**Основные этапы построения машин опорных векторов**

План занятия

1. *Общие положения*.

Основные этапы построения машин опорных векторов:

предобработка исходных данных;

выбор ядра;

определение стратегии формирования выборок;

настройка машины;

обучение настроенной машины;

тестирование построенной машины.

Определяющие этапы этого процесса: настройка машины (выбор модели, сюда можно отнести и выбор ядра) и определение оптимальных параметров настроенной машины (обучение машины).

**Настройка машины** предполагает поиск таких значений гиперпараметров, параметров регуляризации и ядра, которые приводят к наиболее высокому качеству машины.

Необходимо оценивать качество машин, получаемых в процессе настройки, а также качество обученной машины. Для этого надо определить **показатели** (характеристики) **качества обучения**.

Корректно оценить качество данной модели позволяет **технология перевыбора.**

Предобработка исходных данных при построении SVM в первую очередь предполагает **отбор** (feature selection, FS) и **выделение** (конструирование новых признаков из исходных) (feature extraction, FE) **признаков**.

Перед применением методов FS и/или FE целесообразно использовать процедуру **очищения от шума**, т.е. от бесполезных признаков, не влияющих на отклик.

Перед началом анализа все данные необходимо **масштабировать** или **стандартизировать**, что способствует вычислительной устойчивости используемых алгоритмов.

2. *Характеристики качества обучения*.

Качество очередной модели в процессе настройки машины оценивают на основе определенных характеристик, отражающих качество обучения (performance measures, validation function). Эти характеристики неявно зависят от гиперпараметров, поэтому их часто называют **функциями валидации**.

Для сравнения качества обучения различных моделей, полученных при различных значениях гиперпараметров, нужно определить и критерий поиска (поиск может быть и многокритериальным).

Качество SV-классификатора может быть исследовано и с использованием ROC-анализа.

Практическое руководство по корректному использованию ROC-кривых и основных показателей качества классификации можно найти в работе [1].

3.  *Стратегии формирования выборок при настройке машины опорных векторов.*

Корректно оценить качество очередной модели в процессе настройки машины позволяет правильно подобранная **стратегия перевыбора** (resampling technique).

Процедура hold-out:

1) требуется большой объём исходной выборки для получения статистически обоснованных результатов;

2) невозможно определить дисперсию и устойчивость характеристик модели в связи с изменениями в тренировочном множестве.

Для решения этих проблем служат различные методы перевыбора:

процедура *k-кратной перекрёстной проверки*;

процедура *бутстреп* (bootstrap), использующая случайный выбор с возвращением;

процедура *subsampling*, использующая случайный выбор без возвращения.

Методы перевыбора позволяют последовательно генерировать выборки из исходного набора данных и снова обучать машину на каждой выборке, получая таким образом дополнительную информацию об обучаемой модели.

Для того чтобы избежать явления переобучения при построении машин опорных векторов, рекомендуется применять *стратегию вложенного перевыбора* (Рис.1) [2], состоящую в следующем:

в то время, как качество модели оценивается во внешнем цикле, каждая тренировочная последовательность снова подвергается перевыбору во внутреннем цикле для настройки гиперпараметров.

Для успешного построения машины опорных векторов и получения корректной оценки ее качества, следует:

1. Разбить исходные данные на три непересекающихся множества: **тренировочное** для обучения машины; **валидационное** для настройки гиперпараметров; **тестовое**, предназначенное исключительно для оценивания качества уже настроенной модели.
2. Выбрать ядро и зафиксировать значения гиперпараметров.
3. Обучить модель на тренировочном множестве.
4. Определить качество модели на валидационном множестве.
5. Повторить шаги 2 – 4, используя различные значения гиперпараметров (и/или ядра).
6. Выбрать лучшую модель и обучить ее на данных из тренировочного и валидационного множеств.
7. Оценить качество построенной машины на данных из тестового множества.

При использовании *k*-кратной перекрёстной проверки или бутстрепа шаги 3 и 4 нужно повторить для каждого варианта перевыбора.

4. *Настройка машины опорных векторов*.

Настройка машины может проводиться на основе различных методов оптимизации: как методов поиска в пространстве гиперпараметров, так и методов, основанных на использовании градиента функции валидации.

Методы поиска в пространстве гиперпараметров включают:

методы поиска на решетке;

методы случайного поиска;

эволюционные методы на основе генетических алгоритмов.

Общепринятый метод настройки параметров – поиск на решетке (grid search). Сначала используют грубую решетку, затем для полученной области пространства гиперпараметров, отвечающей приемлемому качеству машины, выбирают более мелкое разбиение. Сократить количество вычислительных операций позволяют при таком подходе, например, поиск на квадродереве (quadtree) [3] или на латинском гиперкубе.

В работе [4] исследовано асимптотическое поведение leave-one-out-ошибки для стандартного SV-классификатора с гауссовским ядром. Показано, что в этом случае область подходящих  значений гиперпараметров (*С* и параметра гауссовского ядра σ) имеет достаточно определенные очертания (Рис.2). Здесь – наилучшее значение параметра регуляризации для линейного SV-классификатора, значение зависит от соотношения в тренировочной последовательности (ТП) образцов разных классов.

В литературе предложен ряд подходов для настройки SVM на основе различных методов оптимизации, среди них:

Метод имитации отжига (simulated annealing). Применение метода не гарантирует нахождения глобального экстремума целевой функции, но при корректной генерации случайных точек во время поиска, как правило, происходит улучшение ее значения.

Стохастический метод поиска с чередующимися окрестностями (Variable Neighborhood Search, VNS) [5], быстро и эффективно определяющий значения гиперпараметров, улучшающих способность SV-классификатора к обобщению.

Метаалгоритм SPO [6] – последовательная оптимизация параметров (Sequential Parameter Optimization, SPO).

Поиск по образцу (Pattern Search, метод Хука – Дживса,) [7]. Применим как для классификации, так и для задач регрессии.

Метод глобальной оптимизации Efficient Parameter Selection via Global Optimization (EPSGO) [8]. Применим как для классификации, так и для задач регрессии.

Методы настройки, основанные на генетических алгоритмах [9,10]. Моделируют эволюционный процесс: на существующую популяцию (набор значений гиперпараметров), которая размножается, воздействуют мутации, затем происходит "естественный отбор" на основе функции валидации.

Подход к настройке машины опорных векторов, использующий градиент функции валидации по гиперпараметрам, предложен в работе [11]. Для использования градиентного метода обычно приходится сглаживать функцию валидации. Удобно использовать сигмоидальное сглаживание функций, однако, при этом возникает необходимость оценивания дополнительных параметров.

5. *Некоторые экспериментальные результаты*. (Иллюстрация вышесказанного)

Для решения задачи бинарной классификации применен стандартный SVM-подход, использовано гауссово ядро и петлевая функция потерь.

Функция валидации – доля ошибок классификации.

Для построения SV-машин использован алгоритм SMO (реализованный в среде MatLab).

Для настройки машин использованы два альтернативных подхода:

стохастический метод поиска с чередующимися окрестностями (VNS) [5];

метод, основанный на вычислении градиента функции валидации (Grad) [11].

Исследование проведено на эталонных данных различной природы:

Splice – данные о границах сплайсинга нуклеотидных последовательностей (3175 образцов, 60 признаков); Splice (1000 + 2175)

German – данные кредитования немецкого банка (1000 образцов, 24 признака). German (700 + 300).

[Такое разбиение выборок обусловлено возможностью сравнения полученных результатов с результатами, приведенными в литературе [ 5,11] ].

Чтобы избежать переобучения и получить (почти) несмещенную оценку характеристики качества машины, использована стратегия вложенного перевыбора (Рис.1) с двумя вложенными процедурами перекрестной проверки: 5-ти кратной во внешнем цикле и 4-х кратной во внутреннем.

В таблице 1 приведены результаты работы SV-классификаторов до и после очистки данных от шумящих признаков. Для очистки данных применялась процедура, основанная на использовании линейного дискриминанта Фишера. После очистки данных Splice осталось 33 признака, после очистки данных German – 15 признаков. Представленные результаты подтверждают устойчивость метода опорных векторов к шуму. При этом исключение шумящих признаков приводит к существенному сокращению времени вычислений.

В таблице 2 приведены результаты экспериментов для данных Splice с целью определения объема тренировочного множества, при котором получаемые значения гиперпараметров еще остаются конкурентно способными. С уменьшением размера ТП все еще достигается приемлемый результат, при этом время расчетов по отношению ко времени расчетов по полной ТП заметно сокращается.

Таблица 1. Влияние процедуры очистки от шума на качество классификации

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C | σ | Оценка доли ошибок после настройки | Доля ошибок при тестировании | Число признаков |
| VNS | Splice | | | | |
| 51.0797 | 6.5225 | 0.1750 0.0490 | 0.0989 | 60 |
| 36.1479 | 2.8694 | 0.1520 0.0326 | 0.0846 | 33 (60) |
| German | | | | |
| 1090.9233 | 36.7607 | 0.2114 0.0487 | 0.2067 | 24 |
| 42.0485 | 10.1265 | 0.2086 0.0442 | 0.2000 | 15 (24) |
| Grad | Splice | | | | |
| 1.2040 | 1.3236 | 0.1350 0.0352 | 0.0970 | 60 |
| 1 | 1 | 0.1060 0.0325 | 0.0851 | 33 (60) |
| German | | | | |
| 2.6096 | 2.7211 | 0.2643 0.0311 | 0.2067 | 24 |
| 2.7183 | 2.7183 | 0.2571 0.0303 | 0.2233 | 15 (24) |

Таблица 2. Влияние объема ТП на результаты настройки машины

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Объем ТП | С | σ | Оценка доли ошибок после настройки | Доля ошибок при тестировании | Относительное время расчетов |
| Splice (градиентный метод) | | | | | |
| 1000 | 1.4933 | 1.6159 | 0.1390 0.0439 | 0.0984 | 1 |
| 800 | 1.5236 | 2.0120 | 0.1462 0.0432 | 0.1025 | 0.7918 |
| 600 | 2.8805 | 2.0044 | 0.1750 0.0283 | 0.1053 | 0.4747 |
| 400 | 2.2378 | 1.8995 | 0.1800 0.0326 | 0.1264 | 0.2555 |
| Splice (метод VNS) | | | | | |
| 1000 | 22.4270 | 5.3391 | 0.1520 0.0358 | 0.0864 | 1 |
| 800 | 11.4291 | 4.6458 | 0.1387 0.0366 | 0.0832 | 0.5576 |
| 600 | 7.3313 | 2.6268 | 0.1783 0.0304 | 0.1103 | 0.5225 |
| 400 | 34.3733 | 4.4700 | 0.1900 0.0609 | 0.1149 | 0.2709 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Данные | Поиск на решетке | VNS ([5]) | VNS | Grad([11]) | Grad |
| german | 0.2133 | 0.2133 | 0.2067 | - | 0.2067 |
| splice | 0.1053 | 0.0993 | 0.0956 | 0.0817 | 0.0970 |

Литература

1. *Fawcett T.* ROC Graphs: Notes and Practical Considerations [for](http://www.basegroup.ru/glossary_ajax/definitions/hypersurface) Researchers. Kluwer Academic Publishers, 2004. P. 1.

2. *Bernd Bischl, Olaf Mersmann, and Heike Trautmann*. Resampling Methods in Model Validation. In WEMACS 2010, held in conjunction with PPSN 2010, 11 (2010), pp.14-31.

3. *Monica Beltrami and* *Arinei Carlos Lindbeck da Silva*. Grid - Quadtree Algorithm for Support Vector Classification Parameters Selection. Applied Mathematical Sciences, Vol. 9, 2015, no. 2, 75 - 82 .

4. Keerthi S. S., Lin C. J. Asymptotic behaviors of support vector machines with Gaussian kernel //Neural computation. – 2003. – Т. 15. – №. 7. – С. 1667-1689.

5. *Carrizosa E., Martín-Barragán B., Morales D. R*. A nested heuristic for parameter tuning in support vector machines //Computers & Operations Research. – 2014. – Т. 43. – С. 328-334.

6. *Koch P. et al.* Tuning and evolution of support vector kernels //Evolutionary Intelligence. – 2012. – Т. 5. – №. 3. – С. 153-170.

7. *Momma M., Bennett K. P*. A Pattern Search Method for Model Selection of Support Vector Regression //SDM. – 2002. – Т. 132. – С. 261-274.

8. *Fröhlich H., Zell A*. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization //Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. – IEEE, 2005. – Т. 3. – С. 1431-1436.

9. *Lessmann S., Stahlbock R.*, Crone S. F. Genetic algorithms for support vector machine model selection //Neural Networks, 2006. IJCNN'06. International Joint Conference on. – IEEE, 2006. – С. 3063-3069.

10. *Perolini A.* Genetic Algorithms and Kernel Matrix-based Criteria Combined Approach to Perform Feature and Model Selection for Support Vector Machines //Word Academy of Science, Engineering and Technology. – 2012.

11. *Keerthi S. S., Sindhwani V., Chapelle O*. An efficient method for gradient-based adaptation of hyperparameters in SVM models //Advances in neural information processing systems. – 2006. – С. 673-680.

Рисунки

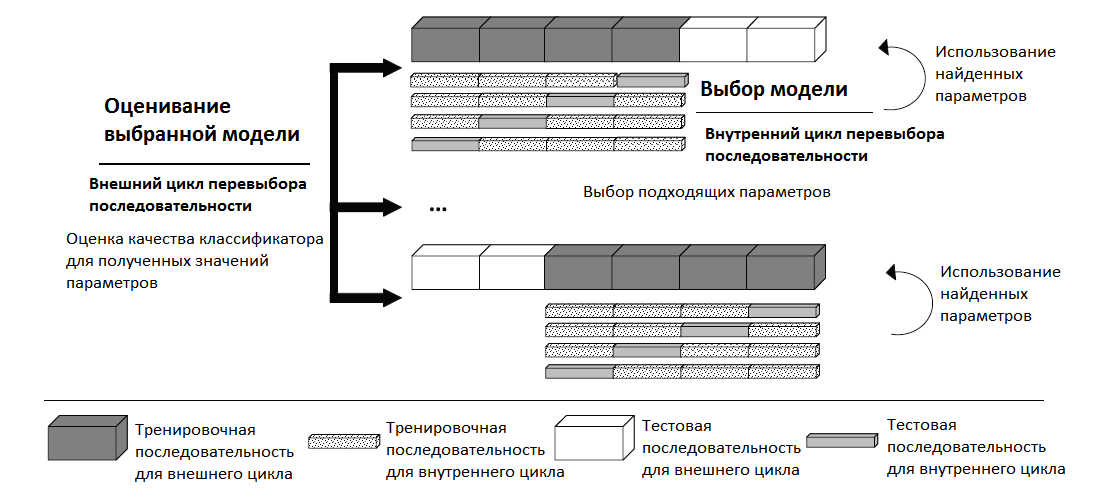


Рис. 1. Схема вложенного перевыбора при настройке SVM.

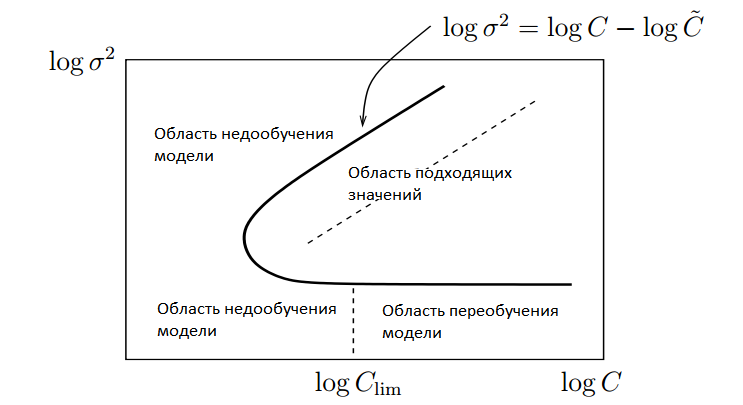


Рис.2. Разбиение пространства поиска гиперпараметров при настройке машины.

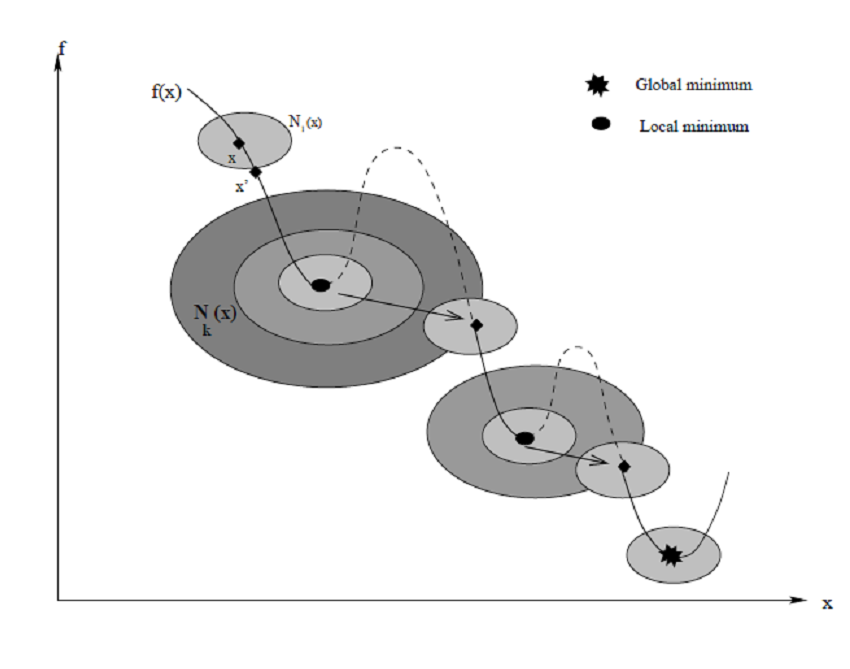


Рис.3 Метод поиска с чередующимися окрестностями.