Индекс УДК: 004.85

«Рекурсивное исключение признаков в задаче определения типа лейкемии»

Ефимова В. А., Университет ИТМО

Научный руководитель: Сметанников И. Б., к. т. н., доцент, факультет информационных технологий и программирования, Университет ИТМО

Введение

ДНК-микрочипы позволяют ученым одновременно анализировать тысячи генов и определять активные, гиперактивные или неактивные гены в нормальной или раковой ткани. Так как ДНК-микрочипы генерируют огромное количество информации, нужно было разработать методы анализа, которые бы определили, имеет ли раковая ткань какие-то сигнатуры экспрессии генов отличные от нормальных тканей или других типов раковых тканей.

В этой работе рассмотрим задачу выбора небольшого подмножества генов из обширного разнообразия экспрессии генов, записанного в ДНК-микрочипе. Используя обучающую выборку, составленную из примеров данных, полученных от больных раком и здоровых пациентов, можно построить классификатор, способный проводить диагностику и поиск лекарств. Ранее эта задача решалась с помощью корреляции. Мы же рассмотрим метод выбора генов с помощью рекурсивного исключения признаков для метода опорных векторов (Recursive Feature Elimination, RFE). Экспериментально показано, что на генах, выбранных с помощью этой техники, достигается высокая точность классификации, кроме того, эти гены биологически более близки к раковым. С помощью данного метода можно автоматически уменьшить избыточность генов и получить более компактные их подмножества. Данный метод позволил выделить 2 гена, классификация на которых дает нулевую ошибку на отдельных объектах.

Целью работы является эффективный выбор признаков для анализа экспрессии генов на основе ДНК-микрочипов.

Метод

Для решения данной задачи был выбран линейный метод опорных векторов (Linear Support Vector Machine, Linear SVM), так как он наилучшим образом способен описать природу данных. В общем случае рекурсивное исключение признаков работает следующим образом:

1. Обучить классификатор.
2. Вычислить ранги для всех признаков.
3. Исключить признак с наименьшим рангом.

В качестве критерия ранжирования признаков используются веса классификатора SVM. Алгоритм в рассмотренном варианте в течение шага исключает один признак, но может быть обобщен без потерь на случай исключения нескольких признаков, что может быть эффективней с точки зрения вычислений.

Результат

Результаты представлены для двух наборов данных, которые состоят из матриц экспрессии генов, полученных из ДНК-микрочипов для некоторого числа здоровых и больных двумя типами лейкемии пациентов (ALL и AML). Метод достигает высоких результатов по эффективности, кроме того с помощью рекурсивного исключения признаков он отобрал некоторое подмножество генов, имеющих реальную связь с раком, в отличие от остальных методов.

Список литературы

1. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), pp.389-422.
2. Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. Artificial intelligence, 97(1-2), pp.273-324.
3. Fodor, S.P., 1997. DNA sequencing: Massively parallel genomics. Science, 277(5324), pp.393-395.

Автор Ефимова Валерия Александровна

Руководитель Сметанников Иван Борисович