электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.93'11

Модификация метода классификации в условиях неполного вектора значений признаков

Гаврилов П.А., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»

Научный руководитель: Майков К.А., д.т.н., профессор
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
maikov@bmstu.ru

Машинное обучение — активно развивающаяся область, методы которой позволяют устанавливать неочевидные связи между наблюдаемыми и скрытыми значениями признаков объекта. В задачах классификации множество объектов разбивается на классы, и решающая функция должна ответить на вопрос «какому классу принадлежит объект (из обучающей выборки или новый)?». В некоторых источниках классы называют образами, а задачу классификации — задачей распознавания образов (pattern recognition) [1].

Согласно [2] приведём математическую постановку задачи классификации. Пусть задано множество объектов X, множество допустимых ответов Y, целевая функция $\dot{y}: X \to Y$, значения которой $y_i = \dot{y}(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \dots, x_l\} \subset X$. Пары объект-ответ (x_i, y_i) называются прецедентами. Совокупность пар $X^l = (x_i, y_i)_{i=1}^l$ называют обучающей выборкой (training sample). Необходимо построить решающую функцию (decision function) $a: X \to Y$, которая приближала бы целевую функцию $\dot{y}(x)$, причём на всём множестве X. Признак (feature) объекта x– это результата измерения некоторой характеристики объекта. Пусть имеется набор признаков $f_1, \dots, f_n$. Вектор $(f_1(x), \dots, f_n(x))$ называют признаковым описанием объекта $x \in X$. Совокупность признаковых описаний всех объектов выборки X^l, записанная в виде матрицы размера $l \times n$, называют матрицей объектов-признаков

$$F = \|f_j(x_i)\|_{l\times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}.$$

Для простоты записи далее будем обозначать матрицу объектов-признаков в виде

$$F = \|x_{ij}\|_{l\times n}.$$

В задачах классификации на C непересекающихся классов $Y = \{1, \dots, C\}$.

Известные методы классификации можно представить в виде следующих групп [2]:

– метрические методы [2];

– логические методы [2];

– линейные методы [1, 2];

– методы регрессионного анализа [1, 2];

– байесовские методы [2];

– нейросетевые методы [1, 2, 3, 4].

При решении задач классификации характерна проблема частичного отсутствия данных об объекте. Задачу классификации с отсутствующими данными можно разбить на две подзадачи:

– обработка отсутствующих данных;

– классификация.

Иерархическая схема методов для классификации объектов в условиях неполноты вектора значений признаков показана на рис. 1.

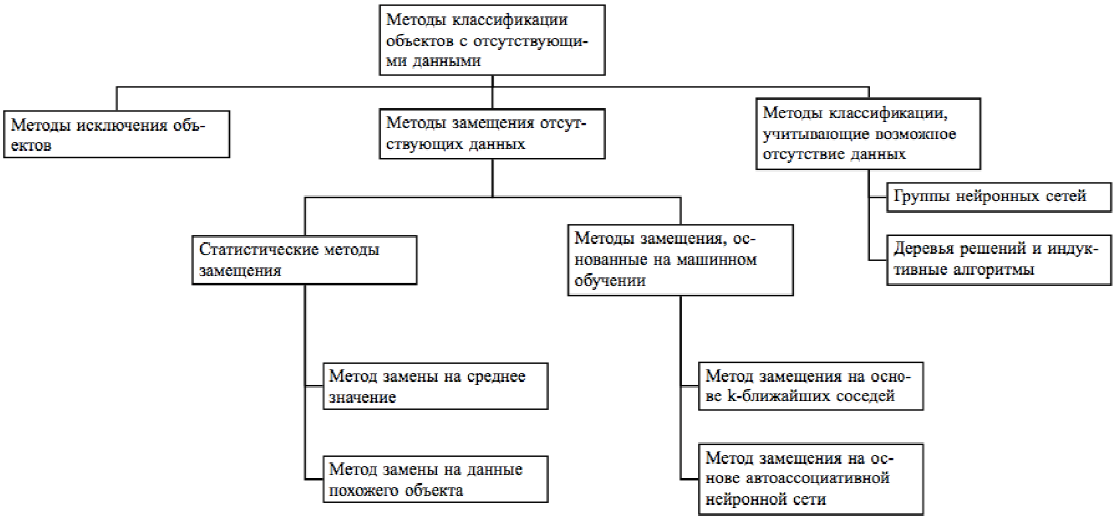

Рис. 1. Методы классификации объектов в условиях неполноты вектора значений признаков

Введем бинарную матрицу, обозначающую отсутствующие данные

$$M = \|g(x_{ij})\|_{l\times n} = \|m_{ij}\|_{l\times n}, \text{ где } g(x_{ij}) = \begin{cases} 1, & \text{если значение } x_{ij} \text{ известно;} \\ 0, & \text{иначе.} \end{cases}$$

Рассмотрим методы, представленные на рис. 1.

Методы исключения объектов.

Метод исключения объекта целиком (exclude cases listwise) заключается в исключении из выборки объектов, у которых отсутствует значение хотя бы одного признака [2].

Согласно описанию преимуществами данного метода являются:

– возможность непосредственного использования известных методов классификации для полных данных;

– простота понимания и реализации;

– скорость работы.

Среди недостатков данного метода можно выделить:

– невозможность использования объектов для обучения классификатора, если у них отсутствует хотя бы одно значение признака;

– невозможность использования обученного классификатора для классификации объектов с неполным вектором признаков.

Для получения желаемых результатов при использовании данного метода необходимо иметь выборку, в которой количество объектов с отсутствующими данными не превышает 5% [1].

Метод попарного исключения признаков (exclude cases pairwise) игнорирует только те объекты, у которых во время классификации оказалось пропущенным значение хотя бы одного из признаков, участвующих в текущем вычислении [2].

Преимуществом данного метода является возможность использования всех известных данных, однако, учитывая специфику его применения, последующая классификация известными методами становится невозможной.

Методы замещения, основанные на статистическом анализе.

Метод замены на среднее значение. Данный метод заменяет отсутствующие компоненты вектора признаков на средние значения соответствующих компонент всех известных случаев. При расчёте среднего значения можно учитывать значения не всех объектов, для которых известны значения данного признака, а только тех, которые попадают в класс объекта с искомой компонентой [2]:

$$x_{ij}^{m} = \frac{1}{N_{obs}} \sum_{j=1}^{L} (1 - m_{ij})\, x_{ij}, \text{ где } N_{obs} = \sum_{j=1}^{L} m_{ij}.$$

Очевидным недостатком данного метода является игнорирование корреляций между значениями различных признаков [2].

Методы замены отсутствующих данных на данные похожего объекта. Метод горячей замены заменяет отсутствующие значения признаков на значения признаков

максимально похожего объекта из имеющейся выборки. Метод холодной замены позволяет сделать то же самое с тем отличием, что похожий объект должен находится не в имеющейся выборке, а в другом источнике данных. Метод учитывает только значения признаков наиболее похожего объекта, а не целого набора данных [6].

Методы замещения, основанные на машинном обучении.

При использовании данной группы методов, строится предсказывающая модель, замещающая отсутствующие данные на основе имеющейся выборки. После замещения, к новой выборке применяется любой из известных алгоритмов классификации [6].

Метод замещения K ближайших соседей. Для объекта с отсутствующим значением признака ищутся K ближайших соседей с известным значением данной компоненты. На основе известных значений рассчитывается отсутствующее значение. Процесс поиска K ближайших соседей осуществляется на основе измерения расстояния между объектами (distance) [2]:

$$D(x_a, x_b) = \sqrt{\sum_{i=1}^{l} D_i(x_{ia}, x_{ib})^2}, \text{ где}$$

$$D_i(x_{ia}, x_{ib}) = \begin{cases} 1, & (1 - m_{ia})(1 - m_{ib}) = 0; \\ D_0(x_{ia}, x_{ib}), & x_i - \text{дискретныйпризнак}; \\ D_N(x_{ia}, x_{ib}), & x_i - \text{численныйпризнак}. \end{cases}$$

В первом случае обрабатывается возможное отсутствие одного из значений.D_0возвращает 0, если данные различны, иначе 1. Нормализованную дистанцию вычисляет:

$$D_N(x_{ia}, x_{ib}) = \frac{|x_{ia} - x_{ib}|}{max(x_i) - min(x_i)}.$$

Данный метод хорошо зарекомендовал себя по сравнению с методом замены на среднее и методом горячей замены. Главным недостатком является количество вычислений, необходимых для нахождения K ближайших соседей.

Методы классификации, учитывающие возможное отсутствие данных.

Выше приведены методы, позволяющие либо игнорировать объекты с неполным набором данных, либо вычислить недостающие данные для последующей классификации стандартными методами классификации. Методы, приведенные далее, позволяют осуществить классификации, не вычисляя недостающие значения признаков и не исключая объект с неполным вектором признаков из выборки [5, 6].

Группы нейронных сетей. В источнике [4] предложен метод нейросетевой редукции. При использовании данного метода создаётся набор многоуровневых перцептронов, каждый из которых классифицирует объекты, используя различные

комбинации существующих данных, с целью учёта отсутствующих значений признаков.

Главным недостатком данного метода является требование большого количества нейронов при большом количестве разнообразных комбинаций отсутствующих данных [6].

Деревья решений. Методы данной группы учитывают отсутствие данных как в обучающей выборке, так и в тестовой [6]. ID3 — дерево решений, алгоритм, обрабатывающий отсутствующие данные с помощью добавления дополнительных ребер для недостающих значений признаков. Таким образом, неизвестное ребро рассматривается как новое возможное значение для каждого атрибута и обрабатывается, как и все остальные значения. C4.5 — расширение ID3, предложенное в [5]. Данный метод использует вероятностный подход к обработке отсутствующих значений. Во время обучения каждому атрибуту сопоставляется его вес. Если значения атрибута известно, то весу присваивается значение единица, иначе любому значению данного атрибута присваивается значение относительной частоты атрибута. Во время тестирования при неполном векторе признаков объекта, алгоритм исследует все возможные ветви (ниже текущей вершины) и выбирает какому классу принадлежит объект на основе наиболее вероятного значения.

Для задач медицинской диагностики характерны задачи классификации объектов (пациентов) с неполным вектором признаков. Специфика данных задач заключается в том, что некоторые из недостающих значений признаков можно измерить. Пациенту можно провести дополнительные исследования для получения недостающих данных, что впоследствии увеличит точность диагноза. Часть медицинских исследований является дорогостоящими, поэтому необходим поиск отсутствующих данных, максимально влияющих на результат классификации.

В связи с этим возникает необходимость решения следующих задач:

– разбиение выборки на множество объектов с полным вектором признаков и множество объектов с частичным отсутствием данных;

– выделение потенциально измеримых данных, наиболее важных для результатов классификации;

– дополнение исходной выборки данными, измеренными в ходе исследований;

– классификация полученного набора данных.

Методы, описанные выше, не обладают возможностью оценки влияния отсутствующих потенциально измеримых данных на результат классификации. Таким образом необходим новый метод, включающий в себя предобработку данных с целью выделения отсутствующих данных, оказывающих максимальное влияние на результат

классификации, и последующую классификацию.

Список литературы

1. Bishop C.M., Svensen M. Pattern recognition and machine learning. New York: Springer, 2006. 101 p.
2. Воронцов К. В. Машинное обучение (курс лекций). Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9,_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2) (дата обращения 22.10.2014).
3. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems 25 (South Lake Tahoe, 3-8 december 2012). South Lake Tahoe, 2012. pp. 5-10.
4. Sharpe P., Solly R. Dealing with missing values in neural network-based diagnostic systems // Neural Comput Appl. 1995. Vol. 3. P. 73-77. DOI: 10.1007/BF01421959.
5. Quinlan J. C4.5: programs for machine learning // Machine Learning. 1993. Vol. 16. P. 235-240. DOI: 10.1007/BF00993309.
6. Garcia-Laencina P., Sancho-Gomez J., Rigueiras-Vidal A. Pattern classification with missing data: a review // Neural Comput Appl. 2010. Vol. 19, P. 263-282. DOI: 10.1007/s00521-009-0295-6.