МЕТОДИКА ВЫБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ОБЪЕКТОВ НА ОСНОВЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

И.А. Зикратов, доктор технических наук, профессор.
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики.
С.А. Техтереков, кандидат педагогических наук.
Сибирская пожарно-спасательная академия – филиал Санкт-Петербургского университета ГПС МЧС России.
В.А. Чижов.
Центр научно-исследовательских и опытно-конструкторских разработок Сибирской пожарно-спасательной академии – филиала Санкт-Петербургского университета ГПС МЧС России

Показана возможность применения байесовского подхода для оценки пожарной опасности зданий и сооружений. Предложена методика отбора признаков не только для решения задачи классификации объектов, но и для проведения кластерного анализа, что позволит обеспечить минимум среднего риска ошибочных решений.

Ключевые слова: байесовский подход, классификация объектов, кластерный анализ, пожарная безопасность, оценка пожарного риска

METHODOLOGY OF THE INFORMATIVE FEATURES SELECTION FOR THE OBJECT CLASSIFICATION BASED ON THE PRINCIPAL COMPONENT ANALYSIS

I.A. Zikratov. Saint-Petersburg state university of information technologies, mechanics and optics.
S.A. Tekhterekov. Siberian fire and rescue academy – branch of Saint-Petersburg university of State fire service of EMERCOM of Russia.
V.A. Chizhov. Center of scientific research and experimental design development of the Siberian fire and rescue academy – branch of Saint-Petersburg university of State fire service of EMERCOM of Russia

The possibility of using the Bayesian approach for assessing the fire hazard of buildings and structures is considered. The method of feature selection is offered not only for solving the problem of the object classification, but also for the cluster analysis. That will allow to provide the minimum average risk of erroneous decisions.

Keywords: bayesian approach, object classification, cluster analysis, fire safety, fire risk assessment

Одной из задач, стоящих при определении расчетных величин пожарного риска в зданиях, сооружениях и строениях различных классов функциональной пожарной опасности, является анализ пожарной безопасности зданий [1].

В общем случае процедура классификации осуществляется в многомерном пространстве признаков, наличие которых создает предпосылки для использования методов распознавания образов.

При моделировании технических систем, которые подвержены воздействию стохастических факторов, широкое распространение получили вероятностные модели. Выбор в пользу вероятностных моделей основан на достаточной разработанности методов

теории вероятностей и математической статистики. Известно, что статистические задачи, независимо от методов их решения, обладают следующим свойством: до того как получена реализация наблюдаемой случайной величины (процесса), для описания ситуации в качестве потенциально возможных рассматривается несколько вероятностных моделей. После обработки наблюдаемых значений получают выраженное в некотором виде знание об адекватности этих моделей. Изменение априорной вероятностной информации о состояниях факторов может быть выявлено после получения новых экспертных оценок или в результате наблюдения соответствующих событий. Полученные апостериорные вероятности могут быть использованы для выработки управляющих воздействий, направленных на поддержание системы защиты в требуемом состоянии.

Одним из наиболее распространенных методов в распознавании является статистический метод, основанный на байесовской теории [2]. В нем задача распознавания связана с определением принадлежности исследуемого объекта к одному из заранее выделенных классов.

Байесовский метод обеспечивает высокую вероятность идентификации регистрируемых результатов и, при наличии набора данных, характеризующих исследуемый объект, обеспечивает минимум среднего риска ошибочных решений. Сама процедура идентификации заключается в последовательной проверке гипотезы о подобии полученного образа и заданных образов состояния объекта, содержащихся в библиотеках баз данных.

Основой этого метода является известная теорема Байеса:

$$P(S_i / A) = \frac{P(S_i)P(A / S_i)}{P(A)},$$

где $P(S_i / A)$ – апостериорное значение величины S_i при условии, что произошло событие A; $P(A / S_i)$ – условная вероятность наступления события A при условии, что произошло событие S_i.

Байесовский подход обычно рассматривают как способ переоценки научных представлений с помощью вновь полученных данных, и при решении статистических задач он находит широкое применение. Основное отличие байесовского подхода от других статистических подходов состоит в том, что до того, как будут получены реализации, лицо, принимающее решение, или эксперт рассматривает степени своего доверия к возможным моделям и представляет их в виде вероятностей. После получения апостериорных данных, теорема Байеса позволяет рассчитать новое множество вероятностей, которые представляют пересмотренные степени доверия к возможным моделям.

Привлекательность байесовского подхода состоит в том, что имеющаяся в распоряжении экспертов информация может не отвечать требованиям представительности статистической выборки, что делает использование многих традиционных частотных подходов неправомерным. Более того, ситуация, в которой принимается решение по оценке зданий (сооружений) на предмет их пожарной опасности, может быть вообще новой и никогда ранее не анализируемой. Эти особенности усложняют процесс принятия решений и могут поставить под сомнение какие-либо выводы и заключения. Поэтому байесовский подход может оказаться весьма полезным и эффективным для количественной оценки факторов.

Покажем возможности использования байесовского подхода в сфере оценки пожарного риска. Рассмотрим пример.

Предположим, что необходимо определить влияние некоторого фактора (угрозы) Y на степень некоторого типа противопожарной защиты здания. Пусть в результате решения этой задачи необходимо отнести рассматриваемый объект к одной из нескольких групп. Первая группа – безопасные здания, вторая группа – здания с высокой степенью

противопожарной защиты, третья группа – системы с низкой степенью противопожарной защиты. Таким образом, при анализе конкретной системы защиты имеются три гипотезы θ_i ее принадлежности *i* группе, *i*=1, 2, 3. Пусть из общей статистики воздействия факторов типа Y на аналогичные системы защиты известно, что 50 % таких систем оказались надежными, 30 % систем имеют высокую и 20 % – низкую надежность. Используя эти данные, можно определить априорные вероятности гипотез $P(\theta_1)=0{,}5$; $P(\theta_2)=0{,}3$; $P(\theta_3)=0{,}2$.

Для иллюстрации использования байесовского подхода в рассматриваемом примере из всех возможных показателей надежности системы защиты (СЗ) выберем два – способность СЗ выявить (распознать) воздействие угрозы типа Y (y_1) и способность СЗ нейтрализовать эту угрозу (y_2). Допустим, что из анализа угроз такого типа известно, что при их актуализации они выявлялись надежными СЗ в 60 % случаев. При воздействии на СЗ высокой надежности обнаружение воздействия происходило в 80 % случаев, и для систем с низкой надежностью такой показатель составляет 15 %. Отсюда можно записать условные вероятности $P(y_1/\theta_1)$=0,6; $P(y_1/\theta_2)$=0,8; $P(y_1/\theta_3)$=0,15. Также известно, что имеющиеся методы и/или средства защиты надежных СЗ позволяют нейтрализовать 70 % угроз рассматриваемого типа, для СЗ высокой и низкой надежности такие показатели равны соответственно 90 % и 2 %. Тогда можно записать условные вероятности $P(y_2/\theta_1)$=0,7; $P(y_2/\theta_2)$=0,9; $P(y_2/\theta_3)$=0,02.

Предположим, что достоверно выявлено воздействие угрозы рассматриваемого типа. Учитывая показатель y_1, вычислим апостериорные вероятности гипотез для одного свидетельства:

$$P(\theta_1/y_1)=\frac{P(y_1/\theta_1)P(\theta_1)}{\sum_{i=1}^{3}P(y_1/\theta_i)P(\theta_i)}=0{,}53;$$

$$P(\theta_2/y_1)=\frac{P(y_1/\theta_2)P(\theta_2)}{\sum_{i=1}^{3}P(y_1/\theta_i)P(\theta_i)}=0{,}42;$$

$$P(\theta_3/y_1)=\frac{P(y_1/\theta_3)P(\theta_3)}{\sum_{i=1}^{3}P(y_1/\theta_i)P(\theta_i)}=0{,}05.$$

Из результатов расчета следует, что после того, как y_1 произошло, доверие к гипотезам θ_1 и θ_2 возросло, а к гипотезе θ_3 снизилось.

Очевидно, что если в результате опыта выяснилось, что СЗ не обнаружила факта воздействия угрозы, то необходимо рассматривать противоположные события $P(\bar{y}_1/\theta_i)=1-P(y_1/\theta_i)$. Тогда получим $P(\theta_1/\bar{y}_1)=0{,}47$; $P(\theta_2/\bar{y}_1)=0{,}14$; $P(\theta_3/\bar{y}_1)=0{,}40$. То есть доверие экспертов к гипотезе о низкой надежности исследуемой СЗ существенно возрастает, а доверие к гипотезе о высокой степени надежности резко уменьшается.

В процессе сбора фактов вероятности гипотез будут повышаться, если факты поддерживают их, или уменьшаться, если факты опровергают их. Если одновременно

получены два показателя y_1 и y_2, то при условии их независимости можно воспользоваться формулой:

$$P(\theta_i / y_1, y_2) = \frac{P(y_1 / \theta_i)P(y_2 / \theta_i)P(\theta_i)}{\sum_{i=1}^{3} P(y_1 / \theta_i)P(y_2 / \theta_i)P(\theta_i)}.$$

Вероятности гипотез в этом случае будут равны $P(\theta_1 / y_1, y_2) = 0{,}49$; $P(\theta_2 / y_1, y_2) = 0{,}51$; $P(\theta_3 / y_1, y_2) = 0$. По сравнению с результатами, полученными по одному показателю y_1, доверие экспертов к первой и третьей гипотезе снизилось, а ко второй – возросло. То есть с вероятностью 0,51 исследуемую СЗ можно отнести к категории СЗ высокой надежности по отношению к воздействию угрозы типа Y.

Приведенный пример, не претендуя на глубокий анализ байесовского подхода, иллюстрирует его практическую применимость в задачах оценки противопожарных рисков. Следует отметить, что рассматриваемые в примере факторы (угрозы) считались статистически независимыми величинами, что существенно упростило расчеты и позволило использовать так называемый «наивный» байесовский классификатор.

Однако на практике необходимо оценивать воздействие не одного, а нескольких факторов. Причем очень часто степень их коррелированности эксперту не известна. Поэтому актуальной является задача выбора комбинации информативных признаков и получение правил классификации многомерных наблюдений на основе значений именно этих факторов.

Кроме того, в ряде случаев объекты могут характеризоваться вектором признаков, число n элементов в котором может превышать число m классифицируемых объектов. Такая ситуация, известная как «проклятие размерности», приводит к явлению «переобучения» системы распознавания.

Поэтому, для выявления из всех признаков тех, которые являются наиболее информативными, с точки зрения задачи распознавания, предлагается предварительно провести анализ данным методом главных компонент.

В этом случае методику можно представить следующими основными этапами:

1. Составление исходной матрицы X «объект-признак» размером $m \times n$. Как правило, для исключения влияния отдельных «выбросов» данных выполняется их нормировка.

2. Составление ковариационной матрицы признаков $K_{nxn}=X^TX$.

3. Составление корреляционной матрицы, элементами которой являются:

$$R_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}.$$

4. Сингулярное разложение корреляционной матрицы $R=USV^T$, где U и V – ортогональные матрицы левых и правых сингулярных векторов; S – матрица-вектор сингулярных чисел, расположенных в порядке убывания [3]. С учетом свойств матрицы S, для получения матрицы R требуется не m столбцов матрицы U, а лишь первые min(m,n) столбцов, аналогично, лишь первые min(m,n) строк матрицы V^T. Именно они оказывают наибольшее влияние на результат произведения.

5. Определение k главных компонент, определяемых по соотношению первых сингулярных чисел. В зависимости от размерности задачи и «зашумленности» матрицы ограничиваются несколькими первыми главными компонентами, которые объясняют более 80–90 % дисперсии.

6. Из анализа первых k столбцов (главных компонент) матрицы U определяются компонентные веса, приходящиеся на каждый информационный признак.

7. Проводится кросс-валидация (в случае малой выборки объектов) или «зашумление» путем добавления исходной матрицы X, и пункты 1–6 повторяются.

8. Определяются признаки, которые по результатам двух расчетов имеют наибольшие компонентные веса. Они и отбираются в качестве признаков для дальнейшей процедуры распознавания.

Очевидно, что предлагаемая методика позволит отбирать признаки не только для решения задачи классификации объектов, но и для проведения кластерного анализа.

Таким образом, рассмотренный подход в сочетании с сингулярным разложением матрицы «объект-признак» позволяет выбрать рабочий словарь признаков пожарной опасности, устойчивый к изменениям априорно неизвестного параметра модели данных.

Литература

1. Методика определения расчетных величин пожарного риска в зданиях, сооружениях и строениях различных классов функциональной пожарной опасности: Приложение к Приказу МЧС России от 30 июня 2009 г. № 382 // Электронный фонд равовой и нормативно-технической документации. URL: http//www.docs.cntd.ru (дата обращения: 22.05.2014).

2. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука, 1979. 346 с.

3. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение: пер. с англ.; под ред. Х.Д. Икрамова. М.: Мир, 1998. 575 с.