АДАПТИВНЫЙ АЛГОРИТМ МУЛЬТИКЛАССОВОЙ КЛАССИФИКАЦИИ

Е. К. Корноушенко, А. А. Лобко

Институт проблем управления имени В. А. Трапезникова
Москва, Россия
E-mail: ekorno@mail.ru; alex.lobko@gmail.com.

Проблема классификации, актуальная для анализа данных, в настоящее время активно исследуется. В результате анализа и развития перспективных подходов к ее решению разработан алгоритм, который оригинально решает одну из главных проблем классификации – отбор значимых признаков. Его главной особенностью является метод поиска наиболее информационно значимых значений признака. При этом различные значения одного признака могут иметь разное влияние на принадлежность классу. Для каждого классифицируемого объекта проводится свой анализ значимости, что делает алгоритм адаптивным. Эффективность нашего подхода эмпирически подтверждена на известных мультиклассовых выборках в сравнении, установленными для них результатами.

Ключевые слова: мультиклассовая классификация, информационная значимость, машинное обучение, анализ данных.

ОБОСНОВАНИЕ ПОДХОДА

Многие практические вопросы интеллектуального анализа данных включают в себя процесс классификации. Области применения алгоритмов классификации исключительно разнообразны: от распознавания образов и интернет-поиска до кредитных скоринговых систем и подбора адекватной модели при массовой оценке недвижимости. Проблема классификации активно исследуется; разработаны десятки подходов к ее решению. Важный аспект для всей данной области представляет собой тот факт, что качество работы определенного алгоритма находится в значительной зависимости от характеристик исходных данных, лежащих в основе поставленной задачи. Установлено, что не существует абсолютно наилучшего классификационного алго-

ритма [1]. Данное явление обычно объясняют известной «no free lunch theorem» [2]. С другой стороны, эта неопределенность оставляет значительную возможность для исследования параметров информационных входов, наиболее соответствующих определенным алгоритмам классификации.

Еще одним важным моментом является то, что в настоящее время обширно представлены исследования по бинарной классификации, в то время как именно мультиклассификация является предметом наиболее прогрессивных изысканий.

Одним из ключевых вопросов при разработке эффективных классификаторов является выбор значимых признаков из исходного множества признаков, описывающих классифицируемые объекты. При этом для понятия «значимости» в настоящее время существует несколько альтернативных представлений:

1. Естественным представляется связать значимость признака с величиной корреляции этого признака с номерами классов объектов, содержащих этот признак в своем описании. При этом может рассматриваться не отдельный признак, а группа признаков, значения которых и определяют принадлежность объекта к тому или иному классу [3]. Однако необходимо помнить, что в реальности связи признаков и классов очень часто отличаются от линейных, и не всегда подобный подход приводит к адекватным выводам.

2. При другом подходе, из полной совокупности признаков, описывающих объекты исходной выборки, случайным образом выбирается группа из k признаков (и их значений), входящих в описание некоторого объекта из класса C. Для этой группы выбираются два объекта – один из класса C, а второй из другого класса – со значениями признаков, близкими к значениям в исходной группе. Затем находятся разности значений соответствующих признаков. Подобная процедура повторяется для данного значения рассматриваемого признака m раз, где величина m выбирается исследователем. По разности средних значений за m определяются веса признаков, и значимость признаков связывается с их весами. В этом – суть известного алгоритма RELIEF по выбору значимых признаков, описанного и используемого в ряде работ [4–6].

3. В некоторых подходах используется предположение о существовании некоторого (неизвестного) вероятностного распределения на декартовом произведении признаков и классов, такого, что условная вероятность $p(C \mid F_i)$ отнесения объекта с набором признаков F_i к классу C превышает подобную вероятность для другого набора признаков. Такой подход может показаться менее пригодным на практике, поскольку в общем случае требуется экспоненциальное время для анализа значимости разных групп признаков. Тем не менее в рамках этого подхода были разработаны алгоритмы EUBAFES и FOKUS, описанные в работах [6, 7]. Среди подобных работ особняком стоит работа [8], в которой предполагается унимодальность плотности такого распределения в случае упорядоченных классов. Унимодальность классификатора обеспечивается либо реализацией в нем какого-либо распределения с таким свойством (биномиального, усеченного Пуассоновского), либо использованием соответствующей регрессионной модели.

Рассмотрим разработанный нами адаптивный алгоритм мультиклассификации, в котором присутствуют элементы указанных выше подходов, однако сами процедуры выбора признаков и отнесения объекта к тому или иному классу существенно отличаются от известных. Главная особенность алгоритма (и его основное отличие от других методов) состоит в том, что вместо априорного определения значимости от-

дельных признаков (или групп признаков) проводится свой анализ значимости для каждого классифицируемого объекта. Более того, рассчитываются информационные значимости конкретных значений признаков с использованием анализа «похожих» объектов обучающей выборки. Сам алгоритм кратко можно представить как процесс последовательной фильтрации исходных данных. Далее для краткости текущий объект контрольной выборки обозначим как ТОКВ. Адаптивность предлагаемого алгоритма классификации заключается в том, что и значимость значений признаков, и выбор достаточной для классификации совокупности признаков полностью определяются набором значений признаков ТОКВ, и для другого объекта, как правило, являются совершенно другими (т. е. процедура классификации «подстраивается» под ТОКВ). Первая версия этого алгоритма для случая бинарной классификации описана в [9].

СТРУКТУРА АДАПТИВНОГО АЛГОРИТМА КЛАССИФИКАЦИИ

Задача классификации, рассматриваемая в данной работе, формулируется стандартным образом. Имеется выборка из N объектов, каждый из которых описывается совокупностью из m одних и тех же признаков с возможно различными значениями. Для всех объектов из этой выборки однозначно указаны классы $K_1,..,K_p$, к которым принадлежат соответствующие объекты. Принципиальным является условие $p > 2$. Из этой выборки произвольным образом выбираются n объектов, образующих так называемую обучающую выборку. Единственным ограничивающим условием является естественное требование того, чтобы все выделяемые исследователем p классов были в ней представлены. Оставшиеся $N–n$ объектов исходной выборки образуют контрольную выборку. Задача классификации – по совокупности значений признаков, описывающих каждый объект контрольной выборки (т. е. по описанию объекта), отнести его к тому или иному классу. В данной работе качество классификации описывается процентом правильно классифицированных объектов.

Часто встречается ситуация, когда объекты представлены в разрезе огромного количества признаков. В таких случаях перед началом работы алгоритма классификации необходимо спроецировать описание всех объектов на некое подмножество из меньшего набора признаков. При первоначальном сужении набора признаков рассматриваются лишь объекты обучающей выборки, и применяется подход, очень близкий к методу, используемому в [6].

Введем понятие близости значений признака. Значение a_1 признака a называется d-близким ($d > 0$) к значению a_2, если справедливо $|a_2 - a_1| \le da_2$. Отношение d-близости в общем случае несимметрично. Для значения x_{ij} признака $X_i, 1 \le i \le m$; входящего в описание ТОКВ, находится совокупность $S(x_{ij}, d)$ d-близких значений признака X_i, входящих в описания объектов обучающей выборки. При этом показатель d-близости выбирается таким, чтобы число элементов в совокупности $S(x_{ij}, d)$ было не меньшим задаваемого числа G, для которого необходима тонкая настройка. Возможен вариант тонкой настройки значений числа G на основе работы алгоритма с самой же обучающей выборкой. Качество классификации весьма чувствительно к

выбору значения *G*. При рассматривании качественных признаков *d*-близкими значения считаются только при полном смысловом совпадении.

Каждый элемент совокупности $S(x_{ij}, d)$ относится к тому классу, к какому относится объект, содержащий этот элемент в своем описании. Подобная операция проводится для каждого из значений признаков, входящих в описание ТОКВ. Результатом этих действий является неотрицательная матрица, строки которой соответствуют классам, а столбцы – признакам. Элемент a_{rs} этой матрицы есть суммарное число значений x_{sq} признака X_s из совокупности $S(x_{sq}, d)$, которые попали в класс K_r. Необходимо нормировать строки этой матрицы на число элементов соответствующих классов; результирующую матрицу размера $p \times m$ обозначим как $M(t, m)$, где *t* – порядковый номер классифицируемого ТОКВ, а *m* – первоначальное количество используемых признаков, и назовем информационной матрицей ранга *m*. Выбираем в матрице $M(t, m)$ столбец с наибольшей (по столбцам) разностью между наибольшим и наименьшим элементами столбца и перемещаем этот столбец в формируемую так называемую классифицирующую матрицу, в которой он будет первым столбцом. Целью всей этой последовательности действий является определение наиболее значимого признака (точнее, его значения) текущего этапа, который в себе несет максимум информации релевантной именно для классификации. Таким образом, оказывается, что различные значения одного признака могут иметь разное влияние на принадлежность классу. Описанная процедура на данный момент является исключительно эвристической.

Блок процедур алгоритма включает в себя: построение информационной матрицы соответствующего ранга, выделение столбца с указанными выше свойствами и перенесение его в классифицирующую матрицу, фиксацию признака, соответствующего выделенному столбцу. Предлагаемый алгоритм классификации ТОКВ имеет блочную структуру: операции последующего блока повторяют операции текущего блока с учетом того, что в каждом последующем блоке совокупность рассматриваемых признаков сужается за счет фиксации значений признаков, выделяемых в предыдущих блоках.

После прохождения *H* блоков процедуры классификации выходом алгоритма при классификации ТОКВ являются классифицирующая матрица размера $p \times H$ и вектор V_t размерности *H* последовательно фиксируемых признаков. Каждый из этих признаков является самым значимым на каком-либо из этапов классификации ТОКВ. Критерий классификации связывается с величиной построчной суммы элементов классифицирующей матрицы. ТОКВ относится к тому классу, которому соответствует наибольшая из таких сумм. Параметр *H* также сильно влияет на результат.

После рассмотрения всех *N–n* объектов контрольной выборки получаем матрицу *MV* размера $(N - n) \times H$, строками которой являются вектора $V_t, t = 1, ..., N - n$, при этом для разных векторов V_t и элементы, и порядок следования схожих элементов могут быть разными. Естественно предположить, что порядок следования признаков в векторе V_t определяет порядковую «значимость» этих признаков при классификации *t*-го ТОКВ. В отличие от известных подходов к определению значимости признаков, определим значимость признака по числу его вхождений в матрицу *MV* и отбросим как незначимые признаки с числом вхождений, меньшим некоторого порога. Такое отбрасывание незначимых признаков и повторный прогон описанного ал-

горитма может существенно улучшить качество классификации. Также апостериорный анализ признаков, наиболее часто применявшихся при классификации, может стать основой экономического анализа закономерностей в исследуемой области.

ПРОВЕРКА АДАПТИВНОГО АЛГОРИТМА КЛАССИФИКАЦИИ

Предлагаемый алгоритм классификации проверялся на некоторых выборках, входящих в репозитарий для сравнения различных алгоритмов классификации [10]. В частности, для выборок BostonHousing, Wine и Glass получены результаты, представленные в таблице.

Результаты проверки эффективности алгоритма классификации

Выборка	Число классов	Число объектов классификации	Число исходных признаков	Число используемых признаков	*G*	*H*	Процент успеха
Boston Housing	3	100	13	6	11–12	3	80
Wine	3	89	13	7	6–9	3	91
Glass	6	107	10	9	6–7	4	84

Процент успеха означает отношение количества правильно классифицированных объектов ко всем объектам, предъявленным к классификации. Полученные результаты вполне сравнимы с лучшими результатами, полученными для этих выборок при других общепризнанных подходах к классификации, такими как SVM и деревья решений. Интересным и перспективным является тот факт, что приведенные в таблице выборки относятся к кардинально разным отраслям, что говорит о достаточно широкой применимости. Также стоит напомнить, что предложенный адаптивный алгоритм предъявляет крайне мало требований к характеристикам выборки. Его функционирование практически не меняется в зависимости от количества и мощностей классов, от того, присутствуют ли и сколько качественных признаков и т. д.

Принципы, положенные в основу алгоритма, в своей основе аналогичны здравому смыслу человека, решающему довольно сложную задачу мультиклассификации, а полученные показатели показывают функциональность и перспективность подхода.

ЛИТЕРАТУРА

1. *Kotsiantis, S. B.* Supervised Machine Training: A Review of Classification Techniques / S. B. Kotsiantis // Informatica. 2007. Vol. 31, № 3. P. 249–268.

2. *Wolpert, D. H.* No free lunch theorems for optimization / D. H. Wolpert, W. G. Macready // Evolutionary Computation, IEEE Transactions on. 1997. Vol. 1, № 1. P. 67–82.

3. *Zhao, Z.* Searching for Interacting Features. / Z. Zhao, H. Liu // Proceedings of IJCAI. 2007. P. 1156–1161.

4. *Liu, H.* Feature Selection with Selective Sampling / H. Liu, L. Yu, H. Motoda // Proceedings of the 19th International Conference on Machine Learning. 2002. P. 395–402.

5. *Kononenko, I.* Estimating Attributes: Analysis and Extension of RELIEF / I. Kononenko // Proc. European Conf. on Machine Learning CML. 1994. P. 171–182.

6. *Sherf, M.* Feature Selection by Means of a Feature Weighting Approach / M. Sherf, W. Brauer // Technical Report FKI-221-97, Technische Universitat Munchen.

7. *John, G. H.* Irrelevant Features and the Subset Selection Problem / G. H. John, R. Kohavi, K. Pfleger // Machine Learning: Proc. 11-th Intern. Conf. 1994. P. 121–129.

8. *Pinto da Costa, J. F.* The unimodal model for the classification of ordinal data / J. F. Pinto da Costa, H. Alonso, J. S. Cardoso // Neural Networks. 2008. Vol. 21, № 1. P. 78–91.

9. *Kornoushenko,Ю E. K.* The original classification algorithm for the improvement of regression models for the purpose of taxation / E. K. Kornoushenko, A. A. Lobko // Computer Data Analysis and Modelling – CDAM-2010 Proceedings of the 9th International Conference, Minsk, Belarus. 2010. Vol. 2. P. 119–123.

10. UCI Machine Learning Repository [Electronic resource]. URL: http://archive.ics.uci.edu/ml/ (accessed: 29.06.2011).

---

t

$\geq$