МЕТОД КЛАССИФИКАЦИИ ОБЪЕКТОВ НА ОСНОВЕ ФОРМИРОВАНИЯ ПРИЗНАКОВ-ОТНОШЕНИЙ И ПРИМЕР ЕГО РЕАЛИЗАЦИИ НА ОСНОВЕ МНОГОСЛОЙНОГО ПЕРЦЕПТРОНА

К. П. Коршунова
Кафедра вычислительной техники, филиал федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет «МЭИ» в г. Смоленске
Смоленск, Российская Федерация
E-mail: ksenya-kor@mail.ru

Предложен метод решения задачи классификации сложных объектов, основанный на формировании и учете признаков-отношений между исходными признаками. Предложен пример реализации метода, заключающийся в использовании многослойного перцептрона для построения и применения классификации. Произведена оценка качества работы бинарного классификатора.

Системы, решающие классификационные задачи, имеют следующую типовую функциональную схему (рисунок 1).

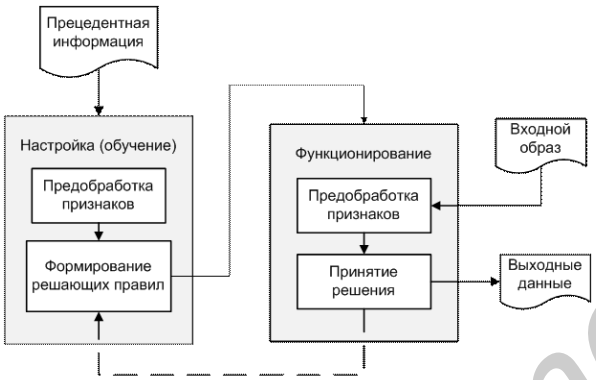

Рис. 1 - Типовая схема решения задачи классификации

Вероятность правильного отнесения объекта к классу зависит от «качества» (информативности) используемых признаков.

На практике возникают ситуации, когда в распоряжении лица, принимающего решение, имеется лишь небольшой набор малоинформативных признаков. Требуется извлечь максимальное количество полезной информации из малого числа признаков. Эта задача должна быть решена на этапе предварительной обработки признаков объектов.

Для увеличения объема полезной информации, используемой для классификации, на этапе предобработки сформируем ряд новых признаков, учитывающих различного рода отношения между между подсистемами (признаками) системы (объекта), – признаки-отношения – и будем использовать их при построении решающих правил. Подробнее о постановке задачи в [1].

В схему решения классификационной задачи внесем дополнительный шаг: выделение n-арных отношений между признаками объектов и определение их значимости с точки зрения рассматриваемой классификации. Тогда предобработка признаков на этапе обучения включает 3 шага:

Шаг 1. Создание формализованного описания объектов: представление исходных признаков в удобном для алгоритма виде (например, дискретизация непрерывных значений).

Шаг 2. Выделение n-арных отношений между признаками объектов. Ограничимся бинарными и тернарными отношениями (n=2,3).

На данном шаге решается специфическая задача кластеризации, состоящая в разбиении n-арных (для учета отношений n-й местности) областей на кластеры одного типа (класса) по имеющейся прецедентной информации. Следует отметить, что данную подзадачу можно решать различными методами, которые окажут влияние на качество решения исходной задачи. Подробнее предлагаемый подход и алгоритм кластеризации, схожий с алгоритмом кластеризации "k ближайших соседей изложены в [2].

Шаг 3. Формирование новой системы признаков: отбор наиболее значимых, ценных признаков с точки зрения рассматриваемой классификации в соответствии с каким-либо критерием (например, дифференциальная информативность, энтропия и пр. [3]).

Возможны 2 подхода к рассмотрению нового набора признаков: исходных и признаков-отношений – на последующих этапах решения классификационной задачи:

1. Формирование решения на основе смешанного набора признаков – рассмотрение новых признаков наравне с исходными: отбор наиболее значимых среди всего множества признаков и их отношений, а также применение единого метода построения классификации.

2. Формирование решения на основе отдельных наборов признаков – рассмотрение множеств исходных и новых признаков по отдельности, то есть применение разных критериев отбора к признакам разного рода и формирование решающих правил отдельно с использованием исходных признаков, а также признаков-

отношений. Подход обусловлен тем, что новые признаки созданы искусственно и не обладают тем физическим смыслом, который имеют исходные, зачастую измерены в другой шкале.

На рисунке представлена общая схема предлагаемого метода решения задачи классификации при формировании решения на основе смешанного набора признаков (рисунок 2).

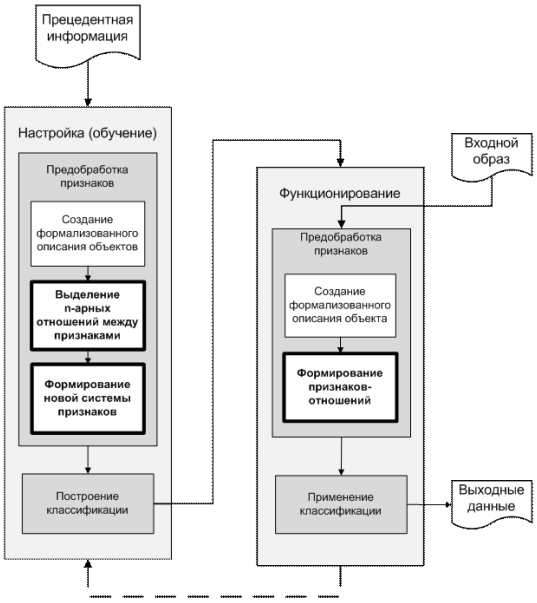

Рис. 2 - Общая схема решения задачи классификации по предложенному методу

Классификационные задачи включают в себя большое разнообразие классов задач и подходов к решению; разные этапы предлагаемой схемы могут быть реализованы различными способами, выбор которых зависит от особенностей конкретной задачи.

Предложим пример реализации метода при использовании первого подхода. Будем решать задачу классификации на основе многослойного перцептрона.

Многослойные перцептроны (МП) представляют собой нейронные сети прямого распространения, которые обучаются с помощью алгоритма обратного распространения ошибки [4]. Выберем для решения задачи простую структуру МП с одним «скрытым» слоем. МП имеет (m+m'+m") входов (значения исходных, бинарных и тернарных признаков соответственно), 2 выхода (вероятности принадлежности входного образа одному из 2 классов) и 2 слоя нейронов. Функции активации нейронов: первого слоя – гиперболический тангенс, второго слоя – softmax-функция.

Рассмотрим пример реальной задачи медицинской диагностики [5]: диагностика рака молочной железы по результатам лабораторного исследования, заключающегося в подсчете клеток определенных типов в образце биологической ткани пациента.

При подаче на вход МП значений как исходных признаков, так и признаков-отношений удалось снизить долю ошибок первого рода с 7% до 6%, а долю ошибок второго рода с 23% до 14,8%. При подаче на вход МП значений только признаков-отношений также удалось снизить долю ошибок первого рода до 6%, а долю ошибок второго рода до 13,5%.

Оценим показатели качества работы бинарного классификатора: точность и полноту [6]. Построим гистограммы, отражающие зависимость основных показателей качества решения от состава учитываемых признаков (рисунок 3).

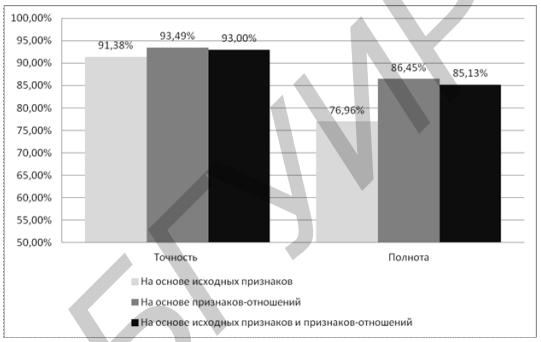

Рис. 3 - Зависимость основных показателей качества решения задачи от состава учитываемых признаков

Как видно из рисунка, наилучшее качество решения задачи классификации для диагностики РМЖ было получено при подаче на вход МП только признаков-отношений.

Таким образом, предложенный способ реализации метода решения задачи классификации на основе формирования признаков-отношений позволил улучшить качество работы бинарного классификатора: точность решения в среднем увеличена с 91,4% до 93,5%, полнота – с 77% до 86,5%.

1. Лямец Л. Л. Подход к формальному описанию объектов в задачах распознавания на основе принципа системности / Математическая морфология. Электронный математический и медико-биологический журнал. - Т. 13. - Вып. 2. - 2014.
2. Коршунова К.П., Борисов В.В. Решение задачи классификации на основе учета бинарных и тернарных отношений между признаками // Сборник трудов IV международной научно-технической конференции «Энергетика, информатика, инновации-2014», Т. 1, 2014. С. 195-201.
3. Биргер И.А. Техническая диагностика. Москва: Машиностроение, 1978.
4. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. Москва: Горячая линия - Телеком, 2006.
5. Абросимов С.Ю. Проверка гипотезы о возможности идентификации стромы биологических тканей в норме, при предопухолевых и опухолевых процессах. [Текст]: научный отчет по проведенному научному исследованию / д.м.н. Абросимов Сергей Юрьевич. – Смоленск, 2006. – 45 с.
6. Manning C., Raghavan P., Schutze H. An Introduction to Information Retrieval. Cambridge: Cambridge UniversityPress, 2009.