УДК 577.29

АНСАМБЛЕВЫЕ МЕТОДЫ В БИОИНФОРМАТИКЕ: ОПЫТ ИХ ПРИМЕНЕНИЯ В ГЕНОМИКЕ И QSAR МОДЕЛИРОВАНИИ

Адылова Ф.Т.,
д.т.н., зав. лабораторией Института математики АН РУз,
тел.: +(99871) 292-98-78, e-mail: fatima_adilova@rambler.ru

Икрамов А.А.
младший научный сотрудник Института математики АН РУз
e-mail: melan44@mail.ru

Сегодня исследования в вычислительной биологии широко используют методы ансамбля из-за их уникальных преимуществ в работе с выборками малых размеров, высокой размерности признаков, и сложных структур данных. Эта статья имеет две цели. Первая,- дать обзор наиболее широко используемых методов обучения ансамбля и их применения в различных задачах биоинформатики, - экспрессии генов, протеомики на основе масс-спектрометрии , идентификации взаимодействия генов и прогнозирования регуляторных элементов из последовательностей ДНК и белков, QSAR моделировании. Вторая цель,- обобщить тенденции будущего развития методов ансамбля в области биоинформатики. Обсуждаются перспективные направления, такие как ансамбль опорных векторов, мета-ансамбль, и ансамбль для отбора признаков.

Ключевые слова: ансамблевое обучение; биоинформатика; микрочипы; протеомика на основе масс-спектрометрии; взаимодействие генов; регуляторные элементы прогнозирования; ансамбль опорных векторов; мета-ансамбль; отбор признаков

ENSEMBLE METHODS IN BIOINFORMATICS: EXPERIENCE OF THEIR APPLICATION IN GENOMICS AND QSAR MODELING

Adilova F.T., Ikramov A.A.

Ensemble learning is an intensively studies technique in machine learning and pattern recognition. Recent work in computational biology has seen an increasing use of ensemble learning methods due to their unique advantages in dealing with small sample size, high-dimensionality, and complexity data structures. The aim of this article is two-fold. First, it is to provide a review of the most widely used ensemble learning methods and their application in various bioinformatics problems, including the main topics of gene expression, mass spectrometry-based proteomics, gene-gene interaction identification from genome-wide association studies, prediction of regulatory elements from DNA and protein sequences and QSAR modelling. Second aim is to identify and summarize future trends of ensemble methods in bioinformatics. Promising directions such as ensemble of support vector machine, meta-ensemble, and ensemble based feature selection are discussed.

Keywords: ensemble learning; bioinformatics; microarray; mass spectrometry-based proteomics; gene-gene interaction; regulatory elements prediction; QSAR modelling; ensemble of support vector machines; meta-ensemble; ensemble feature selection.

БИОИНФОРМАТИКАДА АНСАМБЛЛИ МЕТОДЛАР: ГЕНОМИКА ВА QSAR МОДЕЛЛАШТИРИШДА УЛАРНИ ҚЎЛЛАШ ТАЖРИБАСИ

Адылова Ф.Т., Икрамов А.А.

Бугунги кунда ҳисоблаш биологияси тадқиқодларида кичик ўлчамли танланмалар, юқори ўлчамли белгилар ва маълумотларнинг мураккаб структуралари билан ишлашда яққол устунлиги туфайли ансамблли методлардан қўлланилади. Мақолада асосий икки мақсад кўзланган. Биринчиси ген экспрессияси, масс-спектрометрия асосида протеомикалар, генлар ўзаро таъсир идентификация ДНК ва оқсил кетма кетликларидан регулятор элементларини башоратлаш QSAR моделлаштириш каби биоинформатиканинг турли масалаларида кенг қўлланилаётган ансамблли ўрганиш методларини тахлилини келтириш. Иккинчи биоинформатика соҳасида келажакда ансамблли методларни ривожланиш йўналишларини умумлаштириш. Таянч векторлар ансамбли, мета ансамбл ва белгиларни танлаш учун ансамбл каби перспектив йўналишлар муҳокама қилинади.

Таянч иборалар: ансамблли ўрганиш, биоинформатика, микрочиплар, масс-спектрометрия асосида протеомика, генлар ўзаро таъсири, башоратлашни регулятор элементлари, Таянч векторлар ансамбли, мета ансамбли, белгиларни танлаш.

1. Введение

Современная биология и химия сегодня широко используют вычислительную технику для анализа сложных биологических данных. Различные компьютерные методы, особенно алгоритмы машинного обучения, применяются для выбора генов или белков, классификации в экспрессии генов с данных микрочипов, в анализе данных протеомики на базе масс-спектрометрии, в диагностике заболеваний на генном уровне, в анализе взаимодействия генов между собой и с окружающей средой в исследованиях генома (GWA), в распознавании регуляторных элементов ДНК или белковых последовательностей, в идентификации белок-белковых взаимодействий, фолдинге белка, в QSAR моделировании.

Ансамблевое обучение является эффективным методом, объединяющим несколько алгоритмов машинного обучения для повышения общей точности прогнозирования. Ансамблевые методы имеют преимущество в том, что облегчают решение известной проблемы HDLSS, - небольшого размера выборки при высокой размерности признакового пространства путем усреднения и включения нескольких моделей классификации. За счет этого небольшая обучающая выборка используется более эффективно, что актуально для многих приложений в биоинформатике. В работе даётся обзор наиболее широко используемых ансамблевых методов и их вариантов, применяемых в биоинформатике, и определяются будущие направления их развития. Также представлены три наиболее популярных метода – bagging, boosting и «случайный лес» (random forest) [1] .

Применение ансамблевых методов в биоинформатике показано решением трех различных задач: 1 - экспрессии генов с микрочипов, 2 - идентификации взаимодействия генов в исследованиях GWA, 3 - классификации на SVM и Random Forest в решении задачи QSAR на химических соединениях. В заключении даны несколько расширений ансамблевых методов и приложения ансамблевых методов в выборе признаков.

1.1. Ансамблевые методы

Цель разработки и использования методов ансамбля - добиться более точной классификации на обучении и на контроле. Однако часто это достигается за счет увеличения сложности модели и ухудшения её интерпретируемости [2]. Наилучшее обобщение сути ансамбля объясняется с помощью классического анализа смещения дисперсии [3]. Такие методы, как bagging (бэггинг), улучшают обобщение снижением дисперсии [4], a boosting достигает этого снижением смещения [5].

Идея метода «bootstrap aggregating», или «**bagging**» в том, что при отсутствии большой обучающей выборки можно создавать много случайных выборок из исходной простым выбором с замещением. Хотя элементы в выборках могут пересекаться или дублироваться, на практике результаты объединения по многим выборкам оказываются точнее, чем по одной начальной. Метод объединяет результаты предсказания различных классификаторов, обученных на случайных подмножествах, и оказывается полезен, если малые изменения в начальной выборке приводят к существенным изменениям классификации.

Бустинг (boosting, улучшение) - процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. В течение последних 10 лет бустинг остаётся одним из наиболее популярных методов машинного обучения, наряду с нейронными сетями и машинами опорных векторов. Основные причины этого - простота, универсальность, гибкость (возможность построения различных модификаций), и, главное, высокая обобщающая способность. Бустинг над решающими деревьями считается одним из наиболее эффективных методов с точки зрения качества классификации. Во многих экспериментах наблюдалось практически неограниченное уменьшение частоты ошибок на независимой тестовой выборке по мере наращивания композиции.

К сожалению, теоретические оценки обобщающей способности дают лишь качественное обоснование феномену бустинга. Хотя они существенно точнее общих оценок Вапника - Червоненкиса, всё же они сильно завышены, и требуемая длина обучающей выборки оценивается величиной порядка 104 . . . 106.

Стековое обобщение, или **стекинг,** - еще один способ объединения классификаторов, вводящий понятие мета-алгоритма обучения. В отличие от bagging и бустинга, при стекинге используются классификаторы разной природы. Идея стекинга: 1 - разбить обучающую выборку на два непересекающихся подмножества; 2 - обучить несколько базовых классификаторов на первом подмножестве; 3 - тестировать базовые классификаторы на втором подмножестве; 4 - используя предсказания из предыдущего пункта как входные данные, а истинные классы объектов как выход, обучить мета-алгоритм обучения.

Объединив правила классификации нескольких классификаторов, например, усреднением и большинством голосов, получают лучшее правило классификации. Чтобы обеспечить точность классификации, базовые классификаторы должны быть точными и отличаться друг от друга [6]. Необходимость разнообразия происходит из предположения, что если классификатор делает неправильную классификацию, т.е. еще другой классификатор, то дополняет его правильной классификацией ошибочных ответов. Популярные ансамблевые методы используют деревья решений в качестве базовых классификаторов, потому что деревья решений чувствительны к малейшим изменениям на обучающем множестве и потому подходят для процедуры возмущений

применительно к данным обучения.

Следует отметить, что есть много других известных методов для создания ансамбля классификаторов. Например, стековое обобщение [7] сочетает в себе базовые классификаторы через мета-классификатор, чтобы максимизировать обобщение. Также широко используются такие методы, как выбор базовых классификаторов и построение каскадных классификаторов [8, 9]. Для того чтобы интегрировать базовые классификаторы в консенсус, обычно используют голосование по большинству голосов или простое усреднение. Предполагая, что выходы базовых классификаторов независимы друг от друга (что на практике частично достигается поощрением разнообразия среди базовых классификаторов), уровень ошибок голосования ε_{mv} выражается следующим образом:

$$\varepsilon_{mv} = \sum_{i=M/2+1}^{M} \binom{M}{i} \varepsilon^i (1-\varepsilon)^{M-i} \quad ,$$

где M - количество базовых классификаторов в ансамбле. Учитывая, что $\varepsilon < \varepsilon_{random}$, а ε_{random} является частотой ошибок случайного угадывания, все базовые классификаторы имеют одинаковый уровень ошибок ε, а частота ошибок голосования ε_{mv} монотонно уменьшается и достигает 0 при $M \to \infty$.

2. Материал и методы

В этом разделе опишем применение ансамблевых методов биоинформатики в решении трёх задач:

- классификация данных микрочипов экспрессии генов и данных протеомики на основе масс-спектрометрии (МС);
- идентификация взаимодействия генов с использованием данных единственного полиморфизма нуклеотида (SNP) из исследований GenomWideAssociation (GWA);
- классификация на SVM и Random Forest в решении задачи QSAR на химических соединениях, полученных из базы данных CheMBL.

2.1. Применение методов ансамбля к микрочипам и протеомике на основе масс-спектрометрии (МС)

Многие биологические исследования предназначены отличать больных от здоровых людей, различать типы различных заболеваний, динамику их прогрессирования и т.д., основываясь на профиле экспрессии генов или изобилии белка. Высокопроизводительные методы включают использование микрочипов для измерения экспрессии генов и масс-спектрометрии для измерения большого количества белка. Однако при их использовании эксперименты часто приводят к необходимости оценки огромного количества признаков с ограниченным числом объектов [10]. Эта ситуация широко известна как «проклятие размерности» [11], когда выбор наиболее подходящих признаков [12, 13] и максимальное использование ограниченных выборок [14] являются ключевыми вопросами классификации.

Dettling и Buhlmann [15] предложили алгоритм, называемый LogitBoost, который заменяет функцию экспоненциальной потери, используемую в AdaBoost, на функцию потери типа лог-правдоподобия. Они показали, что LogitBoost является более точным в классификации данных генной экспрессии по сравнению с первоначальным алгоритмом AdaBoost. Long [16] утверждает, что производительность AdaBoost может быть повышена за счет улучшения базовых классификаторов.

Tan и Gilbert [17] показали, что по сравнению с отдельным деревом решений, ансамблевые методы являются более надежными и точными в классификации данных микрочипов. В протеомике на основе МС Qu et al. [18] провели первое исследование, используя ансамбли boosting для классификации масс-спектра профиля сыворотки. Точность классификации, равная 100 %, была получена с использованием стандартного алгоритма AdaBoost, в то время как простейший ансамбль, названный BDSFS (boosted decision stump feature selection) показал меньшую точность классификации (97 %), но дал более интерпретируемые правила классификации.

По сравнению с bagging и boosting, случайные леса имеют уникальное преимущество в том, что метод, используя несколько подмножеств признаков, хорошо подходит для многомерных данных, таких, как генерируемые микрочипы и исследования в протеомике на МС основе. Это показано в ряде исследований [19, 20]. В [20] экспериментальные результаты на десяти наборах микрочипов показали, что случайные леса в состоянии сохранить точность прогноза, получая даже меньшие наборы генов по сравнению с диагональным линейным дискриминантным анализом (DLDA), KNN, SVM, сжатыми центроидами (SC), и KNN с отбором признаков. Другие преимущества «случайных лесов» - устойчивость к шуму, отсутствие зависимости от параметров настройки и скорость вычислений были продемонстрированы Izmirlian [21] при классификации SELDI-TOF протеомических данных.

Из-за эффективности «случайных лесов» в классификации данных высокой размерности разработка вариантов этого алгоритма сегодня является активной областью исследований. Zhang [22] предложил детерминированную процедуру формирования леса классификационных деревьев. Geurts et al. [23] предложили метод «лишних деревьев»: в каждом узле выбирается лучшее среди K случайно сгенерированных разбиений. Этот метод является улучшением «случайных лесов», поскольку базовые деревья в этом случае выращиваются из целых обучающих выборок рандомизацией точек разреза.

2.2. Применение «случайных лесов» для выявления взаимодействия генов

Кроме измерения экспрессий гена и белка, скрининг и сравнение различных генотипов также могут дать

важную информацию о различных заболеваниях и их патогенезе. Что еще более важно, такие исследования, называемые исследованием ассоциации, могут помочь определить восприимчивость различных лиц к различным заболеваниям, а также их реакцию на различные препараты на основе генетических вариаций [24].

Широко используемой схемой для изучения ассоциации являются скрининг общих одиночных нуклеотидных полиморфизмов (SNP) и сравнение различия между выборками конкретных случаев и контроля для идентификации гена, связанного с болезнью, что в масштабе генома называется GWA исследованиями [25]. Принято считать, что многие сложные заболевания, такие как диабет и рак, возникают из комбинации нескольких генов, которые регулируют и часто взаимодействуют друг с другом, чтобы выдать симптомы болезни [26]. Поэтому цель этих исследований заключается в выявлении сложных взаимодействий между несколькими генами, которые вместе с факторами внешней среды могут существенно увеличить риск развития заболеваний. Использование SNP в качестве генетических маркеров обычно формулируется как задача идентификации взаимодействия SNP-SNP и SNP-среда.

Среди многих алгоритмов распознавания алгоритм дерева решений уже давно признан в качестве перспективного инструмента для оценки SNP-SNP взаимодействия [27, 28]. С ростом популярности методов ансамбля, основанных на дереве, они стали в центре внимания многих недавних исследований в проблеме SNP-SNP взаимодействия для комплексного анализа заболевания. Несмотря на различные ансамблевые методы, которые были предложены для выявления SNP-SNP взаимодействия [29, 30], «случайный лес» пользуется наибольшей популярностью [26]. Это происходит в значительной степени из-за его способности обрабатывать множество SNP с учетом нелинейности [31]. Кроме того, «случайные леса» могут быть легко использованы в качестве встроенного алгоритма оценки признаков [32], который очень полезен для болезней, связанных с отбором SNP.

Сегодня исследования сосредоточены на разработке индивидуальных алгоритмов «случайного леса» и их применения для идентификации взаимодействия генов на больших выборках данных, содержащих несколько сотен тысяч кандидатов SNP. Meng et al. [34] изменили «случайные леса», чтобы принять во внимание информацию по неравновесию связи (Linkage Disequilibrium, LD), когда измерялась важность SNP. Jiang et al. [35] разработали последовательную процедуру отбора признаков для улучшения «случайного леса» в идентификации эпистатического взаимодействия. «Случайный лес» был впервые использован для вычисления индекса Джини в общей сложности для 116 204 SNP из набора данных AMD [33], а затем использован в качестве классификатора, чтобы минимизировать ошибку классификации, выбирая подмножество SNP в прямом последовательном порядке с заранее определенным размером окна.

Накопленные данные свидетельствуют о том, что метод ансамбля является одним из наиболее перспективных решений многих биологических проблем. В связи с огромным успехом многих алгоритмов этого класса в приложениях биоинформатики были предложены многочисленные их расширения. Далее приведены некоторые из наиболее перспективных направлений: различные расширения для достижения лучшего предсказания и адаптация ансамблевых методов для отбора признаков.

2.3. Классификация на алгоритмах SVM и Random Forest в решении задачи QSAR на химических соединениях, полученных из базы данных CheMBL

Нами создана программа Random Forest на языке C/C++ на базе MS Visual Studio 2008 и получены выборки (обучающая и тестовая) из базы данных CheMBL. Вещественные значения активности разбили на интервалы, создавая тем самым классы химических соединений. Решающее правило для обучающей выборки разработали с использованием софта LIBSVM [https://www.csie.ntu.edu.tw].

Программа состоит из функций выращивания дерева на обучающей выборке, сортировки слиянием и голосования деревьев на тестовой выборке. В качестве решающего правила выбрано правило энтропии. По каждому признаку из набора выбранных признаков производилась сортировка элементов, находящихся в данной вершине дерева.

Из базы данных было выбрано 4000 записей химических соединений с соответствующими значениями активности. В качестве дескрипторов (признаков) выбраны следующие 23 дескриптора, доступные в БД: mw_freebase | alogp | hba | hbd | psa | rtb | ro3_pass | num_ro5_violations | acd_most_apka | acd_most_bpka | acd_logp | acd_logd | molecular_species | full_mwt | aromatic_rings | heavy_atoms | num_alerts | qed_weighted | mw_monoisotopic | hba_lipinski | hbd_lipinski | num_lipinski_ro5_violations | med_chem_friendly.

Значения активности были разбиты на 18 интервалов нерегулярным образом. «Случайный лес» строился 100 раз на каждую размерность (1000, 1500 и 2000) тестовой выборки. При этом 1000 соединений в обучающей выборке выбирались случайно из общего числа. Таким образом, была усилена статистическая независимость эксперимента.

На первых 1000 соединениях было построено два правила SVM разделения: 1 - с использованием линейного ядра, 2 - с использованием радиального ядра. Было проведено тестирование на трёх тестовых выборках размерами в 1000, 1500 и 2000 записей.

3. Результаты ВЭ и обсуждение

Результаты эксперимента на алгоритме «случайного леса» представлены в таблице.

Точность решения, %	С учётом соседних интервалов			Без учёта соседних интервалов		
	Размер выборки			Размер выборки		
	1000	1500	2000	1000	1500	2000
Средняя	35,28	33,20	32,42	25,39	23,27	22,42
Максимум	38,15	35,56	35,05	28,5	26,67	25,1
Минимум	32,15	29,46	30,12	23,1	20,93	19,9

Соседними считаются интервалы, чей номер отличается на 1 от текущего интервала. В силу нерегулярности разбиения интервалов и достаточно большого числа соединений, находящихся близко друг к другу по значению активности, было логично также включить по 0,5 за каждый ответ в соседний с правильным интервалом. С учётом соседних интервалов максимальное значение точности для SVM равно 12,75 %. Наилучшие результаты алгоритм SVM показал для полиномиального ядра, тем не менее эти результаты значительно слабее «случайного леса».

Также проведено исследование эффективности обоих методов на других типах дескрипторов. Для этого было использовано приложение MOSES Descriptors, http://www.molecular-networks.com/services/mosesdescriptors. Число выбранных дескрипторов было равно 200. Нами исследовано влияние числа случайно выбираемых дескрипторов на точность оценивания: 18 случайно выбранных дескрипторов дают примерно те же результаты, что и 90, и 133 дескриптора. Полученные процентные данные говорят о несущественности добавления большого числа дескрипторов к уже использованным ранее.

Следовательно, с учётом числа классов и достаточно большого количества элементов в выборках простое угадывание должно было дать 5,88 % угадываний класса. В нашем вычислительном эксперименте алгоритм SVM оказался незначительно лучше простого угадывания. В то же время алгоритм случайного леса показал в три раза большее число правильных ответов. Использование нескольких типов дескрипторов позволяет сказать о недостаточной существенности этого выбора, а также о слабости методов для точного предсказания интервала активности.

Расширение ансамблевых методов классификации. ***Ансамбль SVM.*** SVM обычно считают лучшим из имеющихся классификаторов. Простой способ использовать SVM в ансамбле - применение процедуры bagging с базовым SVM классификатором. Этот прием применил Caragea [36], создав ансамбль с базовым классификатором SVM для прогнозирования: при обучении каждого базового классификатора на «сбалансированной» обучающей выборке производительность SVM ансамбля была выше, чем у одного SVM и сбалансированного SVM.

В исследовании Peng [37] изучена концепция генерации и выбора соответствующего подмножества базовых классификаторов. Базовый классификатор SVM и бутстреп на выборке использовались для генерации множества обучающих наборов. По сравнению с деревом решений, SVM является гораздо более стабильным для небольших пертурбаций обучающих выборок. Для получения разнообразия среди базовых классификаторов, процедура отбора базового классификатора на основе кластеризации используется для того, чтобы базовые классификаторы были точны и исключали несогласие друг с другом. Сравнивая один классификатор SVM и ансамбль bagging и boosting, Peng показал, что предложенная кластеризации на основе SVM ансамбля дает лучший результат.

Мета-ансамбль. Появилась идея создать ансамбль ансамблей, т.е. мета-ансамбль. Эта идея была впервые исследована Dettling [38], который предложил объединить алгоритмы bagging и boosting (так называемый BagBoosting) для классификации данных микрочипов. Основная гипотеза - boosting ансамбль имеет более низкую смещенную оценку, но дисперсия является относительно высокой, в то время как bagging ансамбль имеет более низкую дисперсию, но примерно неизменное смещение. Тогда сочетание этих двух методов могло бы привести к способу прогнозирования, который может иметь как небольшое смещение, так и низкий уровень дисперсии. Опыт показывает, что предлагаемый BagBoosting может улучшить прогноз, полученный отдельными процедурами bagging и boosting, что является конкурентным преимуществом по сравнению с некоторыми другими классификаторами - SVM, KNN, DLDA и PAM.

В исследовании Liu and Xu [39] показан другой путь формирования мета-ансамбля классификаторов. Их система основана на генетическом подходе к программированию, который оптимизирует набор небольших ансамблей, называемых суб-ансамблями, состоящих из группы деревьев решений с различными множествами входных признаков. Эксперимент показывает, что система превосходит несколько других алгоритмов эволюционного типа.

Ансамбль из нескольких различных алгоритмов классификации. Еще одно направление расширения идеи ансамбля состоит в том, чтобы получить разнообразие в классификации выборки разными алгоритмами классификации. Другими словами, вместо того, чтобы манипулировать набором данных для обучения различных моделей классификации с использованием заданного алгоритма классификации, например, дерево решений или SVM, эти методы пытаются найти разнообразие базового классификатора с помощью гетерогенных алгоритмов классификации. Например, Bhanot et al. [40] сочетали нейронные сети, SVM, взвешенное голосование, KNN, деревья решений и

логистическую регрессию для классификации данных масс-спектрометрии. Kedarisetti [41] извлекает разные наборы признаков из базы данных белковых последовательностей, чтобы обучить ансамбль классификаторов, использующих KNN, деревья решений, логистическую регрессию и SVM. Ансамбль затем используется для прогнозирования структурных классов белка. Hassan et al. [42] брали ансамбль из 15 классификаторов, таких как KNN, деревья решений, SVM и нейронные сети. Этот ансамбль алгоритмов классификации применяется к данным трех микрочипов, чтобы найти небольшое количество хорошо дифференцируемых экспрессируемых генов.

Общее свойство этого класса ансамблевых методов заключается в том, что разнообразие ансамбля классификаторов получается за счет различных алгоритмов классификации. Тем не менее, можно объединить эти методы с методами пертурбации выборки, чтобы получить мета-ансамбль классификаторов, который потенциально мог бы увеличить общее разнообразие, обеспечивая высокую точность классификации.

Адаптация теории ансамбля к выбору признаков. Идея ансамбля используется и в отборе признаков, возможно, как следствие неустойчивости результатов отбора признаков обычными методами [43].

Один из прямых методов адаптации ансамблей для отбора признаков - делать встроенные алгоритмы отбора признаков. Более общий подход заключается в использовании теории ансамблей для объединения нескольких моделей. В частности, Dutkowski и Gambin [44] комбинировали несколько алгоритмов фильтрации в рамках кросс-валидации для выбора биомаркеров по данным масс-спектрометрии. Несколько алгоритмов классификации используются для оценки выбранных биомаркеров с тем, чтобы получить более стабильные результаты. Zhang et al [45] собрали несколько алгоритмов фильтрации и классификации для повышения точности прогнозирования и стабильности результатов в ранжировании гена на основе генетического алгоритма. Abeel et al. [46] исследовали ансамбль фильтров в рамках бутстрепа. Netzer et al [47] разработали метод выбора признаков с использованием принципа обобщения стека. Алгоритм выбора признаков состоит в ранжировании стека с тем, чтобы определить важные маркеры и повысить точность классификации образца.

Yang [48] интегрировал различные статистические методы, чтобы повысить надежность результатов рейтинга гена на данных микрочипов. Аналогичным образом, Chan et al. [49] комбинировал критерий Уилкоксона с различными процедурами отбора признаков и различными алгоритмами классификации. Основная идея этих методов - гены и белки, имеющие высокую информативность по различным метрикам, скорее всего, должны быть близки к подлинной биологической значимости, чем те, которые проверены по одной метрике [48].

4. Выводы

В классификации и прогнозировании тщательно разработанные ансамблевые алгоритмы обычно дают более высокую точность и стабильность, чем один алгоритм. Кроме того, ансамблевые алгоритмы часто облегчают решение проблемы малого размера выборки и высокой размерности признакового пространства, которые обычно имеют место во многих приложениях биологической и химической информатики. Стоит отметить, что улучшение точности часто сопровождается усложнением модели, ухудшением интерпретируемости её результатов и высокими вычислительными затратами. Тем не менее, теоретические исследования подхода ансамбля и современные вычислительные мощности могут преодолеть эти трудности.

Кроме классификации, многие методы ансамбля с незначительными изменениями также могут быть использованы для выбора и оценки важности признаков. При выборе признаков разработка новых методов, которые руководствуются общей теорией обучения ансамбля, доказали свою плодотворность. Они, вероятно, будут эффективным инструментом для решения постоянно увеличивающегося разрыва между размером выборки и размерностью данных, полученных в биологических экспериментах.

Использование методов ансамбля - одна из современных устойчивых тенденций. Идея ансамбля широко применяется для многих других проблем био- и химической информатики, которые выходят за рамки данной работы [50].

Литература

[1] Dietterich T.G. Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System. Vol. 1857. Springer.- 2000. - Pp. 1-15.

[2] Kuncheva L. Combining Pattern Classifiers: Methods and Algorithms. Wiley. - 2004.

[3] Webb G.I., Zheng Z. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. IEEE Transactions on Knowledge and Data Engineering. - 2004; 16(8):980-991.

[4] Breiman L. Arcing classifiers (with discussion). The Annals of Statistics. - 1998; 26(3):801-849.

[5] Schapire R.E., Freund Y., Bartlett P., Lee W.S. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics. - 1998; 26(5):1651-1686.

[6] Tsymbal A., Pechenizkiy M., Cunningham P. Diversity in search strategies for ensemble feature selection. Information Fusion. - 2005; 6:83-98.

[7] Wolpert D.H. Stacked generalization. Neural Networks. - 1992; 5(2):241-259.
[8] Kuncheva L.I. Switching between selection and fusion in combining classifiers: An experiment. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. - 2002; 32(2):146-156.
[9] Gama J., Brazdil P. Cascade generalization. Machine Learning. - 2000; 41(3):315-343.
[10] Asyali M.H., Colak D., Demirkaya O., Inan M.S. Gene expression profile classification: a review. Current Bioinformatics. - 2006; 1(1):55-73.
[11] Somorjai R.L., Dolenko B., Baumgartner R., Crow J.E., Moore J.H. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics. - 2003; 19:1484-1491.
[12] Saeys Y., Lnza I., Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. - 2007; 23(19):2507-2517.
[13] Hilario M., Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Briefings in Bioinformatics. - 2008; 9(2):102-118.
[14] Braga-Neto U., Dougherty E. Is cross-validation valid for small-sample microarray classification? Bioinformatics. - 2004; 20(3):374-380.
[15] Dettling M.,B Uhlmann P. ¨Boosting for tumor classification with gene expression data. Bioinformatics. - 2003; 19(9):1061-1069.
[16] Long P. Boosting and Microarray Data. Machine Learning. - 2003; 53:31-44.
[17] Tan A., Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics. - 2003; 2(3 Suppl):S75-S83.
[18] Qu Y., Adam B.,Yasui Y.,Ward M., Cazares L., Schellhammer P., et. al. Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles Discriminates Prostate Cancer from non-cancer Patients. Clinical Chemistry. - 2002; 48(10):1835-1843.
[19] Lee J., Lee J., Park M., Song S. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics &Data Analysis. - 2005; 48:869-885.
[20] Dıaz-Uriarte R., de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. - 2006; 7:3.
[21] Izmirlian G. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Annals of the New York Academy of Sciences. - 2004; 1020:154-174.
[22] Zhang H., Yu C., Singer B. Cell and tumor classification using gene expression data: Construction of forests. Proceedings of National Academy Science. - 2003; 100(7):4168-4172.
[23] Geurts P., Fillet M., Seny D., Meuwis M., Malaise M., Merville M., et al. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics. - 2005; 21(15):3138-3145.
[24] Montana G. Statistical methods in genetics. Briefings in Bioinformatics. - 2006; 7(3):297-308.
[25] Hirschhorn J., Daly M. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. - 2005; 6(2):95-108.
[26] Cordell J.H. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics. - 2009; 10:392-404.
[27] Zhang H., Bonney G. Use of classification trees for association studies.Genetic Epidemiology. - 2000; 19(4):323-332.
[28] Huang J., Lin A., Narasimhan B., Quertermous T., Hsiung C.A., Ho L.T., et.al. Tree-structured supervised learning and the genetics of hypertension. Proceedings of the National Academy of Sciences. - 2004; 101(29):10529-10534.
[29] Ye Y., Zhong X., Zhang H. A genome-wide tree-and forest-based association analysis of comorbidity of alcoholism and smoking. BMC Genetics. - 2005; 6(Suppl. 1):S135.
[30] Zhang Z., Zhang S., Wong M.Y., Wareham N.J., Sha Q. An ensemble learning approach jointly modeling main and interaction effects in genetic association studies. Genetic Epidemiology. - 2008; 32(4):285-300.
[31] McKinney B.A., Reif D.M., Ritchie M.D., Moore J.H. Machine learning for detecting gene-gene interactions: a review. Applied Bioinformatics. - 2006; 5(2):77-88.
[32] Bureau A., Dupuis J., Hayward B., Falls K., Van Eerdewegh P. Mapping complex traits using Random Forests. BMC genetics. - 2003; 4(Suppl. 1):S64.
[33] Klein R.J., Zeiss C., Chew E.Y.,Tsai J.Y., Sackler R.S., Haynes C., et. al. Complement factor H polymorphism in age-related macular degeneration. Science. - 2005; 308(5720):385.
[34] Meng Y., Yu Y., Cupples L.,Farrer L., Lunetta K. Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics. - 2009; 10:78.
[35] Jiang R., Tang W., Wu X., Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics. - 2009; 10(Suppl. 1):S65.
[36] Caragea C., Sinapov J., Silvescu A., Dobbs D. Glycosylation site prediction using ensemble of SVM classifiers. BMC Bioinformatics. - 2007; 8:438.
[37] Peng Y. Anovel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine. - 2006; 36:553-573.
[38] Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics. - 2004; 20(18):3583-3593.

[39] Liu K.H., Xu C.G. A genetic programming-based approach to the classification of multiclass microarray datasets. Bioinformatics. - 2009; 25(3):331-337.
[40] Bhanot G., Alexe G., Venkataraghavan B., Levine AJ. A robust meta-classification strategy for cancer detection from MS data. Proteomics. - 2006; 6(2):592-604.
[41] Kedarisetti K.D., Kurgan L., Dick S. Classifier ensembles for protein structural class prediction with varying homology. Biochemical and Biophysical Research Communications. - 2006; 348(3):981-988.
[42] Hassan M.R., Hossain M.M., Bailey J., Macintyre G., Ho J., Ramamohanarao K. A voting approach to identify a small number of highly predictive genes using multiple classifiers. BMC Bioinformatics. - 2009; 10 (Suppl. 1):S19.
[43] Boulesteix A.L., Slawski M. Stability and aggregation of ranked gene lists.Briefings in Bioinformatics. - 2009; 10(5):556-568.
[44] Dutkowski J., Gambin A. On consensus biomarker selection. BMC Bioinformatics. - 2007; 8(Suppl 5):S5.
[45] Zhang Z., Yang P., Wu X., Zhang C. An agent-based hybrid system for microarray data analysis. IEEE Intelligent Systems. - 2009; 24(5):53-63.
[46] Abeel T., Helleputte T., VandePeer Y., Dupont P., Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. - 2010; 26(3):392-398.
[47] Netzer M., Millonig G., Osl M., Pfeifer B., Praun S., Villinger J., et al. A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry. Bioinformatics. - 2009; 25(7):941-947.
[48] Yang Y.H., Xiao Y., Segal M.R. Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics. - 2005; 21(7):1084-1093.
[49] Chan D., Bridges S.M., Burgess S.C. In: An Ensemble Method for Identifying Robust Features for Biomarker Discovery. Chapman& Hall; - 2007. - Pp. 377-392.
[50] Pengyi Yang, Yee Hwa Yang, Bing B. Zhou1, and Albert Y. Zomaya A review of ensemble methods in bioinformatics Current Bioinformatics. - 2010. 5, (4):296-308.