УДК 519.237

О ранжировании показателей по степени важности и их влиянии на кластерную структуру множества

А.С. Сазонова, С.В. Дронов
АлтГУ, г. Барнаул

Рассматривается задача определения информационной важности статистических показателей объектов некоторого множества.

Пусть имеется n объектов, каждый из которых имеет p числовых показателей $X_1,\ldots,X_p$, и q качественных категорированных показателей Y_i с s_i категориями соответственно, $i=1,\ldots,q$. Предположим, имеется некоторое значимое с точки зрения практики разбиение рассматриваемых объектов на m кластеров. Нам не важно, как именно построен каждый из кластеров, но нам известен его «объективный» ранг, который мы временно примем за числовую метку соответствующего кластера.

Подобно тому, как в [1], определим для каждого из объектов значение кластерной переменной. А именно, поставим каждому из объектов в соответствие номер того кластера, в который он отнесен. Т.о., построено отображение f из набора номеров объектов $\{1,\ldots,n\}$ на множество всех имеющихся кластеров, и, тем самым, каждому объекту придана новая числовая характеристика. Значение $f(j)$ этой характеристики для j-го объекта будем называть кластерной переменной.

Рассмотрим задачу определения информационной важности показателей $X_1,\ldots,X_p,Y_1,\ldots,Y_q$, а также их ранжирования в соответствии со степенью важности. Для решения такой задачи предлагается предварительно произвести оцифровку качественных показателей $Y_1,\ldots,Y_q$, т.е. присвоить категориям качественных показателей цифровые метки, которые будут отражать истинные различия между категориями. Потребуем, чтобы задаваемые метки были согласованы с совместными частотами встречаемости каждого из сочетаний категорий признаков. Такие метки назовем частотно-согласованными, следуя терминологии, предложенной в [2].

Способ построения меток, согласованных с таблицами сопряженности, известен под названием анализ соответствий. Подробное изложение этого способа можно найти, например, в [3]. В результате работы анализа соответствий каждая из категорий показателей может получить векторную метку размерности до $s=\min\limits_{1\le i\le m}\{s_i\}-1$ включительно.

Поскольку координаты векторных меток формируются в порядке степени их разброса, то мы, как и в [2], выберем в качестве числовых меток первые координаты получающихся векторных меток (как наиболее информативные). После этого у каждого из рассматриваемых объектов будет иметься $p+q$ числовых показателей $X_1,\ldots,X_p,X_{p+1},\ldots,X_{p+q}$. Здесь через $X_{p+1},\ldots,X_{p+q}$ обозначены «числовые варианты» первоначально заданных категорированных качественных показателей $Y_1,\ldots,Y_q$.

Наиболее часто применяемым показателем взаимозависимости двух случайных величин является парный коэффициент корреляции. Воспользуемся этой характеристикой для решения нашей задачи. Вычислим коэффициенты корреляции ρ_j между показателем X_j и кластерной переменной f, $j=1,\ldots,p+q$. Составим убывающий ряд из модулей найденных коэффициентов корреляции. Будем ранжировать значимость показателей по убыванию $|\rho_j|$, т.е. будем считать, что чем раньше в данном ряду встречается коэффициент, соответствующий какому-либо показателю, тем более важную роль в построении кластеров играет этот показатель.

Для верификации результата исследования можно применить метод оценки степени влияния числового показателя на вид кластерной структуры, предложенный в [2]. Показатели ранжируются там по величине коэффициента кластерных различий разбиений, получаемых по полному набору показателей и после удаления из этого набора изучаемого показателя. Изучаемый показатель оказывается тем важнее, чем больше вычисленный коэффициент отличается от единицы. В случае не подтверждения результата исследования можно предположить, что такое влияние существенно нелинейно. Тогда, сохраняя установленный экспертом порядок следования кластеров, откажемся от равномерной шкалы их меток. В качестве метки для j-го кластера будем использовать значение $f(j)$, $j=1,\ldots m$. Назовем функцию $f(j)$ функцией перехода. Если при выборе какой-то конкретной функции перехода f модуль коэффициента корреляции показателя X_i окажется статистически значимым, это укажет на линейный характер влияния $f^{-1}(X_i)$ на номер кластера.

Итак, пусть нам удалось найти строго монотонно возрастающую функцию с наибольшим по модулю коэффициентом корреляции $\rho=\rho(X,f)$ между показателем X и кластерной переменной, на j-м кластере равной значению $f(j)$. Тогда

$$f(j\_A) = \rho \frac{S_f}{S_X}\left(X_A - \overline{X}\right) + \overline{f},$$

где j_A – номер кластера, к которому относится объект A, X_A – значение показателя X на этом объекте. Поэтому для нахождения по значению X_A номера того кластера, к которому относится объект A, следует вычислить величину

$$Z_X = f^{-1}\left(\rho \frac{S_f}{S_X}\left(X_A - \overline{X}\right) + \overline{f}\right).$$

Естественно, это можно сделать и для каждого из $p+q$ показателей. Полная прогностическая функция строится суммированием отдельных таких Z_X. Для более высокой точности можно учесть абсолютные величины коэффициентов корреляций, например, строя прогностическую функцию по формуле

$$\delta = \sum_{k=1}^{p+q} |\rho_k| Z_{X_k},$$

где ρ_k – соответствующий максимальный по модулю коэффициент корреляции для k-го показателя. Таким образом, мы получили некоторое число, с помощью которого после его нормировки (для попадания в интервал между $f(1)$ и $f(m)$) и округления до ближайшего целого, можно интерпретировать результат, т.е. мы получим номер кластера, к которому относится рассматриваемый объект.

Таким образом, используя априорную информацию о порядке следования кластеров, был предложен метод ранжирования определяющих показателей объектов. Поскольку было заранее известно, что каждый из исследуемых показателей существенно влияет на формирование кластеров, то предполагается существование «достойной» дискриминационной функции перехода, правильно разделяющей объекты по имеющимся кластерам. Нами разработан алгоритм, позволяющий для каждого показателя определить вид функции перехода, посредством которой его влияние выделяется наиболее «правильным» образом.

Работа выполнена в рамках программы стратегического развития ФГБОУ ВПО «Алтайский государственный университет» на 2012-2016 годы «Развитие Алтайского государственного университета в целях модернизации экономики и социальной сферы Алтайского края и регионов Сибири» (мероприятие «Конкурс грантов-2014», № 2014.312.1.4).

Библиографический список

1. Дронов С.В., Герасимова А.С. К проблеме оцифровки кластерной переменной // Анализ, геометрия и топология : труды всероссийской

молодежной школы-семинара. – Барнаул: ИП Колмогоров И.А., 2013. – С. 54-58.

2. Герасимова А.С. Кластеризация объектов с качественными признаками и её использование для оценки силы их связи. // Известия Алтайского государственного университета. – 2013. – Вып. 1/2(77). – С. 66-69.

3. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.

УДК 519.237

МГК для анализа данных успешности обучения

Л.Л. Смолякова
АлтГУ, г. Барнаул

Проблема анализа функциональной зависимости по эмпирическим данным встает перед многими исследователями в различных отраслях науки. В настоящее время активно развиваются методы углубленного анализа данных, в результате которых выявляются не только разнородные наблюдения, ошибки регистрации данных, статистические закономерности, но и более глубокие характеристики исследуемых процессов, такие как скрытые (латентные) факторы, существенно определяющие параметры функционирования исследуемых систем.

Одним из перспективных и широко исследуемых методов детального анализа данных является метод главных компонент (МГК). Судя по литературе, использование данного метода, позволяет выявить главные факторы, определяющие исследуемые процессы. Они выступают агрегатами исходных наблюдений, и прямое воздействие на которые позволяет обосновывать управленческие решения по повышению качества исследуемых процессов.

В данной работе описывается опыт применения данного метода к анализу процессов обучения бакалавров математического факультета Алтайского государственного университета.

В качестве основных факторов наблюдения были рассмотрены следующие показатели, которые, по мнению автора, влияют на успешность обучения, это:

1. Данные по базовой подготовке ЕГЭ (математика, физика или информатика и русский язык) (баллы).