УДК 519.216.22

Д.А. Бражник

ОЦЕНКА ИНФОРМАЦИИ В ЗАДАЧЕ СБОРА ИНФОРМАЦИИ

Введение. Теория информации является мощным средством, позволяющим проектировать и оценивать системы передачи и обработки информации [1]. Однако в последнее время проявилось противоречие между потребностями практики и применимостью теории, так как теория информации, прекрасно описывающая канал связи оказаласьмалопригодной для описания функционирования информационно управляющих систем [2]. Данное положение объясняется тем, что удачная интерпретация формул Хартли и Шеннона была принята как аксиома [3], хотя в основе вывода меры Хартли лежит слово «предположим» [4]. То есть мера Хартли справедлива настолько и там, где «предположим» соответствуют действительности.

Состояние вопроса. На сегодняшний день существует множество различных походов к определению количества информации, содержащейся в сообщении, генерируемой источником или передаваемой каналом [1]. Однако есть существенная особенность, объединяющая эти подходы – все эти соотношения называются мерами. Следует отметить, что в последних работах появляется стремление строго определить методы измерения информации. Так в работе [1] вполне резонно информация определяется как мера связи, в работе [5] вводится норма информационного пространства. При более тщательном анализе проблемы измерения информации оказалось, что предварительно необходимо ввести множество, операции над элементами этого множества и правила получения результата – то есть необходимо конкретизировать пространство, как множество, наделённое структурой, и только после этого рассматривать правила измерения величины элементов этого множества и правила измерения расстояния между элементами. Собственно, имея пространство можно определить его норму и метрику. И только имея норму и метрику можно сочинить меру, построение меры просто из эвристических соображений нецелесообразно.

Постановка задачи исследования Существенный интерес представляет задача сбора информации от множества слабо информированных и несвязанных между собой источников, собственно, – это ситуация всевозможных опросов или анализа открытых источников в прессе. Учитывая, что данная задача встречается достаточно часто, а использование меры Хартли и Шеннона предполагает доверие к истинности сообщений, что в данном случае нецелесообразно, определим норму и метрику в информационном пространстве [6].

Рассматривая множество событий Ω, каждое из которых генерируется отдельным независимым источником и имеет определенную вероятность P, следуя [6] и предполагая аналитичность зависимости вероятности от информации, запишем в окрестности реализованного событияпервые члены разложения вероятности события в ряд Тейлора. Предполагаем, согласно [6], что вероятность события зависит от содержащейся в событии информации

$$P(I) = P_0 + \frac{1}{1!}\frac{\partial P}{\partial I}_{|I_0} \Delta I + \frac{1}{2!}\frac{\partial^2 P}{\partial I^2}_{|I_0} \Delta I^2 + \cdots + R \tag{1}$$

где: I_0 информация реализованного события, P_0 – вероятность реализованного события.
Собственно, таким образом, принимается, что информация есть причина действия, что позволяет описывать широкий круг явлений

$$P = P(I). \tag{2}$$

В таком случае для определения линейной нормы естественно использовать первое приближение (1). Учитывая, что производная вероятности по информации это по определению плотность распределения, запишем

$$f(I) = \frac{\partial P}{\partial I}$$

Естественно, выражение (2) предполагает, что возможно рассматривать зависимость от информации не только вероятности, но и вполне детерминированных величин.

Так как постановка задачи, в силу центральной предельной теоремы, приводит к нормальному распределению, можем записать (2) в виде:

$$\frac{dP}{dI} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(I-mI)^2}{2\sigma^2}} \tag{3}$$

Где: *mI* - ожидаемая информация, σ - среднеквадратическое отклонение.

Собственно, данное утверждение это только гипотеза и, естественно, все дальнейшее правильно, если гипотеза верна, но гипотеза вполне оправдана и реализуема.Исходя из полученной связи (3) для получения явной оценки информации по измеряемой величине – вероятности необходимо решить дифференциальное уравнение

$$dP = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(I-mI)^2}{2\sigma^2}} dI \tag{4}$$

Естественно получаем функцию Лапласа, как связь между вероятностью и информацией, в случае нормального распределения

$$P = \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(I-mI)^2}{2\sigma^2}} dI \ . \tag{5}$$

Используя стандартную замену переменных

$$t = \frac{I - mI}{\sigma_I \sqrt{2}},$$

Получаем

$$P(\alpha < I < \beta) = \int_{\frac{\alpha - mI}{\sigma_I \sqrt{2}}}^{\frac{\beta - mI}{\sigma_I \sqrt{2}}} e^{-t^2} dt, \tag{6}$$

можем записать (5) в виде определённого интеграла, что даётпри нулевом математическом ожидании и σ^2=1/2функцию Лапласа в стандартной форме, определяющую вероятность попадания информации в α интервал

$$\Phi(I_\alpha) = \frac{2}{\sqrt{\pi}} \int_0^{I_\alpha} e^{-t^2} dt = P(-I_\alpha < I < I_\alpha) = P_\alpha. \tag{7}$$

Для получения зависимости информации от измеряемых величин, в данном случае вероятности, используем обратную функцию Лапласа. Для обозначения задачи используем индекс N, подчеркивая нормальное распределение

$$\left\| I \right\|_N = \Phi^{-1}(P_\alpha). \tag{8}$$

Таким образом, в информационном пространстве [7], в рассматриваемом случае сбора информации, нормой является величина доверительного интервала для вероятностиP_α.

Естественно для несимметричного интервала следует использовать две оценки информациинижнюю I_αи верхнюю I_β

$$P(I_\alpha < I < I_\beta) = \frac{1}{2}\left[\Phi\left(\frac{I_\beta - mI}{\sigma_I \sqrt{2}} \right) - \Phi\left(\frac{I_\alpha - mI}{\sigma_I \sqrt{2}} \right) \right]$$

Однако целесообразно использовать соотношение (8), как определение нормы. При этом, учитывая, что информация в данном случае неотрицательна, выполняются условия для нормы

$$\begin{aligned} &1.\ \left\| I \right\|_N = \Phi^{-1}(P_\alpha) \ge 0;; \\ &2.\ \left\| I \right\|_N = \Phi^{-1}(P_\alpha) = 0 \leftrightarrow I_\alpha = 0 \\ &3.\ \left\| I_z \right\|_N \le \left\| I_x \right\|_N + \left\| I_y \right\|_N \leftrightarrow \Phi^{-1}(P_{x+y}) \le \Phi^{-1}(P_x) + \Phi^{-1}(P_y) \ . \end{aligned} \tag{9}$$

Однако в этом случае информационное пространство не линейно

$$\left\| \xi I \right\|_N \ne \left| \xi \right| \left\| I \right\|_N .$$

При этом маловероятные сообщения несут мало информации, а наиболее ценны события с большей вероятностью

$$\begin{aligned} &P_\alpha = 0 \Leftrightarrow \left\| I \right\|_N = 0;; \\ &P_\alpha = 1 \Leftrightarrow \left\| I \right\|_N = \infty. \end{aligned} \tag{10}$$

Собственно это соответствует особенностям задачи, в которой ставится под сомнение компетентность источника.

Таким образом, величина информации в событии определяется доверительной вероятностью нахождения информации I в заданном интервале (I_α, I_β) для этого события, и чем больше мы можем доверять полученному результату, тем выше его информативность. Естественно, в этом случае справедливо правило сигм.

Однако для организации процедуры сравнения и принятия решения необходимо определить метрику, как методику измерения расстояния между элементами в информационном пространстве.
Особенностью информационного пространства, в задачах принятия решения, является наличие априорной информации I^*, порождающей гипотезу и расстояние определяется между пришедшим сообщением и сообщением порожденным гипотезой.

Рассмотрим задачу принятия решения о соответствии гипотезы и получаемой информации I. Данная ситуация возникает в задачах распознавания и идентификации, где система принимающая решение обладает информацией и сравнивает свою гипотезу с явлениями внешней среды.

В данном случае индуцирование метрики неправомерно, так как в данной задаче нет симметрии. Следуя [7] запишем условие связи (2) для условной плотности с математическими ожиданиями mI и mI^*, среднеквадратическими отклонениями σ_I и σ_{I^*}, и взаимной корреляцией ρ_{II^*}

$$\frac{dP_{I/I^*}}{dI} = f(I/I^*) = \frac{1}{\sqrt{2\pi\sigma_I(1-\rho^2_{II^*})}}\exp\left\{-\frac{1}{2\pi\sigma_I(1-\rho^2_{II^*})}\left(I - mI - \rho\sqrt{\frac{\sigma_I}{\sigma_{I^*}}}(I^* - mI^*)\right)^2\right\}. \quad (11)$$

Переобозначим дисперсию и математическое ожидание

$$m_I = mI - \rho_{II^*}\sqrt{\frac{\sigma_I}{\sigma_{I^*}}}(I^* - mI^*); \qquad \sigma_I = \sigma_I(1-\rho^2_{II^*}).$$

Получаем запись для условной плотности нормального распределения в виде

$$dP_{I/I^*} = \frac{1}{\sqrt{2\pi}\sigma_I}e^{-\frac{(I-m_I)^2}{2\sigma^2}}dI$$

Следовательно, метрика в данной задаче определяется как взаимный доверительный интервал для условной доверительной информации. При этом учет требований к метрике диктует использование не доверительной информации, а вероятности недоверия $1\text{-}P_{I/I^*}$ [7]

$$a(I,I^*)_N = \Phi^{-1}(1-P_{I/I^*}).$$

При этом выполняются следующие условия

$$1.\ a(I,I^*)_N = \Phi^{-1}(1-P_{I/I^*}) \ge 0;$$
$$2.\ a(I,I^*)_N = \Phi^{-1}(1-P_{I/I^*}) = 0 \leftrightarrow I = I^*; \quad (12)$$
$$3.\ a(I_x,I_y)_N \le a(I_x,I^*)_N + a(I_y,I^*)_N.$$

При этом выполняются требования к метрике, неотрицательность, аксиома тождества, аксиома треугольника. Однако аксиома симметрии, в данном случае, не применима, так как гипотеза выдвигается только одной стороной.

Естественно, в данном случае сохраняется возможность оценки сигмовых интервалов. Действительно при оценке вероятности события P_a=0.9972 информация равна I= 3σ, следовательно, зная оценку дисперсии можно легко оценивать информативность сообщения.

Следует отметить, что желание использовать среднеквадратическое отклонение для оценки информативности и расстояния между элементами пространства образов [8] существует давно.

Таким образом, в задаче сбора информации, где существенно учитывать несвязность и малую информированность источников, принятие решения осуществляется в нормированном метрическом пространстве – информационном пространстве, где величина информации определяется доверительной вероятностью сообщения. Расстояние между ожидаемым и пришедшим сообщениями, в информационном пространстве, определяется отклонением условной доверительной информации от её максимального значения. Существенным моментом является, что мы имеем «мгновенную» оценку, собственно это оценка события. Для получения оценки источника необходимо определить математическое ожидание оценки информации. Естественно, это будет энтропийная оценка

$$H(I) = M\{\|I\|\};$$
$$H(I/I^*) = M\{a(I,I^*)\}.$$

Если при постановке задачи основным является наличие внутренней связи в объекте, вызывающей изменение информации при изменении состояния объекта, то гипотеза (2) меняется на уравнение органического роста, и мы получаем меру Хартли

$$\frac{dP}{dI} = -\beta P, \Leftrightarrow I = -\frac{1}{\beta}\ln P \quad .$$

Однако при использовании меры Хартли наиболее информативным оказывается маловероятное событие, что в рассматриваемом случае равносильно наибольшему доверию к явно ложной информации.

Выводы:

1. В задачах сбора информации от множества несвязанных, малоинформированных источников нормой информации является доверительный интервал,соответствующий вероятности события.
2. Метрика определяется как доверительный интервал, соответствующий условной вероятности события по отношению к гипотезе.
3. Оценка источника сообщений производится с использованием энтропийной оценки, в смысле математического ожидания нормы и метрики.
4. Гипотеза о первичности информации позволяет адекватно задаче определять норму и метрику информационного пространства.

ЛИТЕРАТУРА:

1. Коротаев С.М. Энтропия и информация – универсальные естественнонаучные понятия Росийский гуманитарный научный фонд http://www.chronos.msu.ru/RREPORTS/korotaev_entropia
2. Могилев А.В. Технологии поиска и хранения информации / А.В. Могилев, Л.В. ЛистроваТехнологии автоматизации управления БХВ-Петербург, 2012. -320 с.
3. Лекции по теории информации МГУ / доступ mindspring.narod.ru/math/it/
4. Хартли Р. Передача информации // Р. Хартли. Теория информации и её приложения. М.: Физматгиз. 1959. – С. 5-35.
5. Бражник Д.О. Розпізнавання методом компенсації інформаційних потоків / Д.О. Бражник, Т.І. Тернова, Л.О. Фаніна //Матеріали восьмої всеукраїнської міжнародної конференції з оброблення сигналів і зображень та розпізнавання образів. 28-31 серпня 2006р., м. Київ, Україна: Київ, 2006. – С.43.
6. Боровиков А.А. Теориявероятностей: Учеб.пособие для вузов.-2-е узд., перераб. и доп. – М.: Наука. Гл. Ред.. физ.-мат. Лит. 1986.-432 с.
7. Бражник Д.О. Анализ сходи мости алгоритма компенсации информационных потоков / Д.О.Бражник, О.М. Бражник, Ф.Б Рогальский // Матеріали міжнародної наукової конференції “Інтелектульні системи прийняття рішень та проблеми обчислювального інтелекту ISDMCI’2009” 18-22 травня 2009р. м. Євпаторія, Україна: Євпаторія, 2009. – т.2 С. 253.
8. Васильев И.В. Распознающие системы / Васильев И.В.[Справочник. Изд. 2-е, перераб. и доп.] - К.:Наукова думка, 1983.- 423 с. ил.

БРАЖНИК Дмитрий Александрович – к.т.н., старший преподаватель кафедры технической кибернетики ХНТУ.

Научные интересы:

– системы распознавания образов.