Алгоритм отбора информативных признаков на основе оценки разделяющей способности пространства признаков с использованием дискриминантного анализа

А.В. Мухин[1], Р.А. Парингер[1,2], Н.Ю. Ильясова[1,2]

[1]*Самарский национальный исследовательский университет им. академика С.П. Королёва, Московское шоссе, 34, Самара, Россия, 443086*
[2]*Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001*

Аннотация

Для решения задач интеллектуального анализа данных применяют различные методы отбора признаков, которые используются для оптимизации процедуры классификации. Такие признаки также называют информативными. Обычно, наборы данных содержат в себе большое число различных признаков и как правило заранее не известно, какие признаки лучше всего подходят для решения определенной задачи. Анализ такого большого пространства признаков является трудной задачей. Лишь некоторые комбинации таких признаков способны решать сложные задачи анализа данных, например таких как задача сегментации изображений. В настоящее время существует ряд алгоритмов, способных сокращать размерность пространства признаков или находить комбинации самых информативных признаков. Однако их применение часто оказывается проблематичным, в силу их медлительности или невысокой точности. В данной работе предлагается новый метод поиска информативных признаков, который основан на алгоритме дискриминантного анализа. Найденные предложенным алгоритмом информативные признаки в совокупности позволяют получить более точные результаты по сравнению с использованием других алгоритмов. Подтверждение эффективности разработанного метода демонстрируется на задаче семантической сегментации изображений глазного дна.

Ключевые слова

Информативные признаки, дискриминантный анализ, алгоритм, сегментация, глазное дно

1. Введение

Поиск информативных признаков, в настоящее время, является популярным методом в области интеллектуального анализа данных [1-2], особенно в тех задачах, для которых сформированы большие наборы данных с большим набором признаков. Основная цель поиска информативных признаков заключается в нахождении такого подмножества признаков, которое наилучшим образом способно описать распределение данных [3]. При этом, данное оптимальное подмножество не должно включать шумные и неинформативные для данной задачи признаки.

Отбор информативных признаков способен помочь исследователям в создании лёгкой и быстрой модели классификатора или в оптимизации уже существующих моделей. Так, классификаторы, построенные на основе отобранных комбинаций информативных признаков хорошо показывают себя в задачах обработки текста и изображений [4-5].

В настоящее время чаще всего применяют один из двух подходов к поиску информативных признаков [6]. Первый из них заключается в исследовании индивидуальной информативности признаков за счет использования специальных метрик. Такой метод является быстрее остальных, однако не учитывает групповой информативности признаков, поэтому полученные таким способом наборы признаков не всегда являются оптимальными. Другой метод

заключается в максимизации функции ошибки жадным алгоритмом. Такой подход позволяет найти оптимальное подмножество признаков. Однако его вычислительная сложность велика, так как для проверки каждой группы признаков необходимо заново обучать модель на выбранных признаках.

Предлагаемый в статье алгоритм был разработан с целью сохранить высокую точность и снизить вычислительную сложность описанных выше подходов. В его основе лежит использование критериев дискриминантного анализа для определения как индивидуальной, так и групповой информативности признаков. Поиск оптимального набора признаков для решения задачи бинарной классификации происходит прямым перебором. Решение задачи многоклассовой классификации происходит за счет представления оной в виде ансамбля бинарных классификаторов. А именно, поиск наборов признаков происходит для каждого класса, где задача представляется в виде нахождения таких множеств признаков, которые разделяют распределение одного класса от остальных.

2. Заключение

Производительность и точность предлагаемого в статье алгоритма была продемонстрирована на задаче семантической сегментации изображений глазного дна. Так для набора данных, состоящего из 1637 признаков и разделенного на 4 класса, с помощью предложенного алгоритма, было получено 3 набора признаков, которые использовались для обучения бинарных классификаторов. Значения F1-метрики для каждого бинарного классификатора равны соответственно: 0,88, 0,89, 0,60.

3. Благодарности

Результаты исследования были получены при частичной финансовой поддержке РФФИ в рамках научного проекта № 19-29-01135.

4. Литература

[1] Cai, J. Feature selection in machine learning: A new perspective / J. Cai, J. Luo, S. Wang, S. Yang // Neurocomputing. – 2018. – Vol. 300. – P. 70-79.

[2] Huang, J. Joint feature selection and classification for multilabel learning / J. Huang, G. Li, Q. Huang, X. Wu // IEEE transactions on cybernetics. – 2017. – Vol. 48(3). – P. 876-889.

[3] Dash, M. Feature selection for classification / M. Dash, H. Liu // Intelligent data analysis. – 1997. – Vol. 1(3). – P. 131-156.

[4] Rehman, A. Feature selection based on a normalized difference measure for text classification / A. Rehman, K. Javed, H.A. Babri // Information Processing & Management. – 2017. – Vol. 53(2). – P. 473-489.

[5] Zaatour, R. Impact of Feature Extraction and Feature Selection Techniques on Extended Attribute Profile-based Hyperspectral Image Classification / R. Zaatour, S. Bouzidi, E. Zagrouba // VISIGRAPP (4: VISAPP). – 2017. – P. 579-586.

[6] Deng, X. Feature selection for text classification: A review / X. Deng, Y. Li, J. Weng, J. Zhang // Multimedia Tools and Applications. – 2019. – Vol. 78(3). – P. 3797-3816.