О МЕТОДЕ ВЫЧИСЛЕНИЯ ОЦЕНОК И КЛАСТЕР–АНАЛИЗЕ

Мгеладзе А.[1], Гоциридзе Г.[2]
1-Грузинский Технический Университет
2-Государственный Университет им. Ак. Церетели

Резюме

Алгоритмы распознавания, основанные на методе вычисления оценок, реализуются, обычно, как оптимизационные процедуры. В связи с интенсивной разработкой теоретического (алгебраического) подхода к синтезу алгоритмов, основные усилия были направлены на создание наиболее общих процедур такого рода. В настоящее время конструируются совершенно новые алгоритмы, которые отличаются большой содержательной наглядностью, имеющей в приложениях не менее значения, чем количественные критерии эффективности. Приведенные в статье модификации представляются полезными, что способствует совершенствованию рассмотренного метода.

Ключевые слова: Вычисления оценок. Кластер.

1. Введение

В монографии Ю. И. Журавлева по методу вычисления оценок подчеркивается, что имеется два принципиально разных способа введения системы Ω опорных множеств [1].

Первый способ состоит в том, что заранее, независимо от анализируемых эмпирических данных, фиксируется определенное свойство А выделяемых множеств. Если подмножество ω обладает этим свойством, то это подмножество включается в выделяемую систему. Чтобы подчеркнуть, что системы выделяются с помощью заранее заданного, постоянного, независящего от обрабатываемых данных, свойства А, оно обозначается через ΩА. Для того, чтобы задать свойство такого типа, обычно достаточно знать только число признаков в таблице T_{nme} обрабатываемых объектов, (n-число классов, на которые делятся объекты, причем для каждого объекта из T_{nme} известно, какому из этих классов он принадлежит).

Второй способ основан на использовании такого свойства, наличие которого у данного подмножества ω существенно зависит от конкретного наполнения таблицы T_{nme} и рассматриваемой классификации K строк этой таблицы (описываемых ею объектов). Поэтому второй способ позволяет строить такие системы Ω опорных множеств, которые специально приспособлены для разделения объектов на разные классы именно данной классификации и именно данной таблицы T_{nme}. В качестве примеров реализации этого способа можно отметить выделение систем так называемых тестов и тупиковых тестов.

2. Основная часть

Подмножество ω признаков относится к семейству $\Omega\left(T_{nme}, K\right)$ тестов классификации K заданных на таблице T_{nme}, если для любых двух различных классов K_q и $Kp(q \neq p)$ из $\left(K_q, K_p \in K\right)$ выполняется условие

$$r_\omega(S, S') = 0, \text{ где } S \in K_q, S' \in K_p \tag{1}$$

Пусть M – множество рассматриваемых объектов (строк таблицы T_{nme}, /M/=м). Введем на множестве всех подмножеств множества P признаков характеристическую функцию δ(ω, M) теста:

$$\delta\left(\omega, M\right) = \begin{cases} 1, \textit{если } \omega - \textit{тест}; \\ 0, \textit{в противном случае}. \end{cases} \tag{2}$$

где символом M подчеркивается, что определение этой характеристической функции зависит не только от ω, но и от множества M объектов, на котором изучаются признаки.

Очевидно, что если δ(ω,M)=1, то

1) δ(ω,M/S)=1,
2) δ(ω′,M)=1 для всех ω′ таких, что ω≤ω′≤P.

Первое из этих свойств теста говорит о том, что сужение исходного множества M объектов сохраняет все тесты, найденные на охватывающем множестве. Из второго свойства следует, что наибольший интерес представляют маломощные тесты. Из него также следует целесообразность выделения специального подкласса тестов, названных тупиковыми. Характеристическое свойство тупикового теста заключается в том, что никакое его собственное подмножество не является тестом, то есть, если ω – тупиковый тест, то для всех ω′⊂ω, δ(ω′,M). Выделение тупикового теста означает выделение такого подпространства ω∈P, которое является безизбыточным по отношению к заданным классификации K и таблице T_{nme}.

С точки зрения качества анализа выбор системы Ω опорных подмножеств в виде множества всех тестов или, что еще лучше, множества тупиковых тестов является высокоэффективным. Однако, с точки зрения оценки сложности выполнения такой выбор является затруднительным или даже невозможным во многих практически интересных случаях (когда /P/~20÷40,/M/~100÷300). Построение такого рода систем требует осуществления слишком большого объема вычислений. Более того, алгоритм реализации такого выбора имеет принципиально переборный характер (экспоненциальную сложность). Особенно трудоемкими оказываются алгоритмы определения семейства тупиковых тестов. Практический интерес могут представлять процедуры выделения не всех тестов, а только тестов малой мошности. При этим, целесообразно строить системы Ω так, чтобы в них не включались тривиальные тесты, которые охватывают уже найденные подмножества – тесты еще меньшей мошности.

Главный недостаток процедур выделения тестов малой мощности состоит в том, что получившаяся система Ω сама может оказаться маломощной, а иногда и вовсе пустой. Чтобы как-то ограничить этот недостаток, прибегают к статистическим алгоритмам поиска тестов. В этом случае разрешается поиск тестов разной мошности, но число поисковых проб ограничено. Такие алгоритмы, не снижая объема вычислений на проверку свойства «тестовости» отдельного подмножества признаков, резко упрощают организацию их перебора. Чтобы добиться снижения объема вычислений на проверку свойства «тестовости» отдельного подмножества, используют следующие три модификации понятия теста:

Подмножество ω называется (q, p)- тестом, если в классификации K найдется такая пара различных классов Kq и Kp, что на части таблицы T_{nme}, выделяемой объектами этих классов, ω является тестом в определенном выше обычном смысле. В этой модификации вместо тестов в качестве элементов системы опорных подмножеств предлагается искать (q, p)-тесты;

Подмножество ω называется (α,β) – квазитестом, если, во-первых, вместо проверки условия (1) для всех S из M, эта проверка осуществляется на заранее выбранной доле α мощности /M/ этого множества, причем выбыраемая α /M/ часть объектов распределяется по M не случайно, а в каждом из ℓ классов заданной классификации K отбирается число, равное α – части мощности этих классов; во-вторых, вместо того, чтобы для каждого выбранного S (пусть $S \in K_q$) осуществлять проверку (1) для всех $S' \in M / K_q$, проводится проверка лишь на числе $\beta \cdot \left| M \setminus K_q \right|$ таких объектов;

Подмножество ω называется (γ, q) – представительным тестом для класса Kq, если, во – первых, ω –это тест в обычном (указанном выше) смысле, и, во – вторых, ω- часть не менее, чем у $\gamma \left| \text{Kq} \right|$ числа объектов класса K_q совпадает.

Указанные модификации понятия теста хотя и дают практическое снижение объема вычислений при проверке данного подмножества (является ли оно тестом), все же не делают это снижение гарантированным. Кроме того, получающееся снижение не оказывается очень большим (сокращает объем вычислений менее, чем в два раза). Наконец, это снижение достигается обычно снижением информационной ценности выделяемых подмножеств: условия для выделения (q, p) и (α, β) тестов – это смягчение условий (1).

3. Заключение

Уже этих обстоятельств достаточно, чтобы заключить, что целесообразно создавать новые процедуры поиска контекстнозависимых систем опорных подмножеств, не связанных с понятием теста. Накопление определенного «запаса» таких процедур представляет, конечно, и самостоятельный интерес: каждая такая процедура строится на выявлении особой связи между

классификацией, индуцированной заданной эмпирической матрицей T_{nme}, и структурой исходного множества признаков, в которой эта классификация функционирует и определяется конкретно, как эти процедуры должны работать для анализа данных об оргсистемах. Расплывчатость и неопределенность, присущая данным об оргсистемах, может быть наиболее просто компенсирована, если строить кластерный анализ наличной косвенной информации об этих системах именно методами вычисления оценок.

Литература:

1. Журавлёв Ю. И. Распознование, классификация, прогноз (Математические методы и их применение) Москва: «Наука», 1989, 302 стр.

THE METHOD OF CALCULATION OF ESTIMATIONS AND THE CLUSTER - ANALYSIS

Mgeladze Anton, Goziridze Gurdzim
Georgian Technical University. Tbilisi
Ak. Tsereteli State University, Kutaisi

Summary

The algorithms of recognition based on a method of calculation of estimations, are usually applied like the optimization procedures. In connection with intensive development of the theoretical (algebraic) approach to synthesis of algorithms, the basic efforts has been directed on creation the most general procedures of such type. Absolutely new algorithms, different by the significant presentation, apply value not less than quantitative criteria of efficiency. The modifications presented in the article are useful thus facilitating the improvement of the given method.

შეფასებათა გამოთვლების მეთოდებისა და კლასტერ ანალიზის შესახებ

ანტონ მგელაძე, გურძიმ გოცირიძე
საქართველოს ტექნიკური უნივერსიტეტი,
აკ. წერეთლის სახელმწიფო უნივერსიტეტი (ქუთაისი)

რეზიუმე

ამოცნობის ალგორითმები, რომლებიც ემყარება შეფასებათა გამოთვლის მეთოდს, რეალიზდება ჩვეულებრივ, როგორც ოპტიმიზაციური პროცედურები. ალგორითმების სინთეზის ინტენსიური თეორიული (ალგებრული) მიდგომის დასამუშავებლად ყურადღება მიმართულია ასეთი სახის ზოგადი პროცედურების შექმნაზე. ეს ალგორითმები გამოირჩევა დიდი შინაარსობრივი თვალსაჩინოებით. ამდენად სტატიაში მოტანილი მოდიფიკაციები სასარგებლოა – წყდება პრობლემა კონკრეტული შემთხვევისათვის და შეფასებათა გამოთვლის მეთოდი ღრმავდება თეორიულად და ხდება უფრო სრულყოფილი.