УДК 519.2

Е. А. Бурков, Е. А. Толкачёва, П. И. Падерно
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Ф. Э. Сатторов
Национальный исследовательский университет ИТМО

Стохастическая модель классификации объектов

Предложены стохастическая модель процедур классификации и способы ее возможного применения. Рассмотрены общая постановка задачи классификации объектов и представление результатов решения этой задачи в форме вероятностной матрицы. Использование подобной стохастической модели классификации позволяет проводить сравнительный анализ результативности различных процедур классификации и осуществлять выбор наиболее подходящей из них по критерию близости к эталонной процедуре. Приведен набор базовых метрик, которые могут быть применены для оценки степени близости вероятностной матрицы анализируемой процедуры классификации к матрице эталонной процедуры. На примере двукратной процедуры классификации рассмотрено агрегирование стохастических моделей отдельных процедур и получение модели комплексной процедуры классификации с помощью произведения Адамара. В заключение приводятся возможные направления развития стохастической модели процедур классификации.

Вероятностная матрица, двукратная процедура классификации, классификация объектов, коэффициент аутентификации, коэффициент идентификации, стохастическая модель

Задачи классификации имеют широкое распространение, а от умения их успешно решать иногда зависит очень многое. Постановка медицинского диагноза, предсказание месторождений полезных ископаемых, оценка кредитоспособности заемщиков и множество других важных задач, с которыми сталкиваются люди каждый день в самых разных сферах своей профессиональной деятельности, – все это по своей природе задачи классификации. Поэтому вполне очевидно, что задачам этого типа уделяется столь значительное внимание: создаются, анализируются и развиваются разнообразные математические и компьютерные модели и методы, призванные обеспечить эффективное – т. е. в первую очередь точное и быстрое – решение задач классификации. В частности, на решение подобных задач ориентированы многие алгоритмы машинного обучения, например алгоритмы распознавания образов.

Однако, несмотря на актуальность этой темы на сегодняшний день, данная статья посвящена не новому способу или подходу к решению задач классификации, а вопросу, который тем не менее косвенно связан с данной темой и, если вдуматься, не менее важен. Речь идет о сравнении эффективности и оценке целесообразности применения уже разработанных и апробированных процедур (методов, моделей и т. д.) классификации объектов в тех или иных предметных областях. На практике решение любой сложной задачи требует затрат тех или иных ресурсов: финансовых, временных, технических и т. д. И если сначала первостепенно важно найти хоть какой-нибудь способ решения задачи, то в дальнейшем по мере исследования проблемной области и накопления знаний на первый план начинает выходить вопрос о том, какой из имеющихся подходов наиболее эффективен и отвечает определенным ресурсным требованиям и ограничениям. А это уже позволяет говорить о задаче выбора или, если более формально, задаче оптимизации, для решения которой, в свою оче-

редь, также требуются определенные математические модели и методы, позволяющие формально описать как сравниваемые альтернативы (в данном случае в их роли выступают процедуры классификации), так и критерии выбора.

Широко известны классические перестановочные методы и алгоритмы классификации объектов [1]. В силу того, что матрицы перестановок не только порождают двоякостохастические матрицы, но и, согласно теореме Биркгофа, любая двоякостохастическая матрица есть выпуклая комбинация конечного числа матриц перестановок [2], возможен подход к описанию процедур классификации на языке стохастических и двоякостохастических матриц, что вполне естественно указывает на возможность создания вероятностных моделей процедур распределения объектов по категориям. В данной статье авторы предлагают к рассмотрению стохастическую модель процедур классификации, которая в общем случае не зависит от их специфики или природы подлежащих классификации объектов и позволяет выполнять сравнительный анализ эффективности различных процедур классификации.

Начнем описание с формальной постановки задачи классификации объектов, которая затем даст возможность ввести в рассмотрение предлагаемую стохастическую модель процедуры классификации.

Общая постановка задачи классификации объектов. Пусть имеется множество объектов V, которое некоторым образом разбито на классы (категории) V_1, V_2, …, V_n. Задача классификации может быть описана следующим образом: задано подмножество объектов $S = \{s_1, s_2, ..., s_m\}$, $S \subset V$, каждый из которых после проведения некоторого комплекса проверок (реализации процедуры классификации) должен быть отнесен к определенной категории V_i, $i = \overline{1,n}$.

Не нарушая общности рассуждений, можно также сделать следующие допущения:

– выработаны признаки описания категорий объектов и существует, пусть и неизвестное нам, эталонное разбиение множества V на классы V_1, V_2, …, V_n, причем $\forall r = \overline{1,m} : s_r \in V_i$, $i = \overline{1,m}$;

– к рассматриваемому множеству объектов S можно применить процедуру классификации, которая рассматривается как некоторое отображение из S на V, т. е. $\forall r = \overline{1,m}\ \exists! i = \overline{1,n} : M(s_r) \in V_i$.

Построение стохастической модели процедуры классификации объектов. Не принимая пока в расчет возможные ошибки методологического плана, но имея в виду ошибки процедуры классификации, т. е. ошибки, возникающие при практическом применении процедуры к конкретным объектам, будем полагать следующее: процедуре классификации M, заданной на множестве объектов V, разбитом на классы (категории) V_1, V_2, …, V_n, может быть поставлен в соответствие набор вероятностных векторов $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{ii}, ..., p_{in})$. Компоненты каждого из них $p_{ij} = P(M(s_r) \in V_j \mid s_r \in V_i)$ представляют собой вероятности отнесения объекта $s_r \in V_i$ после применения к нему процедуры M к классам V_1, V_2, …, V_n соответственно, при этом $p_{ii} = P(M(s_r) \in V_i \mid s_r \in V_i)$ есть вероятность правильной (безошибочной) классификации объекта, т. е. отнесения его к тому классу, к которому он в действительности и принадлежит.

Тогда совокупность подобных вероятностных векторов, сведенных в единую вероятностную матрицу (1), будет представлять собой стохастическую модель некоторой процедуры классификации объектов M:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2n} \\ . & . & . & . & . & . \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots & p_{in} \\ . & . & . & . & . & . \\ p_{n1} & p_{n2} & \cdots & p_{nj} & \cdots & p_{nn} \end{pmatrix}. \qquad (1)$$

Предложенная в форме вероятностной матрицы (1) стохастическая модель процедуры классификации имеет достаточно простое и лаконичное представление, поскольку, как уже ранее отмечалось, на данном этапе нас не интересуют ни специфика методического наполнения или реализации рассматриваемой процедуры классификации, ни природа подлежащих классификации объектов. Соответственно, не рассматриваются и вопросы предварительного определения признаков классов (категорий), числа этих классов, эталонной классификации, а также не решается задача определения критериев допуска диагностических методик, реализуемых в определенном порядке в процессе классификации, по результатам которо-

го каждый объект должен быть отнесен к одной из известных категорий.

Следует заметить, что развитие и расширение области применения рассматриваемой модели допускает и даже предполагает ее усложнение введением в рассмотрение дополнительных понятий и элементов. Так, с целью отражения моделью определенных свойств как предметной области, так и принадлежащих ей объектов вероятностная матрица (1) может быть дополнена вероятностным вектором

$$\mathbf{h} = (h_1, ..., h_n),\ h_i = P(s_r \in V_i),\ \sum_{i=1}^{n} h_i = 1, \qquad (2)$$

который представляет собой распределение вероятностей принадлежности объектов множества S к классам (категориям) V_1, V_2, …, V_n, т. е. априорное распределение вероятностей того, что случайный объект является представителем i-го класса. Данное распределение может быть получено, в частности, на основе предыдущего опыта классификации объектов, например с использованием обучающей выборки объектов. Если же информации для определения априорного распределения недостаточно, то можно воспользоваться гипотезой о форме этого распределения, которая в самом худшем случае, когда у нас нет никакой достоверной информации, будет заключаться в предположении о равномерном распределении объектов по классам.

Матричная форма стохастической модели позволяет достаточно быстро получить представление об эффективности процедуры классификации по критерию безошибочности: для этого следует изучить главную диагональ вероятностной матрицы (1). Элемент матрицы, стоящий на пересечении i-й строки и i-го столбца, будем называть *коэффициентом идентификации* объектов класса V_i рассматриваемой процедурой: $\beta_i = P(M(s_r) \in V_i \mid s_r \in V_i) = p_{ii}$. Соответственно, чем этот коэффициент ближе к единице, тем с большей вероятностью процедура правильно выявляет объекты i-го класса.

Однако эффективность процедуры классификации определяется не только тем, насколько хорошо процедура выявляет объекты конкретного класса, но и тем, насколько результат применения процедуры заслуживает доверия, т. е. насколько он надежен. Например, если анализируемая процедура все объекты относит к одному и тому же классу, то польза от безошибочной идентификации объектов какого-то конкретного класса теряется. Целесообразно в качестве дополнительной характеристики эффективности процедуры классификации ввести *коэффициент аутентификации* объектов класса V_i рассматриваемой процедурой

$$\theta_i = P(s_r \in V_i \mid M(s_r) \in V_i) = h_i p_{ii} \Big/ \sum_{j=1}^{n} h_j p_{ji}.$$

Коэффициент аутентификации показывает, насколько вероятно, что объект, отнесенный процедурой к i-му классу, действительно этому классу принадлежит. Как и в предыдущем случае, чем этот коэффициент ближе к единице, тем эффективнее использование процедуры для выявления объектов i-го класса. При этом можно заметить, что введение в рассмотрение коэффициентов аутентификации неизбежно влечет за собой усложнение исходной стохастической модели, поскольку для данного шага требуется располагать информацией (иметь возможность сформулировать гипотезу) об апостериорном распределении объектов по классам (2).

Несложно понять, что вероятностная матрица эталонной (идеальной) процедуры классификации E, т. е. процедуры, которая безошибочно выявляет классовую принадлежность любого объекта, будет представлять собой $\mathbf{I}_n$ – единичную матрицу порядка n. В свою очередь, все коэффициенты идентификации и аутентификации эталонной процедуры будут иметь единичные значения: $\forall i = \overline{1,n}:\ \beta_i^{(E)} = \theta_i^{(E)} = 1$.

Оценка эффективности процедуры классификации на основе сравнения с эталоном. Решение о приемлемости применения некоторой процедуры классификации объектов, описываемой вероятностной матрицей $\mathbf{P}$, может приниматься с помощью критерия оценки расстояния между диагональными элементами матриц $\mathbf{P}$ и $\mathbf{I}_n$. Задача определения расстояния между двумя точками многомерного пространства достаточно известна и изучена, ее решение предполагает использование различных видов метрик [3]. Оценить расстояние между главными диагоналями матриц $\mathbf{P}$ и $\mathbf{I}_n$ можно, прибегнув, в частности, к следующим способам:

– стандартное евклидово расстояние между диагоналями:

$$\rho_1(\mathbf{P},\mathbf{I}_n)=\sqrt{\sum_{i=1}^{n}(1-p_{ii})^2}=\sqrt{\sum_{i=1}^{n}\left(1-\beta_i^{(M)}\right)^2};$$

– взвешенное расстояние между диагоналями:

$$\rho_2(\mathbf{P},\mathbf{I}_n)=\sqrt{\sum_{i=1}^{n}\alpha_i(1-p_{ii})^2}=\sqrt{\sum_{i=1}^{n}\alpha_i(1-\beta_i^{(M)})^2};$$

– расстояние городских кварталов между главными диагоналями:

$$\rho_3(\mathbf{P},\mathbf{I}_n)=\sum_{i=1}^{n}\left|1-p_{ii}\right|=n-\sum_{i=1}^{n}p_{ii}=n-\sum_{i=1}^{n}\beta_i^{(M)};$$

– взвешенное расстояние городских кварталов между диагоналями:

$$\rho_4(\mathbf{P},\mathbf{I}_n)=\sum_{i=1}^{n}\alpha_i\left|1-p_{ii}\right|=1-\sum_{i=1}^{n}\alpha_i p_{ii}=1-\sum_{i=1}^{n}\alpha_i\beta_i^{(M)}.$$

Весовые коэффициенты α_i удовлетворяют условиям $\forall i=\overline{1,n}:\ \alpha_i\ge 0,\ \sum_i\alpha_i=1$ и позволяют учесть при оценке эффективности рассматриваемой процедуры классификации различную важность в расхождении между соответствующими диагональными элементами вероятностных матриц, т. е. различную стоимость ошибок. В одних случаях цена ошибки в определении класса объекта может быть незначительной, а в других – критичной.

Агрегирование стохастических моделей на основе произведения Адамара. Одно из возможных направлений развития предложенного подхода к построению и использованию стохастической модели процедур классификации заключается в формализации и решении задачи получения единой модели для комплекса совместно используемых процедур классификации. Иными словами, речь идет о способе или алгоритме агрегирования нескольких вероятностных матриц, соответствующих отдельным процедурам классификации, в единую матрицу, позволяющую адекватно описать значимые характеристики всего комплекса рассматриваемых совместно процедур. В данном случае будет рассмотрен способ агрегирования, базирующийся на использовании произведения Адамара.

Для наглядности изложения обратимся к случаю, когда классификация объектов проводится двукратно (повторная классификация), т. е. независимо используются две различные по своим стохастическим характеристикам процедуры классификации к одному и тому же множеству объектов. Будем полагать, что заданы вероятностные матрицы этих процедур классификации:

$$\mathbf{P}_1=\begin{pmatrix} p_{11}^1 & p_{12}^1 & \cdots & p_{1j}^1 & \cdots & p_{1n}^1\\ p_{21}^1 & p_{22}^1 & \cdots & p_{2j}^1 & \cdots & p_{2n}^1\\ . & . & . & . & . & .\\ p_{i1}^1 & p_{i2}^1 & \cdots & p_{ij}^1 & \cdots & p_{in}^1\\ . & . & . & . & . & .\\ p_{n1}^1 & p_{n2}^1 & \cdots & p_{nj}^1 & \cdots & p_{nn}^1 \end{pmatrix};$$

$$\mathbf{P}_2=\begin{pmatrix} p_{11}^2 & p_{12}^2 & \cdots & p_{1j}^2 & \cdots & p_{1n}^2\\ p_{21}^2 & p_{22}^2 & \cdots & p_{2j}^2 & \cdots & p_{2n}^2\\ . & . & . & . & . & .\\ p_{i1}^2 & p_{i2}^2 & \cdots & p_{ij}^2 & \cdots & p_{in}^2\\ . & . & . & . & . & .\\ p_{n1}^2 & p_{n2}^2 & \cdots & p_{nj}^2 & \cdots & p_{nn}^2 \end{pmatrix}. \qquad (3)$$

При различном содержательном наполнении первой и второй процедур классификации, но сравнимых результатах (проведение процедур разными экспертами, если задача классификации решается экспертной группой [4], [5], или использование разных диагностических методик классификации и т. д.) матрицы (3) могут быть существенно различны. Если бы речь шла о том, какую из этих двух процедур более целесообразно использовать с точки зрения эффективности и надежности классификации, то следовало бы выполнить сравнение матриц (3) с вероятностной матрицей эталонной процедуры, как это было рассмотрено ранее в данной публикации, и остановить выбор на той из двух процедур, матрица которой будет ближе к единичной.

Однако не всегда задача выбора наилучшей процедуры классификации тривиальна. Подобный выбор может быть затруднительным в том случае, например, когда априорное распределение подлежащих классификации объектов может иметь разный вид, и в зависимости от его вида эффективность процедур будет варьироваться, т. е. лучше будет то одна процедура классификации, то другая. Отсюда и вытекает необходимость комплексной процедуры классификации.

Таким образом, возможны случаи, когда необходимо учитывать обе процедуры классификации. Матрица двукратной процедуры классификации может быть получена на основе вероятностных матриц (3): $\mathbf{P}_{1,2} = \mathbf{P}_1 \otimes \mathbf{P}_2$, где символ $\otimes$ означает поэлементное перемножение матриц, т. е. произведение Адамара. Необходимо заметить, что матрица $\mathbf{P}_{1,2}$ не вероятностная, так как сумма элементов каждой строки этой матрицы в общем случае имеет значение меньше единицы. В развернутой форме матрицу двукратной процедуры классификации можно представить так:

$$\mathbf{P}_{1,2} = \begin{pmatrix} p_{11}^1 p_{11}^2 & p_{12}^1 p_{12}^2 & \cdots & p_{1j}^1 p_{1j}^2 & \cdots & p_{1n}^1 p_{1n}^2 \\ p_{21}^1 p_{21}^2 & p_{22}^1 p_{22}^2 & \cdots & p_{2j}^1 p_{2j}^2 & \cdots & p_{2n}^1 p_{2n}^2 \\ . & . & . & . & . & . \\ p_{i1}^1 p_{i1}^2 & p_{i2}^1 p_{i2}^2 & \cdots & p_{ij}^1 p_{ij}^2 & \cdots & p_{in}^1 p_{in}^2 \\ . & . & . & . & . & . \\ p_{n1}^1 p_{n1}^2 & p_{n2}^1 p_{n2}^2 & \cdots & p_{nj}^1 p_{nj}^2 & \cdots & p_{nn}^1 p_{nn}^2 \end{pmatrix}. \quad (4)$$

Будем полагать, что в случае двукратной проверки классификация объекта принимается правильной и завершенной, если результаты обеих независимых процедур классификации совпали. При этом необходимо заметить, что результат применения каждой из этих процедур может быть ошибочным. В таком случае матрица (3), исходя из ее природы и способа получения, обладает следующими свойствами:

– вероятность правильной классификации объекта i-го класса равна $p_{ii}^1 p_{ii}^2$;

– вероятность неправильной классификации объекта i-го класса равна $\sum_{j \neq i} p_{ij}^1 p_{ij}^2$;

– вероятность действительной принадлежности объекта к i-му классу при условии отнесения объекта к этому классу равна $p_{ii}^1 p_{ii}^2 \Big/ \sum_{j=1}^{n} p_{ji}^1 p_{ji}^2$;

– вероятность того, что полученное отнесение объекта к i-му классу ошибочно, равна $\sum_{j \neq i} p_{ji}^1 p_{ji}^2 \Big/ \sum_{j=1}^{n} p_{ji}^1 p_{ji}^2$;

– вероятность того, что двукратная процедура классификации не даст результата (при условии равномерного распределения объектов по категориям), равна $1 - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij}^1 p_{ij}^2$.

Под тем, что двукратная процедура классификации не дала результата, подразумевается, что отдельные процедуры классификации отнесли объект к разным классам, вследствие чего возникает неопределенность класса принадлежности объекта. Чтобы избавиться от этой неопределенности, матрица (4) должна быть дополнена *решающим правилом* (принципом), которое позволит определить, какая из двух процедур классификации в случае возникновения противоречия между ними заслуживает большего доверия. Однако вопросы создания и применения решающих правил целесообразно рассмотреть отдельно, поэтому им будет посвящена отдельная публикация в продолжение и развитие данной статьи.

Следует также отметить, что предложенный способ агрегирования стохастических моделей отдельных процедур классификации с помощью произведения Адамара хотя и обладает достоинствами наглядности и удобства использования, однако порождает проблему, которая состоит в том, что получаемая в результате матрица комплексной процедуры классификации не является вероятностной, а значит, не обладает всеми базовыми свойствами предложенной в начале статьи стохастической модели процедур классификации. Следовательно, возможным направлением развития предложенного подхода (будет рассмотрено в продолжении данной статьи) становится либо расширение определения стохастической модели процедуры классификации таким образом, чтобы оно включало в себя более широкий спектр матриц, либо разработка иного способа получения матрицы комплексной процедуры классификации, позволяющего получить на выходе непосредственно вероятностную матрицу с заданными свойствами.

Помимо задачи агрегирования вероятностных матриц отдельных процедур классификации может быть сформулирована и обратная задача – декомпозиция матрицы комплексной процедуры классификации. Матрица (4), будучи положительно полуопределенной, может быть подвергнута разложению по ортогональной системе комплексных векторов [2], что позволяет говорить об

определении такого набора базовых процедур классификации, который даст возможность получить при их агрегировании комплексную процедуру, отвечающую заданным требованиям.

Важным достоинством предлагаемого в статье подхода к стохастическому моделированию процедур классификации может считаться его независимость от специфики процедур классификации объектов и области применения данных процедур. Благодаря этому можно проводить сравнительный анализ и оценивать эффективность процедур классификации, имеющих существенные различия на методологическом, техническом и иных уровнях.

Использование рассмотренной стохастической модели процедур классификации объектов позволяет формализовать постановку и решение широкого спектра научно-исследовательских и практических задач, в частности:

– выбирать оптимальные по заданным критериям процедуры классификации объектов;

– прогнозировать результаты применения моделируемых процедур классификации к заданному множеству объектов;

– конструировать и анализировать комплексные процедуры классификации, представляющие собой наборы более простых процедур.

СПИСОК ЛИТЕРАТУРЫ

1. Do we need hundreds of classifiers to solve real world classification problems? / M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim // J. Mach. Learn. Res. 2014. Vol. 15. P. 3133–3181.

2. Horn R. A., Johnson C. R. Matrix analysis, 2nd ed. Cambridge: Cambridge University Press, 2013.

3. Borg I., Groenen P. J. F. Modern multidimensional scaling. Theory and applications. New York: Springer-Verlag, 2005. Series: Springer Series in Statistics XXII, 2nd ed.

4. Paderno P. I., Burkov E. A., Lavrov E. A. Issues of organization of expertise and problems of expert assessments // J. Phys.: Conf. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1703/1/012047/pdf (дата обращения 01.04.2021).

5. Burkov E. A., Lyubkin P. L., Paderno P. I. Quantitive estimation of extent of coincidence of expertise's objects models // Proc. of the 20th IEEE Intern. Conf. on Soft Computing and Measurements, SCM'2017. Saint Petersburg, 2017. P. 43–45.

E. A. Burkov, E. A. Tolkacheva, P. I. Paderno
Saint Petersburg Electrotechnical University

F. E. Sattorov
National Research University ITMO

STOCHASTIC MODEL OF OBJECT CLASSIFICATION

The article proposes a stochastic model of classification procedures and ways of its possible application. The general formulation of the object classification problem and the presentation of the results of solving this problem in the form of a probabilistic matrix are considered. The use of such a stochastic classification model makes it possible to carry out a comparative analysis of the effectiveness of various classification procedures and to select the most suitable of them according to the criterion of proximity to the reference procedure. A set of basic metrics is presented that can be applied to solve the problem of assessing the degree of proximity of the probabilistic matrix of the analyzed classification procedure to the matrix of the reference procedure. On the example of a double classification procedure, the aggregation of stochastic models of individual procedures and obtaining a model of a complex classification procedure using the Hadamard product are considered. In conclusion, the possible directions of development of the stochastic model of classification procedures are presented.

Probability matrix, double classification procedure, object classification, authentication coefficient, identification coefficient, stochastic model