ИМИТАЦИОННАЯ МОДЕЛЬ ДЛЯ ТЕСТИРОВАНИЯ АЛГОРИТМОВ ОТБОРА ПРИЗНАКОВ

Волков А. В., Яцков Н. Н., Гринев В. В.
Кафедра системного анализа и компьютерного моделирования, кафедра генетики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: andrei@cybergizer.com, yatskou@bsu.by, grinev@bsu.by

Предложен алгоритм имитационного моделирования для генерации многомерных наборов данных, учитывающий информативность групп признаков. Представлены результаты исследования алгоритмов отбора признаков на смоделированных данных.

Введение

Исследование алгоритмов отбора признаков [1] на тестовых наборах данных позволяет определять достоинства и недостатки алгоритмов. В настоящее время в научном сообществе наиболее популярным является тестирование существующих и разрабатываемых алгоритмов отбора признаков на реальных наборах данных. В качестве примера следует выделить репозиторий тестовых наборов данных [2] для различных предметных областей, включающих анализ текстовой информации, изображений и биологических данных. Недостатком данного подхода является невозможность апробирования алгоритмов в широком спектре экспериментальных условий. Кроме того, оптимальный набор признаков для подобного рода наборов данных зачастую является неизвестным. Задачу выбора наиболее подходящего алгоритма можно упростить, используя имитационное моделирование и синтетические данные [3,4]. Преимуществами данного подхода являются:

1. возможность варьирования параметров исследуемых наборов данных (таких как число информативных, шумовых и избыточных признаков, число объектов наблюдения, задание уровня шума для меток классов, регулирование степени информативности признаков) облегчает получение обоснованных выводов по результатам проведенного анализа;
2. априорное знание оптимального набора признаков позволяет осуществлять надежную и автоматизированную оценку степени близости полученного результата отбора признаков к истинному.

Существующие имитационные модели генерации кластеров многомерных данных имеют ряд ограничений [5], среди которых наиболее существенным является невозможность явной задачи степени информативности признака или групп признаков.

Цель работы – разработка имитационной модели учитывающей информативность признаков объектов данных.

I. Методология

Входные параметры имитационной модели: число кластеров в генерируемом наборе данных; число групп признаков с заданной степенью разделимости между кластерами; число шумовых признаков; число избыточных признаков; уровень шума в метках классов; размеры кластеров. В рамках разработанной модели доступны два алгоритма генерации кластеров данных с заданной степенью разделимости:

1. на основе размещения кластеров данных в вершинах гиперкуба, где величина длины ребра определяется заданным индексом разделимости [6].
2. на базе вычисления метрики свободного пространства между двумя произвольными кластерами [7,8];

Алгоритм осуществляет генерацию гауссовых кластеров со случайными ковариационными матрицами, определяющими форму кластеров.

Пример сгенерированного набора данных из трех кластеров представлен на рис. 1.

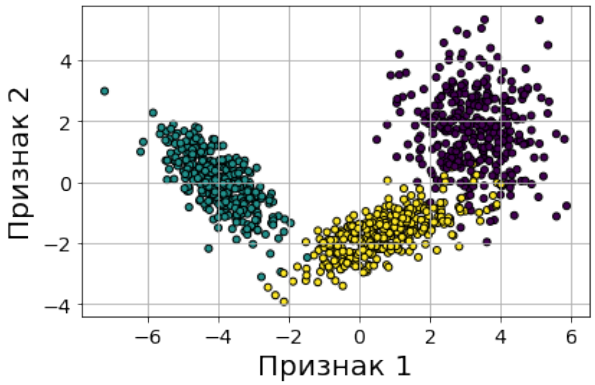

Рис. 1 – Пример сгенерированного набора данных из трех кластеров

Число атрибутов в рамках выделенной группы признаков является меньшим либо равным числу кластеров минус единица [9].

Блок схема алгоритма имитационного моделирования представлена на рис. 2.

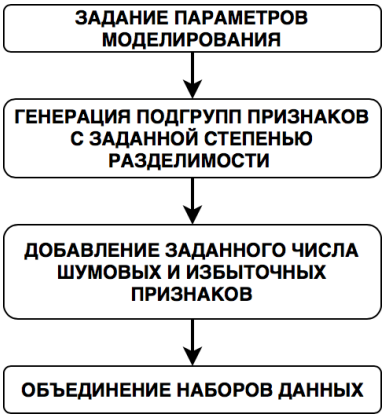

Рис. 2 – Блок схема алгоритма имитационного моделирования

Для проверки применимости предложенного подхода выбраны релевантные представители фильтрующих алгоритмов отбора признаков: алгоритм счета Фишера [10] и алгоритм ReliefF [11].

II. Результаты и обсуждение

Исследована применимость предложенного подхода имитационного моделирования. В ходе анализа сгенерированы наборы данных с различными степенями разделимостей. Работоспособность разработанного алгоритма имитационного моделирования подтверждена на примерах сравнительного анализа алгоритмов отбора признаков. На рис.3 представлен пример значений счета Фишера для набора параметров:

1. подход на основе гиперкуба для генерации трех групп признаков со значениями индекса разделимости 1.2, 0.8 и 0.3 соответственно;
2. 300 наблюдений в каждом кластере;
3. отсутствие дополнительных избыточных и шумовых признаков, а также шума в метках классов;
4. 15 классов;
5. 14 признаков в каждой группе;

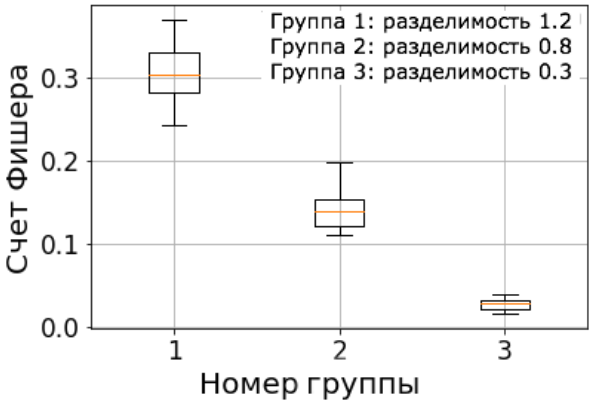

Рис. 3 – Пример диаграммы значений счета Фишера

Наличие статистически значимого различия средних значений счета Фишера для каждой из групп в приведенном примере подтверждается результатом однофакторного дисперсионного анализа: значение p-величины составляет 7.5×10^{-23}).

III. Заключение

В настоящей работе предложена и исследована имитационная модель для задач тестирования алгоритмов отбора признаков. Уникальной особенностью модели является генерация многомерных наборов данных с заданной степенью разделимости в рамках групп признаков. Результаты исследования алгоритмов отбора признаков на смоделированных данных демонстрируют адекватность разработанной модели.

IV. Список литературы

1. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. arXiv preprint arXiv:1601.07996, 2016
2. Feature selection data set repository [Electronic resource] / Jundong Li et al. // Arizona State University. Mode of access: http://featureselection.asu.edu/datasets.php. Date of access: 25.09.2018.
3. V. Bolon-Canedo, N. Sanchez-Maroño, A. Alonso-Betanzos, “An ensemble of filters and classifiers for microarray data classification,”Journals Pattern Recognition, vol. 45, pp. 531-539, 2012.
4. Belanche L.A., Gonzalez FF, Review and evaluation of feature selection algorithms in synthetic problems. http://arxiv.org/abs/1101.2320
5. P. Ghosh, Arka, Ranjan Maitra and D. Anna Peterson, “A Separability Index for Distance-based Clustering and Classification Algorithms”, 2012.
6. I. Guyon, “Design of experiments for the NIPS 2003 variable selection benchmark”, 2003.
7. W. Qiu and H. Joe. Generation of random clusters with specified degree of separation. Journal of Classification, 23:315–334, 2006.
8. W. Qiu and H. Joe. Separation index and partial membership for clustering. Computational Statistics and Data Analysis, 50:585–603, 2006.
9. McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1.
10. P. E. H. R. O. Duda and D. G. Stork. Pattern Classification. Wiley-Interscience Publication, 2001.
11. M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. Machine Learning, 53(1-2):23–69, 2003.