ИМИТАЦИОННАЯ МОДЕЛЬ КЛАСТЕРОВ МНОГОМЕРНЫХ ДАННЫХ С УЧЕТОМ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ ОБЪЕКТОВ

В. Н. Гребенник

Белорусский государственный университет, Минск
leragrebennik@yandex.com
науч. рук. – П.П. Коржуков, ст. преп.

В настоящей работе предложен и исследован алгоритм для генерации многомерных наборов данных, которые учитывают информативность подгрупп признаков. Выполнен сравнительный анализ алгоритма счета Фишера, Relief-F и «минимальная избыточность – максимальная релевантность» на смоделированных данных, которые показывают работоспособность модели в контексте данной работы. Рассмотренные методы способны отбирать информативные признаки в многомерных данных, которые имеют большое количество шумовых признаков. Работа позволяет снизить объем обрабатываемых данных, что дает возможность получить значительно более простые прогностические модели, что повышает их интерпретируемость.

Ключевые слова: отбор признаков, имитационное моделирование, шумовые признаки.

ВВЕДЕНИЕ

Активно развивающиеся информационные технологии и оборудование, позволяют получать многомерные наборы данных. Многомерные наборы данных представляют собой некоторые объекты исследования, характеризуемые набором признаков [1]. Основная часть признаков в таких данных неинформативна, что приводит к снижению точности анализа и к высокому потреблению вычислительных ресурсов [1].

Для выбора информативных признаков используются методы автоматического отбора признаков [2]. Существующие методы достаточно просты и направлены на анализ несложных данных. Однако проблема заключается не в методах, их можно улучшить, зная точно, что требуется найти в наборах данных, а в использовании адекватных наборов экспериментальных или смоделированных данных, позволяющих тщательно исследовать алгоритмы методов

Цель работы – разработка имитационной модели кластеров многомерных данных с учетом информативности признаков для тестирования алгоритмов автоматического отбора признаков.

МЕТОДОЛОГИЯ

Входными параметрами имитационной модели являются: число кластеров в генерируемом наборе данных; число не шумовых, шумовых и избыточных признаков; степень разделимости кластеров; значение или диапазон значений размеров кластеров. На рис. 1 представлена блок-схема имитационного алгоритма.

В основе модели лежит мера или степень разделимости близлежащих кластеров. В качестве метрики используется индекс разделимости Цю и Джо [3].

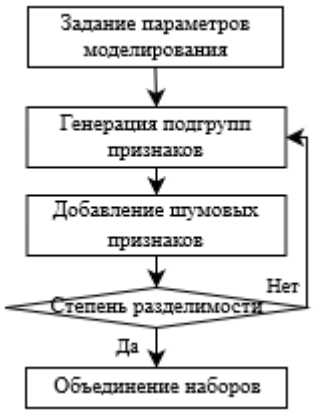

Рис. 1. Блок-схема имитационного алгоритма

Рассмотрены кластеры данных, характеризующиеся различной степенью разделимости. Индекс разделимости для хорошо разделяющихся кластеров положен равным 0,64, для удовлетворительно разделяющихся кластеров – 0,34, и для плохо разделяющихся кластеров – 0,1 (за основу взята экспертная оценка).

Для проверки применимости предложенного подхода выбраны следующие методы отбора признаков:

- алгоритм счёта Фишера [4],
- алгоритм Relief-F [5],
- алгоритм минимальная избыточность - максимальная релевантность [6],
- алгоритм быстрого фильтр на основе корреляции (FCBF) [7].
- Данные алгоритмы позволяют осуществлять ранжирование признаков по значимости и являются вычислительно эффективными.

Метод классификации *k*-ближайших соседей не имеет каких-либо внутренних механизмов отбора признаков, поэтому и выбран в качестве оценки точности классификации на заданном наборе признаков [8].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

С целью, проведения сравнительного анализа алгоритмов отбора признаков в условиях большого числа шумовых признаков был сгенерирован набор данных, содержащий 40 информативных и 10 шумовых признаков. На рис. 2 представлена матрица рассеивания 2 кластеров, индекс разделимости равен 0,64, атрибуты x13 и x15 являются шумовыми.

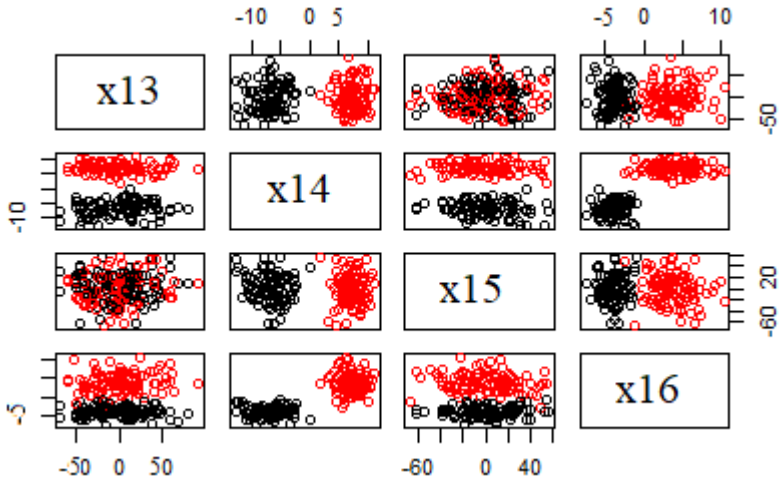

Рис. 2. Матрица рассеивания четырех признаков, кластеры имеют хорошую разделимость. X13 и X15 шумовые признаки

В таблице 1 представлена частота отбора шумового признака исследуемыми методами в условиях разной разделимости кластеров.

Таблица 1

Частота отбора шумовых признаков методами

Индекс разделимости	Счет Фишера	Relif F	mRMR	FCBF
0,64	0	0,4	0,3	0
0,34	0	0,8	0,4	0
0,1	0	0,8	0,4	0

Из таблицы можно сделать вывод, что методы отбора: счет Фишера и FCBF за все время проведения эксперимента показали хорошие результаты, не отобрав ни одного шумового признака. Плохой результат показал метод Relif F, так как вероятность отбора шумовых признаков достаточно высока. Из этого можно сделать вывод, что критерий отбора признаков не справляется с поставленной задачей. Неплохой результат показал метод mRMR. В отличие от Relif F, вероятность отбора шумового признака при удовлетворительном разделении кластеров достаточно мала.

ЗАКЛЮЧЕНИЕ

В данной работе предложена и исследована имитационная модель для тестирования алгоритмов отбора признаков. Уникальной особенностью модели является генерация многомерных наборов данных с заданной степенью разделимости и различной информативности признаков. Результаты исследования алгоритмов отбора признаков на смоделированных данных демонстрируют применимость модели.

Исследованы подходы [4], [5], [6] и [7] для генерации. Методы демонстрируют свою применимость в контексте разработанной имитационной модели.

Библиографические ссылки

1. *Hanan Samet* Foundations of Multidimensional and Metric Data Structures // Morgan Kaufmann Publishers – 2006. ISBN 13: 978-0-12-369446-1
2. *L.Ladha* Feature Selection Methods And Algorithms // International Journal on Computer Science and Engineering Vol. 3 May 2011 - IJCSE11-03-05-051, ISSN : 0975-3397
3. *Qiu,W.L.* Generation of Random Clusters with Specified Degree of Separation / W.L. Qiu, H. Joe //Journal of Classification. – Vol. 23(2). – 2005. – P.315–334. DOI: 10.1007/s00357-006-0018-y
4. *Quanquan Gu, Zhenhui Li, Jiawei Han* Generalized Fisher Score for Feature Selection // Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011
5. *Marko Robinik-Sikonja, Igor Kononenko* Theoretical and Empirical Analysis of ReliefF and RReliefF // Kluwer Academic Publishers. Manufactured in The Netherlands – Machine Learning, 53, 23–69, 2003.
6. *Zhenyu Zhao, Radhika Anand, Mallory Wang* Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. DOI: 10.1109/DSAA.2019.00059
7. *Lei Yu, Huan Liu* Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003
8. *Oliver Sutton* Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction // February, 2012. Corpus ID: 18781499