РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

1. Основные узловые моменты разведочного анализа

Слайд 2

Цель разведочного анализа – представить наблюдаемые данные компактной и простой форме, позволяющей выявить имеющиеся в них закономерности и связи. Разведочный анализ включает преобразование данных и способы наглядного их представления, выявление аномальных значений, грубую оценку типа распределения, сглаживание.

Термин разведочный анализ применяется также в более широком смысле, чем предварительная обработка данных. Например, в многомерных процедурах, таких как факторный анализ, многомерное шкалирование данных, цель разведочного анализа, кроме анализа первичных данных, заключается в определении минимального числа факторов, которые удовлетворительно воспроизводят ковариационную (корреляционную) матрицу или матрицу близостей наблюдаемых переменных

Слайд 3

Считаем, что у исследователя имеются наблюдения в виде матрицы «объект-признак» или вектора признака и частичное или полное отсутствие априорной информации о причинно-следственном механизме этих данных. При анализе обычно возникают следующие вопросы

1. Какой обработке подвергнуть наблюдения?
2. Какую модель выбрать?
3. Какие заключения можно сделать?

Для выбора способа обработки необходима модель наблюдаемых данных. Прежде чем произвести наблюдение необходимо указать природу и свойства измеряемой величины, т.е. использовать априорную информацию. Чем полнее априорная информация, тем точнее и с меньшими затратами можно получить необходимые результаты. Поэтому большое значение имеет формализация методов сбора, обработки и использования априорной информации. На основе анализа этой информации строится модель исследуемого явления, выбирается аппаратура, разрабатывается методика проведения эксперимента.

Слайд 4

Для получения более полной информации об изучаемом явлении проводится первичный анализ данных, получивший название *разведочного анализа (Exploratory data analysis).* Разведочный анализ необходим во всех случаях, за исключением лишь очень простых задач. Например, выбору семейства моделей исследуемого явления в большинстве случаев должен предшествовать предварительный и графический анализ данных. Для иллюстрации сказанного рассмотрим модель простой одномерной линейной регрессии. В соответствии с этой моделью предполагается, что наблюдения n пар (x_1,Y_1), …, (x_n,Y_n) можно описать уравнением

$$\mathbf{M}(Y_i)=\beta_0+\beta_1 x_j, \quad i=1,\mathrm{K}\,,n \qquad (1)$$

В качестве минимального предварительного анализа можно рассматривать график рассеяния точек (x_j,Y_j). В результате анализа графиков можно сделать заключение о постоянстве дисперсии Y_i, о целесообразности преобразования переменных, выявить наличие аномальных наблюдений, для исключения которых необходимы специальные исследования. После такой обработки данных, предполагая, что верна модель (1), необходимо оценить параметры β_0, β_1 и провести графический анализ остатков между наблюдаемыми и оцененными значениями Y_i. На основе этого анализа можно подтвердить или предложить другую модель.

Слайд 5

Рассмотрим простейшие процедуры разведочного анализа, относящиеся к *предварительной обработке данных*. Поясним необходимость проведения разведочного анализа на конкретных вопросах оценивания.

Оценка среднего. Рассмотрим простейший пример оценки $\hat{m}$ истинного среднего m независимой случайной величины x по выборке объема n. Если вычислена оценка среднего, то возникает вопрос: «насколько сильно отличается оценка от ненаблюдаемого истинного значения?» Так как истинное значение m недоступно, то определяется доверительный интервал $\hat{m} \pm tS_{\hat{m}}$, который с заданной вероятностью накрывает истинное значение.

Отношение $t = \hat{m} / s_{\hat{m}}$ имеет t-распределение Стьюдента[1]. Очень часто строят 95%-е доверительные интервалы, считая, что величина t распределена нормально. Для нормального распределения величина t будет равна 1,96, тогда как для t-распределения при числе степеней свободы v ($v = n - 1$), равных 1; 3 и 12, величина t, соответственно, равна 12,7; 4,3 и 2,18. Поэтому *при малых объемах выборок* использование нормального распределения вместо t-распределения приводит к большим ошибкам в интервальной оценке. Большое различие интервальных оценок связано с различием t-распределения от нормального *в хвостах распределения.*

Слайд 6

Хвосты реальных распределений имеют, как правило, больший разброс, чем у нормального распределения. Природа отличия реального распределения от нормального может быть различной:

---

[1] Распределение впервые предложено в 1908 г. У. С. Госсетом (Gosset) (13.6.1876 – 16.10.1937) (псевдоним Student – Стьюдент) и затем более строго обосновано Фишером.

1. Большинство измерений проводится в конкретных единицах, например, в миллиграммах, микронах, и их значения ограничены. Для нормального же закона распределения значения изменяются от $-\infty$ до $+\infty$.

2. Резкая асимметрия некоторых распределений (например,χ^2, F) при малых выборках, обрывистые края у равномерного распределения.

3. Поведение на «хвостах» распределения. Одно или несколько резко выделяющихся значений от основной массы наблюдений могут существенно изменить среднее и катастрофически дисперсию. Неправдоподобные значения почти неизбежны в экспериментальных данных. Количество таких значений в медицинских данных достигает до 30%, а в специально поставленных экспериментах оно составляет около 1% от всех данных.

Оценка среднего среднеарифметическим имеет большие достоинства: несмещенность для генеральных совокупностей, имеющих математическое ожидание, достаточность, полнота и, соответственно, полная эффективность для нормального, пуассоновского, гамма-распределений и при достаточно широких условиях удобное асимптотически нормальное распределение, которое во многих случаях приближенно достигается уже при средних объемах выборок n. Имеются и недостатки такой оценки: эффективность ее равна нулю для равномерного распределения, а для некоторых выборок уже одно неправдоподобно большое наблюдение может сделать среднеарифметическую оценку бесполезной.

Слайд 7

Если нормальность распределения нарушается резко выделяющимися данными, то желательно применять *робастные* (robust – крепкий, здоровый, дюжий) *оценки*. Примером робастной оценки среднего, терпимой к отклонению хвостов распределения от нормального является *медиана* распределения. Она, как срединное значение наблюдений, не зависит от одного или нескольких неправдоподобно больших измерений.

Медиана, как робастная, не является эффективной оценкой относительно среднеарифметической оценки для нормального распределения.

Слайд 8

Мера разброса. На практике для характеристики величины разброса данных используются следующие меры: среднеквадратическое отклонение σ или его квадрат – дисперсия σ^2, а также размах R. Оценки этих величин обозначают соответственно S, S^2, R. Оценка разброса по S широко применяется, и оно полезно

при линейных преобразованиях типа $Y = \beta + \alpha X$. Для некоторых распределений $\sigma^2 = \infty$, а размах применим; неправдоподобно большие отклонения в наблюдениях также могут сделать оценку дисперсии очень большой, что приводит к типу распределения, отличному от истинного.

Оценка разброса по выборочному размаху относится к быстрым процедурам. В связи с появлением быстродействующих ЭВМ вычислительные преимущества *R* по сравнению с *S* становятся все менее важными, но остаются преимущества, связанные с простотой вычисления *R* и возможностью для неспециалистов применять эту статистику. Так, размах практически совсем вытеснил *S* из систем контроля качества, в которых выборки малых объемов берутся через короткие интервалы времени и по средним значениям и размахам строятся контрольные карты.

Следует отметить, что размах можно использовать для распознавания больших неправдоподобных ошибок в вычислениях *S* для выборок из любой генеральной совокупности. Это следует из ограниченности отношения *S/R*.

Слайд 9

Подводя итог рассмотренным оценкам, необходимо сделать вывод, что имеются причины, чтобы не обрабатывать все данные одинаково. Прежде чем приступить к обработке наблюдений, необходимо проверить однородность выборки и, если она неоднородна, то разделить на слои. Наличие резко выделяющихся наблюдений также нарушает однородность выборки. В этом случае один из подходов базируется на обнаружении и удалении этих выделяющихся данных.

Удаление резко выделяющихся наблюдений обеспечивает безопасность оценки, однако обеспечивает эффективность только в случае определения четкой границы между удаленными и не удаленными данными. К явным резко выделяющимся данным примыкает зона «сомнительных» данных (рис. 1), которые не всегда можно распознать.

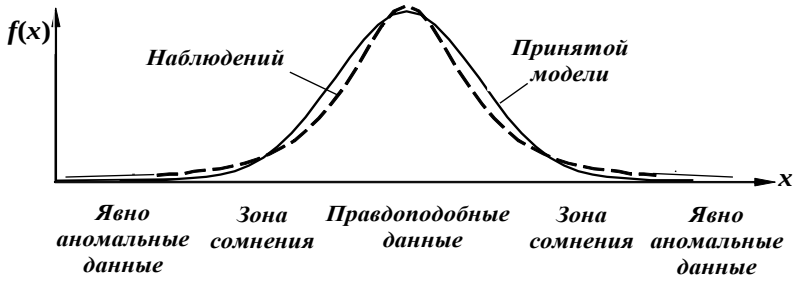

Рис. 1. Плотность распределения. Разбиение данных на три группы

Здесь легко допустить неправильные удаления и необоснованные сохранения, полной эффективности ожидать не приходится даже в идеале после удаления. Эти трудности можно преодолеть, применяя робастные методы оценивания. Робастные алгоритмы обеспечивают безопасность и эффективность оценивания при наличии резко выделяющихся и сомнительных данных.

Слайд 10

О качестве результатов Цель исследования – дать ответ на вопрос: можно полученные результаты применять на практике. Пригодность полученных результатов можно оценить методами перепроверок. Наиболее часто используются методики простой и двойной перепроверок.

Простая перепроверка. Проверка полученной модели проводится на данных, отличных от тех, по которым рассчитаны параметры модели. В этом случае можно выборку наблюдений делить на две (или больше) части. Одну часть используют для обработки, а другую – для проверки. После этого части можно менять местами, что может дать несколько больше информации, хотя здесь имеются определенные трудности, вытекающие из-за связи между двумя оценками качества модели.

Такую перепроверку можно осуществить и для многократного деления данных, например, можно выборку разделить на 10 равных частей. На любых 9 из них провести оценку модели, а на оставшейся одной части осуществить проверку. После этого повторить процедуру 9 раз, беря каждый раз новые 9 частей. В ряде случаев процедуру усложняют. Расчет осуществляют по всем данным без одного наблюдения, а проверку – на отброшенном значении. Расчеты повторяют для каждого из наблюдений выборки. Не следует обольщаться результатами простой проверки, так как контрольная выборка всегда будет больше похожа на рабочую, чем на выборку объектов, для которой будут использоваться результаты исследований.

Двойная перепроверка. Производится проверка на данных отличных, как от тех, по которым строилась модель, так и от тех, которые были использованы для расчета параметров модели. Медики такой метод проверки называют «дважды слепым». «Свежие данные» для перепроверки можно собирать после выбора модели и расчета параметров. Если получение таких данных невозможно, то можно обратиться к архивным данным при условии, что они оставались неизвестными, пока строилась модель и рассчитывались параметры этой модели. При двойной перепроверке важно, чтобы данные, используемые для проверки, явля-

лись отличными от тех, по которым проводились оценки. Можно использовать данные разных лет, если они могут быть отнесены к одному времени, или данные других исследователей.

Слайд 11

2. Неоднородные выборки

Стандартные методы оценивания любой статистики выборочных данных построены на предположении, что выборка взята из однородной совокупности с простой структурой закона распределения. Между тем на практике выборки часто формируются под влиянием различных причин и условий, и они могут быть представлены в виде объединения некоторого множества однородных выборок, каждая из которых имеет простую структуру. *Например, нельзя считать однородными доходы богатых и других граждан государства, так как они имеют различную экономическую основу; объекты различной стоимости, отличающиеся по народнохозяйственным последствиям*. Примерами могут служить неоднородные последовательности динамических моделей в задачах анализа вибраций в машиностроении; сейсмограмм в геофизике; кардиограмм с нарушениями частоты биения сердца.

Природа неоднородности может быть различной. Например, возможны объединения из совокупностей с различными средними и дисперсиями или с одинаковыми средними, но с различными дисперсиями. Важный класс неоднородных выборок образуют также выборки, содержащие одно или несколько *неправдоподобно больших или малых измерений*. Обработка неоднородных выборок теми же методами, какие используются для однородных, недопустима, так как она может привести не только к большим ошибкам, но и к бессмысленным результатам. Подтвердим последнее суждение примером из регрессионного анализа (рис. 2).

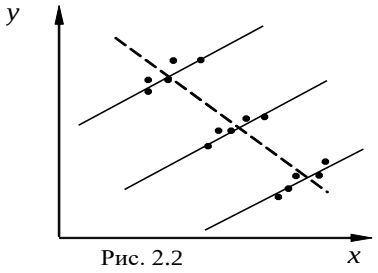

Рис. 2.2

Пусть наблюдения состоят из трех однородных слоев, каждый из которых можно описать простой одномерной регрессией. Эти зависимости показаны на

рис. 2, где прямые – линии регрессий каждой совокупности. Если обработать объединенную выборку этих совокупностей, то получим регрессионную зависимость, изображенную на рис. 2 пунктирной прямой. Очевидно, что регрессия по объединенным данным лишена всякого смысла.

Для определения однородности выборки необходим подробный содержательный анализ исследуемой совокупности. Этот анализ должен базироваться на существенном не случайном признаке, по которому исходная совокупность может быть представлена в виде объединения нескольких однородных совокупностей. *Например, налоговые декларации можно разбить на группы по объемам доходов; учреждения – по числу служащих; фермы – по общей площади земель и валовым доходам.* При разделении выборки на слои требуется ответить на вопросы, по какому признаку лучше производить расслоение, как определить границы между слоями, сколько должно быть слоев.

Слайд 12

Разделение неоднородной совокупности на однородные

Пусть выборка изучаемой совокупности $x_1, \ldots, x_n$, содержит элементы двух независимых случайных величин с плотностями распределений $f(x,\theta_1)$ и $f(x,\theta_2)$. Обозначим через A – множество элементов выборки, принадлежащих к первой случайной величине, B – множество элементов выборки из второй совокупности. Требуется найти оценки $\hat{\theta}_1$, $\hat{\theta}_2$ неизвестных параметров θ_1, θ_2 и множества A и B. Для оценки этих четырех неизвестных используем метод максимума правдоподобия. Неизвестные θ_1, θ_2 и A и B найдем из условия покоординатной максимизации функции правдоподобия

$$L(x_1,\mathrm{K}\,, x_n/\hat{\theta}_1,\hat{\theta}_2, A, B)=\prod_{x_i\in A} f(x_i,\hat{\theta}_1)\prod_{x_i\in B} f(x_i,\hat{\theta}_2)=\max\,.$$

На каждом шаге максимизируется величина функции правдоподобия по одному из неизвестных [18].

Шаг 1. Задаемся произвольным разделением элементов наблюдаемой выборки на A и B.

Шаг 2. Определяем наиболее правдоподобные оценки параметров $\hat{\theta}_1$, $\hat{\theta}_2$, для A и B, для чего решаем две задачи максимизации функций правдоподобия

$$\prod_{x_i\in A} f(x_i,\hat{\theta}_1)=\max\,;\quad \prod_{x_i\in B} f(x_i,\hat{\theta}_2)=\max\,.$$

Шаг 3. По оценкам $\hat{\theta}_1$ и $\hat{\theta}_2$ находим наиболее правдоподобное разделение выборки на множества элементов A и B. Для этого каждый элемент выборки x_i поочередно относим к сомножителям функций правдоподобия. Если при

этом окажется, что $f(x_i, \hat{\theta}_1) > f(x_i, \hat{\theta}_2)$, то наиболее правдоподобным будет отнести x_i к множеству A, и к множеству B, если $f(x_i, \hat{\theta}_1) < f(x_i, \hat{\theta}_2)$,. Если $f(x_i, \hat{\theta}_1) = f(x_i, \hat{\theta}_2)$, , то оба варианта одинаково правдоподобны, что для непрерывных распределений является маловероятным событием. Далее берем следующий элемент и относим его в то или иное множество. Полученные множества сравниваем с множествами на предыдущем шаге. Если они отличаются, то переходим к шагу 2, в противном случае алгоритм останавливается, и задача считается решенной.

Недостатком алгоритма является то, что он останавливается на первом локальном максимуме функции правдоподобия. Частично этого недостаток можно избежать, решая задачу при различных начальных разбиениях на подмножества A и B. Если конечные результаты для нескольких начальных условий различны, то берется то решение, для которого значение функции правдоподобия больше. Отсюда следует, что приведенный алгоритм применим и для выборок, содержащих более двух слоев.

Слайд 13

2.2. Обнаружение аномальных наблюдений

Практика обработки экспериментальных данных показывает, что они наряду с основной однородной массой типичных измерений, представляющих выборку из некоторой генеральной совокупности, как правило, содержат аномальные (неправдоподобные, резко выделяющиеся, «дикие») наблюдения. Аномальные наблюдения в выборке появляются из-за грубых ошибок при регистрации измерений, случайных импульсных помех, сбоев оборудования, измерения в ошибочных единицах и т.д. Если данные резко выделяются на фоне обычных наблюдений, то они могут быть исключены из выборки на предварительном этапе анализа измерений с учетом физической сущности измеряемой величины. Например, легко обнаруживаются наблюдения, которые содержат ошибку в порядке величин. Менее грубые данные, находящиеся вблизи зоны сомнения (рис. 1.6), распознаются сложнее и требуют применения специальных статистических процедур по обнаружению аномальных наблюдений. После обнаружения аномальных наблюдений нельзя считать анализ завершенным и правдивым, если не дано объяснение полученным результатам.

Автоматическое удаление аномальных наблюдений без установления причин их возникновения оправдано лишь тогда, когда исследуемая модель хорошо «обкатана» и доказала право на существование в качестве приближения долгим применением в целевых исследованиях.

Рассмотрим наиболее теоретически обоснованный критерий, предназначенный для обнаружения аномальных наблюдений в одномерных данных. Статистики, применяемые в этих критериях, хорошо изучены и для них имеются таблицы процентных точек.

Слайд 14

Одномерные данные. Обнаружение аномальных наблюдений в одномерных выборках является актуальной задачей при вычислении параметров сдвига, масштаба и при выявлении по остаткам плохо влияющих данных в задаче регрессионного анализа.

Пусть наблюдения $x_1, \ldots, x_n$ являются реализациями независимых случайных величин, подчиняющихся одинаковому нормальному $N(\mu,\sigma^2)$ распределению. Основная гипотеза H_0 состоит в том, что $\mathbf{M}x_i = \mu$, $\mathrm{D}x_i = \sigma^2$, $i = 1, \ldots, n$. Альтернативная гипотеза H_1 заключается в том, что одна или несколько величин имеют среднее $\mu + d$. Это означает, что часть наблюдений описывается тем же нормальным распределением, но со сдвинутым на d средним значением, возможно, сопровождаемому изменениями дисперсии. Если величина сдвига положительна, то говорят о максимальном аномальном наблюдении (гипотеза H_1^+) и о минимальном аномальном наблюдении при отрицательном сдвиге (гипотеза H_1^-).

Одно аномальное наблюдение. Пусть априори неизвестен ни факт наличия аномальных наблюдений, ни место их нахождения. В этом случае для обнаружения аномального наблюдения удобно использовать методы порядковых статистик. Построим вариационный ряд $x_{(1)} \le \ldots \le x_{(n)}$. Проверим нуль-гипотезу против альтернативной H_1^+ для случая одного максимального аномального наблюдения

$$x_{(n)} = \max_i x_i .$$

При построении критерия возможны варианты, зависящие от степени информации о μ и σ. Рассмотрим только случай, когда значения μ и σ неизвестны. В этом случае критериальная статистика вычисляется по формуле

$$D_n = (x_{(n)} - \bar{x})/S ,$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $S^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

Распределение величины D_n получены К. Пирсоном[1] и Н. В. Смирновым[2] (1941). Критические значения D_n, рассчитанные Н. В. Смирновым и Ф. Граббсом (1950), приведены в соответствующих таблицах.

---

[1] Пирсон (Pearson) Карл (27.3.1857 – 27.4.1936) – английский математик, биолог, философ.

Слайд 15

Теперь можно сделать общие выводы об удалении аномальных наблюдений.

1. Любой способ действий с существенно выделяющимися данными, кроме случая, когда он совершенно неприемлем, предотвращает наихудший случай. Особенно чувствительны к таким данным оценки, основанные на МНК. Этим оценкам должны предшествовать проверки на наличие аномальных данных с объективными правилами удаления, а также последующий тщательный анализ остатков. Для данных с неправдоподобными наблюдениями вместо МНК необходимы *робастные процедуры* оценивания.

2. Существенно выделяющиеся данные требуется обнаруживать, преобразовывать и удалять, а также необходимо их интерпретировать, привлекая знания, не относящиеся к статистической природе.

3. Процедуры удаления резко выделяющихся и подозрительно больших наблюдений с последующим оцениванием близки к робастным оценкам. Однако они проигрывают в сравнении с другими методами робастного оценивания.

---

[2] Смирнов Николай Васильевич (17.10.1900 – 2.6.1966) – советский математик, один из создателей непараметрических методов статистики.