СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ

УДК 004.832.32

БАЙЕСОВСКАЯ СТРАТЕГИЯ ОЦЕНКИ ДОСТОВЕРНОСТИ ВЫВОДОВ

Т. В. ЗАЙЦЕВА, Е. В. НЕСТЕРОВА
С. В. ИГРУНОВА, О. П. ПУСНАЯ
Н. П. ПУТИВЦЕВА
Н.Н. СМОРОДИНА

Белгородский государственный национальный исследовательский университет

e-mail:
zaitseva@bsu.edu.ru
nesterova@bsu.edu.ru
igrunova@bsu.edu.ru
pusnaya@bsu.edu.ru
putivzeva@bsu.edu.ru
smorodina@bsu.edu.ru

Байесовская стратегия оценки выводов все чаще применяется в промышленности, экономике и науке.

В данной статье рассматривается пример оценки достоверности гипотезы на основе байесовской стратегии. Приведено обоснование применения формулы Байеса для экспертных систем.

Формула Байеса позволяет уточнять вероятность гипотез с учетом влияния различных факторов или обновления информации.

Известный математический аппарат, простота обработки, возможность компьютерной реализации – все это делает использование Байесовской стратегии оценки выводов перспективным и актуальным.

Ключевые слова: формула Байеса, теория вероятности, стратегия оценки достоверности выводов, экспертная система.

Введение.

Байесовская стратегия оценки выводов - одна из стратегий, применяемых для оценки достоверности выводов (например, заключений продукционных правил) в ЭС. Основная идея байесовской стратегии заключается в оценке вероятности некоторого вывода с учетом фактов, подтверждающих или опровергающих этот вывод.

Теорема Байеса.

Формулировка теоремы Байеса, известная из теории вероятностей, следующая.

Пусть имеется n несовместных событий H1, H2,...,Hn. Несовместность событий означает, что никакие из событий H1, H2,...,Hn не могут произойти вместе (другими словами, вероятности их совместного наступления равны нулю). Известны вероятности этих событий: P(H1), P(H2),...,P(Hn), причем P(H1)+P(H2)+...+P(Hn)=1; это означает, что события H1, H2,...,Hn образуют полную группу событий, т.е. одно из них происходит обязательно. С событиями H1, H2,...,Hn связано некоторое событие E. Известны вероятности события E при условиях того, что какое-либо из событий H1, H2,...,Hn произошло: P(E/H1), P(E/H2),..., P(E/Hn). Пусть известно, что событие E произошло. Тогда вероятность того, что какое-либо из событий Hi (i=1,...,n) произошло, можно найти по следующей формуле (формула Байеса):

$$P(H_i/E) = \frac{P(E/H_i)\,P(H_i)}{P(E/H_1)P(H_1) + P(E/H_2)\,P(H_2) + \ldots + P(E/H_n)\,P(H_n)} = \frac{P(EH_i)}{P(E)}$$

События H1, H2,...,Hn называются гипотезами, а событие E - свидетельством. Вероятности гипотез P(Hi) без учета свидетельства (т.е. без учета того, произошло событие E или нет) называются доопытными (априорными), а вероятности P(Hi/E) - послеопытными (апостериорными). Величина P(EHi) - совместная вероятность событий E и Hi, т.е. вероятность того, что произойдут оба события вместе. Величина P(E) - полная (безусловная) вероятность события E.

Формула Байеса позволяет уточнять вероятность гипотез с учетом новой информации, т.е. данных о событиях (свидетельствах), подтверждающих или опровергающих гипотезу.

В ЭС формула Байеса может применяться для оценки вероятностей заключений продукционных правил на основе данных о достоверности их посылок. Заключения (выводы) в этом случае соответствуют гипотезам в теореме Байеса, а посылки - свидетельствам. Обычно посылка правила в ЭС содержит несколько условий. Вероятности P(Hi) и P(E/Hi) определяются на основе статистических данных с использованием формул теории вероятностей. Основные из этих формул следующие.

Формула умножения вероятностей (вероятность того, что произойдет и событие A, и событие B):

P(AB) = P(A)P(B/A) = P(B)P(A/B),

где P(A), P(B) - вероятности событий A и B;

P(B/A) - условная вероятность события B, т.е. вероятность события B при условии, что произошло событие A;

P(A/B) - условная вероятность события A, т.е. вероятность события A при условии, что произошло событие B.

Если события A и B независимы (т.е. вероятность одного события не зависит от того, произошло ли другое событие), то формула умножения вероятностей записывается следующим образом:

P(AB) = P(A)P(B).

Формула умножения вероятностей для нескольких событий (вероятность того, что произойдут все указанные события вместе):

P(A1A2 ...An) = P(A1) P(A2/A1) P(A3/A1,A2) ... P(An/A1,A2,...,An-1).

Формула сложения вероятностей (вероятность того, что произойдет хотя бы одно из событий):

P(A+B) = P(A) + P(B) - P(AB).

Если события A и B несовместны (т.е. не могут произойти вместе), то P(AB)=0, и формула сложения вероятностей принимает следующий вид:

P(A+B) = P(A) + P(B).

Формула сложения вероятностей для нескольких событий обычно записывается следующим образом: $P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A}_1 + \overline{A}_2 + \ldots + \overline{A}_n)$,

где $P(\overline{A}_1 + \overline{A}_2 + \ldots + \overline{A}_n)$ - вероятность того, что не произойдет ни одного из событий A1, A2,...An. Эту величину можно найти, например, по формуле умножения вероятностей.

Пример оценки достоверности гипотезы на основе байесовской стратегии.

В качестве примера рассмотрим ЭС, которая оказывает помощь в оценке условий труда на предприятии. Имеются статистические данные о 5000 человек (из них у 315 возникло профессиональное заболевание) (см. таблицу).

Таблица

Пример оценки достоверности гипотезы

Условие	Значение	Количество случаев возникновения проф. заболеваний	Количество работников, у которых проф. заболевание не обнаружено
Работа с вредными веществами	Постоянно	262	237
	Часто	37	517
	Эпоизодически	14	1011
	Нет	2	2920
Физические нагрузки	Большие	168	927
	Средние	111	1847
	Отсутствуют	36	1911
Нервное напряжение	Есть	202	1532
	Нет	113	3153
Общие заболевания	Есть	196	2011
	Нет	119	2674

Оценить вероятность возникновения профессионального заболевания у работника, работающего с вредными веществами почти ежедневно; при этом его работа связана со средними физическими нагрузками, не связана с нервным напряжением. Общих заболеваний работник не имеет.

Приведенные в таблице данные означают, например, следующее: среди наблюдаемых (5000 человек) постоянно с вредными веществами работают 499 человек (499=262+237). У 262 человек при этом возникло проф. заболевание, а у 237 человек – проф. заболевание не обнаружено.

Здесь в качестве гипотез рассматриваются состояние здоровья у сотрудников: H1 – проф. заболевание обнаружено, H2 - проф. заболевание не обнаружено. Свидетельством здесь является сочетание четырёх факторов, характеризующих работу сотрудников: работа с вредными веществами, наличие физических нагрузок, нервное напряжение, общие заболевания (можно сказать, что в данном случае используются четыре свидетельства); обозначим эти факторы как E1, E2, E3, E4. Обозначим наблюдаемое сочетание факторов (частая работа с вредными веществами, средние физические нагрузками, отсутствие нервного напряжения, отсутствие общих заболеваний) как событие E.

Интерпретация вероятностей, получаемых в ходе расчетов на основе байесовской стратегии.

Определим вероятности, необходимые для расчетов по формуле Байеса. Априорные вероятности гипотез (т.е. вероятности обнаружения проф. заболеваний и отсутствие проф. заболеваний без учета условий работы):

P(H1)=315/5000=0,063;

P(H2)=120/173=0,937.

Наблюдаемое свидетельство (отсутствие/присутствие проф. заболевания) представляет собой сочетание четырёх событий, наблюдаемых вместе: частой работы с вредными веществами, средней физической нагрузки, отсутствия нервного напряжения, отсутствия общих заболеваний. Считая эти события независимыми (т.е. считая, например, что работа с вредными веществами не зависит от отсутствия общих заболеваний, и т.д.), можно найти условные вероятности свидетельства по формуле умножения вероятностей:

P(E/Hi) = P(E1,E2,E3/Hi) = P(E1/Hi) P(E2/Hi) P(E3/Hi), i=1,2.

Найдем величины, необходимые для применения формулы умножения вероятностей:

P(E1/H1)=37/315=0,117; P(E2/H1)=111/315=0,352; P(E3/H1)=113/315=0,359; P(E4/H1)=119/315=0,378;

P(E1/H2)=517/4685=0,11; P(E2/H2)=1847/4685=0,394; P(E3/H2)=3153/4685=0,673; P(E4/H2)=2674/4685=0,571.

Здесь, например, P(E1/H1) – вероятность того, что работник часто работает с вредными веществами, при условии, что в будущем будет найдено проф заболевание. Эта величина показывает, насколько часто у рабочих, часто работающих с вредными веществами, обнаруживают проф заболевание.

Подставляя найденные величины в формулу умножения вероятностей, получим:

P(E/H1) = 0,117·0,352·0,359·0,378 = 0,006;

P(E/H2) = 0,11·0,394·0,673·0,571 = 0,017.

Здесь, например, величина P(E/H1) - вероятность условий работы (частая работа с вредными веществами, средняя физическая нагрузка, отсутствие нервного напряжения, отсутствие общих заболеваний) при условии, что в будущем возникнет проф заболевание.

Найдем вероятность возникновения профессионального заболевания при наблюдаемых условиях работы (апостериорную вероятность):

$$P(H_1/E) = \frac{P(E/H_1)P(H_1)}{P(E/H_1)P(H_1) + P(E/H_2)P(H_2)} = \frac{0{,}006 \cdot 0{,}063}{0{,}006 \cdot 0{,}063 + 0{,}017 \cdot 0{,}937} = 0{,}022.$$

Эта величина является более точной оценкой вероятности возникновения профессионального заболевания, чем априорная вероятность P(H1), рассчитанная на основе статистических данных без учета условий работы.

Следует также отметить, что полученная апостериорная вероятность (0,022) меньше, чем априорная (0,063). Это означает, что наблюдаемые свидетельства (частая работа с вредными веществами, средняя физическая нагрузка, отсутствие нервного напряжения, отсутствие общих заболеваний) подтверждают гипотезу о том, что профессиональное заболевание не возникнет.

Заключение.

Согласно приведенному примеру применения Байесовской стратегии оценки достоверных выводов можно сделать выводы о том, что при использовании данной методики могут быть получены результаты, учитывающие влияние различных факторов (например, учет условий работы).

Главными достоинствами рассмотренной стратегии являются простота обработки, возможность компьютерной реализации с минимальными временными затратами, возможность накапливать и учитывать новые данные и знания, получать актуальные результаты в зависимости от существующей статистики, а также хорошо известный математический аппарат.

Список литературы

1. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПБ: Питер, 2006. – 384с.
2. Тельнов Ю.Ф. Интеллектуальные информационные системы в экономике/ 2–изд. доп. М.: СИНТЕГ, 1999.-214с.
3. Зайцева Т.В. Алгоритм перевода коэффициентов нечеткой логики в коэффициенты уверенности при разработке экспертной системы в среде GURU / Т.В. Зайцева, Е.В. Нестерова, С.В. Игрунова и др. // Вопросы радиоэлектроники. – Серия ЭВТ. – Выпуск 1. – 2012. – С. 112-119.

BAYESIAN STRATEGY EVALUATION OF THE RELIABILITY OF CONCLUSIONS

T.V. ZAITSEVA, E.V. NESTEROVA
S.V. IGRUNOVA, O.P.PUSNAYA
N.P. PUTIVZEVA
N.N. SMORODINA

Belgorod National Research University

e-mail:
zaitseva@bsu.edu.ru
nesterova@bsu.edu.ru
igrunova@bsu.edu.ru
pusnaya@bsu.edu.ru
putivzeva@bsu.edu.ru
smorodina@bsu.edu.ru

Bayesian strategy assessment findings are increasingly being used in industry, business and science.

This article demonstrates the validation of the hypothesis based on the Bayesian strategy. The rationale of the Bayesian formula for expert systems.

Bayes' formula allows you to specify the probability of hypotheses with the influence of different factors or update information.

The well-known mathematical tools, ease of handling, the ability to implement a computer - all this makes the use of Bayesian estimation strategy promising findings and relevant.

Keywords: Bayesian formula, probability theory, strategy evaluation credibility findings, the expert system.