УДК 681.3

А.Е. Архипов, С.А. Архипова, С.А. Носок

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ СТРУКТУРИРОВАНИЯ ДАННЫХ ЭКСПЕРТНОГО ОПРОСА

Введение

Экспертный опрос – информационная технология, традиционно применяемая в сферах деятельности, характеризующихся с одной стороны высокой сложностью объектов исследований (так называемые сложные системы), а с другой – недостаточной проработанностью формализованного базиса, не позволяющего описать единой логико-математической структурой имеющее место в реальности многообразие свойств и особенностей изучаемого объекта.

Нередко подобная ситуация характерна для новых, динамично развивающихся и теоретически недостаточно оформившихся предметных областей, которые еще “не обросли” необходимой прослойкой специалистов-профессионалов, по уровню своей компетентности адекватных понятию “эксперт”. Это обуславливает появление противоречия между идеей применяемой информационной технологии (экспертный опрос) и составом исполнителей, которому приходится реализовать её практически.

Следствием возникшего противоречия является необходимость внесения изменений в сложившиеся методики обработки экспертных данных.

Постановка задачи

Традиционные методики обработки результатов экспертного опроса базировались на гипотезе однородности данных экспертизы, что обеспечивалось возможностью подбора относительно ровного по компетентности состава экспертов. В наиболее критических ситуациях предполагалось, что можно тем или иным способом оценить профессиональную подготовленность экспертов (например, методом само- и взаимооценки уровня компетентности экспертов [1]), и, построив на основе этой информации систему весов, учитывающих компетентность экспертов, обеспечить формирование взвешенных оценок результатов экспертизы, имеющих приемлемо высокую точность.

В ситуации, когда обеспечение достаточно ровного состава экспертов оказывается невозможным, данные экспертизы могут содержать аномальные результаты. Во избежание существенных искажений в итоговых экспертных оценках эти аномальные результаты должны быть выявлены и исключены из обработки [3]. Из-за небольшого количества экспертов (обычно не более $N = 10$), статистические процедуры выявления аномальных результатов оказываются неработоспособными, по-

этому требуется разработка принципиально новых методов их выявления. В [4] показано, что аномальные результаты в данных экспертизы являются следствием аномального поведения эксперта, причем можно выделить как минимум три стратегии его поведения, приводящие к появлению аномалий (три модели “аномальных” экспертов). Выявление аномальных экспертов базируется на задании системы признаков, отличающих аномальное поведение эксперта от нормального, идентификации по этим признакам соответствующего типа аномалий в результатах опроса, а затем устранении из исходной выборки данных, полученных “аномальными” экспертами и проверки справедливости этих действий с помощью специального критерия.

Подобная процедура предварительной обработки данных, являясь довольно трудоемкой и продолжительной, по сути представляет собой реализацию системы поддержки принятия решения, позволяющей лицу, проводящему обработку данных, сформировать гипотезу об аномальности того или иного эксперта, проверить её и принять окончательное решение, содержащее определенную долю субъективизма. Поэтому актуальна проблема объективизации решения задачи классификации экспертов по степени их “аномальности”. Одним из путей преодоления этой проблемы может быть классификация данных экспертного опроса для выявления среди них аномальных и последующая идентификация с их помощью аномальных экспертов (либо дополнительная проверка справедливости уже выдвинутых гипотез аномальности определенных экспертов).

Модель распределения плотности вероятности экспертных данных

Полагаем, что данные опроса N экспертов можно представить прямоугольной матрицей вида:

$$[x_{ij}] = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{M1} & x_{M2} & \dots & x_{MN} \end{vmatrix} = |X_1, X_2, \dots\dots, X_N|, \qquad (1)$$

столбец которой – вектор, содержащий M ответов эксперта на соответствующие пункты опросного листка, каждый ответ – некоторая количественная оценка, выставленная экспертом в L-балльной шкале.

В [4] для эффективного осуществления предварительной обработки данных введена типизация экспертов в зависимости от стратегии их поведения во время экспертного опроса: компетентный эксперт, случайный, либеральный и жесткий эксперты. Оценки, продуцируемые компетентными экспертами, в первом приближении можно представить моделью вида:

$$x_{ij} = x_{ijucm} + e_{ij}, \qquad (2)$$

 ISSN 1562-9945

содержащей наряду с информативной составляющей x_{ijucm} шум оценивания e_{ij}, являющийся реализацией случайной величины E_{ij}. Для компетентных экспертов в первом приближении можно считать, что все случайные величины E_{ij} имеют практически одинаковое одномодовое распределение E с математическим ожиданием $m\{E\} = 0$ и дисперсией $d\{E\} = \sigma_e^2$. Для случайного эксперта $x_{ij} = e_{ij}^{сл}$, где $e_{ij}^{сл}$ – реализация случайной величины $E^{сл}$ с плотностью вероятности, достаточно равномерно распределенной вдоль L-балльной шкалы с дисперсией $\sigma_{сл}^2 \gg \sigma_e^2$. Для либерального эксперта $x_{ij} = e_{ij}^{либ}$, причем случайная величина $E^{либ}$, реализацией которой является значение $e_{ij}^{либ}$, имеет математическое ожидание $m\{E^{либ}\}$, смещенное в область высоких балльных оценок, т.е. в правую часть L-балльной шкалы. Аналогично для жесткого эксперта $x_{ij} = e_{ij}^{ж}$, величина $m\{E^{ж}\}$ смещена в левую часть шкалы оценок, в область низких баллов.

С учетом изложенного выше модель, описывающую генерацию экспертных данных для каждой строки матрицы $[x_{ij}]$ можно представить смесью вероятностных распределений вида:

$$\varphi(x) = p_{комп} f(x_{iucm}, e) + p_{сл} f(e^{сл}) + p_{либ} f(e^{либ}) + p_{ж} f(e^{ж}), \qquad (3)$$

где $P = [p_{комп}, p_{сл}, p_{либ}, p_{ж}]$ – вектор удельных весов (априорных вероятностей) соответствующих компонентов смеси, $f(\ldots)$ – одномодовый закон распределения плотности вероятностей значений оценок соответствующего типа эксперта. Задача предварительной обработки данных экспертного опроса исходя из общей модели (3) сводится к задаче расщепления смеси распределений, а в узко прикладном смысле – к извлечению из общей выборки данных совокупности оценок, принадлежащих компетентным экспертам. Модель (3) предполагает существование приведенных выше типажей экспертов в чистом виде, что существенно упрощает процедуру оценивания её параметров. Например, составляющие вектора удельных весов P могут быть рассчитаны по формуле:

$$p_0 = \frac{V_0}{V}, \qquad (4)$$

где V_0 – количество экспертов соответствующего типа, а $V = V_{комп} + V_{сл} + V_{либ} + V_{ж}$ – общее количество экспертов.

Однако на практике рассмотренные выше явно выраженные типажи экспертов встречаются редко, являясь в известном смысле идеализацией реальной ситуации. Для описания поведения некомпетентного эксперта более характерна смесевая модель вида (3), объединяющая первый компонент с одним из трех последующих, например:

$$\varphi(x) = (1 - p_{либ}) f(x_{iucm}, e) + p_{либ} f(e^{либ}), \qquad (5)$$

В соответствии с этой моделью аномальное поведение эксперта носит фрагментарный, эпизодический характер. Гистограмма распределения относительных частот ω_t балльных оценок, проставленных таким эк-

спертом, изображена на рис. 1 “Либеральная” составляющая этой гистограммы соответствует резкому всплеску относительной частоты появления максимальной 9-балльной оценки. Гистограмма, рассчитанная по данным опроса компетентного эксперта, приведена на рис. 2.

Подводя итог изложенному, отметим, что в реальной ситуации, приступая к проведению предварительной обработки экспертных данных, нельзя рассчитывать на наличие априорной информации о распределении этих данных. С этих позиций целесообразно для классификации экспертных данных и, в частности, для выделения “аномальных” экспертов, использовать методы кластерного анализа, не требующие для своего применения знания формальной модели изучаемого явления [4].

Применение агломеративных процедур кластеризации для обработки экспертных данных

Очевидно, что каждый вектор X_j может быть представлен в N-мерном пространстве S, а совокупность из N точек образует некоторое облако в этом пространстве, плотность и форма которого зависит от характеристик группы экспертов.

При отсутствии “аномальных” экспертов это облако достаточно локализовано и компактно, образуя один общий кластер, характеризующийся некоторым центром с координатами $X^0 = \left[x_1^0, x_2^0, \ldots, x_M^0\right]^T$, определяемыми отношениями:

$$X^0 = \underset{X_j, X^O \in S}{\arg\min} \sum_{j=1}^{N} r(X_j, X^0),$$

где $r(X_j, X^0)$– мера близости векторов X_j, X^0в пространстве S. Появление “аномальных” экспертов ведет к размыванию и деформации исходной формы облака, формированию новых кластеров.

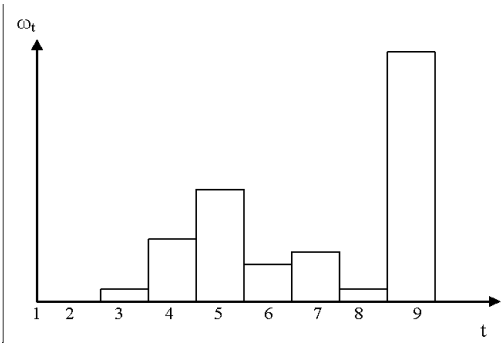

Рис. 1 – Гистограмма относительных частот либерального эксперта

Рассмотрим возможности применения процедуры кластеризации данных экспертного опроса для выявления аномальных экспертов. Исполь-

 ISSN 1562-9945

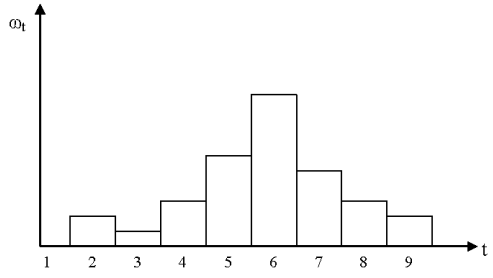

Рис. 2 – Гистограмма относительных частот компетентного эксперта

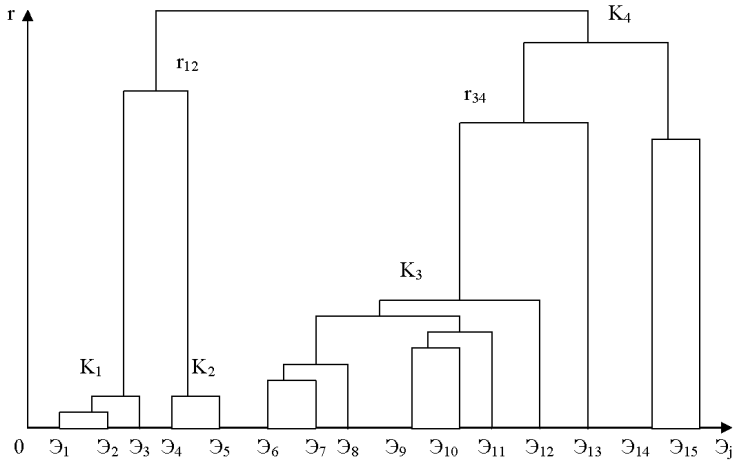

Рис. 3 – Дендрограмма экспертных оценок

зуем иерархический агломеративный алгоритм кластеризации, позволяющий проследить динамику процесса формирования и объединения кластеров из групп оценок, принадлежащих каждому эксперту, а также интерпретацию структуры образующейся при этом дендрограммы (рис.3), ось абсцисс которой образована порядковыми номерами экспертов $Э_1, Э_2, \ldots$, а на оси ординат фиксируются значения расстояний r между элементами S-мерного пространства, группируемыми в кластер, либо объединяемыми вместе кластерами.

Первый шаг построения дендрограммы соответствует объединению оценок экспертов $Э_1, Э_2$ в кластер K_1, к которому на втором шаге присоединяется экспертная оценка $Э_3$. Почти при таком же расстоянии между оценками $Э_4, Э_5$ формируется кластер K_2. Полученные кластеры $K_1, \boldsymbol{K}_2$ принадлежат либеральным и жестким экспертам. В идеализированной ситуации, при выставлении оценок жесткими (либеральными) экспертами, соответствующие парные значений расстояний внутри каждого из кластеров должны быть нулевыми, а межкластерное расстояние r_{12} зависело бы от балльности L шкалы оценок. После образования кластеров $K_1, \boldsymbol{K}_2$ начинается формирование кластера оценок компетентных экспертов K_3, объединяющего экспертные оценки $Э_6 - Э_{12}$. Расстояния между элементами этого кластера выше, чем для кластеров $K_1 - \boldsymbol{K}_2$, т.к. здесь объединяются экспертные оценки, допускающие определенный уровень случайного разброса. Завершающий этап построения дендрограммы – формирование кластера K_4 оценок случайных экспертов, разброс которых намного выше всех ранее рассмотренных оценок, что и определяет максимальные значения расстояний r между элементами внутри K_4. Кроме того, на этом этапе возможно присоединение оценок отдельных случайных экспертов ($Э_{13}$) к кластеру K_3, но расстояние, получаемое для этой ситуации, будет намного большим, чем между элементами кластера K_3.

Проверить правильность сделанного разделения оценок на кластеры $K_1 - \boldsymbol{K}_4$ можно, вычислив для них внутрикластерные расстояния:

$$r(K_l) = \frac{1}{V_l} \sum_{Э_g, Э_j \in \boldsymbol{K}_l} r(Э_g, Э_j), l = \overline{1,4},$$

где V_l - сумма всех возможных пар элементов, входящих в кластер K_l, $l = \overline{1,4}$. При правильном выделении кластеров значение $r(K_l)$ будет минимальным для кластеров идеализированных либеральных и жестких экспертов (т.е. $l = 1, 2$), максимальным для случайных экспертов ($l = 4$) и иметь промежуточную величину для кластера компетентных экспертов ($l = 3$): $r(K_1) \approx r(K_2) \leq r(K_3) << r(K_4)$. Эту проверку целесообразно дополнить сопоставлением межкластерных расстояний $r(K_1, K_2)$, $r(K_1, K_3)$, $r(K_2, K_3)$, $r(K_3, K_4)$ с внутрикластерными $r(K_1)$, $r(K_2)$, $r(K_3)$, которые должны быть существенно ниже межкластерных. Последние в свою очередь должны быть близки внутрикластерному расстоянию $r(K_4)$.

 ISSN 1562-9945

Учитывая, что результаты применения агломеративных алгоритмов могут существенно зависеть от способа вычисления расстояний r, целесообразно прибегнуть к применению “батареи” [4] алгоритмов агломеративного типа с последующей комплексной оценкой совокупности полученных результатов.

Кроме того, принимая во внимание изложенное выше относительно фрагментарности, спонтанности проявления аномалий в работе эксперта (смесевая модель (5)), эффективным может оказаться комбинированный подход к разделению совокупности экспертных оценок на кондиционные и аномальные. Первый этап этого подхода состоит в построении гистограммных моделей (вида рис. 1, 2), дающих интегральное описание поведения каждого эксперта, второй – в кластеризации гистограммных моделей в L-мерном пространстве с выделением кластеров компетентных экспертов, третий – отбор из общей совокупности данных той части, которая получена от компетентных экспертов.

Заключение

При анализе и обработке данных экспертного опроса на этапе разделения их на информативные (полученные от компетентных экспертов) и неинформативные (результат аномального поведения экспертов) наиболее эффективным является кластерный анализ, позволяющий с помощью стандартных агломеративных алгоритмов достаточно оперативно и объективно решить поставленную задачу и интерпретировать полученное разделение.

Литература

1. Архипов О.Є., Архіпова С.А. Математичне моделювання соціальних систем і процесів: Навч.-метод. посіб.- К.: ІВЦ „Видавництво „Політехніка”, 2002.- 60с.
2. Коваленко И.И., Бидюк П.И., Баклан И.В. Системный анализ и информационные технологии в управлении проектами. – К.: “Экономика и право”, 2001. -270с.
3. Архипов А.Е., Архипова С.А., Носок С.А., Пишко И.В. Применение методов кластеризации в задаче обработки данных экспертного опроса // Радіоелектроніка, інформатика, управління, 2003, 2(10). – с. 104-108.
4. Дубров А.Н., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. –М.: Финансы и статистика. 1998. -352с.