BIG DATA, CLUSTER ANALYSIS AND OPTIMIZATION IN SYSTEM ANALYSIS

Filipchik E.[1], Perskevich D.[2], German O.[3]

BIG DATA, КЛАСТЕРНЫЙ АНАЛИЗ И ОПТИМИЗАЦИЯ В СИСТЕМНОМ АНАЛИЗЕ

Филипчик Е. Ф.[1], Перскевич Д. Т.[2], Герман О. В.[3]

[1]Филипчик Егор Федорович / Filipchik Egor – магистрант;
[2]Перскевич Денис Тадеушевич / Perskevich Denis – магистрант,
кафедра информационных технологий автоматизированных систем, направление: системный анализ, управление и обработка информации;
[3]Герман Олег Витольдович / German Oleg – кандидат технических наук, доцент,
Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Аннотация: *в данной работе нами предложена техника оптимизации для некоторого диапазона практических задач большой размерности. Идея этого подхода состоит в том, чтобы выполнять расчеты на эталонных представителях кластеров, на которые разбиваются входные экземпляры, а не на индивидуальных многомерных объектах. Число кластеров мы делаем как можно большим, но удовлетворяющим некоторым априорным ограничениям. Эта идея позволяет снять ограничения на размерность решаемых задач, например, в EXCEL (Поиск решения).*

Abstract: *in this paper, we proposed the optimization technique for a range of practical large scale problems. The idea of this approach is to perform calculations on the reference representatives of the clusters, which split the input instances, and not on the individual multidimensional objects. The number of clusters we make as large as possible, but satisfy some a priori constraints. This idea allows to remove restrictions on the dimensions of tasks, for example, in EXCEL (solver).*

Ключевые слова: *BIG DATA, кластерный анализ, системный анализ, метод Саати.*
Keywords: *BIG DATA, cluster analysis, system analysis, method of Saaty.*

ВВЕДЕНИЕ.

Обработка больших массивов данных в задачах оптимизации может быть сопряжена с весьма большими трудностями. Например, задача линейного программирования с десятками и сотнями тысяч неравенств может даже не быть формализована в таких широко используемых пакетах как EXCEL PROBLEM SOLVER. Вместе с тем, для практических целей вполне можно использовать не оптимальное, но близкое к нему решение. Будем рассматривать следующую задачу в качестве иллюстрации. Пусть дана обучающая таблица очень большого размера вида

Таблица 1. Исходная обучающая таблица

№№	X1	X2	X3		X*m*	Y
1	x_{11}	x_{12}	x_{13}		x_{1m}	Y_1
2	x_{21}	x_{22}	x_{23}		x_{2m}	Y_2
3	x_{31}	x_{32}	x_{33}		x_{3m}	Y_3
4	x_{41}	x_{42}	x_{43}		x_{4m}	Y_4
…						
N	x_{N1}	x_{N2}	x_{N3}		x_{Nm}	Y_N

Здесь X1, X2, …, X*m* – критерии (факторы, атрибуты) входных многомерных объектов; Y – выходной (результирующий) признак. Для ясности, пусть значение $Y \geq 0$ соответствует классу A, а везде, где $Y < 0$ объект не относится к классу A. Для удобства пусть первые *t* строк таблицы соответствуют объектам класса A, а остальные строки *t*+1, …, *N* соответствуют другим классам, отличным от A. Задача формулируется так:

$$L = p_1 \cdot (a_1)^2 + p_2 \cdot (a_2)^2 + \ldots + p_m \cdot (a_m)^2 \rightarrow \max$$

$$a_1 \cdot x_{11} + a_2 \cdot x_{12} \quad \ldots. + \quad a_m \cdot x_{1m} \geq 0$$
$$a_1 \cdot x_{21} + a_2 \cdot x_{22} \quad \ldots. + \quad a_m \cdot x_{2m} \geq 0$$

$$a_t \cdot x_{t1} + a_2 \cdot x_{t2} \quad \ldots. + \quad a_m \cdot x_{tm} \geq 0$$
$$-a_1 \cdot x_{t+11} - a_2 \cdot x_{t+12} \quad \ldots. - a_m \cdot x_{t+1m} \geq -\varepsilon$$
$$-a_1 \cdot x_{N1} - a_2 \cdot x_{N2} \quad \ldots. - a_m \cdot x_{Nm} \geq -\varepsilon$$
$$\varepsilon \geq 0. \qquad (1)$$

Обсудим эту систему. Здесь p_1, p_2, …, p_m - приоритеты критериев. Приоритет тем выше, чем больше значение неизвестного коэффициента по модулю. Действительно, коэффициенты, вносящие незначительный вклад, должны иметь малый приоритет. Для выбора приоритетов критериев мы используем метод Т. Саати, как определено далее в этой работе.

Если целевую функцию вообще исключить, то, как нетрудно видеть, задача сводится к отысканию коэффициентов линейного распознавателя с *m* входами. Очевидно, что даже в этом случае система может не

иметь решения. Таким образом, два обстоятельства не позволяют в общем случае применить «стандартную» технику решения:

- число ограничений очень велико;
- система неравенств несовместна.

Описываемый здесь подход к решению заключается в том, чтобы разбить исходное множество многомерных объектов на кластеры следующим образом:

- в кластер попадают похожие объекты;
- число кластеров максимально возможно, но обеспечивает отыскание решения.

Последовательно рассмотрим реализацию нашего плана [3].

РАЗБИЕНИЕ НА КЛАСТЕРЫ.

Техника разбиения на кластеры достаточно богата [1-3]. Можно использовать любой известный метод из отмеченных. Нам нужно разбиение на заданное число кластеров K > 1. Начинаем с K=2. Этот случай как раз соответствует исходной таблице. Поэтому два кластера мы гарантированно имеем. Находим типичных (эталонных) представителей в каждом кластере. Получаем упрощенную систему:

$$L = p_1\cdot(a_1)^2+p_2\cdot(a_2)^2+\ldots+p_m\cdot(a_m)^2 \rightarrow \max$$
$$a_1 \cdot x_{1э} + a_2 \cdot x_{2э} \ldots + a_m \cdot x_{mэ} \geq 0$$
$$a_1 \cdot y_{1э} - a_2 \cdot y_{2э} \ldots - a_m \cdot y_{mэ} \geq -\varepsilon$$
$$\varepsilon \geq 0 \qquad (2)$$

Здесь класс A представлен эталоном ($x_{1э}, x_{2э}, \ldots, x_{mэ}$).

Далее переходим к трем кластерам: K=3. Для этого класс с наибольшим разбросом (внутриклассовой дисперсией) разбиваем на два подкласса и снова решаем задачу оптимизации. Если задача имеет решение, то продолжаем разбиение классов по аналогии: всегда делим класс с наибольшей внутриклассовой дисперсией на два кластера. Каждый из полученных кластеров представляется объектом-эталоном (т.е. эталон получается как объект с усредненными значениями признаков (критериев) в пределах класса). Этот процесс дихотомического деления ведем до тех пор пока выполняются два условия:

- число кластеров максимально возможно, но обеспечивает отыскание решения;
- эталонные представители кластеров отличаются не менее, чем на заданное пороговое значение. Данное требование означает, что нет смысла дробить кластеры на подкластеры с близкими значениями эталонных представителей. Выбор порогового значения можно из практических соображений определить на уровне 10–15% [1].

ВЫБОР ПРИОРИТЕТОВ КРИТЕРИЕВ.

Критерии имеют тем больший приоритет, чем больше они коррелируют с выходным признаком. Таким образом, приоритеты критериев оцениваются как коэффициенты корреляции ρ_{xy}. Вместе с тем, коэффициенты корреляции могут иметь отрицательные значения. Поэтому мы несколько изменим технику, введя в рассмотрение матрицу критериев Саати вида [2].

Таблица 2. Матрица критериев Саати

	X1	X2		X*m*
X1	1	$abs(\rho_{1,Y}/\rho_{2,Y})$		$abs(\rho_{1,Y}/\rho_{m,Y})$
X2	$abs(\rho_{2,Y}/\rho_{1,Y})$	1	…..	
…			…..	
X*m*	$abs(\rho_{m,Y}/\rho_{1,Y})$	$abs(\rho_{m,Y}/\rho_{2,Y})$		1

Отношения $abs(\rho_{w,Y}/\rho_{u,Y})$ приводятся к метрической шкале Т. Саати.

ИСПОЛЬЗОВАНИЕ ПОЛУЧЕННОГО РЕШЕНИЯ.

Для произвольного объекта ($z_1, z_2, \ldots, z_m$) определяем, к какому кластеру он наиболее близок. Затем используя формулу

$$a_1 \cdot z_1 + a_2 \cdot z_2 \ldots + a_m \cdot z_m = Y, \qquad (3)$$

находим значение выходной величины, соответствующей этому объекту.

ПРИМЕР.

Нас будет интересовать процесс расхода топлива и постановка прогноза, когда при данном графике расхода запасы топлива полностью иссякнут. Существует большое количество методов прогнозирования, но с помощью этих методов возможно контролировать процесс лишь по прошедшему периоду без учёта факторов, влияющих на формирование прогноза.

Таблица 3. Обучающая таблица

Дневной расход (x1)	Запас на складе (x2)	Дисперсия (x3)	Ожидаемый срок выработки (x4)	Реальный срок выработки Y
2	100	2	50	58
3	120	2	40	48
3	30	1,5	10	11
4	200	3	50	45
4	120	2	30	32
4	40	1	10	10
2	50	2	25	40
4	20	2	5	8

Пусть текущая ситуация характеризуется следующими данными:
Дневной расход – 3;
Запас на складе – 45;
Дисперсия – 2.
Ожидаемый срок выработки – 15.

Требуется получить оценку реального срока выработки ресурса. Заметим, что ожидаемый срок выработки вычисляется как запас на складе деленный на величину дневного расхода. В реальных вычислениях таблица может содержать тысячи и более строк. Согласно нашей концепции, разобьем записи на два кластера (сначала). В первый кластер попадут записи с реальным сроком выработки, превосходящим ожидаемый срок выработки на 5 и более дней. Пусть этот кластер будет A0. В кластер B0 попадут записи, где реальный срок отклоняется от ожидаемого не более, чем на 5 дней. Для расчетов ограничимся усеченной таблицей:

Таблица 4. Обучающая усеченная таблица

дневной расход (x1)	Дисперсия (x2)	ожидаемый срок выработки (x3)	реальный срок выработки Y
2	2	50	58
3	2	40	48
3	1,5	10	11
4	3	50	45
4	2	30	32
4	1	10	10
2	2	25	40
4	2	5	8

Составляем систему

$$\begin{gathered} L = a_1^2+a_2^2+a_3^2 \rightarrow \max \\ a_1 \cdot x_{1э} + a_2 \cdot x_{2э} + a_3 \cdot x_{3э} \geq 0 \\ -a_1 \cdot y_{1э} - a_2 \cdot y_{2э} - a_3 \cdot y_{3э} \geq -0.01 \end{gathered} \quad (4)$$

Координаты эталонных объектов для кластеров A0 и B0 находим как средние значения в соответствующих разрядах. В нашем случае имеем:

$$\begin{gathered} L = a_1^2+a_2^2+a_3^2 \rightarrow \max \\ a_1 \cdot 2.75 + a_2 \cdot 2.25 + a_3 \cdot 41.25 \geq 0 \\ -a_1 \cdot 3.75 - a_2 \cdot 1.67 - a_3 \cdot 13.75 \geq -0.01 \end{gathered} \quad (5)$$

Находим решение (ограничив значение L=100):
a_1 = -5.9; a_2= 8.0; a_3 = 0.78.
Подставляя исходные данные, найдем
-5.9*3+2*8+15*0.78=10>0.

Вывод: для исходных данных задачи реальный срок превзойдет ожидаемый более чем на 5 дней. Чтобы получить более точную оценку, нужно найти запись, где реальный срок превзошел ожидаемый не менее чем на пять дней, причем такую, которая «наиболее точно» соответствует исходным данным.

В таблице подходящей записью будет: 2, 50, 2, 25. Реальный срок выработки ресурса может составить порядка 40 дней.

Ясно, что в этом примере мы ограничились только двумя кластерами. Можно было бы продолжить «дробление» кластеров по схеме, описанной в этой статье.

ЗАКЛЮЧЕНИЕ

Нами предложена техника оптимизации для некоторого диапазона практических задач большой размерности. Идея подхода состоит в том, чтобы выполнять расчеты на эталонных представителях кластеров, на которые разбиваются входные экземпляры, а не на индивидуальных многомерных объектах. Число кластеров мы делаем как можно большим, но удовлетворяющим некоторым априорным ограничениям. Эта идея позволяет снять ограничения на размерность решаемых задач, например, в EXCEL (Поиск решения).

Литература

1. *Динг С., Хе Х.* K-средства кластеризации с помощью анализа главных компонентов: Труды двадцать первой Международной конференции по вопросам машинного обучения, 2004. С. 1–9.
2. *Маккуин Ж. Б.* Некоторые методы классификации и анализа многомерных наблюдений: Труды симпозиума по Беркли математической статистике и теории вероятности, 1967. С. 281–297.
3. *Эстер М., Кригель Н. Р., Сандер Ј., Виммер М., Ксю Х.* Инкременатальная кластеризация для добывания полезных ископаемых в среде хранилищ данных: Труды Международной конференции по очень большим базам данных, 1998. С. 323–333.