К ВОПРОСУ О ПОДГОТОВКЕ ДАННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧ DATA MINING

Е.Н. Живицкая
Проректор по учебной работе БГУИР, кандидат технических наук, доцент

А.Т. Кусаинова[1]
Докторант Евразийского национальный университет имени Л.Н. Гумилева, магистр технических наук

В.А. Пархименко
Заведующий кафедрой экономики БГУИР, кандидат экономических наук, доцент

М.М. Татур
Профессор кафедры электронных вычислительных машин БГУИР, доктор технических наук, профессор

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
[1]Евразийский национальный университет им. Л.Н. Гумилева, Республика Казахстан
Email: jivitskaya@bsuir.by, ainurkussainova89@gmail.com, parkhimenko@bsuir.by, tatur@bsuir.by

Abstract. In the article, the problem of data preparation for data mining process has been described. Usually this stage of data preprocessing is skipped contrary to theoretical recommendations. Much more attention usually is paid to sophisticated methods and algorithms. This could lead to outcomes without any applicable meaning. In the article, authors have offered several ideas on dealing with main problems within data preparation step.

Введение. В учебной литературе по интеллектуальному анализу данных, как правило, излагаются классические формальные алгоритмы, а в научной литературе – их модификации и глубокое исследование различных аспектов их применения. При этом вопросам подготовки данных несправедливо уделяется мало внимания. Однако несложно показать, что подготовка (качество подготовки) данных может оказывать значительно более кардинальное влияние на конечный результат, нежели выбранный метод или модификация алгоритма.

Если рассмотреть (рис. 1) уже ставшее хрестоматийным изложение процесса Data Mining в рамках методологии CRISP-DM (Cross Industry Standard Process for Data Mining), то можно отметить, что в теории этап непосредственного применения конкретных алгоритмов Data Mining (на рисунке отмечено как «Modelling») является лишь одним из 6 этапов и носит соподчиненный характер по отношению к пониманию прикладных проблем из предметной области («Business understanding») и реализации принятых по итогам анализа решений («Deployment»).

В то же время, как показывает опыт, на практике в рамках многих проектов по Data Mining акцент переносится, напротив, сугубо на использование конкретных алгоритмов, а другие этапы (в том числе важный этап подготовки данных – Data preparation) опускаются.

В настоящей работе авторы затрагивают некоторые важные вопросы именно этого этапа, в частности проблему подготовки логических признаков и процедуру взвешивания и нормализации данных.

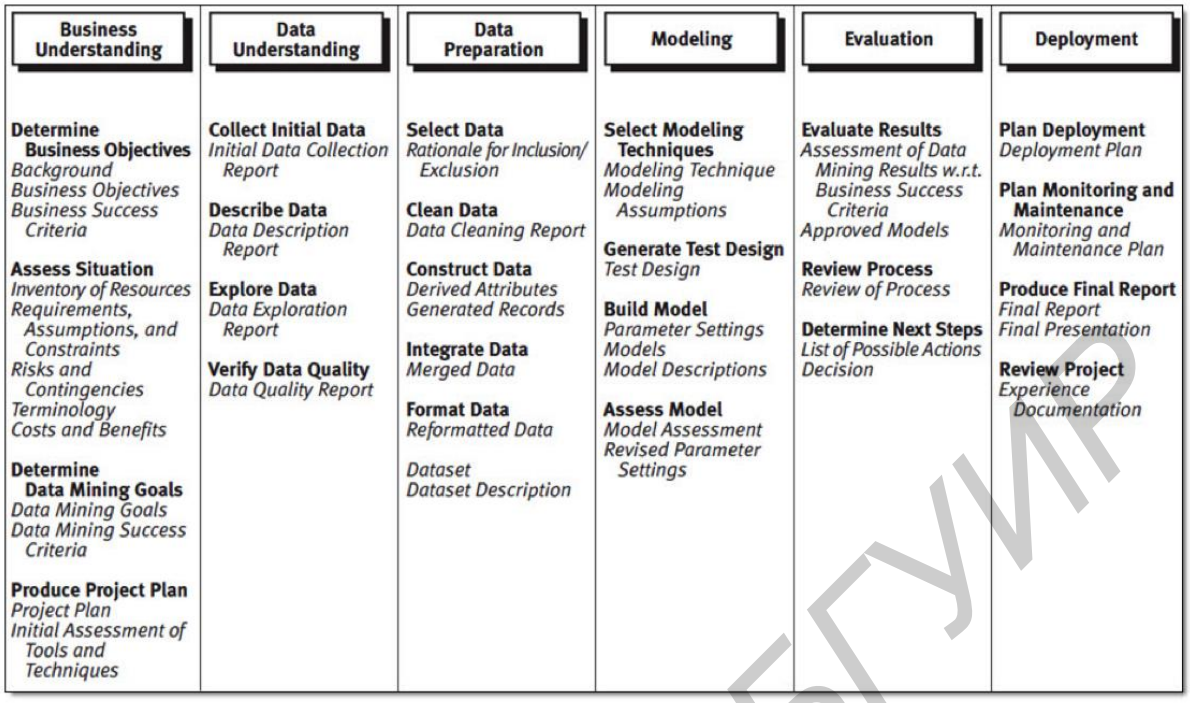

Рис. 1. Этапы, задачи и результаты Data Mining в соответствии с методологией CRISP-DM [1]

1. О совместном использовании логических и количественных признаков. Общепринято в анализируемых данных выделять количественные, логические, категориальные, порядковые и некоторые другие типы. Однако при рассмотрении конкретных алгоритмов Data Mining (например, кластеризации k-средних, классификации k-ближайших соседей и т.п.) речь идет исключительно о количественных данных. Как быть, если в задачах присутствуют данные различных типов [2], в частности логические?

Для ответа на этот вопрос рассмотрим формальный пример. Пусть имеется 3 объекта (образа), каждый из которых представлен двумя информативными признаками, при этом оба признака логические (признак либо присутствует, либо отсутствует у конкретного объекта).

Таблица 1 – Исходные данные

x_1	x_2	№ образа
0	0	–
0	1	1
1	0	3
1	1	2

$$O_1 = \overline{x_1}\, x_2$$
$$O_2 = x_1\, x_2$$
$$O_3 = x_1\, \overline{x_2}$$

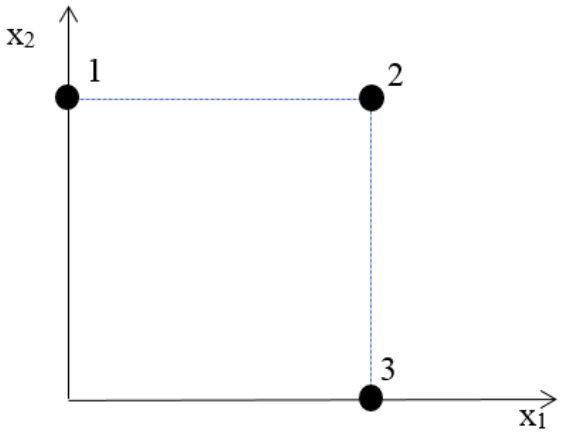

Рис. 2. Распределение образов с «чистыми» логическими признаками

Исходные данные по всем трем образам приведены в табл. 1. При их геометрической

интерпретации (рис.2) образы окажутся расположенными в вершинах квадрата (в общем случае, при наличии большого числа признаков – в вершинах гиперкуба). Максимальное число различимых образов – 2n, где n – число информативных признаков или размерность признакового пространства.

В такой постановке положение каждого из образов может быть описано булевым выражением, а кодовое расстояние между ними, при необходимости, может быть выражено в метрике Хэмминга.

Например, в коробке имеются детали, различающиеся по размеру – x_1 (большие и малые) и форме – x_2 (круглые и продолговатые). Необходимо записать правило принятия решения для автомата-сортировщика.

В данном случае признаки с точки зрения принятия решений являются «взаимно нейтральными»: большие круглые детали не лучше, не важнее, не значимее, чем, например, малые и продолговатые. Поэтому придание указанным признакам весовых значений лишено смысла, а булевы описания образов O_i и будут являться правилами принятия решений, чисто логических решений.

Но как только мы устанавливаем между этими признаками некоторое количественное соотношение, влияющее на принятие решения, тогда возникает подтип логических признаков, для которых правомерно оценивать геометрическое расстояние между образами.

Например, на ферме разводят кроликов. Имеющиеся кролики различаются по цвету (белый, серый) – x_1 и полу – x_2. Необходимо их сравнить в плане коммерческой (селекционной или др.) выгоды. Пусть признак x_1 при принятии решений в три раза важнее, чем признак x_2, тогда их взаимное расположение в пространстве (в частности, на плоскости) будет выглядеть, как представлено на рис.3. А значит, мы имеем все основания, чтобы вычислить евклидовы расстояния между образами и решать задачи кластеризации, ранжирования и др. известными алгоритмами.

Рис. 3 Распределение образов со взвешенными логическими признаками

Продолжим пример. Пусть один из признаков (x_1) будет количественным, и мы к имеющимся трем добавим пятый и шестой образы, с параметрами, указанными в табл. 2. Так как признак x_1 остается в соответствии с условием важнее в три раза признака x_2, то распределение признаков графически можно будет представить таким, как показано на рис.4.

Если поменять соотношение весов признаков на противоположное (признак x1 станет менее весомым, чем признак x2 в три раза), то получим распределение, как показано на рис.5.

Таблица 2 – Дополнительные образы

x_1	x_2	№ образа
0	0	–
0	1	1
1	0	3
1	1	2
0,3	1	4
0,5	0	5

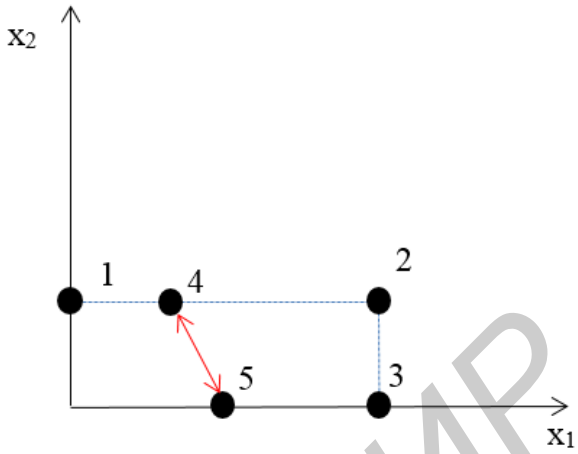

Рис. 4. Распределение образов с количественным и взвешенным логическим признаками, в соотношении 1:3

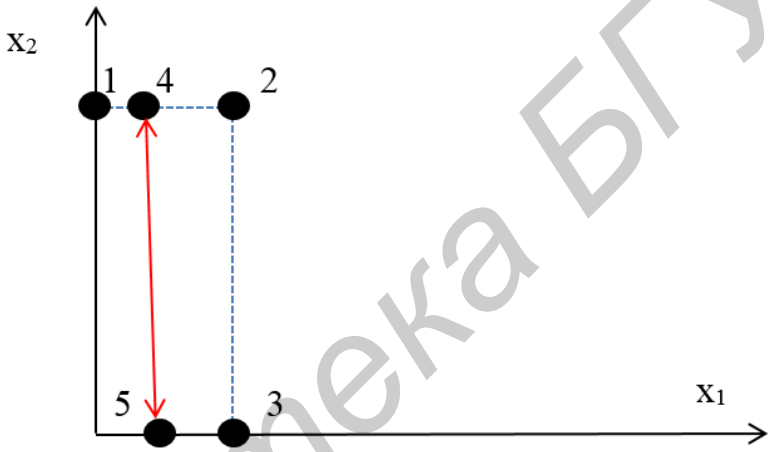

Рис.5. Распределение образов с количественным и взвешенным логическим признаками, в соотношении 3:1

Очевидно, что соотношения евклидовых расстояния между образами на рис. 4 и рис. 5 диаметрально противоположные:

$$\rho E\,(4,5) < \rho E\,(3,5),\ \rho E\,(4,5) < \rho E\,(2,4);$$

$$\rho E\,(4,5) > \rho E\,(3,5),\ \rho E\,(4,5) > \rho E\,(2,4),$$

а, следовательно, при решении задач Data Mining будут получены противоположные результаты.

2. Методические рекомендации по взвешиванию и нормализации информативных признаков. Давать строгие рецепты по взвешиванию и нормализации данных в виде методик и алгоритмов, по всей видимости, не возможно и даже не желательно [3]. Дело в том, что часть этапов подготовки данных, по своей сути, вообще не поддается формализации, а оставшаяся изобилует различными оговорками и частными случаями.

Поэтому заявленные «методические рекомендации» по взвешиванию и нормализации информативных признаков следует понимать, как некоторые эвристические правила, которые необходимо держать в поле зрения аналитику данных, особенно начинающему свою профессиональную карьеру.

Подобные правила авторы формулируют следующим образом:

1 Ранжировать признаки и установить веса означает, что необходимо сопоставить значимость всех используемых признаков в контексте конкретной решаемой задачи. Простейшим примером метода определения величины веса – является метод экспертных оценок. Однако не стоит забывать, что в некоторых случаях (при наличии априорной информации, репрезентативной выборки прецедентов и т.п.) значения весов могут быть «вычислены» в результате решения задачи Data Mining.

2 Чтобы унифицировать пользовательский интерфейс вычислительных систем при решении задач с различными диапазонами весов информативных признаков рекомендуется нормализовать область допустимых значений весов, т.е. вес максимально-значимого признака принять за единицу, а, минимально-значимого за ноль.

3 Изначально каждый из признаков представлен в оригинальной системе отсчета, исчисления, условных единицах и т.п. Для его использования в алгоритмах Data Mining в общем случае необходимо нормализовать одним из способов:

3.1 $x_i/(x_{max}-x_{min})$, т.е. текущее значение приводится к диапазону, вычисляемому из выборки;

3.2 $x_i/\Delta x$, т.е. текущее значение приводится к априори заданному диапазону Δx.

В зависимости от способа нормализации (3.1, 3.2) могут быть получены формально противоречивые результаты. Поэтому осуществлять нормализацию следует в контексте решаемой прикладной задачи.

Заключение. Логические признаки следует разделять на чисто логические и взвешиваемые, в зависимости от специфики решаемой задачи. Те и другие имеют фиксированное число кодовых комбинаций – 2^n, определяющее максимальное число распознаваемых образов.

Взвешиваемые логические признаки могут наряду с количественными информативными признаками использоваться для описания образов и без ограничений использоваться в алгоритмах Data Mining.

В ходе подготовки данных все признаки рекомендуется ранжировать и нормализовать как внутри собственной шкалы (области допустимых значений) так и глобально.

Некорректное соотнесение (взвешивание) информативных признаков, а также обработка ненормализованных данных может кардинально изменить результаты анализа, порой в большей степени, нежели выбор модификации примененного алгоритма Data Mining.

Вопросы геометрической интерпретации образов с категориальными признаками и их совместного использования с количественными станет предметом обсуждения в следующей работе. Еще одной важной проблемой на этапе подготовки данных является проблема выбора информативных признаков и связанная с ней проблема «очистки данных». В совокупности они напрямую влияют на результаты интеллектуального анализа и будут рассмотрены в дальнейших публикациях.

Литература

[1]. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth. CRISP-DM 1.0: Step-by-step data mining guides [Электронный ресурс]. – Режим доступа: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[2]. S. Ray. Simple Methods to deal with Categorical Variables in Predictive Modeling. [Электронный ресурс]. – Режим доступа: https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling

[3]. L. A. Shalabi, Z. Shaaban, B. Kasasbeh. Data Mining: A Preprocessing Engine // Journal of Computer Science. – 2006. – №2. – P. 735-739.