0,7963	0,7902	0,8006
3	0,7904	0,8001	0,7903	0,7986
4	0,9192	0,8331	0,9173	0,8306
5	0,9181	0,8330	0,9173	0,8289
6	0,9185	0,8347	0,9177	0,8308
7	0,9178	0,8357	0,9188	0,8299
8	0,9186	0,8357	0,9184	0,8302

Из представленных результатов видно, что изменения признаков в целом улучшают точность модели. Наибольшее улучшение в точности было достигнуто за счет добавления метрических признаков на шаге 4. Данный прием применим и для более широкого класса задач, где существует разумное определение расстояния между объектами и где объекты обучающей выборки плотно перемежаются с объектами тестовой выборки. Примечательно, что при настройке метрики веса, соответствующие различным географическим признакам, в рассматриваемой задаче оказались сильно неравномерными, что свидетельствует о разной способности географических признаков различать отличные точки на местности (табл. 2).

Таблица 2

Веса признаков в метрике расстояния между объектами

Признак в метрике	Вес признака
Горизонтальное расстояние до ближайшей дороги	0,0912
Горизонтальное расстояние до ближайшего пожара	0,0778
Горизонтальное расстояние до источника воды	0,1134
Вертикальное расстояние до источника воды	0,2852
Высота над уровнем моря	0,4324

Таким образом, на примере задачи классификации типов лесных массивов было показано, как вычислять расстояния между объектами по косвенным географическим признакам, даже не располагая явными географическими координатами на карте, а также как расширять признаковое пространство, используя новые метрические признаки. В указанной задаче добавление метрических признаков дало наибольший прирост в точности, но указанный прием применим и для более широкого класса задач, где возможно разумное определение расстояния и объекты обучающей и контрольной выборки находятся рядом друг с другом. Одним из направлений будущих исследований может быть применение указанного приема к другим задачам моделирования и построения прогнозов.

Также в статье метрика настраивалась в классе взвешенных евклидовых метрик. Интересно рассмотреть оптимизацию метрики в более широком классе, например, в классе метрик Махаланобиса, что может привести к большей точности итогового классификатора.

Список литературы

1. *Chih-Fong Tsai, Chia-Ying Lin*. A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection // Pattern Recognition. – 2010. – Vol. 43 (1). – P. 222–229.

2. *Chih-Fong Tsai, Wei-Yang Lin, Zhen-Fu Hong, Chung-Yang Hsieh*. Distance-Based Features in Pattern Classification // EURASIP Journal on Advances in Signal Processing. – 2011. – N 62. – URL: http://asp.eurasipjournals.com/content/2011/1/62 (accessed 25.05.2015).

3. *Greeshma K., Merin Meleet*. Detecting Network Intrusions Using New Feature Representations // International Journal of Advanced Research in Computer Science and Software Engineering. – 2013. – N 3 (6). – P. 645–649.

4. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – URL: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf (accessed 25.05.2015).

References

1. Chih-Fong Tsai, Chia-Ying Lin. A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection. *Pattern Recognition*, 2010, Vol. 43 (1), pp. 222–229.

2. Chih-Fong Tsai, Wei-Yang Lin, Zhen-Fu Hong, Chung-Yang Hsieh. Distance-Based Features in Pattern Classification. *EURASIP Journal on Advances in Signal Processing*, 2011, No. 62. Available at: http://asp.eurasipjournals.com/content/2011/1/62 (accessed 25.05.2015).

3. Greeshma K., Merin Meleet. Detecting Network Intrusions Using New Feature Representations. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2013, No. 3 (6), pp. 645–649.

4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Available at: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf (accessed 25.05.2015).