УДК 681.513.7

О ДОСТИЖЕНИИ КОМПРОМИССА МЕЖДУ ТОЧНОСТЬЮ И УСТОЙЧИВОСТЬЮ КЛАССИФИКАТОРОВ В ЗАДАЧЕ ВЫБОРА НАИЛУЧШЕЙ ЯДРОВОЙ ФУНКЦИИ ПРИ БАЙЕСОВСКОМ ОБУЧЕНИИ[1]

Ветров Д.П.[*], **Кропотов Д.А.**[**], **Пташко Н.О.**[*]
[*] ВМиК МГУ им. М.В. Ломоносова, г. Москва
[**] ВЦ РАН, г. Москва

Поступила в редакцию 25.05.2009, после переработки 26.06.2009.

Рассматривается задача подбора ядровой функции в методе релевантных векторов (RVM). В части 1 данной работы был сформулирован принцип устойчивости и на его основе определен коэффициент ядровой пригодности KV, максимизация которого позволяет подбирать значение параметра ширины ядровой функции в RVM. Часть 2 данной работы описывает алгоритм обучения и содержит результаты экспериментов по применению предложенного подхода для модельных и реальных задач.

In the paper we show that RBF kernel selection in relevance vector machines (RVM) classifier requires extension of classifiers model. In new model integration over posterior probability becomes computationally unavailable. We propose a method of local evidence estimation which establishes a compromise between accuracy and stability of classifier.

Ключевые слова: распознавание образов, байесовский подход, выбор модели, метод релевантных векторов.
Keywords: machine learning, bayesian framework, model selection, relevance vector machine.

1. Алгоритм обучения

В части 1 данной работы был сформулирован *принцип устойчивости* и на его основе определен *коэффициент ядровой пригодности* KV (ч. 1, 14). Максимизация данной величины позволяет подбирать значение параметра σ - ширины ядровой функции в RVM. В качестве семейства ядровых функций используется параметрическое семейство гауссиан. Процедура подбора σ может быть сформулирована следующим образом:

1) Выбираем некоторое значение σ.

2) Полагаем $\vec{z} = \vec{x}$.

[1]Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-01-00211-а, 08-01-00405, 08-01-90016, 08-01-90427).

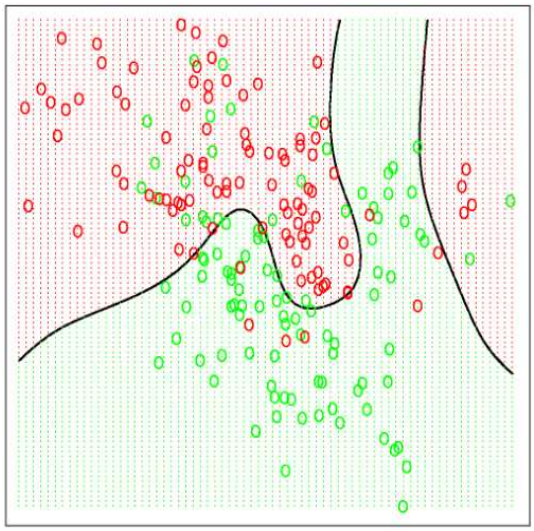

Рис. 1: Обучающая выборка искусственной задачи. Черная линия - оптимальная по Байесу граница между классами.

3) Запускаем итеративный процесс обучения RVM. На каждом шаге сначала ищем точку максимума $w_{MP}^{\vec{}}$ функции $L_{\sigma,\vec{\alpha}}(\vec{w},\vec{x})$. Затем используем приближение Лапласа (ч. 1, 5) и с помощью уравнений (ч. 1, 7), (ч. 1, 8) вычисляем новые значения $\vec{\alpha}$. Шаги повторяются по тех пор, пока процесс не сойдется.

4) В точке $w_{MP}^{\vec{}}$ подсчитываем величину коэффициента ядровой пригодности (ч. 1, 14), где выражения A_{ij} взяты из (ч. 1, 12), а эффективные веса γ_i из (ч. 1, 8).

Значение σ, отвечающее наибольшему значению коэффициента ядровой пригодности, полагается наилучшим.

Следующим шагом является проверка предложенной процедуры подбора ядровой функции. Очевидно, что и слишком узкие, и слишком широкие гауссианы будут иметь низкий показатель коэффициента ядровой пригодности. Узкие ядровые функции ведут к чрезвычайно неустойчивым классификаторам относительно сдвига центров, а классификаторы с широкими гауссианами сильно проигрывают в точности распознавания. Проверим, наблюдается ли данный эффект в экспериментах на искусственной задаче и реальных данных.

2. Результаты экспериментов

2.1 Искусственная задача

Для начала проведем эксперименты на легко интерпретируемой задаче, взятой из [1]. Также найти ее можно по адресу *http://www-stat.stanford.edu/ElemStatLearn.* Рассматривается двухклассовая задача классификации с нелинейной границей

между классами. Размерность признакового пространства равна двум. Обучающая выборка, состоящая из 200 объектов, и оптимальная по Байесу граница представлены на фиг. 1[2]. В качестве тестовой выборки были сгенерированы 5000 объектов с тем же распределением вероятности. Ошибка на столь большой тестовой выборке может рассматриваться как ошибка на генеральной совокупности. Важность выбора правильного значения параметра ширины ядровой функции проиллюстрирована на фиг. 2, где приведены ошибки на обучении и на тестовой выборке для его различных значений. Для того, чтобы проверить качество предлагаемого алгоритма, было проведено сравнение метода с популярным альтернативным подходом - кросс-валидацией (использовалась 5-fold cross-validation). На фиг. 2 представлены ошибка на кросс-валидации и значение коэффициента ядровой пригодности (вычисленное описанным выше способом). Легко видеть, что ядровая функция, выбранная на основе максимума коэффициента ядровой пригодности, лучше подходит для данной задачи (хотя на ней и не достигается минимум тестовой ошибки). Причина относительной неудачи кросс-валидации заключается в том, что, несмотря на свою несмещенность [2], оценка, получаемая с помощью процедуры кросс-валидации, имеет большую дисперсию, особенно для малых выборок. Оба метода - и кросс-валидация, и показатель ядровой пригодности используют лишь обучающую выборку, которая, вообще говоря, отличается от генеральной совокупности. При этом в отличие от кросс-валидации для подсчета коэффициента ядровой пригодности требуется лишь один цикл обучения. Таким образом, предлагаемый метод определения наилучшей ядровой функции работает во много раз быстрее.

2.2 Реальные данные

Для сравнения качества различных методов определения ядровой функции было проведено 180 экспериментов на основе 9 задач из UCI-репозитория [3]. Эксперименты проводились следующим образом. Для каждой задачи данные были случайным образом разбиты на обучающие (33%) и тестовые (67%) выборки. Далее для разных значений параметра ширины ($\sigma = 0.01, 0.1, 0.3, 1, 2, 3, 4, 5, 7, 10$), был обучен и протестирован метод релевантных векторов, подсчитаны кросс-валидационные ошибки (используя 5-fold cross-validation) и показатели ядровой пригодности. После этого тестовые ошибки, соответствующие ядровым функциям с максимальной ядровой пригодностью и с наименьшей кросс-валидационной ошибкой были усреднены по 20 парам обучения/контроля для каждой задачи. Усредненные результаты вместе со своими стандартными отклонениями представлены в таблице 1. Также был рассмотрен популярный метод опорных векторов (SVM). Столбец RVM CV показывает результаты подбора параметра ширины по процедуре кросс-валидации для RVM. В следующем столбце представлены аналогичные результаты, полученные применением кросс-валидации к SVM. Столбец RVM MV результаты подбора ядровой функции по максимуму коэффициента ядровой пригодности. Столбец SVM MV показывает качество работы SVM *с теми же* ядровыми функциями, как и в столбце RVM MV. Заметим, что для выбора ядровой функции в SVM не применялась процедура максимизации ядровой пригодности,

[2]Рисунок взят из [1] с разрешения авторов.

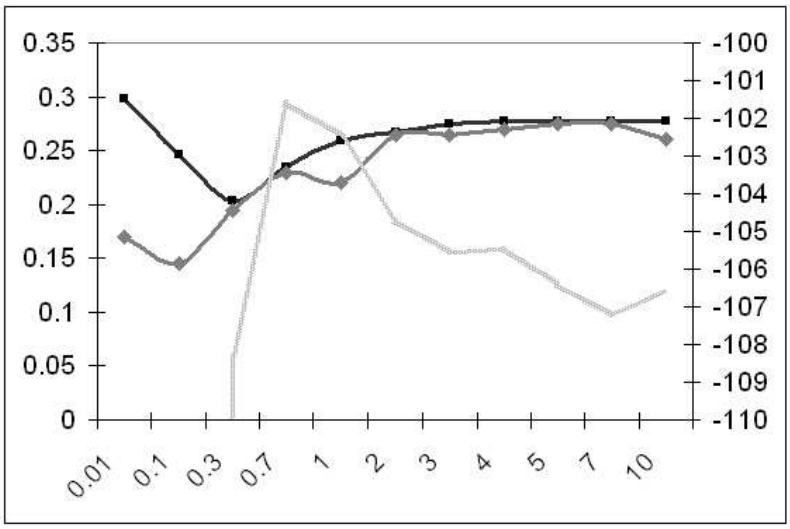

Рис. 2: Различные показатели качества ядровых функций для искусственной задачи. Черная линия показывает тестовую ошибку. Серая линия обозначает ошибку на кросс валидации. Логарифм коэффициента ядровой пригодности показан светло-серой линией. Как можно видеть, узкие гауссианы приводят к переобучению, в то время как широкие ядровые функции не позволяют получить хорошего качества распознавания даже для обучающей выборки. Заметим, что ни кросс-валидация, ни коэффициент ядровой пригодности не выбрали лучшую из возможных ($\sigma = 0.3$) ширину ядровой функции. Это может быть объяснено тем фактом, что оба показателя были вычислены на основе ограниченной обучающей выборки.

а просто бралась та ядровая функция, которая оказалась лучшей для RVM (в смысле пригодности). Столбец SVM MV позволяет оценить, определяется ли оптимальная ядровая функция только самой задачей или зависит также и от алгоритма обучения. Наконец последний столбец содержит минимально возможные ошибки RVM, усредненные по 20 парам обучающих/тестовых выборок.

Для интерпретации данных из таблицы 1 результаты оценивались следующим образом. Для каждой задачи наименьшей тестовой ошибке давалась оценка 1, следующей за ней - 2, и т.д. Худшему результату присваивалась оценка 4. Далее оценки были просуммированы по всем девяти задачам из UCI-репозитория. Итоговые результаты показаны в последней строке таблицы (ИТОГО). Основываясь на них, можно сделать несколько заключений. Во-первых, можно утверждать, что RVM и SVM показывают примерно одинаковые результаты, хотя сами алгоритмы существенно различны, как и положение релевантных (соответственно опорных) векторов. Эксперименты подтвердили, что RVM, вообще говоря, намного разреженней SVM, и в среднем использует в 5-8 раз меньше ядровых функций для классификации. Еще одно важное замечание состоит в том, что предложенная в данной работе мера ядровой пригодности работает не хуже, чем альтернативная процедура скользящего контроля (кросс-валидации). Более того, она требует лишь одного цикла обучения и, следовательно, работает в несколько раз быстрее. Довольно интересным эффектом стало низкое качество SVM при использовании

Таблица 1: Результаты экспериментов для различных методов выбора модели в RVM и SVM. В таблице представлены тестовые ошибки вместе со значениями стандартных отклонений для задач, взятых из UCI репозитория. В столбцах RVM CV и SVM CV приведены результаты для RVM и SVM соответственно, с ядровыми функциями, полученными с помощью кросс-валидациии. RVM MV содержит результаты работы RVM, использующей показатель ядровой пригодности. В столбце SVM MV представлен SVM с той же ядровой функцией, как и в RVM MV. MinTestError указывает минимально возможную тестовую ошибку RVM для каждой выборки. В строке ИТОГО представлены суммарные оценки каждого метода.

Задача	**RVM CV**	**SVM CV**	**RVM MV**	**SVM MV**	**MinTest Error**
AUSTRALIAN	15.5 ± 1.2	16.5 ± 1.9	18.6 ± 4.35	21 ± 3.6	13.4
BUPA	41 ± 0.4	37.5 ± 2.5	39 ± 3.6	37.6 ± 3.8	31
CLEVELAND	18.6 ± 1.8	21 ± 2.7	20 ± 3.5	28 ± 5.6	17
CREDIT	17.3 ± 2.7	18 ± 1.6	16.9 ± 2.4	20 ± 2.9	14.5
HEPATITIS	43 ± 5.6	39.17 ± 3.8	39 ± 3.9	39.21 ± 4.6	36
HUNGARY	22 ± 4.4	20 ± 2.3	24 ± 5.3	26 ± 4	18
LONG BEACH	25.25 ± 0.5	25.18 ± 0.9	27 ± 4.7	26 ± 4.6	24.5
PIMA	34 ± 2.7	30 ± 2	27 ± 2.5	29.6 ± 2.9	23
SWITZERLAND	6.4 ± 1.6	8 ± 1.8	7 ± 2	7.6 ± 2.3	5.8
ИТОГО	**21**	**20**	**20**	**29**	

ядровых функций, являющихся лучшими (в смысле ядровой пригодности) для RVM. Это доказывает, что качество ядровой функции определяется не только топологией выборки, но и сильно зависит от самого метода обучения. Также следует заметить, что ни кросс-валидация, ни процедура максимизации ядровой пригодности не позволяют достичь минимально возможной тестовой ошибки.

3. Обсуждение и выводы

Результаты экспериментов позволяют сделать следующие выводы. Во-первых, идея устойчивости может быть использована для обобщения принципа максимальной обоснованности. В отличие от структурной минимизации риска [2], ограничивающей излишнюю гибкость классификаторов, и принципа минимальной длины описания [4], штрафующего алгоритмическую сложность, концепция Байесовской регуляризации (и ее модификация, описанная выше) основана на поиске модели, в которой решение устойчиво относительно изменений параметров классификатора. Различные реализации принципов Байесовской регуляризации при выборе модели, как и проведенные выше эксперименты, подтверждают, что такой подход в машинном обучении является перспективным. В этой работе была представлена другая, нестатистическая интерпретация обоснованности модели. Байесовский подход был модифицирован путем концентрации непосредственно на идее устойчивости, а не на применении принципа максимума правдоподобия для моделей (т.е. приближения обоснованности). Такая модификация позволяет эксплуатировать этот подход для нелинейных моделей, используя один классификатор вместо интегрирования

по всему пространству параметров с весами, определяемыми апостериорным распределением $P(\vec{w}|\vec{\alpha}, D_{train})$. Предложенный показатель ядровой пригодности не показывает, насколько хороша конкретная ядровая функция для данной задачи, он лишь может служить показателем степени ее применимости в случае фиксированной процедуры обучения (в нашем случае RVM). Качество классификатора, полученного, например, с помощью логистической регрессии или SVM с той же самой ядровой функцией может значительно отличаться. Это происходит из-за того, что оценивается не пригодность всей модели (т.к. используется только один классификатор с $\vec{w} = w_{\vec{MP}}$), а рассматривается лишь локальная устойчивость $Q(\vec{w})$ в точке $w_{\vec{MP}}$. Данный метод является довольно общим и, возможно, может быть применен к другим сложным алгоритмам машинного обучения для настройки параметров модели.

Стоит отметить, что факт влияния алгоритма на обобщающую способность указан многими авторами [2], [5].

Список литературы

[1] Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer (2001)

[2] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York (1995)

[3] Murphy, P.M., Aha, D.W.: UCI Repository of Machine Learning Databases [Machine Readable Data Repository]. Univ. of California, Dept. of Information and Computer Science, Irvine, Calif. (1996)

[4] Rissanen J.: Modelling by the shortest data description. Automatica 14 (1978)

[5] Vorontsov, K.V.: Combinatorial substantiation of learning algorithms. Journal of Comp. Maths Math. Phys. 44(11) (2004) 1997-–2009 http://www.ccas.ru/frc/papers/voron04jvm-eng.pdf