УДК 519.256

Анализ пространства параметров в задачах выбора мультимоделей

А.А. Адуенко[1], В.В. Стрижов[1,2]
[1]Московский физико-технический институт (государственный университет)
[2]Вычислительный центр им. Дороницына РАН

В работе рассматривается задача выбора мультимоделей при построении моделей в задачах двухклассовой классификации. Мультимодели являются интерпретируемым обобщением случая одной модели, позволяющим учитывать неоднородности в данных. Признаковые пространства моделей в мультимодели могут не совпадать. Кроме того, мультимодель может содержать большое число близких моделей, что ведет к низкому качеству прогноза и отсутствию интерпретируемости. Для решения этой проблемы предлагается метод статистического сравнения моделей для прореживания мультимодели. Вводится понятие адекватной мультимодели, то есть мультимодели, все модели в которой являются попарно статистически различимыми.

Для статистического сравнения моделей предлагается ввести функцию близости между апостериорными распределениями параметров моделей. Такая функция должна быть определена для случая пары распределений с несовпадающими носителями, а также не различать два распределения, одно из которых является малоинформативным. Показано, что дивергенция Кульбака-Лейблера, расстояния Дженсона-Шеннона, Хеллингера, Бхаттачарайа не удовлетворяют этому требованию. Предлагается функция близости для пары распределений, которая удовлетворяет этим требованиям. Доказаны асимптотические свойства распределения введенной функции близости в условиях истинности гипотезы о совпадении моделей. Использование предлагаемой функции близости распределений для сравнения моделей проиллюстрировано на синтетических данных. С помощью статистических свойств распределения введенной функции близости получены оценки на максимальное количество попарно различимых моделей в мультимодели для выборки фиксированного размера. Доказана оценка снизу на количество различимых моделей путем построения набора различимых моделей. Для использования в задаче сравнения новой модели с базовой предложена несимметричная версия введенной функции близости.

Литература

1. *Bishop C.M.* Pattern recognition and machine learning. // Springer, 2006.
2. *Bishop C.M., Nasrabadi N.M.* Pattern recognition and machine learning. // Journal of electronic imaging, 2007. Vol. 16. No. 4.
3. *Gelman A., Hill J.* Data analysis using regression and multilevel/hierarchical models // Cambridge University Press, 2006.
4. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring // Wiley, 2006.
5. *Hosmer D.W., Lemeshow S. Applied logistic regression // A Wiley-Interscience Publication, 2000.*
6. *Hastie T., Tibshirani R., Friedman J.H.* The Elements of Statistical Learning // Springer, 2001.
7. *Motrenko A., Strijov V., Weber G.W.* Bayesian sample size estimation for logistic regression.
8. *Van den Noortgate W., De Boeck P., Meulders M.* Cross-classification multilevel logistic models in psychometrics // Journal of Educational and Behavioral Statistics, 2003. Vol. 28. No. 4. Pp. 369--386.
9. *Moerbeek M., Van Breukelen G.J.P., Berger M.P.F.* Optimal experimental designs for multilevel logistic models // Journal of the Royal Statistical Society: Series D (The Statistician), 2001. Vol. 50. No. 1. Pp. 17--30.
10. *Link W.A., Barker R.J.* Model weights and the foundations of multimodel inference // Ecology, 2006. Vol. 87. No. 10. Pp. 2626--2635.