УДК 519.21

ФИЛЬТРАЦИЯ СОСТОЯНИЙ ЧАСТИЧНО НАБЛЮДАЕМОЙ СТОХАСТИЧЕСКОЙ СЕТИ

К.В. Семенихин

Институт радиотехники и электроники им. В.А. Котельникова РАН

Россия, 125009, Москва, ул. Моховая 11, корп. 7

E-mail: siemenkv@mail.ru

Ключевые слова: частично наблюдаемая система, стохастическая сеть, точечный процесс, стохастическая фильтрация, система массового обслуживания.

Аннотация: Описано решение задачи фильтрации состояний частично наблюдаемой стохастической сети. Динамика сети с двумя типами узлов (наблюдаемыми и скрытыми) описывается набором точечных процессов, чьи интенсивности зависят от состояния узлов. Представлены уравнения оптимальной нелинейной фильтрации загрузки узлов вместе с алгоритмом, реализующим их аппроксимацию. Теоретические результаты и их алгоритмическая реализация применены к задаче оценивания числа абонентов колл-центра.

1. Введение

Мартингальная теория фильтрации получила множество применений к задачам оценивания, управления и оптимизации в стохастических системах, описываемых скачкообразными марковскими процессами [5, 6, 9–11]. Однако в области систем массового обслуживания (СМО) работ по применению теории фильтрации остается мало [2, 3, 8, 14, 16]. Тем не менее, проблемы восстановления скрытых состояний на основе частично наблюдаемой динамики образуют важный класс обратных задач в теории массового обслуживания [1]. В области управления перегрузками в сетях передачи данных уделяется значительное внимание задаче оценивания текущей пропускной способности (bandwidth estimation), трактуемой в виде задачи нелинейной фильтрации, для отслеживания изменений во входящем потоке данных на основе оперативных данных о заполнении буферов [7, 15].

В данной работе рассматривается стохастическая сеть типа Джексона [13] с наблюдаемыми и скрытыми узлами. Количество заявок в каждом скрытом узле должно отслеживаться по изменениям в загрузке наблюдаемых узлов. Вместо бесконечномерной дифференциальной системы относительно условных вероятностей используется метод замены меры [5] для вывода уравнений, описывающих условные математические ожидания и условные ковариации. Для практической реализации данного метода оценивания предлагается численная схема, основанная на упрощении уравнений оптимальной фильтрации. Эта схема апробирована на модели колл-центра, для которого требовалось оперативно отслеживать два скрытых состояния (число недозвонившихся и число неудовлетворенных абонентов) по наблюдениям за длиной очереди основной системы.

2. Описание модели и постановка задачи

Рассмотрим стохастическую сеть с множеством узлов $S = \{1, 2 \ldots, d\}$, каждый из которых принимает заявки от других узлов и извне. Фиктивный узел 0 используется как источник входящего трафика, так и финальный узел, означающий завершение обслуживания в сети. Динамика сети описывается процессом $X(t) = \{X_\alpha(t)\}_{\alpha \in S}$, определенном на вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$, где $X_\alpha(t)$ равно числу заявок на узле α в момент $t \geqslant 0$.

Возможны только три типа событий, когда заявка а) переходит с узла $\alpha \in S$ на другой узел $\beta \in S$, б) заканчивает обслуживание в сети на узле $\alpha \in S$, в) прибывает в сеть через узел $\beta \in S$. Эти потоки событий описываются точечными процессами $N_{\alpha,\beta}(t)$, $N_{\alpha,0}(t)$ и $N_{0,\beta}(t)$, которые имеют непрерывные справа траектории и представление $N_{\alpha,\beta}(t) = \int_0^t \nu_{\alpha,\beta}(s)\, ds + M_{\alpha,\beta}(t)$ через квадратично интегрируемые $\mathbf{F}$-мартингалы $M_{\alpha,\beta}$ и $\mathbf{F}$-предсказуемые интенсивности $\nu_{\alpha,\beta} \geqslant 0$, где $\mathbf{F} = \mathcal{F}_{t t \geqslant 0}$ – пополнение фильтрации, порожденной указанными точечными процессами и начальным состоянием сети $X(0)$. Тогда состояние узла $m \in S$ принимает вид $X_m(t) = X_m(0) + \sum_\alpha N_{\alpha,m}(t) - \sum_\beta N_{m,\beta}(t)$, где α, β пробегают $S \cup \{0\}$.

Частично наблюдаемая стохастическая сеть задается разбиением $S = J \sqcup H$ на непосредственно наблюдаемые узлы $j \in J$ и узлы со скрытой динамикой $k \in H$. Узел 0 также считается ненаблюдаемым, т.е. нет точной информации о моментах прихода заявок извне или их ухода из сети.

Введем фильтрацию $\mathbf{Y} = \{\mathcal{Y}_t\}_{t \geqslant 0}$, порожденную процессом $Y(t) = \{X_i(t)\}_{i \in J}$, определяющим загрузку наблюдаемых узлов, и начальным состоянием всей сети $X(0)$. *Задача оптимальной фильтрации* состояний частичной наблюдаемой стохастической сети состоит в том, чтобы определить условное математическое ожидание $\hat{Z}(t) = \mathsf{E}\{Z(t) \,|\, \mathcal{Y}_t\}$ загрузки скрытой части сети $Z(t) = \{X_k(t)\}_{k \in H}$ по наблюдениям, доступным к текущему моменту времени t. Поскольку эта задача решается без нахождения апостериорного распределения, для характеристики точности оценки понадобится условная ковариация $Q(t) = \mathsf{cov}\{Z(t), Z(t) \,|\, \mathcal{Y}_t\}$.

3. Оценка состояния скрытых узлов

Оценки оптимальной фильтрации состояний скрытых узлов $\hat{X}_k(t) = \mathsf{E}\{X_k(t) \,|\, \mathcal{Y}_t\}$, а также условные ковариации $Q_{k,l}(t) = \mathsf{E}\{\varepsilon_k(t)\varepsilon_l(t) \,|\, \mathcal{Y}_t\}$, $k, l \in H$ ошибок оценивания $\varepsilon_k(t) = X_k(t) - \hat{X}_k(t)$, будучи согласованы с фильтрацией $\mathbf{Y}$, выражаются только через начальное состояние сети $X(0)$ и три типа точечных процессов

$$N_{i,j}, \quad N_j^a = \sum_{k \notin J} N_{k,j} \quad \text{и} \quad N_i^d = \sum_{k \notin J} N_{i,k} \quad (i, j \in J).$$

Процессы $\{N_{i,j}\}$ описывают переходы внутри наблюдаемой части сети J, а N_j^a и N_i^d определяют, соответственно, число поступлений на узел $j \in J$ с любого ненаблюдаемого $k \notin J$ и число уходов с узла $i \in J$ на любой ненаблюдаемый $k \notin J$. Соответствующие интенсивности принимают форму $\nu_j^a = \sum_{k \notin J} \nu_{k,j}$ и $\nu_i^d = \sum_{k \notin J} \nu_{i,k}$.

Сделаем несколько предположений о рассматриваемой модели сети:

(i) скачки точечных процессов $N_{\alpha,\beta}$ не происходят одновременно (иначе говоря, моменты всех событий в сети различны);

(ii) для $i \in J$, $\beta \in S \cup \{0\}$ интенсивности $\nu_{i\beta}$ $\mathbf{Y}$-предсказуемы, т.е. имеется прямая информация о скорости, с которой заявки уходят с наблюдаемых узлов;

(iii) $\exists\, C = \text{const}\colon\ \sum_{\alpha,\beta} \nu_{\alpha,\beta} \leqslant C \sum_{\alpha,\beta} N_{\alpha,\beta}$, что означает не более чем линейный рост интенсивностей относительно числа событий в сети.

Теорема. *В условиях* (i)–(iii) *справедливы следующие утверждения*:

1) *для узла $k \in H$ оценка его состояния $\hat{X}_k(t)$ описывается системой уравнений*

$$d\hat{X}_k = \Big\{ \hat{\nu}^a_k - \hat{\nu}^d_k - \sum_{j \in J} \hat{c}_{k,j} \Big\}\, dt + \sum_{i \in J} \xi^d_{i,k}\, dN^d_i + \sum_{j \in J} \xi^a_{k,j}\, dN^a_j, \tag{1}$$

$$\hat{\nu}^a_k = \sum_{m \notin J} \hat{\nu}_{m,k}, \qquad \hat{\nu}^d_k = \sum_{m \notin J} \hat{\nu}_{k,m}, \qquad \xi^d_{i,k} = \frac{\nu_{i,k}}{\nu^d_i}, \qquad \xi^a_{k,j} = \frac{\hat{c}_{k,j} - \hat{\nu}_{k,j}}{\hat{\nu}^a_j}, \tag{2}$$

$$\hat{c}_{k,j} = \mathsf{cov}\{X_k(t-), \nu^a_j \mid \mathcal{Y}_{t-}\} = \widehat{X_k \nu^a_j} - \hat{X}(t-)\hat{\nu}^a_j, \tag{3}$$

2) *для узла $k \in H$ условная дисперсия ошибки $Q_{k,k}(t)$ удовлетворяет уравнению*

$$\begin{aligned} dQ_{k,k} = \Big(\hat{\nu}^a_k + \hat{\nu}^d_k + 2\hat{b}_{k,k} - \sum_{j \in J} \hat{\tau}_{k,k,j} \Big)\, dt + \sum_{i \in J} \big(1 - \xi^d_{i,k}\big) \xi^d_{i,k}\, dN^d_i + \\ + \sum_{j \in J} \Big\{ \frac{1}{\hat{\nu}^a_j} \big(\hat{\tau}_{k,k,j} + \hat{\nu}_{k,j} - 2\hat{\varkappa}_{k,k,j} \big) - \big(\xi^a_{k,j}\big)^2 \Big\}\, dN^a_j, \end{aligned} \tag{4}$$

3) *для пары узлов $k \neq l$ условная ковариация ошибки $Q_{k,l}(t)$ имеет вид*

$$\begin{aligned} dQ_{k,l} = \big(\hat{b}_{k,l} + \hat{b}_{l,k} - \hat{\nu}_{k,l} - \hat{\nu}_{l,k} - \sum_{j \in J} \hat{\tau}_{k,l,j} \big)\, dt - \sum_{i \in J} \xi^d_{i,k} \xi^d_{i,l}\, dN^d_i + \\ + \sum_{j \in J} \Big\{ \frac{1}{\hat{\nu}^a_j} \Big(\hat{\tau}_{k,l,j} - \hat{\varkappa}_{k,l,j} - \hat{\varkappa}_{l,k,j} \Big) - \xi^a_{k,j} \xi^a_{l,j} \Big\}\, dN^a_j, \end{aligned} \tag{5}$$

где использованы $\mathbf{Y}$-предсказуемые версии соответствующих условных ковариаций:

$$\begin{aligned} &\hat{\varkappa}_{k,l,j} = \mathsf{cov}\{X_k(t-),\, \nu_{l,j} \,|\, \mathcal{Y}_{t-}\}, \quad \hat{\tau}_{k,l,j} = \mathsf{cov}\{\varepsilon_k(t-)\varepsilon_l(t-),\, \nu^a_j \,|\, \mathcal{Y}_{t-}\}, \\ &\hat{b}_{k,l} = \mathsf{cov}\{X_k(t-),\, \nu^a_l - \nu^d_l \,|\, \mathcal{Y}_{t-}\} = \sum_{m \notin J} \big(\hat{\varkappa}_{k,m,l} - \hat{\varkappa}_{k,l,m} \big). \end{aligned}$$

Доказательство теоремы опубликовано в [12].

Поясним структуру оценки (1). Отношения $\nu_{i,k}/\nu^d_i$ и $\hat{\nu}_{k,j}/\hat{\nu}^a_j$ определяют, соответственно, долю переходов с наблюдаемых узлов на данный скрытый узел k и переходов с узла k в наблюдаемую часть сети. Единственное отличие между этими двумя членами – дополнительное корректирующее слагаемое $\hat{c}_{k,j}$, которое добавляется к скачкам dN^a_j и вычитается из соответствующего им сноса.

Для получения конечномерной реализации фильтра рассмотрим сеть, в которой интенсивность переходов со скрытых узлов линейно зависит от их загрузки, т.е. $\nu_{k,\beta} = \mu_{k,\beta} X_k(t-)$ для всех $k \in H$ и $\nu_{0,\beta} = \lambda_\beta$, где $\mu_{k,\beta}, \lambda_\beta$ – $\mathbf{Y}$=предсказуемые коэффициенты. Тогда получается, что $\hat{c}_{k,j} = \sum_{m \in H} Q_{k,m}(t-)\mu_{m,j}$ и $\hat{\nu}_{k,\beta} = \mu_{k,\beta}\hat{X}_k(t-)$. Другие коэффициенты (2), (3) также выражаются через сами оценки $\{\hat{X}_k\}_{k \in H}$ и условные ковариации их ошибок $\{Q_{k,l}\}_{k,l \in H}$, в частности, $\hat{\varkappa}_{k,l,\beta} = Q_{k,l}(t-)\mu_{l,\beta}$.

Для упрощения уравнений (4)–(5) исключим слагаемые, содержащие условные моменты третьего порядка, так как они порождают центрированный мартингал $dM^{\tau}_{k,l} = \sum_{j\in J} \hat{\tau}_{k,l,j}\big(dN^a_j/\hat{\nu}^a_j - dt\big)$. Это упрощение имеет смысл операции проекции.

Теперь система (1)–(5) определяет конечномерный фильтр (он далее называется *субоптимальным* и обозначается кратко SF). Между скачками процессов N^d_i, N^a_j фильтр SF описывается системой линейных дифференциальных уравнений:

$$\dot{\hat{Z}} = \Lambda^{\top}\hat{Z} + \lambda - Q\gamma,$$
$$\dot{Q} = (Q - \mathrm{diag}[\hat{Z}])\Lambda + \Lambda^{\top}(Q - \mathrm{diag}[\hat{Z}]) + \mathrm{diag}[\Lambda^{\top}\hat{Z} + \lambda].$$

Здесь $\hat{Z}$ – вектор-столбец оценок, а Q – матрица условных ковариаций ошибок, причем $\lambda = \{\lambda_k\}_{k\in H}$ и $\gamma = \{\gamma_k\}_{k\in H}$ – вектор-столбцы, а $\Lambda = \{\Lambda_{k,l}\}_{k,l\in H}$ – квадратная матрица, где $\gamma_k = \sum_{j\in J}\mu_{k,j}$ и $\Lambda_{k,l} = \mu_{k,l} - \delta_{k,l}\sum_{m\notin J}\mu_{k,m}$.

4. Оценивание числа абонентов колл-центра

Рассмотрим модель колл-центра в виде частично наблюдаемой сети, состоящей из трех узлов (см. рис. 1). Наблюдаемый узел 1 – многоканальная СМО, моделирующая процесс обработки простейшего потока вызовов интенсивности λ с помощью m операторов с экспоненциальным временем обслуживания $\mathcal{E}(\mu_1)$. Максимальное число удерживаемых абонентов равно K. Скрытые узлы 2 и 3 включают абонентов, не получивших обслуживания из-за занятости всех операторов и потому делающих повторный вызов через время $\mathcal{E}(\mu_2)$ и, соответственно, абонентов, оставшихся неудовлетворенными и желающих получить дополнительную информацию через время $\mathcal{E}(\mu_3)$. Вероятность повторного вызова известна и равна $r_{1,3}$. Начальное состояние сети предполагается нулевым.

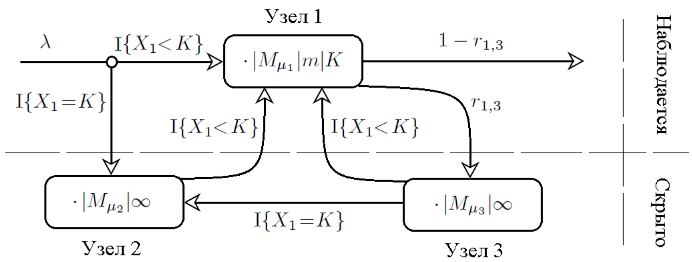

Рис. 1. Модель колл-центра как частично наблюдаемой стохастической сети

Для сравнительного анализа качества субоптимального фильтра SF реализованы два алгоритма: TF – усеченный фильтр, получаемый из уравнений (1) путем обнуления корректирующего члена $\hat{c}_{k,j}$; DF – фильтр, использующий только снос $\nu_{\alpha,\beta}\,dt$ в уравнениях точечных процессов $dN_{\alpha,\beta}$. Фильтр SF описывается пятью уравнениями (включая условные ковариации), а фильтры TF и DF – только двумя.

На рис. 2 приведены траектории состояний узлов и их субоптимальные оценки на двух промежутках времени. По результатам обработки 1000 траекторий с.к. ошибка (RMSE) субоптимального фильтра SF числа неудовлетворенных абонентов оказалась ниже, чем для альтернативных схем на 10–15%. То же соотношение ошибок сохранилось при оценивании числа необслуженных абонентов фильтром DF, но для фильтра TF с.к. ошибка оказалась в 2,5 раза больше, чем для субоптимального.

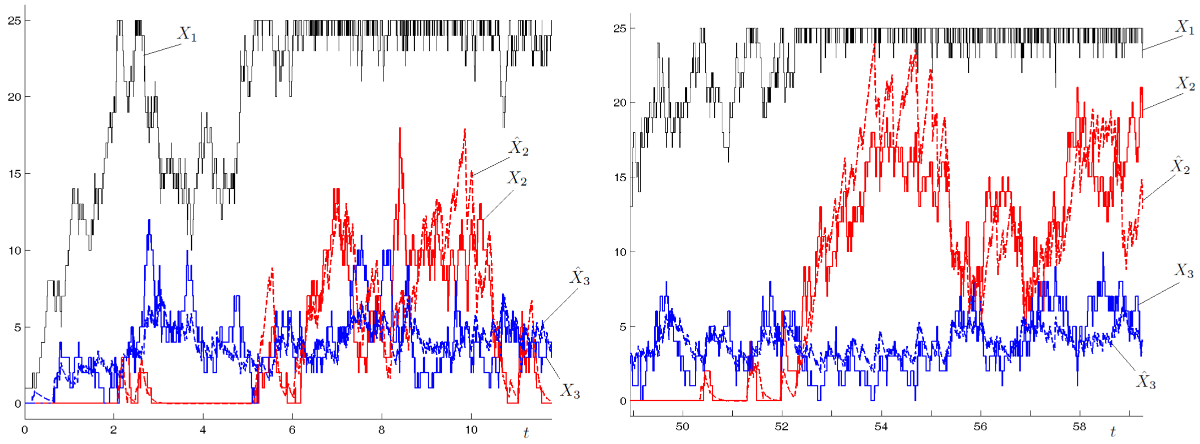

Рис. 2. Траектории состояний узлов (сплошные линии) и их оценки SF (штриховые линии); цвет: черный – узел 1, красный – узел 2, синий – узел 3

Список литературы

1. Baccelli F., Kauffmann B., Veitch D. Inverse problems in queueing theory and Internet probing // Queueing Systems. 2009. Vol. 63. P. 59–107.
2. Bensoussan A., Cakanyildirim M., Sethi S.P., Shi R. An incomplete information inventory model with presence of inventories or backorders as only observations // J. Optimization Theory Appl. 2010. Vol. 146, No. 3. P. 544–580.
3. Borisov A.V. Application of optimal filtering methods for on-line of queueing network states // Automation and Remote Control. 2016. Vol. 77. P. 277–296.
4. Bremaud P. On the output theorem of queueing theory, via filtering // Journal of Applied Probability. 1978. Vol. 15, No. 2. P. 397–405.
5. Elliott R.J., Aggoun L., Moore J.B. Hidden Markov models. Estimation and control. New York: Springer, 2008.
6. Elliott R.J., Dufour F., Malcolm W.P. State and mode estimation for discrete-time jump Markov systems // SIAM J. Contr. Optimization. 2005. Vol. 44, No. 3. P. 1081–1104.
7. Li X., Yousefi'zadeh H. Robust EKF-based wireless congestion control // IEEE Trans. Communications. 2013. Vol. C-61, No. 12. P. 5090–5102.
8. Lukashuk L.I., Semenchenko Y.A. Filtering of a semi-Markov queueing system with retrials // Cybernetics and Systems Analysis. 1991. Vol. 27, No. 4. P. 627–631.
9. Miller B.M., Avrachenkov K.E., Stepanyan K.V., Miller G.B. Flow control as a stochastic optimal control problem with incomplete information // Problems of Information Transmission. 2005. Vol. 41, No. 2. P. 150–170.
10. Miller B.M., Miller G.B., Semenikhin K.V. Optimal channel choice for lossy data flow transmission // Automation and Remote Control. 2018. Vol. 79. P. 66–77.
11. Rieder U., Winter J. Optimal control of Markovian jump processes with partial information and applications to a parallel queueing model // Math. Meth. Oper. Res. 2009. Vol. 70. P. 567–596.
12. Semenikhin K.V. State Estimation in Partially Observed Stochastic Networks with Queueing Applications // Piunovskiy A., Zhang Y. (eds) Modern Trends in Controlled Stochastic Processes: Emergence, Complexity and Computation, vol. 41. Cham: Springer, 2021.
13. Serfozo R. Introduction to Stochastic Networks. New York: Springer, 1999.
14. Solodyannikov Yu.V. Control and Observation for Dynamical Queueing Networks. I// Automation and Remote Control. 2014. Vol. 75. P. 422–446.
15. Stuckey N., Vasquez J., Graham S., Maybeck P. Stochastic control of computer networks // IET Control Theory and Applications. 2012. Vol. 6, No. 3. P. 403–411.
16. Walrand J., Varaiya P. Flows in queueing networks: A martingale approach // Mathematics of Operations Research. 1981. Vol. 6, No. 3. P. 387–404.
17. Wong E., Hajek B. Stochastic processes in engineering systems. New York: Springer, 1985.