بسمه تعالی

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گزارش مطالعاتی

مدل مخفی مارکوف و الگوریتمهای آموزش

Hidden Markov Model and Training Algorithms

تهیه کننده:

محمد حسین معطر

moattar@ce.aut.ac.ir

اردیبهشت ۱۳۸۵

فهرست مطالب

۱- مقدمه

یکی از مسائلی که در پردازش سیگنال توجهات را به خود معطوف نموده است، مدلسازی سیگنال است. انتخابهای مختلفی برای مدل کردن سیگنال و خصوصیات آن وجود دارد. از یک دیدگاه می توان مدلهای سیگنال را به دو دسته مدلهای معین[1] و مدلهای آماری[2] تقسیم بندی نمود. مدلهای معین عمدتا برخی خواص شناخته شده سیگنال را مورد استفاده قرار می دهند. در این حالت تشکیل مدل سیگنال سرراست است و تنها کافی ست مقادیر پارامترهای مدل تخمین زده شود. در مدلهای آماری سعی در ایجاد مدل با استفاده از خواص آماری سیگنال است. مدلهای گاوسی، زنجیره مارکوف و مدل مخفی مارکوف از جمله این روشها هستند. فرض اساسی در مدلهای آماری این است که می توان خواص سیگنال را به شکل یک فرآیند تصادفی پارامتری مدل نمود.

مدل مخفی مارکوف در اواخر دهه ۱۹۶۰ میلادی معرفی گردید و در حال حاضر به سرعت در حال گسترش دامنه کاربردها می باشد. دو دلیل مهم برای این مساله وجود دارد. اول اینکه این مدل از لحاظ ساختار ریاضی بسیار قدرتمند است و به همین دلیل مبانی نظری بسیاری از کاربردها را شکل داده است. دوم اینکه مدل مخفی مارکوف اگر به صورت مناسبی ایجاد شود می تواند برای کاربردهای بسیاری مورد استفاده قرار گیرد.

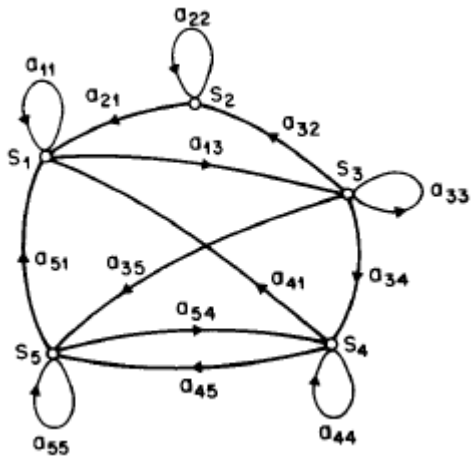
[بازگشت به فهرست]

۲- فرآیند مارکوف گسسته

یک سیستم مانند شکل زیر را که در هر لحظه در یکی از حالت متمایز $S_1,\ldots,S_N$ است در نظر بگیرید. در زمانهای گسسته و با فواصل منظم، حالت سیستم با توجه به مجموعه ای از احتمالات تغییر می کند. برای زمانهای $t=1,2,\ldots$ حالت در لحظه t را با q_t نشان می دهیم. برای یک توصیف مناسب از سیستم فعلی نیاز به دانستن حالت فعلی در کنار تمام حالات قبلی می باشد. برای یک حالت خاص از زنجیره مارکوف مرتبه اول، توصیف احتمالاتی تنها با حالت فعلی و حالت قبلی مشخص می شود.

[1] Deterministic Model
[2] Statistical Model

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \mid q_{t-1} = S_i) \qquad (۱)$$



شکل ۱: یک زنجیره مارکوفی با ۵ حالت [Rabiner 1989]

حال تنها فرآیند هایی را در نظر می گیریم که در آنها سمت راست رابطه فوق مستقل از زمان است و به همین دلیل ما مجموعه ای از احتمالات انتقال بین حالتها را خواهیم داشت.

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i) \qquad 1 \le i, j \le N \qquad (۲)$$

که در آن احتمال انتقال بین حالات دارای خواص زیر است.

$$a_{ij} \ge 0$$
$$\sum_{j=1}^{N} a_{ij} = 1 \qquad (۳)$$

فرایند تصادفی فوق را مدل مارکوف قابل مشاهده[3] می گویند زیرا خروجی مدل مجموعه ای از حالات است که قرار گرفتن در آنها متناظر با یک مشاهده می باشد. ما می توانیم دنباله مشاهدات مورد انتظار خود را تولید کنیم و احتمال وقوع آن در زنجیره مارکوف را محاسبه نماییم. برای مثال با داشتن دنباله مشاهدات $O = \{q_1, \ldots, q_t\}$ احتمال وقوع آن به صورت زیر بیان می شود.

$$P(O \mid Model) = P(\{q_1, q_2, \ldots, q_2\} \mid Model\} = P(q_1).P(q_2 \mid q_1).P(q_3 \mid q_2)\ldots P(q_t \mid q_{t-1}) \qquad (۴)$$

[3] Observable

یکی دیگر از مواردی که مطرح می شود این است که اگر سیستم در حالت q_i باشد با چه احتمالی به حالت q_j می رود و با چه احتمالی در همان حالت q_i باقی می ماند.

[بازگشت به فهرست]

۳- مرتبه مدل مارکوف

۳-۱- مدل مارکوف مرتبه صفر

یک مدل مارکوف از مرتبه صفر هیچ حافظه ای ندارد و برای هر t و t' در دنباله سمبلها، $pr(x_t = S_i) = pr(x_{t'} = S_i)$ خواهد بود.

مدل مارکوف از مرتبه صفر مانند یک توزیع احتمال چند جمله ای می باشد. چگونگی تخمین پارامترهای مدل مارکوف مرتبه صفر و همچنین پیچیدگی مدل در Wallace and Boulton [1968] آماده است.

[بازگشت به فهرست]

۳-۲- مدل مارکوف مرتبه اول

یک مدل مارکوف مرتبه اول دارای حافظه ای با طول ۱ می باشد. توزیع احتمال در این مدل به صورت زیر مشخص می شود.

$$pr(x_t=S_i \mid x_{t-1}=S_j), \text{ for } i = 1..k \ \& \ j = 1..k$$

تعریف فوق مانند این است که k مدل مارکوف در مرتبه صفر برای هر S_j داشته باشیم.

[بازگشت به فهرست]

۳-۳- مدل مارکوف مرتبه m ام

مرتبه یک مدل مارکوف برابر است با طول حافظه ای که مقادیر احتمال ممکن برای حالت بعدی به کمک آن محاسبه می شود. برای مثال، حالت بعدی در یک مدل مارکوف از درجه ۲ (مدل مارکوف مرتبه دوم) به دو حالت قبلی آن بستگی دارد.

مثال ۱: برای مثال اگر یک سکه معیوب A داشته باشیم که احتمالات شیر یا خط آمدن برای آن یکسان نباشد، می توان آن را با یک مدل مارکوف درجه صفر با استفاده از احتمالات pr(H) و pr(T) توصیف نمود.

$$pr(H)=0.6,\ pr(T)=0.4$$

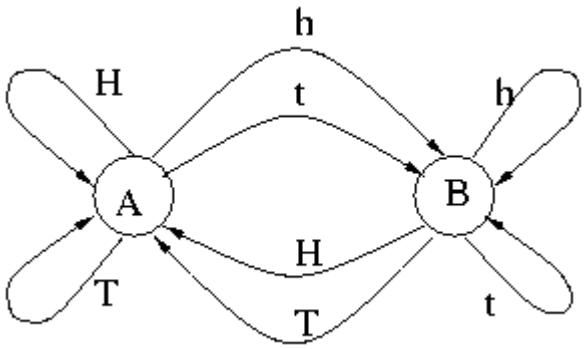
مثال ۲: حال فرض کنید که سه سکه با شرایط فوق در اختیار داریم. سکه ها را با اسامی A، B و C نام گذاری می نماییم. آنگاه برای توصیف روال زیر به یک مدل مارکوف مرتبه اول نیاز داریم:

۱) فرض کنید سکه X یکی از سکه های A و یا B باشد.
۲) مراحل زیر را تکرار می کنیم.
a) سکه X را پرتاب می کنیم و نتیجه را می نویسیم.
b) سکه C را نیز پرتاب می کنیم.
c) اگر سکه C خط آمد، آنگاه سکه X را تغییر می دهیم (A را با B یا B را با A جایگزین می کنیم.) و در غیر اینصورت تغییری در سکه ها نمی دهیم.

انجام روال فوق مدل مارکوف مرتبه اول زیر را نتیجه خواهد داد.

	x_{t-1}			
	H	T	h	t
$pr(x_t=H)$	0.6×0.8	0.4×0.8	0.3×0.2	0.7×0.2
$pr(x_t=T)$	0.6×0.8	0.4×0.8	0.3×0.2	0.7×0.2
$pr(x_t=h)$	0.6×0.2	0.4×0.2	0.3×0.8	0.7×0.8
$pr(x_t=t)$	0.6×0.2	0.4×0.2	0.3×0.8	0.7×0.8

یک پردازش مارکوفی مانند نمونه فوق در طول پیمایش احتمالات، یک خروجی نیز خواهد داشت. یک خروجی نمونه برای پردازش فوق می تواند به شکل HTHHTHHttthtttHHTHHHHttthtthttht باشد.



شکل ۳: مدل مخفی مارکوف برای پرتاب سکه طبق راول فوق

مدل مارکوف فوق را می توان به صورت نموداری از حالات و انتقالها نیز نشان داد. کاملا مشخص است که اینگونه بازنمایی از مدل مارکوف مانند بازنمایی یک ماشین انتقال حالت محدود است که هر انتقال با یک احتمال همراه می باشد.

[بازگشت به فهرست]

۴- مدل مخفی مارکوف (HMM)

تا اینجا ما مدل مارکوف، که در آن هر حالت متناظر با یک رویداد قابل مشاهده بود را معرفی نمودیم. در این بخش تعریف فوق را گسترش می دهیم، به این صورت که در آن، مشاهدات توابع احتمالاتی از حالتها هستند. در این صورت مدل حاصل یک مدل تصادفی با یک فرآیند تصادفی زیرین است که مخفی است و تنها توسط مجموعه ای از فرآیند های تصادفی که دنباله مشاهدات را تولید می کنند قابل مشاهده است.

برای مثال فرض کنید که شما در یک اتاق هستید و در اتاق مجاور آن فرد دیگری سکه هایی را به هوا پرتاب می کند و بدون اینکه به شما بگوید این کار را چگونه انجام می دهد و تنها نتایج را به اطلاع شما میرساند. در این حالت شما با فرآیند مخفی انداختن سکه ها و با دنباله ای از مشاهدات شیر یا خط مواجه هستید. مساله ای که اینجا مطرح می شود چگونگی ساختن مدل مارکوف به منظور بیان این فرآیند تصادفی است. برای مثال اگر تنها مشاهدات حاصل از انداختن یک سکه باشد، می توان با یک مدل دو حالته مساله را بررسی نمود. یک مدل مخفی مارکوف را می توان با تعیین پارامترهای زیر ایجاد نمود:

- تعداد حالات ممکن: تعداد حالتها در موفقیت مدل نقش به سزایی دارد و در یک مدل مخفی مارکوف هر حالت با یک رویداد متناظر است. برای اتصال حالتها روشهای متفاوتی وجود دارد که در عمومی ترین شکل تمام حالتها به یکدیگر متصل می شوند و از یکدیگر قابل دسترسی می باشند (مدل ارگودیک[4]).

[4] Ergodic Model

- تعداد مشاهدات در هر حالت: تعداد مشاهدات برابر است با تعداد خروجیهایی که سیستم مدل شده خواهد داشت.
- تعداد حالتهای مدل N
- تعداد سمبلهای مشاهده در الفبا، M. اگر مشاهدات گسسته باشند آنگاه M یک مقدار نا محدود خواهد داشت.

$$A = \{a_{ij}\}$$

- ماتریس انتقال حالت $A = [a_{ij}]$: یک مجموعه از احتمالات انتقال در بین حالتها

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}, \quad 1 \le i, j \le N,$$

که در آن q_t بیانگر حالت فعلی می باشد. احتمالات انتقال باید محدودیتها طبیعی یک توزیع احتمال تصادفی را برآورده نمایند. این محدودیتها شامل موارد زیر می گردند.

$$a_{ij} \ge 0, \quad 1 \le i, j \le N$$

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N$$

برای حالات مدل ارگودیک برای تمام i و j ها مقدار a_{ij} بزرگتر از صفر است و در موردی که اتصالی بین حالات وجود ندارد $a_{ij} = 0$.

- توزیع احتمال مشاهدات: یک توزیع احتمال برای هر یک از حالتها

$$B = \{b_j(k)\}$$

.

$$b_j(k) = p\{o_t = \nu_k | q_t = j\}, \quad 1 \le j \le N, \quad 1 \le k \le M \qquad (۵)$$

که در آن ν_k بیانگر k^{th} سمبل مشاهده شده در الفبا است و o_t بیانگر بردار پارامترهای ورودی فعلی می باشد. در مورد مقادیر احتمال حالتها نیز شرایط موجود در نظریه احتمال باید رعایت گردند.

$$b_j(k) \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M$$

$$\sum_{k=1}^{M} b_j(k) = 1, \quad 1 \le j \le N$$

اگر مشاهدات به صورت پیوسته باشند، باید به جای احتمالهای گسسته از یک تابع چگالی احتمال پیوسته استفاده شود. معمولا چگالی احتمال به کمک یک مجموع وزندار از M توزیع نرمال $\mathcal{N}$ تخمین زده می شود.

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mu_{jm}, \Sigma_{jm}, o_t)$$

که در آن c_{jm} ، μ_{jm} و Σ_{jm} به ترتیب ضریب وزندهی، بردار میانگین و ماتریس کواریانس می باشند. در رابطه فوق مقادیر c_{jm} باید شرایط زیر را ارضا نمایند:

$$c_{jm} \geq 0, \quad 1 \leq j \leq N, \quad 1 \leq m \leq M$$

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1 \leq j \leq N$$

- توزیع احتمال حالت آغازین $\pi = \{\pi_i\}$ که در آن

$$\pi_i = p\{q_1 = i\}, \quad 1 \leq i \leq N$$

به این ترتیب ما می توانیم یک مدل مخفی مارکوف با توزیع احتمال گسسته را با استفاده از سه گانه زیر مشخص نماییم.

$$\lambda = (A, B, \pi) \qquad (۶)$$

همچنین یک مدل مخفی مارکوف با توزیع احتمال پیوسته به صورت زیر نشان داده می شود.

$$\lambda = (A, c_{jm}, \mu_{jm}, \Sigma_{jm}, \pi)$$

[بازگشت به فهرست]

۵- یک مثال واقعی

فرض کنید دوست دارید که دور از شما زنگی می کند و شما با او در مورد اینکه هر روز چه کاری انجام می دهد از طریق تلفن صحبت می کنید. دوست شما تنها به سه کار علاقه مند است: پیاده روی در پارک، خرید و نظافت آپارتمان خود. انتخاب او کاملا با وضعیت هوایی هر روز در ارتباط است. شما هیچ اطلاعی از آب و هوای محلی که دوست شما در آن زندگی می کند ندارید اما بر حسب آنچه که او هر روز از کارهای خود تعریف می کند شما سعی می کنید که آب و هوای محل زندگی دوستتان را حدس بزنید.

شما قبول دارید که هوا مانند یک زنجیره مارکوف (Markov chain) گسسته عمل می کند. دو وضعیت ممکن است وجود داشته باشد: هوا بارانی (rainy) باشد و یا هوا آفتابی (sunny) باشد. اما شما نمی توانید آنها را مستقیما مشاهده

کنید زیرا آنها از شما مخفی هستند. هر روز این شانس وجود دارد که دوست شما یکی از عملیت ''walk''، ''shop'' و یا ''clean'' را با توجه به وضعیت هوا انجام دهد. دوست شما در مورد فعالیتی که انجام می دهد به شما توضیحاتی می دهد که به آنها مشاهدات می گوییم. اینگونه سیستمها را مدل مخفی مارکوف می گویند.

شما وضعیت کلی هوا و اینکه دوستتان تمایل دارد چه کاری را انجام دهد می دانید. به بیان دیگر پارامترهای مدل HMM مشخص است. می توان مدل HMM مورد نظر را به صورت نمادین زیر بیان نمود.

```
states = ('Rainy', 'Sunny')
observations = ('walk', 'shop', 'clean')
start_probability = {'Rainy': 0.6, 'Sunny': 0.4}
transition_probability = {
  'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
  'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
  }
emission_probability = {
  'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
  'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
  }
```

در این قسمت کد start_probability بیانگر عدم قطعیت شما در مورد این است که در زمانی که دوست شما با شما تماس می گیرد، مدل HMM در کدام حالت است (همه چیزی که شما می دانید این است که آب و هوا به صورت پیش فرض بارانی است). مقدار transition_probability چگونگی تغییرات آب و هوا را در زنجیره مارکوف مشخص می نماید. در مدل بالا تنها ۳۰ درصد شانس وجود دارد که اگر هوای امروز بارانی است هوای فردا آفتابی باشد. مقدار emission_probability تعیین می کند که دوست شما به چه میزان علاقه مند به انجام یک فعالیت خاص در یک روز است. اگر هوا بارانی باشد ۵۰ درصد شانس وجود دارد که او آپارتمان خود را تمیز کند و اگر آفتابی باشد، به احتمال ۶۰ درصد دوست شما برای پیاده روی از منزل خارج می شود.

[بازگشت به فهرست]

۶- سه مساله اصلی

برای اینکه مدل HMM در دنیای واقعی قابل استفاده باشد باید سه مساله مهم حل شود. این سه مساله به قرار زیرند:

۱- مساله ارزیابی(Evaluation Problem)

با داشتن دنباله مشاهدات $O=\{O_1,...,O_t\}$ و مدل $\lambda=\{A,B,\pi\}$ چگونه $P(O|\lambda)$، احتمال تولید دنباله مشاهدات توسط λ را محاسبه نماییم؟

۲- مساله کدگشایی (Decoding problem)

با داشتن دنباله مشاهدات $O=\{O_1,...,O_t\}$ و مدل $\lambda=\{A,B,\pi\}$ چگونه دنباله حالات بهینه $Q=\{q_1,...,q_t\}$ برای تولید $O=\{O_1,...,O_t\}$ را بدست آوریم؟

۳- مساله آموزش (Learning problem)

چگونه پارامترهای مدل، A,B,π، را بدست آوریم؟

در کاربردی مانند شناسایی گفتار، مساله ارزیابی برای شناسایی کلمات جدا (isolated word recognition) استفاده می شود. مساله کدگشایی با کاربردهایی مانند شناسایی گفتار پیوسته و تقطیع سر و کار دارد. مساله آموزش نیز برای اینکه ما بتوانیم مدل HMM را در کاربردهای مختلف شناسایی استفاده نماییم، باید حل شود.

برای حل مسأله اول روالهای پیشرو و پسرو[5] پیشنهاد شده‌اند. در این روش تمام دنباله حالتهای با طول t در نظر گرفته می شود. برای مثال احتمال تولید دنباله مشاهدات $O=\{O_1,...,O_t\}$ از دنباله حالات $Q=\{q_1,...,q_t\}$ از مدل λ به صورت زیر محاسبه می شود.

$$P(O|Q,\lambda)=\prod_{i=1}^{t}P(O_i|q_i,\lambda) \qquad (۷)$$

که می توان آن را به شکل زیر نیز نوشت:

$$P(O|Q,\lambda)=b_{q_1}(O_1).b_{q_2}(O_2)...b_{q_t}(O_t). \qquad (۸)$$

احتمال وقوع دنباله حالتهای $Q=\{q_1,...,q_t\}$ نیز به شکل زیر محاسبه می شود:

$$P(Q|\lambda)=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}...a_{q_{t-1}q_t}. \qquad (۹)$$

حال احتمال اینکه هر دو رویداد O و Q همزمان رخ دهد را با $P(O,Q|\lambda)$ نشان می دهیم:

$$P(O,Q|\lambda)=P(O|Q,\lambda)P(Q,\lambda). \qquad (۱۰)$$

[5] Forward Backward Procedure

در اینجا احتمال وقوع O با جمع احتمال برای تمام دنباله حالات ممکن بدست می آید.

$$P(O \mid \lambda) = \sum_{allQ} P(O \mid Q, \lambda) P(Q, \lambda) \qquad (۱۱)$$

مسأله دوم که بدست آوردن رشته حالات بهینه است توسط الگوریتم جستجوی ویتربی[۶] انجام می‌شود و در فاز بازشناسی بکار می‌آید و مسأله سوم نیز مسأله تخمین بهینهٔ پارامترهای مدل یا همان مسأله آموزش مدل مارکف می‌باشد و به یکی از دو روش ویتربی و یا بام- ولش[۷] انجام می‌گردد. البته آموزش مدل می‌تواند ترکیبی از این دو روش نیز باشد. در عمل و در فاز پیاده‌سازی، روشهایی برای حل مسائلی چون ناکافی بودن داده‌های آموزشی، مقدار دهی اولیه مدل برای شروع آموزش و کم کردن خطای محاسباتی پیشنهاد شده است، که باید در پیاده‌سازی عملی یک سیستم مبتنی بر مدل مارکف لحاظ شوند. بررسی دقیق موارد فوق نیاز به فرصت جداگانه دارد به همین دلیل تا این اندازه اکتفا می شود. برای بررسی راه حلهای ارائه شده برای مسائل فوق می توانید به [Rabiner 1989] مراجعه نمایید.

[بازگشت به فهرست]

۷- انواع مدلهای مخفی مارکوف و HMM پیوسته

همانطور که گفته شد نوع خاصی از HMM وجود دارد که در آن تمام حالات موجود با یکدیگر متصل هستند. لیکن مدل مخفی مارکوف از لحاظ ساختار و اصطلاحا توپولوژی انواع مختلف دارد. همانطور که گفته شد برای مدل ارگودیک برای تمام i و j ها $a_{ij} > 0$ است و ساختار مدل مثل یک گفتار کامل است که راسها در آن دارای اتصالات بازگشتی نیز می باشند. لیکن برای کاربردهای متفاوت و با توجه به پیچیدگی فرآیند نیاز به ساختار متفاوتی وجود دارد. از جمله این ساختارها که به شکل گسترده ای در کاربردهای شناسایی گفتار مبتنی بر واج و شناسایی گوینده مورد استفاده قرار می گیرد، مدل چپ به راست[۸] یا مدل بکیس[۹] است. این مدل که ساختار آن را در شکل ۲ نیز می بینید، دارای اتصالات چپ به راست است و برای مدل کردن سیگنالهایی که خواص آنها با زمان تغییر می کند مورد استفاده قرار میگیرد. در مدل چپ به راست تنها یک حالت ورودی وجود دارد که همان حالت اول است و به این ترتیب:

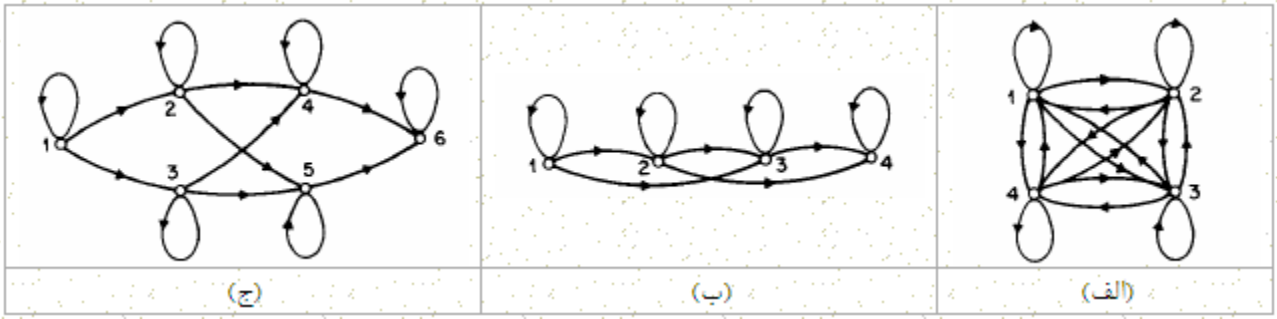
$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \qquad (۱۲)$$

مدلهای ارگودیک و چپ به راست مدلهای HMM پایه هستند و در پردازش گفتار نیز بیشترین کاربرد را دارا می باشند. هرچند می توان با اتصال چندین مدل و یا تغییر در ساختار اتصالات آن مدلهایی با انعطاف پذیری بیشتری ایجاد

---

[6] Viterbi Search
[7] Baum Welch
[8] Left to Right
[9] Bakis

نمود[Rabiner 1989]. شکل ۲-ج یک نمونه از مدل موازی چپ به راست، که شامل دو مدل چپ به راست است، را نشان می دهد.



شکل ۲: سه ساختار برای مدل HMM (الف) مدل HMM ارگودیک
(ب) مدل چپ به راست (ج) مدل موازی چپ به راست [Rabiner 1989]

در قسمتهای قبل مدلهای HMM برای مجموعه مشاهدات گسسته را مورد بررسی قرار دادیم. اگر چه می توان با چندی سازی تمام فرآیندهای پیوسته را به فرآیندهای با دنباله مشاهدات گسسته تبدیل نمود، اما این کار ممکن است باعث افت مدل شود. در مدل HMM پیوسته احتمال قرار گرفتن مشاهدات در یک حالت را با توابع چگالی احتمال نشان می دهند. در این شرایط برای هر حالت i و ورودی O، احتمال مشاهده $b_i(O)$ به صورت یک توزیع شامل M مخلوط نشان داده می شود:

$$b_i(O) = \sum_{m=1}^{M} c_{im} \Re(O, \mu_{im}, U_{im}), \qquad 1 \le i \le N \qquad (۱۳)$$

که در آن c_{im} ضریب مخلوط m ام است و $\Re$ می تواند هر تابع چگالی باشد. معمولا از تابع گاوسی برای این منظور استفاده می شود. ضرایب مخلوط فوق باید محدودیتهای زیر را داشته باشند:

$$\sum_{m=1}^{M} c_{im} = 1 \qquad 1 \le i \le N$$
$$c_{im} \ge 0 \qquad 1 \le i \le N \qquad 1 \le m \le M \qquad (۱۴)$$

[بازگشت به فهرست]

۸- مدل مخلوط گاوسی[۱۰]

[10] Gaussian Mixture Model(GMM)

مدل مخلوط گوسی یکی از مهمترین روشهای مدل کردن سیگنال است که در واقع شبیه یک HMM یک حالته است که تابع چگالی احتمال آن حالت دارای چندین مخلوط نرمال می باشد. احتمال تعلق بردار آزمایشی x به یک مدل مخلوط گاوسی دارای M مخلوط به شکل زیر بیان می شود:

$$P(x \mid GMM) = \sum_{i=1}^{M} c_i . N(\mu_i, \Sigma_i), \qquad (۱۵)$$

که در آن c_i وزن مخلوط و μ_i و Σ_i به ترتیب بردار میانگین و ماتریس کوواریانس توزیع نرمال هستند. ماتریس کوواریانس مدل GMM معمولاً به صورت قطری در نظر گرفته می‌شود، گرچه امکان استفاده از ماتریس کامل نیز وجود دارد. رابطه (۱۵) را می توان با استفاده از فرمول تابع چگالی احتمال نرمال به صورت زیر نیز بیان نمود:

$$P(x \mid GMM) = \sum_{i=1}^{M} c_i \frac{1}{(2\pi)^{\frac{d}{2}} \left|\sum_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i - \mu_i)^T \sum_i (x_i - \mu_i)\right) \qquad (۱۶)$$

که در آن d بعد فضای ورودیها است. برای بدست آوردن پارامترهای مدل GMM، شامل وزن مخلوطهای گاوسی و میانگین وکواریانس توزیعها، از الگوریتم ماکزیمم نمودن امید ریاضی(EM)[11] استفاده می شود. باید توجه داشت که تعداد مخلوطهای گاوسی با تعداد نمونه های موجود آموزشی رابطه مستقیم دارند و نمی توان با مجموعه داده ای ناچیز یک مدل GMM دارای تعداد بیش از حد از مخلوطها را آموزش داد. در تشکیل و آموزش مدل GMM مانند تمام روشهای تشکیل مدل رعایت نسبت میزان پیچیدگی مدل و نمونه های آموزشی الزامی می باشد.

[بازگشت به فهرست]

۹- فرضیات تئوری مدل مخفی مارکوف

برای اینکه مدل مخفی مارکوف از لحاظ ریاضی و محاسباتی قابل بیان باشد فرضهای زیر در مورد آن در نظر گرفته می شود.

۱- فرض مارکوف

با داشتن یک مدل مخفی مارکوف، احتمال انتقال از حالت i به حالت j به صورت زیر تعریف می شود:

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}.$$

[11] Expectation Maximization(EM)

به بیان دیگر فرض می شود که حالت بعدی تنها به حالت فعلی بستگی دارد. مدل حاصل از فرض مارکوف یک مدل HMM مرتبه صفر می باشد.

در حالت کلی، حالت بعدی می تواند با k حالت قبلی وابسته باشد. این مدل که مدل HMM مرتبه k ام گفته می شود، با استفاده از احتمالات انتقال به صورت زیر تعریف می گردد.

$$a_{i_1 i_2 \ldots i_k j} = p\{q_{t+1} = j | q_t = i_1, q_{t-1} = i_2, \ldots, q_{t-k+1} = i_k\}, \quad 1 \leq i_1, i_2, \ldots, i_k, j \leq N.$$

به نظر می رسد که یک مدل HMM از مرتبه بالاتر باعث افزایش پیچیدگی مدل می شود. علی رغم اینکه مدل HMM مرتبه اول متداول ترین مدل است، برخی تلاشها برای استفاده از مدلهای دارای مرتبه بالاتر نیز در حال انجام می باشد.

[بازگشت به فهرست]

۲- فرض ایستایی(stationarity)

در اینجا فرض می شود که احتمال انتقال در بین حالات از زمان واقعی رخداد انتقال مستقل است. در این صورت می توان برا ی هر t_1 و t_2 نوشت:

$$p\{q_{t_1+1} = j | q_{t_1} = i\} = p\{q_{t_2+1} = j | q_{t_2} = i\},$$

[بازگشت به فهرست]

۲- فرض استقلال خروجی

در این حالت فرض می شود که خروجی (مشاهدات) فعلی به صورت آماری از خروجی قبلی مستقل است. می توان این فرض را با داشتن دنباله ای از خروجی ها مانند بیان نمود:

$$O = o_1, o_2, \ldots, o_T$$

آنگاه مطابق با این فرض برای مدل HMM با نام λ خواهیم داشت:

$$p\{O | q_1, q_2, \ldots, q_T, \lambda\} = \prod_{t=1}^{T} p(o_t | q_t, \lambda).$$

اگر چه بر خلاف دو فرض دیگر این فرض اعتبار کمتری دارد. در برخی حالات این فرضیه چندان معتبر نیست و موجب می شود که مدل HMM با ضعفهای عمده ای مواجه گردد.

[بازگشت به فهرست]

۱۰- مساله ارزیابی و الگوریتم پیشرو (forward)

در این حالت مساله این است که با داشتن مدل $\lambda = (A, B, \pi)$ و دنباله مشاهدات $O = \{O_1, O_2, ..., O_T\}$ باید مقدار $P(O \mid \lambda)$ را پیدا نماییم. می توانیم این مقدار را با روشهای آماری مبتنی بر پارامترها محاسبه نماییم. البته این کار به محاسباتی با پیچیدگی $O(N^T)$ احتیاج دارد. این تعداد محاسبات حتی برای مقادیر متوسط t نیز بسیار بزرگ است. به همین دلیل لازم است که راه دیگری برای این محاسبات پیدا نماییم. خوشبختانه روشی ارائه شده است که پیچیدگی محاسباتی کمی دارد و از متغیر کمکی $\alpha_t(i)$ با نام متغیر پیشرو استفاده می کند.

متغیر پیشرو به صورت یک احتمال از دنباله مشاهدات $O = \{O_1, O_2, ..., O_t\}$ تعریف می شود که در حالت i خاتمه می یابد. به بیان ریاضی:

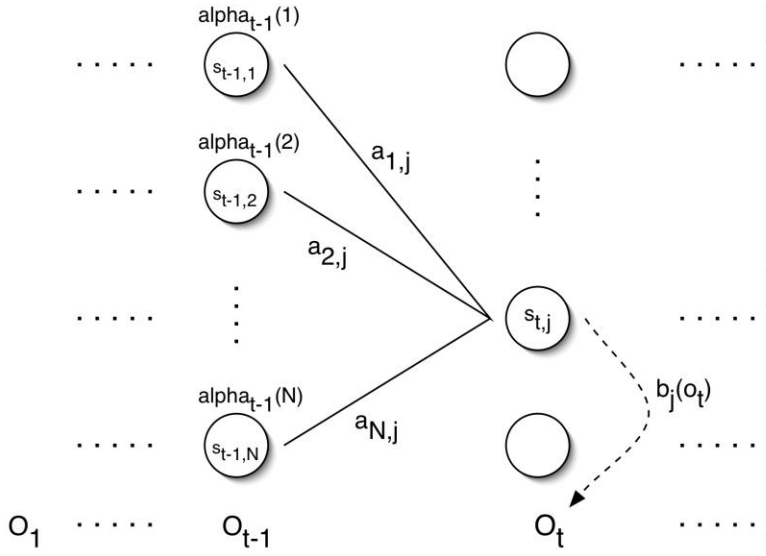
$$\alpha_t(i) = p\{o_1, o_2, \ldots, o_t, q_t = i|\lambda\} \qquad (۱۷)$$

آنگاه به سادگی مشاهده می شود که رابطه بازگشتی زیر برقرار است.

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}, \quad 1 \le j \le N, \quad 1 \le t \le T-1 \qquad (۱۸)$$

که در آن

$$\alpha_1(j) = \pi_j b_j(o_1), \quad 1 \le j \le N$$



شکل ۴: احتمالات پیشرو

با داشتن این رابطه بازگشتی می توانیم مقدار زیر را محاسبه نماییم.

$$\alpha_T(i),\ 1 \le i \le N$$

و آنگاه احتمال $P(O \mid \lambda)$ به صورت زیر محاسبه خواهد شد:

$$p\{O|\lambda\} = \sum_{i=1}^{N} \alpha_T(i). \qquad (۱۹)$$

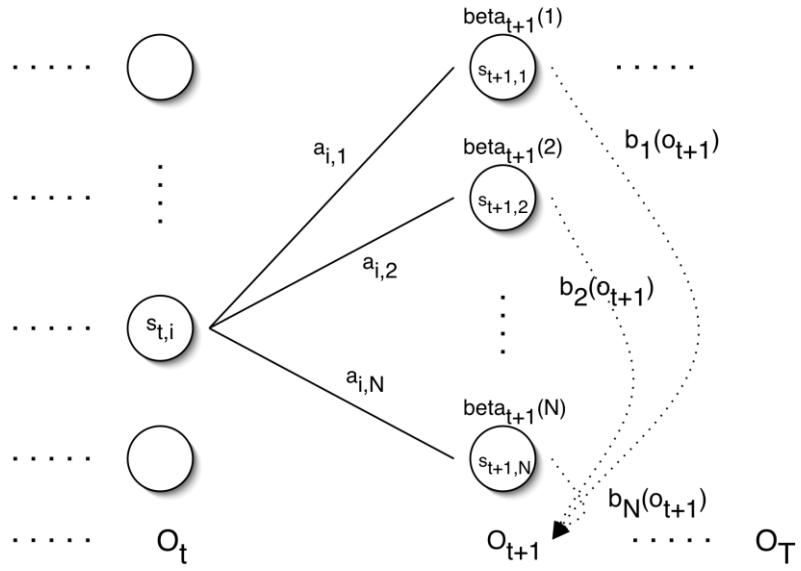
پیچیدگی محاسباتی روش فوق که به الگوریتم پیشرو معروف است برابر با $O(N^2T)$ است، که در مقایسه با حالت محاسبه مستقیم که قبلا گفته شد، و دارای پیچیدگی نمایی بود، بسیار سریعتر است.

روشی مشابه روش فوق را می توان با تعیین متغیر پسرو، $\beta_t(i)$، به عنوان احتمال جزئی دنباله مشاهدات $O = \{O_{t+1}, O_{t+2}, \ldots, O_T\}$ در حالت i تعریف نمود. متغیر پیشرو را می توان به شکل زیر نمایش داد.

$$\beta_t(i) = p\{o_{t+1}, o_{t+2}, \ldots, o_T | q_t = i, \lambda\} \qquad (۲۰)$$

مانند روش پیشرو یک رابطه بازگشتی به شکل زیر برای محاسبه $\beta_t(i)$ وجود دارد.

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}),\ \ 1 \le i \le N,\ \ 1 \le t \le T-1 \qquad (۲۱)$$



شکل ۵: احتمالات پسرو

که در آن

$$\beta_T(i) = 1,\ \ 1 \le i \le N$$

می توان ثابت کرد که

$$\alpha_t(i)\beta_t(i) = p\{\boldsymbol{O}, q_t = i|\lambda\},\ \ 1 \le i \le N,\ \ 1 \le t \le T \qquad (۲۲)$$

آنگاه می توان با کمک هر دو روش پیشرو و پسرو مقدار احتمال $P(O \mid \lambda)$ را محاسبه نمود.

$$p\{\boldsymbol{O}|\lambda\} = \sum_{i=1}^{N} p\{\boldsymbol{O}, q_t = i|\lambda\} = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad (۲۳)$$

رابطه فوق بسیار مهم و مفید است و بخصوص برای استخراج روابط آموزش مبتنی بر گرادیان لازم می باشد.

[بازگشت به فهرست]

۱۱- مساله کد گشایی و الگوریتم ویتربی (Viterbi Algorithm)

در این حالت می خواهیم با داشتن دنباله مشاهدات $O=\{O_1,...,O_t\}$ و مدل $\lambda=\{A,B,\pi\}$ دنباله حالات بهینه $Q=\{q_1,...,q_t\}$ برای تولید $O=\{O_1,...,O_t\}$ را بدست آوریم.

یک راه حل این است که محتمل ترین حالت در لحظه t را بدست آوریم و تمام حالات را به این شکل برای دنباله ورودی بدست آوریم. اما برخی مواقع این روش به ما یک دنباله معتبر و با معنا از حالات را نمی دهد به همین دلیل باید راهی پیدا نمود که یک چنین مشکلی نداشته باشد.

در یکی از این روشها که با نام الگوریتم Viterbi شناخته می شود، دنباله حالات کامل با بیشترین مقدار نسبت شباهت پیدا می شود. در این روش برای ساده کردن محاسبات متغیر کمکی زیر را تعریف می نماییم.

$$\delta_t(i)=\max_{q_1 q_2\cdots q_{t-1}} p\{q_1,q_2,\ldots,q_{t-1},q_t=i,o_1,o_2,\ldots,o_{t-1}|\lambda\},$$

که در شرایطی که حالت فعلی برابر با i باشد، بیشترین مقدار احتمال برای دنباله حالات و دنباله مشاهدات در زمان t را می دهد. به همین ترتیب می توان روابط بازگشتی زیر را نیز بدست آورد.

$$\delta_{t+1}(j)=b_j(o_{t+1})\left[\max_{1\le i\le N}\delta_t(i)a_{ij}\right],\ 1\le i\le N,\ 1\le t\le T-1 \qquad (۲۴)$$

که در آن

$$\delta_1(j)=\pi_j b_j(o_1),\ 1\le j\le N$$

به همین دلیل روال پیدا کردن دنباله حالات با بیشترین احتمال از محاسبه مقدار $\delta_j(i), i\le j\le N$ و با کمک رابطه فوق شروع می شود. در این روش در هر زمان یک اشاره گر به حالت برنده قبلی خواهیم داشت. در نهایت حالت j^* را با داشتن شرط زیر بدست می آوریم.

$$j^*=\arg\max_{1\le j\le N}\delta_T(j),$$

و با شروع از حالت j^*، دنباله حالات به شکل بازگشت به عقب و با دنبال کردن اشاره گر به حالات قبلی بدست می آید. با استفاده از این روش می توان مجموعه حالات مورد نظر را بدست آورد. این الگوریتم را می توان به صورت یک جستجو در گراف که نودهای آن برابر با حالتها مدل HMM در هر لحظه از زمان می باشند نیز تفسیر نمود.

[بازگشت به فهرست]

۱۲- مساله یادگیری

به طور کلی مساله یادگیری به این موضوع می پردازد که چگونه می توان پارامترهای مدل HMM را تخمین زد تا مجموعه داده های آموزشی به بهترین نحو به کمک مدل HMM برای یک کاربرد مشخص بازنمایی شوند. به همین دلیل می توان نتیجه گرفت که میزان بهینه بودن مدل HMM برای کاربردهای مختلف، متفاوت است. به بیان دیگر می توان از چندین معیار بهینه سازی متفاوت استفاده نمود، که از این بین یکی برای کاربرد مورد نظر مناسب تر است. دو معیار بهینه سازی مختلف برای آموزش مدل HMM وجود دارد که شامل معیار بیشترین شباهت (ML) و معیار ماکزیمم اطلاعات متقابل (Maximum Mutual Information (MMI)) می باشند. آموزش به کمک هر یک از این معیارها در ادامه توضیح داده شده است.

[بازگشت به فهرست]

۱۲-۱- معیار بیشترین شباهت (Maximum Likelihood (ML))

در معیار ML ما سعی داریم که احتمال یک دنباله ورودی O^w که به کلاس w تعلق دارد را با داشتن مدل HMM همان کلاس بدست آوریم. این میزان احتمال برابر با نسبت شباهت کلی دنباله مشاهدات است و به صورت زیر محاسبه می شود.

$$L_{tot} = p\{\boldsymbol{O}^{\mathbf{w}} | \lambda_w\}$$

با توجه به رابطه فوق در حالت کلی معیار ML به صورت زیر تعریف می شود.

$$L_{tot} = p\{\boldsymbol{O} | \lambda\} \qquad (۲۵)$$

اگر چه هیچ راه حل تحلیلی مناسبی برای مدل $\lambda = \{A, B, \pi\}$ وجود ندارد که مقدار L_{tot} را ماکزیمم نماید، لیکن می توانیم با استفاده از یک روال بازگشتی پارامترهای مدل را به شکلی انتخاب کنیم که مقدار ماکزیمم بدست آید. روش Baum-Welch و یا روش مبتنی بر گرادیان از جمله این روشها هستند.

[بازگشت به فهرست]

۱۲-۱-۱- الگوریتم بام- ولش

این روش را می توان به سادگی و با محاسبه احتمال رخداد پارامترها و یا با محاسبه حداکثر رابطه زیر بر روی $\bar{\lambda}$ تعریف نمود.

$$Q(\lambda, \bar{\lambda}) = \sum_{q} p\{q|O, \lambda\} \log[p\{O, q, \bar{\lambda}\}]$$

یکی از ویژگیهای مخصوص این الگوریتم این است که همگرایی در آن تضمین شده است. برای توصیف این الگوریتم که به الگوریتم پیشرو- پسرو نیز معروف است، باید علاوه بر متغیرهای کمکی پیشرو و پسرو که قبلا تعریف شده اند، متغیرهای کمکی بیشتری تعریف شود. البته می توان این متغیرها را در قالب متغیرهای پیشرو و پسرو نیز تعریف نمود.

اولین متغیر از این دست احتمال بودن در حالت i در زمان t و در حالت j در زمان t+1 است، که بصورت زیر تعریف می شود.

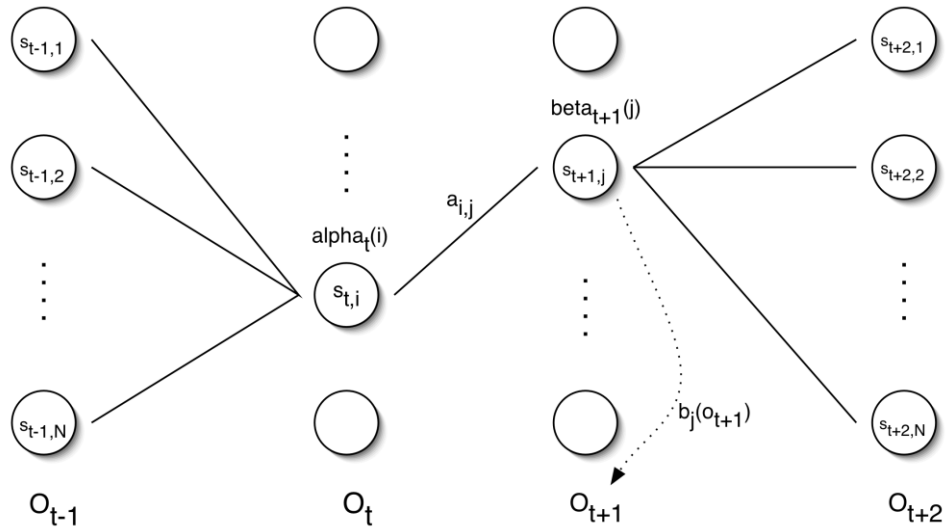
$$\xi_t(i, j) = p\{q_t = i, q_{t+1} = j|O, \lambda\} \qquad (۲۶)$$

این تعریف با تعریف زیر معادل است.

$$\xi_t(i, j) = \frac{p\{q_t = i, q_{t+1} = j, O|\lambda\}}{p\{O|\lambda\}} \qquad (۲۷)$$

می توان این متغیر را با استفاده از متغیرهای پیشرو و پسرو به صورت زیر تعریف نمود.

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})} \qquad (۲۸)$$



شکل ۶: بازتخمین احتمالات انتقال

متغیر دوم بیانگر احتمال پسین حالت i با داشتن دنباله مشاهدات و مدل مخفی مارکوف می باشد و به صورت زیر بیان می شود.

$$\gamma_t(i) = p\{q_t = i|\boldsymbol{O}, \boldsymbol{\lambda}\} \qquad (۲۹)$$

این متغیر را نیز می توان در قالب متغیرهای پیشرو و پسرو تعریف نمود.

$$\gamma_t(i) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}\right] \qquad (۳۰)$$

رابطه بین دو متغیر فوق بصورت زیر بیان می شود.

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j),\ 1 \le i \le N,\ 1 \le t \le M \qquad (۳۱)$$

اکنون می توان الگوریتم آموزش بام – ولش را با ماکزیمم کردن مقدار $P(O|\lambda)$ بدست آورد. اگر مدل اولیه ما $\lambda = \{A, B, \pi\}$ باشد، می توانیم متغیرهای پسرو و پیشرو را به استفاده از روابط (۱۸) و (۲۱) و متغیرهای ξ و γ را با

استفاده از روابط (۲۸) و (۲۹) تعریف نمود. مرحله بعدی این است که پارامترهای مدل را با توجه به روابط بازتخمین زیر بروزرسانی نماییم.

$$\bar{\pi}_i = \gamma_1(i), \quad 1 \le i \le N \qquad (۳۲)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \le i \le N, \quad 1 \le j \le N \qquad (۳۳)$$

$$\bar{b}_j(k) = \frac{\sum_{\substack{t=1 \\ O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \quad 1 \le j \le N, \quad 1 \le k \le M \qquad (۳۴)$$

فرمولهای باز تخمین را می توان به راحتی به شکلی تغییر داد که با توابع چگالی پیوسته نیز قابل استفاده باشند.

[بازگشت به فهرست]

۱۲-۱-۲- الگوریتم حداکثر سازی امید ریاضی (Expectation Maximization)

الگوریتم حداکثر سازی امید ریاضی یا EM به عنوان یک نمونه از الگوریتم بام – ولش در آموزش مدلهای HMM مورد استفاده قرار می گیرد. الگوریتم EM دارای دو فاز تحت عنوان Expectation و Maximization است. مراحل آموزش مدل در الگوریتم EM بصورت زیر است.

۱) مرحله مقدار دهی اولیه: پارامترهای اولیه مدل λ را تعیین می نماییم.

۲) مرحله امید ریاضی (Expectation): برای مدل λ موارد زیر را محاسبه می کنیم.

- مقادیر α با استفاده از الگوریتم پیشرو
- مقادیر β و γ با استفاده از الگوریتم پسرو

۳) مرحله ماکزیمم سازی (Maximization): مدل λ را با استفاده از الگوریتم باز تخمین محاسبه می نماییم.

۴) مرحله بروز رسانی $\lambda \leftarrow \lambda'$

۵) بازگشت به مرحله امید ریاضی

روال فوق تا زمانی که میزان نسبت شباهت نسبت به مرحله قبل بهبود مناسبی داشته باشد ادامه می یابد.

[بازگشت به فهرست]

۱۲-۱-۳- روش مبتنی بر گرادیان

در روش مبتنی بر گرادیان هر پارامتر Θ از مدل λ با توجه به رابطه زیر تغییر داده می شود.

$$\Theta^{new} = \Theta^{old} - \eta \left[\frac{\partial J}{\partial \Theta}\right]_{\Theta=\Theta_{old}} \quad (۳۵)$$

که در آن مقدار J با ید مینیمم شود. در این حالت خواهیم داشت.

$$J = E_{ML} = -\log(p\{\boldsymbol{O}|\lambda\}) = -\log(L_{tot}) \quad (۳۶)$$

از آنجا که مینیمم کردن J معادل است با مینیمم کردن L_{tot}، نیاز است است تا معیار ML بهینه بدست آید. آنگاه مساله، یافتن مقدار مشتق $\frac{\partial J}{\partial \Theta}$ برای تمام پارامترهای Θ از مدل است. این کار را می توان به سادگی با استفاده از مقدار L_{tot} انجام داد. به عنوان که گام کلیدی، با استفاده از روابطه (۲۳) و (۲۵) خواهیم داشت.

$$L_{tot} = \sum_{i=1}^{N} p\{\boldsymbol{O}, q_t = i|\lambda\} = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \quad (۳۷)$$

با مشتق گرفتن از رابطه (۳۶) بر حسب پارامتر Θ خواهیم داشت.

$$\frac{\partial J}{\partial \Theta} = -\frac{1}{L_{tot}} \frac{\partial L_{tot}}{\partial \Theta} \quad (۳۸)$$

از آنجا که در رابطه فوق مقدار $\frac{\partial J}{\partial \Theta}$ بر حسب $\frac{\partial L_{tot}}{\partial \Theta}$ بدست می آید، می توان $\frac{\partial J}{\partial \Theta}$ را به کمک رابطه (۳۷) بدست آورد. در روش مبتنی بر گرادیان، مقدار $\frac{\partial L_{tot}}{\partial \Theta}$ را باید برای پارامترهای a_{ij} (احتمال انتقال) و b_{ij} (احتمال مشاهدات) بدست آورد.

[بازگشت به فهرست]

۱۲-۱-۴- محاسبه گرادیان برحسب پارامترهای احتمال حالات

با استفاده از قانون زنجیره ای داریم:

$$\frac{\partial L_{tot}}{\partial a_{ij}} = \sum_{t=1}^{T} \frac{\partial L_{tot}}{\partial \alpha_t(j)} \frac{\partial \alpha_t(j)}{\partial a_{ij}} \qquad (۳۹)$$

با مشتقگیری از رابطه (۳۷) بر حسب $\alpha_t(j)$ خواهیم داشت:

$$\frac{\partial L_{tot}}{\partial \alpha_t(j)} = \beta_t(j), \qquad (۴۰)$$

همچنین با مشتقگیری از رابطه (۱۸) برحسب a_{ij} داریم:

$$\frac{\partial \alpha_t(j)}{\partial a_{ij}} = b_j(o_t)\alpha_{t-1}(i) \qquad (۴۱)$$

روابط (۳۹)، (۴۰) و (۴۱) مقدار $\frac{\partial L_{tot}}{\partial a_{ij}}$ را نتیجه می دهیم و با جایگزینی آن در رابطه (۳۸) ما به نتیجه مورد نظر دست می یابیم.

$$\frac{\partial J}{\partial a_{ij}} = -\frac{1}{L_{tot}} \sum_{t=1}^{T} \beta_t(j)\, b_j(o_t)\, \alpha_{t-1}(i) \qquad (۴۲)$$

[بازگشت به فهرست]

۱۲-۱-۵- محاسبه گرادیان برحسب پارامترهای احتمال حالات

با استفاده از قانون زنجیره ای داریم:

$$\frac{\partial L_{tot}}{\partial b_j(o_t)} = \frac{\partial L_{tot}}{\partial \alpha_t(j)} \frac{\partial \alpha_t(j)}{\partial b_j(o_t)} \quad (۴۳)$$

با مشتقگیری از رابطه (۱۸) بر حسب $b_j(O_t)$ خواهیم داشت:

$$\frac{\partial \alpha_t(j)}{\partial b_j(o_t)} = \frac{\alpha_t(j)}{b_j(o_t)} \quad (۴۴)$$

آنگاه می توان با استفاده از روابط (۴۴)، (۴۰) و (۴۳) مقدار $\frac{\partial L_{tot}}{\partial b_j(o_t)}$ را بدست آورد.

$$\frac{\partial J}{\partial b_j(o_t)} = -\frac{1}{L_{tot}} \frac{\alpha_t(j)\beta_t(j)}{b_j(o_t)}, \quad (۴۵)$$

رابطه فوق را می توان به صورت زیر نیز بیان نمود.

$$\frac{\partial J}{\partial b_j(o_t)} = -\frac{\gamma_t(j)}{b_j(o_t)}, \quad (۴۶)$$

در نهایت می توان مقدار احتمال مورد نظر را با جایگزینی $\frac{\partial L_{tot}}{\partial b_j(o_t)}$ در رابطه (۳۸) بدست آورد. اگر برای احتمال مشاهدات توابع چگالی پیوسته مورد استفاده قرار گرفته باشند، می توان مقادیر $\frac{\partial J}{\partial \Theta}$ را برای پارامترهای این توابع چگالی بدست آورد.

[بازگشت به فهرست]

۱۲-۲- معیار ماکزیمم اطلاعات متقابل

در معیار ML ما مدل HMM را تنها برای یک کلاس در هر لحظه بروز رسانی می نمودیم و به مدلهای دیگر توجه نمی کردیم. به همین دلیل این روش مفهوم تمایز را که در شناسایی الگو بسیار مورد توجه است در نظر نمی گیرد. به همین دلیل روش یادگیری ML، به خصوص هنگامی که نمونه های فاز آموزش با نمونه های ورودی در فاز شناسایی متفاوت هستند، تمایزات ضعیفی بین مدلها ایجاد می کند. اینگونه ناهمخوانی ها به دو دلیل ایجاد می شوند. اول اینکه داده های آموزشی و آزمایشی ویژگیهای آماری متفاوتی دارند و دوم اینکه در مرحله آموزش نمی توان پارامترهای مدل را به شکل قابل اطمینانی تخمین زد.

در مقابل معیار یادگیری MMI در هر لحظه تمام مدلهای مربوط به کلاسها را مورد آموزش قرار می دهد. در این حالت پارامترهای مدل اصلی برای بازنمایی مناسب داده ها بروز رسانی می شوند در حالی که پارامترهای سایر مدلها به شکلی تغییر می کنند که به میزان کمتری نمونه های آموزشی را بازنمایی نمایند. این روال باعث می شود که تمایز بین مدلها افزایش یابد و به همین دلیل است که روش یادگیری MMI به گروه روشهای یادگیری تمایزی تعلق دارد. حال فرض کنید که یک مجموعه از مدلهای HMM را در اختیار داریم:

$$\Lambda = \{\lambda_\nu,\ 1 \le \nu \le V\}.$$

مساله این است که عدم قطعیت شرطی کلاس ν را با داشتن دنباله مشاهدات O^* حداقل نماییم. این مساله معادل است با کم کردن اطلاعات شرطی کلاس ν بر حسب Λ و به صورت زیر بیان می شود:

$$I(\nu|\boldsymbol{O}^*,\Lambda) = -\log p\{\nu|\boldsymbol{O}^*,\Lambda\} \qquad (۴۷)$$

در چهارچوب تئوری اطلاعات مساله فوق مانند حداقل کردن آنتروپی شرطی است.

$$H(\mathcal{V}|\boldsymbol{O}) = E[I(\nu|\boldsymbol{O}^*)] \qquad (۴۸)$$

که در آن $\mathcal{V}$ بیانگر مجموعه تمام کلاسها است و O دنباله مشاهدات را نشان می دهد. آنگاه اطلاعات متقابل بین کلاسها و دنباله مشاهدات

$$H(\mathcal{V},\boldsymbol{O}) = H(\mathcal{V}) - H(\mathcal{V}|\boldsymbol{O}) \qquad (۴۹)$$

باید حداکثر شود، $H(\nu)$ ثابت است. به همین دلیل است که این روش را روش حداکثر سازی اطلاعات متقابل گویند. این روش با نام حداکثر سازی احتمال پسین یا (MAP) نیز شناخته می شود زیرا در رابطه (۴۷) مقدار احتمال $p\{\nu \mid O^*,\Lambda\}$ باید حداکثر شود. رابطه (۴۷) را می توان به کمک تئوری بیز به شکل زیر نیز بیان نمود.

$$
\begin{aligned}
E_{MMI} &= -\log p\{\nu|\hat{\boldsymbol{O}},\Lambda\} \\
&= -\log \frac{p\{\nu,\hat{\boldsymbol{O}}|\Lambda\}}{p\{\hat{\boldsymbol{O}}|\Lambda\}} \\
&= -\log \frac{p\{\nu,\hat{\boldsymbol{O}}|\Lambda\}}{\sum_{w} p\{w,\hat{\boldsymbol{O}}|\Lambda\}}
\end{aligned}
\qquad (۵۰)
$$

که در آن w یکی از کلاسهای آموزشی را نشان می دهد. رابطه (۵۰) را می توان به کمک روابط زیر نیز بیان نمود.

$$
L_{tot}^{clamped} = p\{\nu,\hat{\boldsymbol{O}}|\lambda\} \qquad (۵۱)
$$

$$
L_{tot}^{free} = \sum_{w} p\{w,\hat{\boldsymbol{O}}|\lambda\} \qquad (۵۲)
$$

$$
E_{MMI} = -log\frac{L_{tot}^{clamped}}{L_{tot}^{free}} \qquad (۵۳)
$$

در این حالت می توان مقدار E_{MMI} را با استفاده از روش مبتنی بر گرادیان یا روش بازتخمین ML حداقل نمود. برای مثال روش مبتنی بر گرادیان را می توان به صورت زیر تعریف نمود. این روش با تلاش برای مینیمم کردن رابطه زیر شروع می شود.

$$
J = E_{MMI},
$$

آنگاه می توان مساله فوق را به صورت مساله محاسبه $\frac{\partial J}{\partial \Theta}$ که در آن Θ یکی از پارامترهای مدلهای Λ است، ساده نمود.

$$
\frac{\partial J}{\partial \Theta} = \frac{1}{L_{tot}^{free}}\frac{\partial L_{tot}^{free}}{\partial \Theta} - \frac{1}{L_{tot}^{clamped}}\frac{\partial L_{tot}^{clamped}}{\partial \Theta} \qquad (۵۴)
$$

تکنیک مشابهی مانند آنچه در یادگیری ML استفاده شد می تواند در اینجا نیز مورد استفاده قرار گیرد. در قدم اول رابطه های (۵۲) و (۵۳) را با کمک متغیرهای پیشرو و پسرو به صورت زیر و با کمک رابطه (۲۳) می نویسیم.

$$
L_{tot}^{clamped} = \sum_{i \in \text{class } \nu} \alpha_t(i)\beta_t(i) \qquad (۵۵)
$$

$$L_{tot}^{free} = \sum_{w} \sum_{i \in \text{class w}} \alpha_t(i)\beta_t(i) \qquad (۵۶)$$

آنگاه مقدار گرادیان مورد نظر از تفاضل دو مشتق (۵۵) و (۵۶) حاصل می شود. مانند روش ML این محاسبات را باید برای بدست آوردن هر دو پارامتر احتمال انتقالات و احتمال مشاهدات انجام داد.

[بازگشت به فهرست]

۱۲-۲-۱- محاسبه گرادیان برحسب احتمالات انتقال

با استفاده از قانون زنجیره ای داریم:

$$\frac{\partial L_{tot}^{(\cdot)}}{\partial a_{ij}} = \sum_{t=1}^{T} \frac{\partial L_{tot}^{(\cdot)}}{\partial \alpha_t(j)} \frac{\partial \alpha_t(j)}{\partial a_{ij}} \qquad (۵۷)$$

با مشتقگیری از روابط (۵۵) و (۵۶) بر حسب $\alpha_t(j)$ و استفاده از نتایج رابطه (۴۱) می توانیم سمت راست رابطه (۵۷) را بدست آوریم. این جایگزینی دو نتیجه جداگانه را برای حالتهای clamped و free خواهد داشت.

$$\frac{\partial L_{tot}^{clamped}}{\partial a_{ij}} = \delta_{kv} \sum_{t=1}^{T} \beta_t(j)\, b_j(o_t)\, \alpha_{t-1}(i), \qquad i \in \text{class k} \qquad (۵۸)$$

در رابطه فوق مقدار δ_{kv} دلتای کرونیکر (Kronecker) است.

$$\frac{\partial L_{tot}^{free}}{\partial a_{ij}} = \sum_{t=1}^{T} \beta_t(j)\, b_j(o_t)\, \alpha_{t-1}(i) \qquad (۵۹)$$

با جایگزینی روابط (۵۸) و (۵۹) در رابطه (۵۴)، با توجه به اینکه $\Theta = a_{ij}$ ، خواهیم داشت:

$$\frac{\partial J}{\partial a_{ij}} = \left[\frac{1}{L_{tot}^{free}} - \frac{\delta_{kv}}{L_{tot}^{clamped}}\right] \sum_{t=1}^{T} \beta_t(j)\, b_j(o_t)\, \alpha_{t-1}(i),$$

$$i \in \text{class k} \qquad (۶۰)$$

[بازگشت به فهرست]

۱۲-۲-۲- گرادیان برحسب احتمالات مشاهدات

با استفاده از قانون زنجیره ای داریم:

$$\frac{\partial L_{tot}^{(\cdot)}}{\partial b_j(o_t} = \frac{\partial L_{tot}^{(\cdot)}}{\partial \alpha_t(j)} \frac{\partial \alpha_t(j)}{\partial b_j(o_t} \qquad (۶۱)$$

با مشتقگیری از روابط (۵۵) و (۵۶) بر حسب $b_j(o_t)$ و استفاده از نتایج رابطه (۴۴) می توانیم سمت راست رابطه (۶۱) را بدست آوریم. این جایگزینی دو نتیجه جداگانه را برای حالتهای clamped و free خواهد داشت.

$$\frac{\partial L_{tot}^{clamped}}{\partial b_j(o_t)} = \delta_{kv} \frac{\alpha_t(j)\,\beta_t(j)}{b_j(o_t)},$$

$$j \in \text{class k} \qquad (۶۲)$$

در رابطه فوق مقدار δ_{kv} دلتای کرونیکر(Kronecker) است.

$$\frac{\partial L_{tot}^{free}}{\partial b_j(o_t)} = \frac{\alpha_t(j)\,\beta_t(j)}{b_j(o_t)} \qquad (۶۳)$$

با جایگزینی روابط (۶۲) و (۶۳) در رابطه (۵۴)، خواهیم داشت:

$$\frac{\partial J}{\partial b_j(o_t)} = \left[\frac{1}{L_{tot}^{free}} - \frac{\delta_{kv}}{L_{tot}^{clamped}}\right] \frac{\alpha_t(j)\,\beta_t(j)}{b_j(o_t)},$$

$$j \in \text{class k} \qquad (۶۴)$$

روابطه فوق را می توان به شکل خوش فرم زیر نیز تبدیل نمود:

$$\gamma_t(j)^{clamped} = \delta_{kv}\frac{\alpha_t(j)\,\beta_t(j)}{L_{tot}^{clamped}}, \quad j \in \text{class k} \qquad (۶۵)$$

$$\gamma_t(j)^{free} = \frac{\alpha_t(j)\,\beta_t(j)}{L_{tot}^{clamped}}. \qquad (۶۶)$$

با استفاده از تعاریف فوق می توان رابطه (۶۴) را بصورت زیر نیز بیان نمود.

$$\frac{\partial J}{\partial b_j(o_t)} = \frac{1}{b_j(o_t)}\left[\gamma_t(j)^{free} - \gamma_t(j)^{clamped}\right] \qquad (۶۷)$$

رابطه زیر به صورت کامل چگونگی بروز رسانی احتمال مشاهدات را نشان می دهد. اگر به جای توابع چگالی احتمال گسسته، توابع چگالی احتمال پیوسته مورد استفاده قرار گیرند، می توان با استفاده از قانون زنجیره ای این مشتقها را در مدل HMM منتشر نمود.

[بازگشت به فهرست]

۱۳- استفاده از مدل HMM در شناسایی گفتار

بحث شناسایی اتوماتیک گفتار را می توان از دو جنبه مورد بررسی قرار داد.

۱- از جنبه تولید گفتار
۲- از جنبه فهم و دریافت گفتار

مدل مخفی مارکوف (HMM) تلاشی است برای مدل سازی آماری دستگاه تولید گفتار و به همین دلیل به اولین دسته از روشهای شناسایی گفتار تعلق دارد. در طول چندین سال گذشته این روش به عنوان موفقترین روش در شناسایی گفتار مورد استفاده قرار گرفته است. دلیل اصلی این مساله این است که مدل HMM قادر است به شکل بسیار خوبی خصوصیات سیگنال گفتار را در یک قالب ریاضی قابل فهم تعریف نماید.

در یک سیستم ASR مبتنی بر HMM قبل از آموزش HMM یک مرحله استخراج ویژگیها انجام می گردد. به این ترتیب ورودی HMM یک دنباله گسسته از پارامترهای برداری است. بردارهای ویژگی می تواند به یکی از دو طریق بردارهای چندی سازی شده یا مقادیر پیوسته به مدل HMM آموزش داده شوند. می توان مدل HMM را به گونه ای

طراحی نمود که هر یک از این انواع ورودیها را دریافت نماید. مساله مهم این است که مدل HMM چگونه با طبیعت تصادفی مقادیر بردار ویژگی سازگاری پیدا خواهد کرد.

[بازگشت به فهرست]

۱۴- استفاده از HMM در شناسایی کلمات جداگانه

در حالت کلی شناسایی واحدهای گفتاری جدا از هم به کاربردی اطلاق می شود که در آن یک کلمه، یک زیر کلمه یا دنباله ای از کلمات به صورت جداگانه و به تنهایی شناسایی شود. باید توجه داشت که این تعریف با مساله شناسایی گفتار گسسته که در آن گفتار به صورت گسسته بیان می شود متفاوت است. در این بین شناسایی کلمات جداگانه کاربرد بیشتری به نسبت دو مورد دیگر دارد و دو مورد دیگر بیشتر در عرصه مطالعات تئوری مورد بررسی قرار می گیرند.

برای این کاربرد راه حلهای مختلفی وجود دارد زیرا معیارهای بهینه سازی متفاوتی را برای این منظور معرفی شده است و الگوریتمهای پیاده سازی شده مختلفی نیز برای هر معیار موجود است. از بین این روشها روش مبتنی بر گرادیان و معیار بهینه سازی MMI به شکل مختصر توضیح داده شده اند. این مساله را از دو جنبه آموزش و شناسایی مورد بررسی قرار می دهیم.

[بازگشت به فهرست]

۱۴-۱- آموزش

فرض می کنیم که فاز پیش پردازش سیستم دنباله مشاهدات زیر را تولید نماید:

$$\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N\}.$$

پارامترهای اولیه تمام مدلهای HMM را با یک مجموعه از مقادیر مشخص مقدار دهی می نماییم.

$$\lambda_i, \ 1 \leq i \leq N$$

این پارامترها را می توان با استفاده از رابطه (۳۵) و در حالی که گرادیان مورد نظر از روابط (۶۰) و (۶۴) بدست می آیند، بروز رسانی نمود. البته برای این کاربرد خاص (شناسایی جداگانه) مقادیر نسبت شباهت در دو رابطه آخر به شکل خاصی محاسبه می شوند.

در آغاز این مساله را برای حالت *clamped* در نظر بگیرید. از آنجایی که ما برای هر کلاس از واحدها یک HMM داریم، می توانیم مدل λ_l از کلاس l را که دنباله مشاهدات فعلی به آن مربوط می شود، را انتخاب نماییم. آنگاه با کمک رابطه (۵۵) خواهیم داشت.

$$L_{tot}^{clamped} = L_l^l = \sum_{i \in \lambda_l} \alpha_t(i)\beta_t(i) = \sum_{i \in \lambda_l} \alpha_T(i) \quad (۶۸)$$

که در رابطه فوق، خط دوم از رابطه (۱۹) نتیجه شده است. برای حالت free نیز به مانند حالت قبل می توان مقدار نسبت شباهت را به کمک رابطه (۵۵) بدست آورد.

$$L_{tot}^{free} = \sum_{m=1}^{N} L_m^l = \sum_{m=1}^{N} \left[\sum_{i \in \lambda_m} \alpha_t(i)\beta_t(i) \right] = \sum_{m=1}^{N} \sum_{i \in \lambda_m} \alpha_T(i) \quad (۶۹)$$

که در آن L_m^l بیانگر میزان شباهت دنباله مشاهدات فعلی به کلاس l در مدل λ_m است. با کمک مقادیر نسبت شباهت (۶۸) و (۶۹) مقدار گرادیان به شکل زیر خواهد بود.

$$\frac{\partial J}{\partial a_{ij}} = \left[\frac{1}{\sum_{m=1}^{N} L_m^l} - \frac{\delta_{kl}}{L_l^l} \right] \sum_{t=1}^{T} \beta_t(j)\, b_j(o_t)\, \alpha_{t-1}(i), \quad i, j \in \lambda_k \quad (۷۰)$$

$$\frac{\partial J}{\partial b_j(o_t)} = \left[\frac{1}{\sum_{m=1}^{N} L_m^l} - \frac{\delta_{kl}}{L_l^l} \right] \frac{\alpha_t(j)\,\beta_t(j)}{b_j(o_t)}, \quad j \in \lambda_k \quad (۷۱)$$

با توجه به موارد فوق، روال آموزش با استفاده از روش مبتنی بر گرادیان و معیار MMI را می توان به شکل زیر خلاصه نمود.

۱) هر یک از مدلهای HMM، $\lambda_i = (A_i, B_i, \pi_i), 1 \le i \le N$، را با یکی از دو روش تصادفی یا خوشه بندی K-means مقداردهی اولیه می کنیم.

۲) با داشتن دنباله مشاهدات

- مقادیر احتمال پیشرو و پسرو هر HMM را با استفاده از روابط (۲۱) و (۱۸) محاسبه می کنیم.
- مقدار نسبت شباهت را با استفاده از روابط (۶۸) و (۶۹) محاسبه می کنیم.
- بااستفاده از روابط (۷۰) و (۷۱) مقدار گرادیان را بر حسب پارامترهای هر مدل محاسبه می نماییم.
- با کمک رابطه (۳۵) پارامترهای مدل را بروز رسانی می کنیم.

۳)اگر همه نمونه های آموزشی استفاده نشده اند به گام ۲ بر می گردیم.

۴) مراحل ۲ و ۳ را تا رسیدن به شرط همگرایی ادامه می دهیم.

این مراحل را می توان به سادگی برای مدلهای HMM دارای چگالی پیوسته نیز تغییر داد. این کار با انتشار گرادیان با استفاده از قانون زنجیره ای انجام می شود.

[بازگشت به فهرست]

۱۴-۲- شناسایی

در مقایسه با آموزش، روال شناسایی بسیار ساده تر است.

۱) الگوریتم دنباله مشاهدات مورد نظر را دریافت می کند.

- مقادیر احتمالات پیشرو و پسرو را برای هر یک از مدلها و با استفاده از روابط (۲۱) و (۱۸) محاسبه می کنیم.
- با استفاده از رابطه (۶۹) مقدار احتمال $L_m^l, 1 \le m \le N$ را برای تمام مدلها محاسبه می کنیم.
- سپس کلاس دنباله مشاهدات ورودی با استفاده از رابطه زیر تعیین می شود.

$$l^* = \arg \max_{1 \le m \le N} L_m^l.$$

در این حالت نرخ شناسایی بصورت نسبت بین واحدهای شناسایی صحیح به کل واحدهای آموزشی حساب می شود.

[بازگشت به فهرست]

۱۵- استفاده از مدل HMM در شناسایی گفتار پیوسته

در حالت شناسایی کلمات جداگانه ما برای هر واحد از یک HMM استفاده نمودیم. در حالت شناسایی گفتار پیوسته این کار قابل انجام نیست زیرا در این حالت باید مجوعه ای جملات پیوسته شناسایی شوند که اگر مدلها بر حسب جملات باشد با تعداد بسیار زیادی مدل احتیاج داریم. علاوه بر این دو مشکل اساسی دیگر نیز وجود دارد.

۱) ما از نقطه پایانی واحدهای گفتاری در هر جمله اطلاعی نداریم.
۲) نمی دانیم که چه تعداد واحد گفتاری در هر جمله وجود دارد.

به دلیل همین دو مشکل شناسایی گفتار پیوسته از شناسایی واحدهای گفتاری جداگانه پیچیده تر است. البته HMM شرایط مناسبی را برای شناسایی گفتار پیوسته بوجود می آورد. در این حالت ما مدلهای HMM هر واحد گفتاری را به هم متصل می کنیم تا مدل HMM یک جمله را تشکیل دهد. اگر اتصال مدلهای HMM به هم و ترتیب آنها از هیچ محدودیتی برخوردار نباشد، اصطلاحا گفته می شود که شناسایی مستقل از گرامر (Grammar free) انجام می شود. در مقابل می توان توالیهای معتبر مدلها را به کمک یک مدل زبانی که گرامر زبان را تعیین می کند، مشخص نمود. در این حالت نیز به مانند شناسایی واحدهای گفتاری چگونگی انجام دو مرحله آموزش و آزمایش باید تعیین گردد.

[بازگشت به فهرست]

۱۵-۱- آموزش مدلهای HMM برای کاربرد شناسایی گفتار پیوسته

می توان یک شناسایی کننده گفتار پیوسته را به هر یک از روشهای ML و یا MMI آموزش داد.

۱۵-۱-۱- آموزش ML

اگر از تکنیک مبتنی بر گرادیان استفاده نماییم روال آموزش به صورت زیر خواهد بود.

۱) هر یک از مدلهای HMM، $\lambda_i = (A_i, B_i, \pi_i), 1 \le i \le N$، را با یکی از دو روش تصادفی یا خوشه بندی K-means مقداردهی اولیه می کنیم.

۲) دنباله مشاهدات ورودی را برای هر جمله آموزشی دریافت می کنیم.

- مدل متناظر با جمله ورودی را با استفاده از مدل HMM واحدهای گفتاری جمله ورودی تشکیل می دهیم.
- مقادیر احتمال پیشرو و پسرو هر HMM را با استفاده از روابط (۲۱) و (۱۸) محاسبه می کنیم.
- مقدار نسبت شباهت را با استفاده از رابطه (۳۷) برای مدل جمله آموزشی محاسبه می کنیم.
- بااستفاده از روابط (۴۲) و (۴۵) مقدار گرادیان را بر حسب پارامترهای مدل محاسبه می نماییم.
- با کمک رابطه (۳۵) پارامترهای مدل را بروز رسانی می کنیم.

۳)اگر همه نمونه های آموزشی استفاده نشده اند به گام ۲ بر می گردیم.

۴) مراحل ۲ و ۳ را تا رسیدن به شرط همگرایی ادامه می دهیم.

[بازگشت به فهرست]

۱۵-۱-۲- آموزش MMI

برای آموزش بر حسب معیار MMI به فرمهای مناسب روابط (۵۵) و (۵۶) نیاز داریم. در آغاز حالت *clamped* را در نظر بگیرید. از آنجا که ما شناسایی را در سطح جمله انجام می دهیم، یک مدل HMM مانند Λ_t را با اتصال مدلهای HMM واحدهای گفتاری تشکیل می دهیم که با دنباله آموزشی ورودی متناظر است. آنگاه با داشتن رابطه (۵۵) خواهیم داشت.

$$L_{tot}^{clamped} = \sum_{i \in \Lambda_t} \alpha_t(i)\beta_t(i) = \sum_{i \in \Lambda_t} \alpha_T(i) \quad (۷۲)$$

خط دوم رابطه فوق از رابطه (۱۹) نتیجه شده است.

در حالت *free* ، ما تنها یک مدل HMM مانند $\mathbf{\Lambda}$ داریم که تمام زبان را بازنمایی می نماید. آنگاه رابطه (۵۶) به صورت زیر تغییر می کند.

$$L_{tot}^{free} = = \sum_{\forall i \in \Lambda} \alpha_t(i)\beta_t(i) = \sum_{\forall i \in \Lambda} \alpha_T(i) \quad (۷۳)$$

با توجه به روابط فوق روال آموزش MMI به صورت زیر خواهد بود:

۱) هر یک از مدلهای HMM، $\lambda_i = (A_i, B_i, \pi_i), 1 \le i \le N$، را با یکی از دو روش تصادفی یا خوشه بندی K-means مقداردهی اولیه می کنیم.

۲) دنباله مشاهدات ورودی را برای هر جمله آموزشی دریافت می کنیم.

- مدل متناظر با جمله ورودی را با استفاده از مدل HMM واحدهای گفتاری جمله ورودی تشکیل می دهیم.
- مقادیر احتمال پیشرو و پسرو هر HMM را با استفاده از روابط (۲۱) و (۱۸) محاسبه می کنیم.

- مقدار نسبت شباهت دنباله مشاهدات به مدل Λ_l را با استفاده از روابط (۷۲) و (۷۳) برای مدل جمله آموزشی محاسبه می کنیم.
- بااستفاده از روابط (۶۴) و (۶۰) مقدار گرادیان را بر حسب پارامترهای مدل برای مدلهای واحدهای گفتاری محاسبه می نماییم. مقدار δ_{kv} در روابط فوق با مقدار زیر جایگزین می شود.

$$\delta_{il} = \begin{cases} 1 & i \in \Lambda_l \\ 0 & \text{otherwise} \end{cases}$$

- با کمک رابطه (۳۵) پارامترهای مدل را بروز رسانی می کنیم.

۳)اگر همه نمونه های آموزشی استفاده نشده اند به گام ۲ بر می گردیم.

۴) مراحل ۲ و ۳ را تا رسیدن به شرط همگرایی ادامه می دهیم.

[بازگشت به فهرست]

۱۵-۲- شناسایی با استفاده از شناسایی کننده گفتار پیوسته

با فرض اینکه مدل تمام واحدهای گفتار موجود باشند، عمل شناسایی گفتار پیوسته به معنای یافتن دنباله واحدهای گفتاری است که با دنباله مشاهدات O برای یک جمله نا مشخص مطابق است. این مساله را می توان به صورت زیر نیز بیان نمود:

$$\boldsymbol{w}^{\star} = \arg\max_{\boldsymbol{w}} p\{\boldsymbol{w}, \boldsymbol{O}|\Lambda\} \qquad (۷۴)$$

که در آن $w = \{w_1, w_2, ..., w_S\}$ یک دنباله از واحدهای گفتاری با طول S است. از آنجا که $p\{w\}$ توسط مدل زبانی تعیین می شود تنها کاری که باید برای تعیین w^* انجام گیرد، محاسبه $p\{O \mid w, A\}$ برای تمام انتخابهای ممکن w است. واضح است که این روال از لحاظ محاسباتی بسیار پیچیده است، زیرا حتی برای یک مجموعه کوچک از کلمات زبان تعداد جملات بسیار زیاد خواهد بود. یک راه حل ساده تر این است که بهترین دنباله حالات را در مدل زبانی بدست آوریم. برای این کار کافی ست رابطه زیر را محاسبه کنیم.

$$\boldsymbol{q}^{\star} = \arg\max_{\boldsymbol{q}} p\{\boldsymbol{q}, \boldsymbol{O}|\Lambda\}. \qquad (۷۵)$$

سپس می توان از طریق دنباله حالات دنباله واحدهای گفتاری را بدست آورد. برای محاسبه q^* می توان مستقیما از الگوریتم ویتربی استفاده نمود یا از روش دیگری تحت عنوان Level building Viterbi استفاده کرد. ازآنجا که

الگوریتم شناسایی ویتربی نیمه بهینه است، روشهای موثری برای محاسبه میزان نسبت شباهت جمله ارائه شده است که الگوریتم N-best از جمله این روشها می باشد. در ادامه به الگوریتم شناسایی ویتربی و الگوریتمهای بهینه سازی شده Level Building و N-best نیز خواهیم پرداخت.

[بازگشت به فهرست]

۱۵-۲-۱- شناسایی مبتنی بر الگوریتم ویتربی

در الگوریتم ویتربی، امتیاز ویتربی، $\delta_t(t)$، برای همه حالتهای مدل زبانی Λ در زمان t محاسبه می شود و آنگاه محاسبه امتیاز ویتربی در زمان t+1 با کمک رابطه (۳۴) ادامه می یابد. این روال از آن جهت که پردازش را در زمان t کاملا انجام می دهد و آنگاه به زمان t+1 می رود، تحت عنوان جستجوی ویتربی همگام با زمان (time synchronous Viterbi search) شناخته می شود. در نهایت برگشت به عقب در جهت حالتهای دارای بیشترین امتیاز، دنباله حالات بهینه را نتیجه می دهد. البته اگر تعداد حالات زیاد باشد جستجوی ویتربی بسیار پر هزینه خواهد بود. در این حالت تنها تعدادی از حالات که دارای بیشترین امتیاز هستند نگهداری می شوند و از نگهداری بقیه حالات صرف نظر می گردد. این روال جستجو با نام جستجوی پرتویی یا beam search نیز نامیده می شود.

[بازگشت به فهرست]

۱۵-۲-۲- الگوریتم ساخت سطح Level Building

در این الگوریتم بر خلاف الگوریتم جستجوی ویتربی، مدلهای HMM هر واحد گفتاری به صورت جداگانه مدنظر قرارمی گیرد. جستجو در سطوح مختلفی با استفاده از الگوریتم ویتربی انجام می شود، که در آن هر سطح با محل یک واحد گفتاری در جملات محتملتر متناظر است. بعد از جستجو در هر سطح، ما بیشترین مقدار امتیاز ویتربی را برای تمام مدلهای واحدهای گفتاری در هر فریم زمانی t بدست می آوریم. جستجو در سطوح بعدی با بیشتین امتیاز سطح قبلی در فریم زمانی قبلی آغاز می شود. بعد از انجام جستجوی در l سطح، دنباله مدلهای گفتاری برنده، گفتار شناسایی شده دارای بیشترین احتمال را تعیین می کند.

[بازگشت به فهرست]

۱۵-۲-۳- جستجوی N-best

الگوریتم جستجوی N-best بسیار به الگوریتم جستجوی همگام با زمان ویتربی شبیه است. از آنجا که هدف جستجوی N-best یافتن دنباله واحدهای بهینه به جای دنباله حالات بهینه است، به جای عملیات یافتن مقدار ماکزیمم باید یک عملیات جمع بندی بر روی دنباله حالات انجام شود. در این حالات علاوه بر عملیات هرسی که برای حذف حالات دارای کمترین امتیاز انجام می شود، یک عملیات هرس نیز برای یافتن بهترین N مسیر موجود با بیشترین امتیاز انجام می گردد. در پایان این اگوریتم N بهترین جمله را نتیجه می دهد.

[بازگشت به فهرست]

۱۶- برخی کاربردها

مدلهای مارکوف می توانند کاربردهای مختلفی در زمینه مدل سازی و یادگیری داشته باشند. چند نمونه از این کاربردها به قرار زیرند:

- شناسایی گفتار
- شناسایی کاراکترهای نوری
- ترجمه ماشینی
- بیوانفورماتیک و ژنشناسی

- K. F. Lee And H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. 31, No. 11, 1989.

در این مقاله مدل مخفی مارکوف برای یک کاربرد شناسایی واج مستقل از گوینده پیشنهاد شده است. از این مقاله از چندین کتاب کد از پارامترهای LPC و مدلهای مخفی مارکوف استفاده شده است. در نهایت این روش با توجه به ویژگیهای اکوستیکی دادگان گفتاری و همچنین مدل زبانی مورد استفاده راندمان شناسایی بین ۵۸.۸ تا ۷۳.۸ داشته است.

- T. Vinar, "Enhancements to Hidden Markov Models for Gene Finding and Other Biological Applications," Thesis presented to the University of Waterloo in fulfilment of requirement for the degree of Doctor of Philosophy in Computer Science, Waterloo, Ontario, Canada, 2005.

در این تز استفاده از مدل مخفی مارکوف برای مساله یافتن ژنها در ساختارهای DNA پیشنهاد شده است. یافتن ژنها یکی از قدمهای اصلی در مساله تحلیل مولکولهای DNA است. در این مطالعه یک روش برای افزایش قابلیتهای مدل مخفی مارکوف به منظور بهبود پارامترهای آماری مدل پیشنهاد شده است. در این پایان نامه اشاره شده است که استفاده از مدلهای HMM دارای ساختارهای پیچیده باعث کاهش دقت مدل می شود و برای حل آن تنها باید روشهای پیش بینی و آموزش مدل پیچیده تری را مورد استفاده قرار داد.

- D. O. Tanguay, "Hidden Markov Models for Gesture Recognition," Department of Electrical Engineering and Computer Science, In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science, August , 1995.

در این مقاله روشی برای درک و دریافت حرکات بدن انسان با استفاده از مدل مخفی مارکوف پیشنهاد شده است. این مساله از جمله ممترین مسائل در بینایی ماشین و همچنین طراحی سیستمهای دارای ارتباط متقابل با کاربر می باشد.

- H. Lee and A. Y. Ng, "Spam Deobfuscation using a Hidden Markov Model".

در این مقاله از مدل مخفی مارکوف برای شناسایی و رفع ابهام از هرزنامه ها استفاده شده است. آزمایشات انجام شده نشان می دهد که این روش نسبت به تمام انواع هرزنامه ها موثر است.

- S. M. Thede and M. P. Harper, "A Second-Order Hidden Markov Model for Part-of-Speech Tagging," School of Electrical and Computer Engineering, Purdue University.
- در این مقاله یک نوسعه بر مدل مخفی مارکوف با استفاده از تخمین مرتبه دوم برای برچسبگذاری جملات مورد استفاده قرار گرفته است. نتایج این مقاله حاکی از بهبود برچسبگذاری به نسبت برخی دیگر از روشهای ارائه شده در این زمینه می باشد.

- T. Yang and Y. Xu, "Hidden Markov Model for Gesture Recognition," The Robotics Institute Carnegie Mellon University, Pittsburgh, Pennsylvania, May 1994.
- در این مقاله روشی مبتنی بر مدل مخفی مارکوف به منظور توسعه یک سیستم مبتنی بر حرکات بدن ارائه شده است. به جای استفاده از ویژگیهای هندسی در این روش از دنباله ای از سمبلها استفاده شده است. برای آموزش این دنباله از سمبلهای ورودی از مدل مخفی مارکوف استفاده شده است. راندمان بدست آمده برای این سیستم برای تشخیص حرکات جداگانه ۹۹.۷۸ درصد بوده است و البته نتایج مناسبی نیز برای کاربردهای شناسایی حرکات پیوسته بدست آمده است.
- بازنمایی و استنتاج گرامر یک زبان ساده

C. S. Wallace & M. P. Georgeff. A General Objective for Inductive Inference. [TR32 (HTML)], March 1983, Department of Computer Science, Monash University, Australia.
Also:
M. P. Georgeff & C. S. Wallace. A General Selection Criterion for Inductive Inference. European Conference on Artificial Intelligence (ECAI, ECAI84), Pisa, pp473-482, September 1984.

- مدلسازی ساختار سریهای زمانی و یا سایر دنباله هایی از داده های چند متغیره

T. Edgoose, L. Allison & D. L. Dowe. An MML classification of protein structure that knows about angles and sequence. Pacific Symposium on Biocomputing '98, pp585-596, Jan. 1998 [paper]

T. Edgoose & L. Allison. Minimum message length hidden Markov modelling. IEEE Data Compression Conf., Snowbird UT, pp169-178, March 1998

T. Edgoose & L. Allison. MML Markov classification of sequential data. Stats. and Comp. 9(4) pp269-278, Sept. 1999

- مدل سازی ارتباط بین جفتهای DNA

L.Allison, C.S.Wallace and C.N.Yee. Inductive Inference over Macro-Molecules. [90/148 (HTML)], CSSE, Monash University, 1990

L.Allison, C.S.Wallace & C.N.Yee. Finite-State Models in the Alignment of Macro-Molecules. J.Molec.Evol. 35(1) pp77-89, 1992

L.Allison. Normalization of Affine Gap Costs Used in Optimal Sequence Alignment. J.Theor.Biol. 161 pp263-269, 1993 [www inc' pdf paper]

C. N. Yee & L. Allison. Reconstruction of Strings Past. CABIOS 9(1) pp1-7 (now J. Bioinformatics) 1993

- انطباق زوج دنباله های از مقادیر غیر تصادفی، مانند مقادیر مدل شده توسط یک مدل با حالات محدود یا یک مدل چپ به راست

D. R. Powell, L. Allison, T. I. Dix, D. L. Dowe. Alignment of low information sequences. Australian Comp. Sci. Theory Symp. (CATS98), Perth, pp215-230, Springer Verlag isbn:981-3083-92-1, Feb. 1998

L.Allison. Information-Theoretic Sequence Alignment. [98/14 (HTML)] CSSE Monash University, 1998

L. Allison, D. Powell & T. I. Dix. Compression and Approximate Matching, Computer Journal, 42(1), pp1-10, 1999

D. R. Powell, L. Allison, T. I. Dix. Modelling-Alignment for Non-Random Sequences, AI2004, Springer Verlag, LNCS/LNAI 3339, pp.203-214, Dec. 2004

- مدلسازی ارتباط بین چندین دنباله از مقادیر با استفاده از درختهای تکاملی

L. Allison and C. S. Wallace. The Posterior Probability Distribution of Alignments and its Application to Estimation of Evolutionary Trees and Optimisation of Multiple Alignments. Jrnl. Molec. Evol. 39 pp418-430, 1994

- کشف الگوهای ضعیف در دنباله های DNA

L. Allison, T. Edgoose, T. I. Dix. Compression of Strings with Approximate Repeats. Intelligent Systems in Molecular Biology ISMB98 pp8-16, Montreal, 28 June - 1 July 1998

L. Allison, L. Stern, T. Edgoose & T. I. Dix. Sequence Complexity for Biological Sequence Analysis. Computers and Chemistry 24(1), pp43-55, Jan. 2000

L. Stern, L. Allison, R. L. Coppel, T. I. Dix. Discovering patterns in Plasmodium falciparum genomic DNA. Molecular and Biochemical Parasitology, 118(2) pp175-186, 2001

[بازگشت به فهرست]

۱۷- برخی مراجع مفید در زمینه مدل مخفی مارکوف و ابزارهای موجود

- Hidden Markov Model (HMM) Toolbox for Matlab (by Kevin Murphy)
- Hidden Markov Model Toolkit (HTK) (a portable toolkit for building and manipulating hidden Markov models)
- Hidden Markov Models (an exposition using basic mathematics)
- GHMM Library (home page of the GHMM Library project)
- Jahmm Java Library (Java library and associated graphical application)
- A step-by-step tutorial on HMMs (University of Leeds)
- Software for Markov Models and Processes (TreeAge Software)
- Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77 (2), p. 257–286, February 1989.
- Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1999. ISBN 0521629713.
- Kristie Seymore, Andrew McCallum, and Roni Rosenfeld. Learning Hidden Markov Model Structure for Information Extraction. AAAI 99 Workshop on Machine Learning for Information Extraction, 1999. (also at CiteSeer: [1])
- http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html

- J. Li, A. Najmi, R. M. Gray, Image classification by a two dimensional hidden Markov model, IEEE Transactions on Signal Processing, 48(2):517-33, February 2000.

[بازگشت به فهرست]