Метрический подход к проблеме оценивания ошибок алгоритмов классификации

Черепнин А. А.
cherepnin@forecsys.ru
Москва, ЗАО «Форексис»

В работах [1, 2, 3] были введены понятия радиусов разрешимости и регулярности задач классификации. При их определении предполагалось, что на пространстве задач с фиксированной системой универсальных ограничений [4, 5, 6] введена метрика. Под радиусом регулярности задачи при этом понимается расстояние от нее до ближайшей нерегулярной задачи, а под радиусом разрешимости, соответственно, — расстояние до ближайшей неразрешимой.

Величины радиусов разрешимости и регулярности позволяют судить (естественно, если метрика на пространстве задач введена достаточно адекватно), например, о том, что задача оказывается разрешимой в силу избыточно точного измерения некоторых признаков.

Основная идея предлагаемого подхода состоит в том, что величины радиусов прежде всего разрешимости можно использовать для получения оценок ошибок алгоритмов классификации и для анализа отдельных признаков в описании объектов.

Для получения оценок ошибок на отдельных объектах или группах объектов предлагается проводить сравнение радиусов разрешимости «редуцированных» задач, получаемых из исходной элиминацией отдельных объектов или равномощных групп объектов. При этом в качестве оценки ошибки может быть использован некоторый монотонно убывающий функционал от радиуса разрешимости задачи, полученной в результате элиминации оцениваемого объекта (или группы). Действительно, если при элиминации определенного объекта радиус разрешимости редуцированной задачи оказывается максимальным среди всех задач, полученных элиминацией одного объекта, то это означает, что именно этот объект делает исходную задачу «максимально близкой» к неразрешимой задаче, то есть к задаче, в которой универсальные и локальные ограничения взаимно противоречивы.

При анализе групп объектов возникают проблемы переборного характера, поскольку приходится рассматривать количество задач, равное количеству сочетаний из исходного числа объектов по количеству элементов в анализируемых группах. В докладе описываются свойства метрик на пространствах задач, обеспечивающие возможность резкого снижения сложности перебора при решении этой проблемы. Отметим, что снижe-

ние сложности перебора достигается за счет решения специальных задач дискретной оптимизации.

Применение метрического подхода к проблеме оценивания ошибок алгоритмов классификации возможно, в частности, путем введения оценок ошибок на объектах как значений монотонно убывающих функционалов от радиусов разрешимости редуцированных задач, полученных элиминацией групп, в которые входит анализируемый объект. После этого получение оценки интегральной ошибки алгоритма сводится к суммированию оценок ошибок объектов, на которых алгоритм принял неправильное решение.

Отдельный интерес представляет использование предлагаемого подхода для анализа результатов использования различных алгоритмов при решении одной и той же задачи. Действительно, даже если два алгоритма допустили на обучении одинаковое количество ошибок, то может оказаться, что один из них допускал ошибки на объектах, имеющих сравнительно низкие оценки, полученные по вышеописанной методике, другой же — наоборот. В этом случае, использование второго алгоритма представляется заведомо более предпочтительным.

Отметим, что основной особенностью предложенного подхода представляется независимость получаемых с его помощью оценок от фиксации какого-либо конкретного семейства алгоритмов классификации. Использование таких семейств обычно сводится к тому, что в качестве оценки объекта выступает доля алгоритмов, давших для этого объекта при той или иной стратегии проведения экспериментов правильный результат. Очевидно, что, варьируя используемое семейство алгоритмов, для любого объекта при такой методике можно получить произвольный наперед заданный результат. Основным элементом произвола при предлагаемом подходе оказывается способ метризации пространства задач, так что в тех случаях, когда такая метризация может быть проведена на основе достаточно реалистичных предположений, можно надеяться на получение достаточно адекватных результатов.

Также следует отметить, что значительная часть способов оценивания с помощью фиксированных семейств алгоритмов классификации также включает в себя этап метризации пространств объектов, так что в этом случае выбор семейства алгоритмов оказывается дополнительным элементом произвола, избежать которого позволяет предлагаемый подход.

Работа выполнена при поддержке РФФИ, проект № 07-07-00711.

Литература

[1] *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 2. — С. 30–35.

[2] *Рудаков К. В.* Полнота и универсальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 3. — С. 106–109.

[3] *Рудаков К. В.* Симметрические и функциональные ограничения для алгоритмов классификации // Кибернетика. — 1987. — № 4. — С. 73–77.

[4] *Рудаков К. В., Черепнин А. А., Чехович Ю. В.* О метрических свойствах пространств задач классификации // Доклады РАН. — 2007. — Т. 416, № 4.

[5] *Черепнин А. А.* О радиусах разрешимости и регулярности задач распознавания // всеросс. конф. ММРО-11, Пущино, 2006. — С. 210–211.

[6] *Черепнин А. А.* Об оценках регулярности задач распознавания и классификации // ЖВМиМФ. — 1993. — № 1. — С. 155–159.