Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ
Заведующий кафедрой

_____ В. В. Шайдуров

«___» _____ 2018 г.

БАКАЛАВРСКАЯ РАБОТА
Направление 02.03.01 Математика и компьютерные науки

ВЫЧИСЛЕНИЕ MIC НА ОСНОВЕ КЛАСТЕРИЗАЦИИ ДАННЫХ

Руководитель _____ канд. филос. наук, доцент Б. В. Олейников
подпись, дата
Выпускник _____ С. Р. Гречищева
подпись, дата

Красноярск 2018

РЕФЕРАТ

Выпускная квалификационная работа на тему «Вычисление *MIC* на основе кластеризации данных» содержит 56 страниц текстового документа, 5 приложений, 26 использованных источника.

ВЗАИМОСВЯЗЬ, НЕЛИНЕЙНАЯ ЗАВИСИМОСТЬ, MIC, КЛАСТЕРИЗАЦИЯ ДАННЫХ, ВЗАИМНАЯ ИНФОРМАЦИЯ, КОРРЕЛЯЦИЯ, ОПТИМИЗАЦИЯ, АЛГОРИТМ, КОЭФФИЦИЕНТ

Цели работы:

- обзор современных коэффициентов, измеряющих нелинейную зависимость признаков;
- подробный обзор *MIC* и существующих алгоритмов его вычисления;
- разработка нового алгоритма вычисления *MIC* с меньшей вычислительной сложностью;
- разработка программного обеспечения алгоритма.

В настоящей работе были рассмотрены существующие коэффициенты взаимосвязи двух признаков, приведена сравнительная таблица некоторых из коэффициентов, подробно рассмотрен *MIC* и его свойства, выявлены его преимущества и недостатки, обзор некоторых алгоритмов его вычисления.

В результате был разработан новый алгоритм подсчета *MIC*, в основе которого лежит кластеризация данных. Его эффективность при большой выборке значительно выше, чем у алгоритма прямого подхода. Полученный алгоритм гибок для модификаций.

ВВЕДЕНИЕ

Проблема измерения взаимосвязи между переменными – она из фундаментальных задач статистики.

В статистике взаимосвязи по своей форме делятся на две группы: линейные и нелинейные (криволинейные). Если первая группа хорошо изучена и традиционно измеряется коэффициентом линейной корреляции, то вторая группа значительно сложнее. Она включает в себя обширные типы связей (экспоненциальная, периодическая, полиномиальные и пр.), и это влечет за собой сложности в распознавании многих типов связей, особенно в тех ситуациях, когда зависимость затруднительно предсказать заранее.

Широко используются коэффициенты Пирсона (линейная корреляция) для шкалы отношений, а для шкал порядка – ранговый коэффициент Спирмана, тау Кендалла, коэффициент Гудмена, меру Сомерса. Они вычислительно эффективны и теоретически понятны, но в то же время очень сильно ограничены в классах рассматриваемых связей, в основном линейные или монотонно возрастающие связи. Нелинейные связи ими либо плохо идентифицируются, либо не идентифицируются вовсе. Поэтому ученые стремятся разработать новые коэффициенты, способные идентифицировать нелинейные взаимосвязи характеристик.

Одним из результатов такой работы служит максимальный информационный коэффициент, или *MIC* (англ. *maximal information coefficient*) разработанный в 2012 году Дэвидом Решефом в сотрудничестве с другими учеными. Решеф продемонстрировал, что *MIC* чувствителен ко многим нелинейным взаимосвязям, например, синусоидальным, экспоненциальным, кубическим [4]. Его разработка показала , что вместо того, чтобы делать предположения о типе связи, можно перейти к дискретной модели и использовать элементы теории информации.

Введенный им коэффициент *MIC* измеряет зависимость между двумя признаками, по сути вычисляя их взаимную информацию, но с оговорками: во-первых, само множество, в котором представлены признаки, подвергается множественным разбиениям; во-вторых, взаимная информация нормализуется величиной, зависящей от начальной выборки.

Поиск всех возможных вариантов разбиения множества – затратная задача при больших объемах данных и экспоненциально возрастает с увеличением выборки. С целью оптимизации этого процесса в настоящей работе предложена идея разбиения множества с использованием предварительной кластеризации данных, а также ее реализация.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Габидулин Э. М., Пилипчук Н. И. Лекции по теории информации – МФТИ, 2007. – 214 с.
2. Федоряева Т.И. Комбинаторные алгоритмы: Учебное пособие / Новосиб. гос. ун-т. Новосибирск, 2011.– 118 с.
3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.
4. Detecting Novel Associations in Large Data Sets : науч. статья / D. N. Reshef [и др.]. – 2012. – 8 с.
5. Pearson correlation coefficient [Электронный ресурс] : опред. коэффициента Пирсона // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.
6. Correlation ratio [Электронный ресурс] : опред. корреляционного отношения // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.
7. Linearization of Local Probabilistic Sensitivity via Sample Re-Weighting : науч. статья / R.M.Cooke, D. Kurowicka, I. Meilijson. – Delft, The Netherlands, 2003. –13 с.
8. Brownian Distance Covariance : науч. статья / G. J. Szekely, M. L. Rizzo. – Bowling Green State University, 2009. – 30 с.
9. Distance correlation [Электронный ресурс] : св-ва корреляционного расстояния // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.
10. Aspiras-Paler, M. E. On Modern measures And Test Of Multivariate Independence : дис. DPh : / Mary Elvi Aspiras-Paler. – 2015. – 124 с.
11. Lopez-Paz, D. The Randomized Dependence Coefficient : науч. статья / D. Lopez-Paz, P. Hennig, B. Scholkopf. – Max Planck Institute for Intelligent Systems, 2013. – 9 с.
12. Cover, T. M. Elements of Information Theory (Wiley ed.) : учеб. пособие / T. M. Cover, J. A. Thomas. – 1991. – 563 с.
13. Mutual Information [Электронный ресурс] : основание логарифма // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.
14. Pointwise mutual information [Электронный ресурс] : опред. поточечной взаимной информации // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.
15. Bouma, G. Normalized (Pointwise) Mutual Information in Collocation Extraction : науч. статья / G. Bouma .– Potsdam, 2009. – 11 с.
16. Entropy(information theory) [Электронный ресурс] : опред. энтропии // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.

17. Conditional entropy [Электронный ресурс] : опред. условной энтропии // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.

18. New methods for finding associations in large data sets: generalizing the maximal information coefficient (MIC) : науч. статья / T. Ignac, N. Sakhanenko, A. Skupin, D. Galas. – Люксембург, 2012. – 5 с.

19. Пат. WO2013067461A2. Identifying associations in data / D. N. Reshef, Y. A. Reshef ; заявитель и патентообладатель ; заявл. 04.11.2011 ; опубл. 03.11.2012. – 26 с.

20. Minepy 0.3.5 documentation [Электронный ресурс] // «Minepy documentation». – Режим доступа: http://minepy.sourceforge.net.

21. RapidMic: Rapid Computation of the Maximal Information Coefficient : науч. статья / D. Tang, M. Wnag, W. Zheng, H. Wang .– 2014 .– 11 с.

22. A Novel Algorithm for the Precise Calculation of the Maximal Information Coefficient : науч. статья / Y. Zhang, S. Jia, H. Huang, J. Qiu, C. Zhou. – Hebei : Departament of Mathematics, 2014. – 5 с.

23. Bolte, A. Optimization simulated annealing schedules with genetic programming / A. Bolte, U. W. Thonemann // European Journal of Operational Research. – 1996. – vol. 92.– №2. 402-416 с.

24. SGMIC | Simulated annealing and Genetic Maximal Information Coefficient [Электронный ресурс] // «OmicTools». – Режим доступа: https://omictools.com.

25. Stirling numbers of the second kind [Электронный ресурс] // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.

26. Binomial coefficient [Электронный ресурс] // Свободная энциклопедия «Wikipedia». – Режим доступа: https://en.wikipedia.org.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ
Заведующий кафедрой

_____ В. В. Шайдуров

«8» июня 2018 г.

БАКАЛАВРСКАЯ РАБОТА
Направление 02.03.01 Математика и компьютерные науки

ВЫЧИСЛЕНИЕ MIC НА ОСНОВЕ КЛАСТЕРИЗАЦИИ ДАННЫХ

Руководитель _____ 08.06.18 канд. филос. наук, доцент Б. В. Олейников
подпись, дата

Выпускник _____ 08.06.18 С. Р. Гречищева
подпись, дата

Красноярск 2018