ЛЕКЦИЯ 9

План лекции:

1. Построение и применение стохастических грамматик.
2. Оценка вероятностей стохастических грамматик.

9.1 Статистический анализ в задачах распознавания

Для определения и описания переменных, представляющих случайную среду должны быть привлечены статистические понятия и методология. В распознавании образов случайность появляется в основном в результате воздействия двух принципиальных факторов: шума, возникающего при измерении характеристик объекта, и неполноты информации о характеристиках классов образов.

Для получения статистического аппарата, используемого в ходе решения задач распознавания, выполняется обобщение основной модели формальной грамматики G распространением ее на случаи статистического характера.

Для придания статистического характера рассмотренным моделям грамматик используют следующий прием: считают недетерминированными правила подстановки и ставят в соответствие каждому из них некоторую вероятностную меру. Исходя из этого, стохастическую грамматику определяют так:

$$G=(Vn,Vt,P,Q,S),$$

где все ее составляющие определяются по-прежнему, а Q – это множество вероятностных мер, заданных на множестве правил подстановки P.

Рассмотрим процесс порождения терминальной цепочки *x*, начинающейся с S: $S \overset{r_1}{\Rightarrow} \alpha_1 \overset{r_2}{\Rightarrow} \alpha_2 \Rightarrow \ldots \overset{r_m}{\Rightarrow} \alpha_m = x,$ где $(r_1, r_2, \ldots r_m)$ представляют любые m правил подстановки из множества P и $\alpha_1, \alpha_2 \ldots \alpha_{m-1}$ - промежуточные цепочки. Пусть различные правила подстановки применяются с вероятностями $p(r_1), p(r_2), \ldots p(r_m).$ Тогда вероятность порождения цепочки x определяется как

$$p(x) = p(r_1)\,p(r_2 \mid r_1)\,p(r_2 \mid r_1 r_2) \ldots p(r_m \mid r_1 r_2 \ldots r_{m-1}),$$

где $p(r_j \mid r_1 r_2 \ldots r_{j-1})$ - условная вероятность, поставленная в соответствие правилу r_j при предварительном применении правил $r_1 r_2 \ldots r_{j-1}$.

Если $p(r_j \mid r_1 r_2 \ldots r_{j-1}) = p(r_j)$, распределение вероятностей, поставленных в соответствие правилу r_j, называется *неограниченным*, множество Q неограниченно, если все составляющие его распределения вероятностей неограниченны. Стохастическую грамматику называют *неоднозначной*, если существует n различных путей порождения цепочки *x*, характеризующихся

вероятностями $p_1(x), p_2(x)...p_{(x)}, n>1.$ Т.о., вероятность порождения цепочки *x* неоднозначной стохастической грамматикой определяется как $p(x)=\sum_{i=1}^{n} p_i(x).$ $\sum p(x)=1.$ Множество Q совместно, если $x\in L(G)$ Стохастический язык *L(G)* – это язык, порожденный стохастической грамматикой *G*. Каждая терминальная цепочка *x* языка *L(G)* должна обладать вероятностью p(x) порождения данной цепочки. Стохастический язык, порожденный стохастической грамматикой *G*, формально можно определить так:

$$L(G)=\left\{[x,p(x)]\,|\,x\in V_T^+, S\overset{*}{\Rightarrow}x, p(x)=\sum_{i=1}^{n} p_i(x)\right\}(1),$$

где V_T^+ - множество всех терминальных цепочек, исключая пустую, порожденных грамматикой G; обозначение $S\overset{*}{\Rightarrow}x$ используется для обозначения выводимости цепочки x из начального символа S посредством соответствующего применения правил подстановки из множества P. Т.е. выражение (1) означает, что стохастический язык – это множество всех терминальных цепочек, каждой из которых поставлена в соответствие вероятность ее порождения, причем все цепочки выводимы из начального символа S. Вероятность порождения p(x) задается суммированием вероятностей всех различных способов порождения цепочки x. При n>1 стохастический язык становится неоднозначным.

Пример. Рассмотрим стохастическую грамматику *G=(Vn,Vt, P,Q,S),* где

$$V_t=(a,b), V_n=(S), P,Q: S\overset{p}{\rightarrow}aSb, S\overset{1-p}{->}ab.$$

Каждому правилу подстановки поставлена в соответствие вероятность его применения. Дважды применив первое правило, а затем один раз второе, получим последовательность *S->aSb->aaSbb->aaabbb*. Обозначив терминальную цепочку aaabbb через *x* и используя (1), имеем *p(x)=(p)(p)(1-p)=p²(1-p).* Язык, порожденный грамматикой *G*, задается в данном случае следующим образом:

$L(G)=\{[a^t b^t, p^{t-1}(1-p)]\,|\,t>=1\}.$ Где каждая цепочку имеет связанную с ней вероятность. Эта стохастическая грамматика не является неоднозначной, так как существует всего одна последовательность правил подстановки, ведущая к каждой терминальной цепочке.

В стохастических языках используются те же методы грамматического разбора, что и в других грамматиках. Однако для облегчения процесса разбора могут привлекаться знания о вероятности применения правил подстановки. Предположим, например, что на определенном шаге процедуры восходящего грамматического разбора имеется несколько правил-кандидатов, одно из которых следует выбрать и применить. Очевидно, что для успешного разбора, следует начинать с того правила, которое имеет большую вероятность

применения для порождения анализируемой терминальной цепочки. Вероятности применения грамматических правил должны использоваться в грамматическом разборе для увеличения скорости распознавания стохастических систем.

9.2 Оценка вероятностей правил подстановки на основе процедур обучения

При необходимости использовать стохастические грамматики требуется располагать механизмом оценки вероятностей, присутствующих в стохастических грамматиках.

Рассмотрим задачу разделения M классов, характеризующуюся стохастическими грамматиками $G_q = (V_{N_q}, V_{T_q}, P_q, Q_q, S_q), q = 1,2,....M$ (1).

Предполагается, что $V_{N_q}, V_{T_q}, P_q, Q_q, S_q$ известны и грамматики однозначны. Требуется оценить вероятности правил подстановки Q_q, $q=1,2..M$, при помощи множества выборочных терминальных цепочек $T=\{x1, x2,...xm\}$, где каждая цепочка принадлежит языку, порожденному одной из стохастических грамматик.

Собрав все цепочки, перенумеруем их и обозначим через $n(x_h)$ количество появлений цепочки x_h. Каждая цепочка подвергается также разбору с помощью каждой грамматики и число $N_{qij}(x_h)$ обозначает, сколько раз при грамматическом разборе цепочки x_h применялось правило подстановки A_i->b_j грамматики G_q. Хотя вероятности правил подстановки грамматик (1) не известны, предполагается, что сами правила подстановки известны. Поэтому грамматический разбор возможен.

Математическое ожидание n_{qij} числа вхождений правила подстановки A_i->b_j грамматики G_q в грамматический разбор данной цепочки можно аппроксимировать следующим выражением: $n_{qij} = \sum_{x_h \in T} n(x_h)\, p(G_q | x_h) N_{qij}(x_h),$ где $p(G_q \mid x_h)$ – вероятность порождения данной цепочки x_h грамматикой G_q. В процессе обучения эта вероятность должна быть определена для каждой цепочки. Вероятность p_{qij} применения правила подстановки A_i->b_j в грамматике G_q может быть аппроксимирована соотношением $\hat{p}_{qij} = \frac{n_{qij}}{\sum_k n_{qik}}$, где $\hat{p}_{qij}$ - оценка вероятности p, а суммирование в знаменателе выполняется по всем правилам подстановки грамматики G_q, имеющим вид A_i->b_k , т.е. для всех правил подстановки грамматики G_q с одинаковой нетерминальной левой частью A_i.

По мере приближения числа цепочек в T к бесконечности оценка вероятности $\hat{p}_{qij}$ приближается к истинной вероятности правила подстановки p_{qij} при выполнении следующих условий:

1. Множество T – репрезентативное подмножество языков $L(G_q)$, $q=1,2,...M$, в том смысле, что $T->L$, где L – объединение языков, т.е. $L = \bigcup_{q=1}^{M} L(G_q)$.

2.Оценка вероятности появления цепочки x_h в множестве T, определяемая соотношением $\hat{p}(x_h) = \frac{n(x_h)}{\sum_{x_k \in L} n(x_k)}$, приближается к истинной вероятности $p(x_h)$.

3. В процессе обучения для каждой цепочки x_h может быть определена вероятность $p(G_q|x_h)$.

Вероятность $p(G_q|x_h)$ того, что данная цепочка x_h принадлежит классу c_q, обычно без проблем может быть установлена в обучающей фазе. Если точно известно, что данная цепочка принадлежит исключительно классу c_q, то $p(G_q|x_h)=1$. Аналогично, если известно, что x_h не может принадлежать c_q, то $p(G_q|x_h)=0$. Однако некоторые цепочки могут принадлежать более чем одному классу. В этом случае оценку вероятности $p(G_q|x_h)$, $q=1,2,...M$ для этих цепочек можно получить, фиксируя относительную частоту, с которой они встречаются в каждом классе. При этом необходимо, чтобы $\sum_{q=1}^{M} p(G_q \mid x_h) = 1.$

Когда невозможно определить относительную встречаемость неоднозначных цепочек в каком-либо определенном классе, наиболее оправданным для них считается допущение $p(G_q|x_h)=1/M$.