Статистический подход к оцениванию зависимых признаков в интеллектуальных системах

Колесникова С. И., Янковская А. Е.
skolesnikova@yandex.ru, yank@tsuab.ru
Томск, Томский госуниверситет систем управления и радиоэлектроники,
Томский архитектурно-строительный университет

Приводятся два метода определения весовых коэффициентов зависимых признаков, используемых в интеллектуальных системах поддержки принятия решений, основанных на теоретико-информационном понятии энтропии и методе главных компонент.

Введение

Одной из наиболее важных проблем при создании интеллектуальных систем выявления закономерностей и поддержки принятия решений является проблема анализа признакового пространства на предмет выделения наиболее значимых признаков и оценивания величины их значимости [1, 2, 4] в виде весовых коэффициентов признаков (ВКП), используемых при принятии решений в интеллектуальных тестовых распознающих системах с матричным представлением данных и знаний [2, 3].

Матрица описаний Q и матрица различений R задают описание объектов в пространстве характеристических признаков объектов и в пространстве классификационных признаков, соответственно. Элемент q_{ij} матрицы Q задает значение j-го признака для i-го объекта. Множество всех неповторяющихся строк матрицы R сопоставлено множеству выделенных образов.

Задача распознавания состоит в определении по матрицам Q и R образа, которому принадлежит заданный совокупностью признаков исследуемый объект, как правило, не входящий в обучающую выборку.

Признаки называются зависимыми, если имеется хотя бы одна пара объектов из разных образов, ими различаемая.

Совокупность признаков, различающих все пары объектов из разных образов (классов), назовем диагностическим тестом (далее по тексту просто тестом).

Строки матрицы тестов T соответствуют тестам, а столбцы — признакам Z, каждый из которых содержится хотя бы в одном тесте. Два объекта считаются различимыми (при $q_{ij} = 0, 1, \Delta$, где $q_{ij} = \Delta$ означает, что признак может принимать как нулевые, так и единичные значения), если хотя бы один признак в описании одного из них принимают значение 1 (0), а в описании другого — инверсное значение 0 (1).

Два предложенных метода базируются на представлении совокупности всех различимых пар объектов из разных образов для каждого

признака в виде мультимножества [4, 5], использовании матриц парных сравнений признаков на основе специальным образом выбранных мер относительной важности признаков, учитывающих их взаимозависимость [4]: $\frac{\delta(|P_i-P_j|)}{\delta(|P_j-P_i|)}, \frac{\delta(/P_i-P_j/)}{\delta(/P_j-P_i/)}$, где $|P_i|$ — мощность i-го мультимножества, соответствующего признаку z_i ; $/P_i/$ — размерность (количество элементов, встречающихся один раз) i-го мультимножества; $P_i - P_j$ — разность мультимножеств, соответствующих признакам z_i и z_j; $\delta(x) = x$, если $x \neq 0$, и $\delta(x) = 1$ иначе.

Постановка задачи

Пусть по матрицам Q и R построены все (или некоторые) безусловные безызбыточные диагностические тесты, представленные матрицей тестов , строки которой сопоставлены тестам, а столбцы — характеристическим признакам, и определено число различающих пар «образ–образ» по каждому (характеристическому) признаку. Требуется определить весовые коэффициенты признаков, входящих в объединение всех (или некоторых) диагностических тестов [2, 5] с учетом их зависимости.

Метод определения ВКП на основе главных компонент

Исходными данными для метода на основе главных компонент является матрица описаний Q, содержащая информацию о частотах встречаемости признаков или вероятностях «проявления» признака для каждого из образов (классов). Под «проявлением» признака здесь понимается различение этим признаком пар типа «образ–образ», образущих соответствующее мультимножество.

Перечислим кратко основные этапы метода:

1) вычисление матрицы ковариации признаков;
2) вычисление собственных значений и собственных векторов матрицы ковариации признаков;
3) вычисление вектора главных компонент;
4) определение относительных дисперсий каждой главной компоненты;
5) определение первых главных компонент, обеспечивающих достаточную вариабельность признаков;
6) интерпретация главных компонент для конкретной задачи.

Метод на основе теоретико-информационного понятия энтропии

Предлагаемый метод включает следующие основные шаги:

1) отдельному тесту $T_1 = (z_1, \ldots, z_M)$, где M — количество признаков в тесте, сопоставляется объединение мультимножеств $\mathbf{P}_{T_1} = \bigcup_{j=1}^{M} P_j$, порожденных признаками, входящих в тест T_1;

2) мультимножество $\mathbf{P}_{T_1}$ представляется в виде объединения попарно непересекающихся мультимножеств $\mathbf{P}_{T_1} = \bigcup_{j=1}^{N} P_j,\ N = 2^M - 1$;
3) определяется дискретное распределение вероятностей $Pr(T_1)$ «проявления» совокупности признаков в тесте и распределение вероятностей «проявления» каждого признака $Pr(z_i)$, входящего в тест, на подмножествах $S_1, \ldots, S_N$;
4) вычисляется содержащаяся информация (энтропия) в тесте $I(T_1)$ и в признаке $I(z_i)$;
5) определяются значения ВКП w_i и теста w_{T_1}, в качестве которых принимаются величины $w_i = I_0 - I(z_i),\ w_{T_1} = I_0 - I(T_1)$, соответственно, где $\log_2 K$ — исходная неопределенность относительно образов, K — количество образов.

Заключение

Отметим, что в реальных данных зависимость между признаками наблюдается очень часто, и в этом случае оценками их индивидуальной «информативности» руководствоваться некорректно. Предложенные методы позволяют: во-первых, представить «вес» теста не в виде суммы ВКП признаков, что не корректно [3, 5] в силу возможной взаимозависимости этих признаков, а в виде суммы «весов» непересекающихся мультимножеств, составляющих тест; во-вторых, учитывать пары «образ–образ» («объект–образ») без дополнительного дублирования (в силу пересечения соответствующих мультимножеств), что вносило искажение в значение ВКП и, соответственно, неточность в решающее правило.

Работа выполнена при поддержке РФФИ, проект № 07-01-00452, и РГНФ, проект № 06-06-12603в.

Литература

[1] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и анализ изображений // Искусственный интеллект в 3-х кн. Кн 2. Модели и методы: Справочник под ред. Д. А. Поспелова. Москва: Радио и связь, 2005. — С. 149–190.

[2] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур. Доклады 3-ей Всероссийской конф. с междунар. участием. Томск: Изд-во СО РАН. 2000. — С. 163–168.

[3] *Yankovskaya A. E.* Test Pattern Recognition with the Use of Genetic Algorithms // Patt. Recog. and Image Anal. 1999. — Vol. 9.— No. 1. — P. 121–123.

[4] *Петровский А. Б.* Упорядочивание и классификация объектов с противоречивыми признаками // Новости искусственного интеллекта. — 2003. — № 4. — С. 34–43.

[5] *Янковская А. Е., Колесникова С. И.* О применении мультимножеств к задаче вычисления весовых коэффициентов признаков в интеллектуальных рас-

познающих системах // Искусственный интеллект, Украина. Донецк: «Наука і освіта», 2004. — № 2. — С. 216–220.