Тип работы: Отчет по практике

Предмет: Переводоведение

-

Текст оригинала
Machine learning approaches to explore digenic inheritance
Atsuko Okazaki and Jurg Ott
Some rare genetic disorders, such as retinitis pigmentosa or Alport syndrome, are caused by the co-inheritance of DNA variants at two different genetic loci (digenic inheritance). To capture the effects of these disease-causing variants and their possible interactive effects, various statistical methods have been developed in human genetics. Analogous developments have taken place in the field of machine learning, particularly for the field that is now called Big Data. In the past, these two areas have grown independently and have started to converge only in recent years. We discuss an overview of each of the two fields, paying special attention to machine learning methods for uncovering the combined effects of pairs of variants on human disease.
Beyond Mendelian genetics
For heritable traits, classical human genetic approaches have investigated one DNA variant at a time for linkage or association with disease, which has been very fruitful for Mendelian traits, that is, heritable diseases that are due to a single variant. Powerful genome sequencing technologies, such as exome sequencing and whole-genome sequencing, have enabled us to identify many novel genes in Mendelian diseases caused by changes in a single gene or locus. However, several large-scale genetic cohorts revealed that more than half of the rare genetic disease patients are yet to be diagnosed. One of the underlying mechanisms for rare genetic diseases are digenic or oligogenic inheritances, where interaction of two or more genes are observed. Digenic inheritance is the simplest form among such inheritances and has been identified in some rare genetic diseases, seemingly in a Mendelian fashion. Many efforts have been made to identify novel digenic inheritances in genetic rare diseases. One of the accomplishments of such efforts is a novel database providing detailed information on genes and genetic variants seen in digenic diseases. Using the database, a study succeeded in identifying digenic disease genes via machine learning. Other than using databases, it is still important to apply suitable methods for identifying novel interactions between variants. To deal with large numbers of variants obtained from genome sequencing analyses, statistical and machine learning methods have been developed and will be discussed in subsequent sections. Here, we mainly discuss machine learning approaches from a user’s perspective for the detection of pairs of variants underlying digenic traits. Combining disease association effects over multiple variants in human genetics, various specialized ways have been developed to combine information on disease association over multiple DNA variants, which we will briefly discuss here before embarking on the main topic of our outline. With genome-wide sequencing, analysts are faced with 100 000s if not millions of variants available for analysis, for example, case-control association analysis. One approach known as the collapsing method assigns an index variable K to each individual, where K = 1 if the individual carries a (rare) minor allele at any of the variants in a given gene and K = 0 otherwise. Thus, analysis proceeds at the level of genes rather than variants, where the latter far outnumber the former. While such approaches appear to be economical, the effect of a single disease-associated variant or genotype might be diluted by non-associated variants in the same gene. A telling example of such a situation in opioid dependency demonstrated a significant pair of genotypes, but that effect vanished in an analysis based on variants rather than genotypes. A collapsing method was recently applied in a large study on epilepsy. Collapsing methods have also been extended to work with two genes at a time for analysis of digenic traits. So-called complex traits are under the influence of a large number of genes, with schizophrenia being a prime example. A common approach for such traits is to compute for each individual a polygenic risk score, that is, the sum of risk allele frequencies over all SNPs; or the risk allele frequency may be replaced by the number of risk alleles (minor alleles). This approach has been fruitful in traits like Alzheimer’s disease and other complex traits and can at least demonstrate genetic effects, but there is still the task of pinpointing which of the potentially large numbers of SNPs are disease-causing. More and more traits emerge as being digenic, that is, determined by two variants; for example, severe immunodeficiency and autoimmunity, non-syndromic hearing impairment, Noonan syndrome, cancer, and familial hypercholesterolemia. In fact, as we outline later, digenic traits may be more common than monogenic traits. While

many digenic traits have been found in a fortuitous manner, a systemic search may be attempted by an exhaustive enumeration of all pairs of genotypes. Such analyses have been carried out for psoriasis and schizophrenia, but these authors had to restrict attention to biologically interpretable variants. However, investigating all pairs of genes is certainly possible. More sophisticated approaches are afforded by machine learning methods, as outlined next.

Machine learning

Machine learning refers to algorithms in artificial intelligence (AI) that ‘learn’ patterns in data, gradually improving the accuracy of classifications or predictions. Six machine learning methods (stochastic gradient boosting, random forest, neural network (NN), support vector machines, adaptive gradient boosting, and elastic-net penalized logistic regression) have recently applied to predict clinical improvement in 442 patients after an invasive procedure in sports medicine. The method listed last showed the best performance. Also, the relative performances of several data mining approaches have been compared for predicting coronary artery disease and cancer. Most recently, a random forest approach has been developed for identifying candidate digenic disease gene pairs based on biological networks, evolutionary history, and functional annotations. Useful overviews of data mining algorithms have been published. In classical statistical analysis, the number of observations should be a multiple of the number of variables. Nowadays, however, the situation is often reversed. For example, the number of DNA variants tends to greatly exceed the number of individuals, a situation known as the curse of dimensionality. While classical approaches like stepwise multiple regression can cope with this situation to some degree, modern machine learning methods do this in a much more sophisticated manner, by iteratively improving successive solutions to a prediction or classification problem. Artificial NNs Among the earliest examples of AI are artificial NNs, first proposed some 80 years ago. An early application of NNs to genetic data, published 25 years ago, worked with 367 variants and an artificial quantitative trait presented at Genetic Analysis Workshop 10. Today’s NNs consist of many layers of ‘neurons’, hence the term deep learning. NNs are used in many areas of human genetics, for example, to perform cancer type classification and gene identification. Next, we focus on methods in human genetics for the genetic mapping of deleterious variants underlying digenic traits.

Multifactor dimensionality reduction (MDR)

First published 20 years ago, the MDR method carries out an exhaustive evaluation of all possible pairs of variants and then ranks variant pairs by balanced accuracy, that is, a compromise between high power and low P value. For the best pair, MDR can compute an empirical significance level by permutation analysis. MDR has been successfully applied to a large number of diseases and is being used to this day, for example, in an investigation of gene– gene interactions on chronic obstructive pulmonary disease.

Предпереводческий анализ практического материала исследования.

1. Сбор внешних сведений о тексте.

Авторы - Ацуко Окадзаки и Юрг Отт

Время создания и публикации – Октябрь 2022г.

Текст взят из научного журнала «Trends in Genetics»

2. Источник текста – генетические исследования в области дигенного наследования, рецепиент – научная аудитория, которая интересуется вопросами генетики, а также люди, связанные с областью наследования генов.

3. Определить состав информации и её плотность.

Текст несет в себе когнитивную информацию.

Когнитивная информация представляет собой объективные сведения о внешнем мире. Такого рода тексты мы условно назовём информационно-терминологическими и отнесём к ним научные, юридические и технические тексты, учебники, инструкции, деловые письма.

4. Коммуникативное задание текста.

Текст несет в себе задачу передать важную информацию о взаимодействии машинного обучения и порядке дигенного наследования различных заболеваний человека.

5. Определение речевого жанра текста.

Жанр текста – научный.

Основными признаками научной коммуникации являются следующие: научная тематика, точное определение понятий, стремление к обобщению, абстракции, логичность и доказательность изложения, объективный характер изложения, насыщенность фактической информацией, сжатость изложения. Научный стиль имеет ряд общих черт, проявляющихся независимо от характера самих наук (естественных, точных, гуманитарных) и жанров высказывания (монография, научная статья, доклад, учебник и т. д.), что

позволяет говорить о специфике стиля в целом. Его характеризуют логическая последовательность изложения, упорядоченная система связей между частями высказывания, стремление авторов к точности, сжатости, однозначности выражения при сохранении насыщенности содержания. Основной функцией научного стиля является не только передача логической информации, но и доказательство ее истинности, новизны и ценности. Вторичная функция научного стиля — активизация логического мышления читателя (слушателя). Научный стиль делится на три основные разновидности: собственно научный, научноучебный, научно-популярный и множество разновидностей, обслуживающих сферу науки. При рассмотрении классификации текстов важно понимать, что внутри каждого из функционально-стилистических типов есть своя иерархия текстов. Так, если мы обратимся к научным текстам, то увидим, что к ним можно отнести и строго академическую статью, и статью в энциклопедии, и научно-популярный очерк и т. п. Все это – разные виды подачи материала.

Текст перевода

Подходы машинного обучения для изучения дигенного наследования

Ацуко Окадзаки и Юрг Отт

Некоторые редкие генетические заболевания, такие как пигментный ретинит или синдром Альпорта, развиваются сонаследованием вариантов ДНК в двух разных генетических локусах (дигенное наследование). Чтобы изучить влияние этих болезнетворных вариантов и их возможные интерактивные эффекты, в генетике человека были разработаны различные статистические методы. Аналогичные разработки произошли в сфере машинного обучения, особенно в области, которая сейчас называется «большими данными». В прошлом, эти две области развивались независимо друг от друга и только в последние годы начали сближаться. Мы сделаем обзор каждой из двух областей, уделяя особое внимание методам машинного обучения для выявления комбинированного воздействия пар вариантов на заболевания человека.

Помимо менделевской генетики

Классические генетические подходы исследовали за раз только один вариант ДНК человека на предмет сцепления или связи с заболеванием, что оказалось очень плодотворным для менделевских признаков, то есть наследственных заболеваний, обусловленных одним вариантом. Мощные технологии секвенирования генома, такие как секвенирование экзома и секвенирование всего генома, позволили нам идентифицировать множество новых генов при менделевских заболеваниях, вызванных изменениями в одном гене или локусе. Однако несколько крупномасштабных генетических когорт показали, что более половины пациентов с редкими генетическими заболеваниями еще не диагностированы. Одним из механизмов, лежащих в основе редких генетических заболеваний, является дигенное или олигогенное наследование, при котором наблюдается взаимодействие двух и более генов. Дигенное наследование является самой простой формой среди таких наследований и было выявлено при некоторых редких генетических заболеваниях, по-видимому, по менделевскому типу. Было предпринято много усилий, чтобы идентифицировать новые дигенные наследования при генетических редких заболеваниях.

Одним из достижений таких усилий является новая база данных, содержащая подробную информацию о генах и генетических вариантах, наблюдаемых при дигенных заболеваниях. Используя базу данных,

удалось идентифицировать гены дигенных заболеваний с помощью машинного обучения. Помимо использования баз данных, по-прежнему важно применять подходящие методы для выявления новых взаимодействий между вариантами. Для работы с большим количеством вариантов, полученных в результате анализа секвенирования генома, были разработаны статистические методы и методы машинного обучения, которые будут рассмотрены в последующих разделах.
В этой статье мы обсуждаем подходы машинного обучения с точки зрения пользователя для обнаружения пар вариантов, лежащих в основе дигенных признаков.
Объединение эффектов ассоциации с заболеванием в нескольких вариантах
В генетике человека были разработаны различные специальные способы объединения информации об ассоциации заболеваний с несколькими вариантами ДНК, которые мы кратко обсудим здесь, прежде чем приступить к основной теме нашего обзора.
При полногеномном секвенировании аналитики сталкиваются со 100 000, а, если и не с миллионами вариантов, доступных для анализа, например, анализ исследования случай-контроль. Один подход, известный как «метод коллапса», присваивает индексную переменную K каждому индивидууму, где K = 1, если индивидуум несет (редкий) минорный аллель в любом из вариантов данного гена, и K = 0 в противном случае. Таким образом, анализ идет на уровне генов, а не вариантов, где последних намного больше, чем первых.
Хотя такие подходы и кажутся экономичными, эффект одного варианта или генотипа, ассоциированного с заболеванием, может быть ослаблен неассоциированными вариантами того же гена. Наглядный пример такой ситуации при опиоидной зависимости продемонстрировал значимую пару генотипов, но этот эффект исчез при анализе, основанном на вариантах, а не на генотипах. Метод «коллапса» был недавно применен в крупном исследовании эпилепсии. Метод «коллапса» был также расширен для работы с двумя генами одновременно для анализа дигенных признаков.
Так называемые сложные черты находятся под влиянием большого количества генов, ярким примером которых является заболевание шизофрения. Обычный подход к таким характеристикам состоит в том, чтобы вычислить для каждого человека оценку полигенного риска, то есть сумму частот аллелей риска по всем однонуклеотидным полиморфизмам (ОНП); или частота аллелей риска может быть заменена количеством аллелей риска (минорными аллелями). Этот подход может быть полезен в изучении таких болезней, как синдром Альцгеймера и другие, и могут, по крайней мере, продемонстрировать генетические эффекты, но перед исследователями все еще остается задача точного определения того, какие из потенциально большого числа ОНП являются болезнетворными.

-