Секция «Математика и механика»

Многоклассовый мультимодельный прогноз сердечно-сосудистых заболеваний

Мотренко Анастасия Петровна
Студент
Московский физико-технический институт, управления и прикладной математики, Долгопрудный, Россия
E-mail: pastt.petrovna@gmail.com

Заболевания сердечно-сосудистой системы могут протекать, не проявляясь клинически. Тем не менее, обнаружение нарушений работы сердца по косвенным признакам вполне возможно [1, 8]. В данной работе в качестве признаков (биомаркеров) используются концентрации белков и их соединений, абсорбированные на поверхности кровяных телец. Разделение пациентов на группы по состоянию здоровья приводит к задаче многоклассового прогнозирования. Эта задача сведена к задаче двуклассовой классификации; используется подход «каждая группа против каждой». В этом случае рассматриваются все возможные пары групп пациентов и решается задача вида «к какой из двух данных групп пациент принадлежит с большей вероятностью?». Данный подход принят в связи с относительно небольшим объемом выборки, на которой проводился вычислительный эксперимент.

Для каждой пары групп решается задача логистической регрессии [5], в основе которой лежит предположение о биномиальном распределении независимой переменной, и оцениваются параметры функции регрессии [2]. Предполагается, что число измеряемых признаков избыточно; требуется отыскать оптимальный набор признаков, эффективно разделяющий классы.

Отбор признаков осуществляется путем полного перебора, т.к. он дает экспертам гарантию, что рассмотрены все возможные сочетания признаков при выборе модели. При этом экспертами вводились ограничения на сложность модели. Задача выбора признаков поставлена с использованием площади под ROC-кривой [4] в качестве внешней функции ошибки.

Задача классификации сопряжена с оценкой минимального объема выборки, достаточного для проведения классификации. Для этого используются метод доверительных интервалов, метод скользящего контроля [3], сравнение предполагаемых распределений на различных подвыборках [6]. При проведении вычислительного эксперимента и прогноза вероятности наступления инфаркта были использованы данные [7], предложенные специалистами парижской лаборатории анализа крови «Иммуноклин».

Литература

1. Azuaje F., Devaux Y., Wagner D. Computational biology for cardiovascular biomarker discovery // Brief Bioinform. 2009. V. 10, No 4. P. 367–377.

2. Bishop C. M. Pattern recognition and machine learning. Springer, 2006. 738 p.

3. Bos S. How to partition examples between cross-validation set and fraining set? / Saitama, Japan: Laboratory for information representation RIKEN. 1995. 4 p.

4. Fawcet T. ROC graphs: notes and practical considerations for researchers // HP Laboratories, 2004. 38 p.

5. Hosmer D., Lemeshow S. Applied logistic regression. N. Y.: Wiley, 2000. 375 p.

6. Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions // IEEE International Symposium on Information Theory, 2008.

7. Standart flow cytometry analysis of nondental patients. Paris: ImmunoClin laboratory. 2007. 1 p.

8. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure / Heidecker [et al.]. Circulation. 2008. V. 118, No 3. P. 238–246.