Анализ и прогнозирование временных рядов

Лекция 4 04.10.2018

- 11.10.2018 - семинар
- 18.10.2018 - консультация + дедлайн ЛР 1
- 20.10.2018 - коллоквиум.

Алгоритм Gatha-Geva

Развитие идей алгоритма Zahn и рассчитан на тот случай, когда привычгые критерии отделения хороших ребер от плохих в осntовном дереве перестают работать. Таких ситуаций существует много, одна из них - chain problem. Соотвесттвенно авторы предложили в добавление к первым двум классическим критериям еще один, который базируется на понятии нечеткой дисперсионно-ковариационной матрицы кластера. Матрица дефинируется след образом

$$F_i = \frac{\sum_{j=1}^{N}(\mu_{ij})^m(x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{N}(\mu_{ij})^m}$$

x_j - j-й вектор наблюдений, v_i - центроид i-го кластера, произведение векторов подразумечается как векторное , μ_{ij} - значение функции принадлежности j-го вектора i-му кластеру, m - параметр, связанный с дефаззификацией/фуззификацией.

Величину $V_i = \sqrt{detF_i}$ быдем называть нечетким обхемом кластера, а $\sum V_i = TFV$ - Total fuzzy value (суммарный нечеткий объем)

В качестве функции качества кластеризации в алгоритме GG выступает TFV. Алгоритм:

Строим минимальное остовное дерево. Строим бинарное дерево, где каждая вершина отвечает некоторой части разбиения на кластера (дивизивный алгоритм). Текущему состоянию листьев дерева отвечает текущий набор кластеров и на каждом шаге подлежит разбиению тот кластер, у которого значение TFV максимально. Вершиной дерева является ..., листьями - получающиеся в процессе разбиения кластера.

На первом шаге каждого этапа пытаемся разделить кластер с помощью 1 критерия из алгоритма Zahn, т.е удалить такое ребро, длина которого максимально отличается от средней длины. Если это не удается (макс длина ребра несильно отличается от среднего), то используем второй критерий, т.е. удаляем ребро, длина которого много больше, чем средняя длина смежных с ним ребер. Если и это не помогает, то применяем 3-й критерий, специфичный для этого алгоритма, а имено, мы перебираем в рамках выбранного кластера все ребра, производим бинарное разбиение и смотрим, насколько уменьшилось значение TFV.

Процесс завершается либо когда изменение TFV становится малым, либо когда число эл-в в кластерах становится меньше предела, либо TFV меньше заданного значения. В конце листьям дерева соответствуют кластера. Обычно ребра более высокого уровня возникают при применении критериев 1 и 2, а низкого - критерия 3.

Недостатком этого алгоритмом является его неробастность, т.е. малые изменения в выборке и параметтрах алгоритма приводят к существенно различным разбиениям. Этот алгоритм хорошо определяет истинное значение кластеров при разумных значениях параметров, поэтому обычно его сочетают с каким-нибудь алгоритмом типа gaussian mixture, где число кластеров и начальное приближение берется из алгоритма GG.

Алгоритмы, основанные на принципе bottle neck principle.

Речь идет о вероятностоном алгоритме. Будем обозначать через x \in X случайную величину, породившую объекты x, подлежащие кластеризации, через y\in Y - сл.в., описывающую релевантные переменные, t\in T - случайная величина, описывающая кластера, которые мы должны получить.

Пример:

- Релев перем - мно-во слов
- X - классифицируемый текст
- $P(x|y) = \frac{n(x|y)}{\sum_{y'} n(x|y')}$, n - число раз, в которых слово y встретилось в слове x.

Взаимная информация между сл.в. X и Y:

$I(X,Y) = \sum_{x\in X, y\in Y} P(x,y) \log \frac{P(x,y))}{P(x)P(y)} = D_{KL}(P(X,Y)||P(X)\cdot P(Y)) =,$

где P(x,y) -совместное распределение, P(x),P(y) - распределения X и Y

Нетрудно заметить, что так определенная величина представляет собой KL дивергенцию между P(x,y) и P(x)\cdot P(y). Удобно записать в след виде:

$= \sum P(y|x)P(x) \log \frac{P(y|x)}{P(y)}$

Основная идея этой группы алгоритмов заключается в следующем. Собственного говоря при кластеризации существует две идеи, которые, вообще говоря, противоречат друг другу:

1. Во-первых, полученные в результате кластеризации кластера должно быть как можно более компактны.
2. С другой стороны, они должны удерживать как можно больше информации по релевантной переменной.

Всякий алгоритм кластеризации представляет собой опредленный компромисс между двумя этими требованиями. Соответственно, принцип “узкого места” подводит научную основу к этому компромиссу. Эти идеи можно переформулировать в формальном виде:

- Компактность = требование максимизации I(X,T)
- Требование удержания макс объема информации = максимизация $\frac{I(X,T)}{I(X,Y)}$

Результатт решения задачи кластеризации в этих терминах выражается в след. случайных величинах:

1. P(t) - априорное распр по кластерам
2. функция принадлежности P(t|x)
3. распределение P(x|y)

В теории bottleneck доказывется, что эти величины должны удовлетворять след замкнутой системе разрешающих соотношений:

1. $P(t|x) = \frac{P(t)}{Z(\beta,X)}\exp[-\beta D_{KL}(P(y|t)||P(y|x))]$
2. $p(y|t) = \frac{1}{p(t)}\sum P(x)P(y|x)P(x,t)$
3. $p(t) = \sum_x p(t|x)p(x)$

Здесь $\beta = \frac{1}{T}$ - параметр - обратная синтетическая температура (simulated annealing), T - аналог физической температуры в градусах Кельвина. $Z(\beta, X)$ - нормировочный коэффициент, аналог фнкции разбиения в стат механике.

В качестве оценки качества разбиения для этого семейства алгоритмов берут F(T)=I(X,T). Они все обозначаются IB (information bottleneck). К примеру, aIB - agglomerative, sIB - sequential.

aIB. Аггломеративный алгоритм - самый простой алгоритм, когда мы стартуем, когда каждый элемент представляет собой отдельный кластер и присоединяем друг к дугу к два элемента таким образом, чтобы добиться максимального изменения функции I(X,T).

sIB. Предполагается, что мы знаем число кластеров и на началльном этапе элементы между ними рапределяются случайным образом. Далее мы случайно выбираем один элемент из некоторого кластера и присоединяем его к тому кластеру, который дает макс увеличение I(X,T). Обычно эту вариацию IB цепляют к какому-то из алгоритмов типа Gatta-Geva, который дает оценку правильного числа кластеров.

Линейные модели прогнозирования

Прогнозирование в моделях регрессии

$y = X\beta + \varepsilon$,

X - детерминированная матрица n x k, k - число компонентов в каждом наблюдении, n - число наблюдений, β - вектор параметров, ε - случайный вектор, удовлетворяющий условиям Гаусса-Маркова.

M\varepsilon=0

D\varepsilon = \sigma^2 I

Кроме этого мы добавляем к этому уравнению еще одно скалярное уравнение:

$y_{n+1} = x_{n+1}^t\beta + \varepsilon_{n+1}$,

Где ε_{n+1} считается некоррелированной с комонентами вектора ε, вектор x_{n+1} - еще одно наблюдение, y_{n+1} - наблюдения зависимой величины.

Ставится задача отыскания оценки y_{n+1}. Здесь различают случаи безусловного, условного прогнозирования и прогнозирования при наличии автокорреляции ошибок. В 1-2 мы предполагаем выполнение условий ГМ , в 3-м случае они нарушаются.

Лит-ра:

- Магнус, Катышев, ... Эконометрика
- Фан, Яо “nonlinear time series”

Безусловное прогнозирование - предполагаем, что x_n+1 известна точно (незашумлена).

Условное - наличествует некотторый шум.

1. β, σ^2 нам известны точно. В этом случае естественным образом $\hat{y}_{n+1} = x^t_{n+1}\beta$

   $M(y_{n+1} - \hat{y}_{n+1}) = 0$,

   значит оценка несмещенная и дисперсия равна

   $M(y_{n+1} - \hat{y}_{n+1})^2 = M\varepsilon^2_{n+1} = \sigma^2$

   Если мы помимо допущений в структуре шума $\eps$, $\eps_{n+1}$ предположим, что они распределены нормально (стандартная регрессионная модель), то легко видеть, чт ов качестве доверительного интервала $(\hat{y_{n+1}} - \alpha z_{\alpha,n}, \hat{y_{n+1}} + \alpha z_{\alpha,n})$.

   где $z_{\alpha,n}$ - квантиль нормального распределения с ур значимости \alpha и n степенями свободы.

Предположение, что мы знаем значения β, σ^2 нереалистична, поэтому нам нужно каким-то образом оценить их, используя выборк. Используя метод наименьших квадратов, получим

$\hat{\beta} = (X^t X)^{-1} X^t y$

$s^2 = e^t e/(n-k)$

Соответственно в этом случае в качестве оценки исползуем:

```
$\hat{y}_{n+1} = x_{n+1}^t \hat\beta$
```

$M(y_{n+1} - \hat{y}_{n+1}) = 0$, - оценка по-прежнему несмещенная,

Однако она обладает еще одним хорошим свойством, а именно ее СКО - минимальное среди всех возможных линейных по y несмещенных оценок.

Теорема:

Еогда, если $\tilde{y} = C^t y$, -произвольная линейная и несмещенная оценка

$M(\tilde{y} - y_{n+1})^2 \geq M(\hat{y}_{n+1} - y_{n+1})^2$

Док-во:

обратим внимание на тот факт, что коль скоро $\tilde{y}$ является несмещенной оценкой, то, с одной стороны M\tilde y = C^t X \beta, а с другой = $x_{n+1}\beta$, что дает выражение, связывающее X и x_{n+1}.

С другой стороны, если мы рассмотрим

$M(\tilde{y} - y_{n+1})^2 = M(\tilde{y} - \hat{y}_{n+1} + \hat{y}_{n+1} - y_{n+1})^2 =$
$M(\tilde{y} - y_{n+1})^2 + M(\hat{y}_{n+1} - \hat{y}_{n+1})^2 + 2M(\tilde{y} - \hat{y}_{n+1})(\hat{y}_{n+1} - y_{n+1})$

Упражнение: третье слагаемое = 0.

Что, собственно, и доказывает утверждение теоремы.

□

Оценки такого рода носят название best linear unbiased estimator (BLUE).

Замечание 1: теорема имеет место только в классе линейных по y оценок. Мы можем найти нелинейную оценку, которая даст лучший результат по сравнению с оценкой МНК.

Замечание 2: мы накладывали условие несмещенности, что довольно естественно. Но в прогнозных задачах поведение случайных составляющих оказывается настолько серьезно, что приходится отказываться от свойства несмещенности в пользу эффективности.

Дисперсия будет дававться выражением

$My_{n+1}^2 = \sigma^2(1 + X_{n+1}^t(X^tX)^{-1}X_{n+1})$

Если для класифческого регрессионного анализа матрицы обычно хорошо обусловленны, потому что векторы столбцов независимы, в прогнозных задачах часто встречается мультиколлинеарность. Соответственно, можно показать, что в качестве интервальной оценки здесь выступает:

$(\hat{y}_{n+1} - t_{\alpha,n-k}s, \hat{y}_{n+1} + t_{\alpha,n-k}s)$

где $t_{\alpha,n-k}$ - квантиль распредлеления Стьюдента, а $s = \sqrt{\hat{\sigma}^2(1 + X_{n+1}^t(X^tX)^{-1}X_{n+1})}$